Lie Methods for Nonlinear Dynamics with Applications to Accelerator Physics

Alex J. Dragt

University of Maryland, College Park

http://www.physics.umd.edu/dsat/

19 November 2020

Alex J. Dragt
Dynamical Systems and Accelerator Theory Group
Department of Physics
University of Maryland
College Park, Maryland 20742

http://www.physics.umd.edu/dsat

Work supported in part by U. S. Department of Energy Grant DE-FG02-96ER40949.

Contents

List of Figures

List of Tables

Preface

John Wallis (1616-1703), Savilian Professor of Geometry at Oxford, was a mathematician and predecessor of Isaac Newton. His most important book, published in 1656, was *Arithmetica Infinitorum*. It introduced, among others, the concepts of negative and fractional exponents, and considered the problem of finding the areas under curves described by functions involving such exponents. He also introduced the symbol ∞. In 1685 he published *Algebra*.

His contemporary Thomas Hobbes (1588-1679), a philosopher and political theorist, read (or perhaps only paged through) this 1656 book, and described it as a "a scab of symbols as if a hen had been scraping there." Apparently taken by this simile, on another occasion he wrote of Wallis: "And for (your book) on *Conic Sections*, it is covered over with a scab of symbols that I had not the patience to examine whether it be well or ill demonstrated." He goes on to say: "Symbols, though they shorten the writing, yet do not make the reader understand it sooner than if it were written in words. $\cdots$ (with the use of symbols) there is a double labour of the mind, one to reduce your symbols to words, which are also symbols, another to attend to the ideas which they signify."

But, according to Leibniz (1646-1716), "In symbols one observes an advantage in discovery which is greatest when they express the exact nature of a thing briefly and, as it were, picture it; then indeed the labor of thought is wonderfully diminished."[1] Laplace (1749-1827) was even more enthusiastic when he wrote "Such is the advantage of a well-constructed language that its simplified notation often becomes the source of profound theories." And, according to Whitehead (1861-1947), "Civilization advances by extending the number of

[1] Leibniz invented much of modern calculus notation. He also introduced the term *dynamick* for what Newton (1642-1727) had previously called *rational mechanics*. But Newton objected to this name, not because of its "inadequacy to describe the subject matter", but rather because Leibniz had "set his mark upon the whole science of forces calling it Dynamick, as if he had invented it himself & is frequently setting his mark upon things by new names & new Notations". Leibniz was kinder to Newton when he wrote "Taking mathematics from the beginning of the world to the time of Newton, what he has done is much the better half." For a history of how Leibnizian notation came to be used in Great Britain, see the Web site https://en.wikipedia.org/wiki/Analytical_Society.

To Descartes (1596-1650) we owe the use of the symbols $a, b, c\cdots$ as constants, the symbols $x, y, z\cdots$ as variables, writing xx as x^2 etc., and, of course, honor for forging the connection between algebra and geometry (to create analytic geometry) by the use of Cartesian coordinates including making graphs of functions. To add to the list: Robert Recorde in 1540 introduced the $+$ and $-$ symbols for addition and subtraction and in 1557 introduced the equal sign $=$, William Oughtred in 1631 introduced the multiplication sign $\times$ and the trigonometric function symbols sin and cos, Johann Rahn in 1659 introduced the division sign $\div$ and the therefore sign $\therefore$, and William Jones in 1706 introduced use of the Greek letter π to denote the value that is the ratio of the circumference to the diameter for any circle and use of a dot above a letter to denote differentiation with respect to time.

important operations which we can perform without thinking of them.”

The purpose of this book is to explore and illustrate how Lie-algebraic/map methods and Lie-algebraic concepts/symbols are broadly applicable to many areas of Nonlinear Dynamics including Accelerator Physics.

Reference

J. Mazur, *Enlightening symbols: a short history of mathematical notation and its hidden power*, Princeton University Press (2014).

Acknowledgments

U.S. Department of Energy, University of Maryland, Institut des Hautes Études Scientifiques (IHÉS), Institute for Theoretical Physics Santa Barbara, Los Alamos National Laboratory, Lawrence Berkeley National Laboratory, Stanford Linear Accelerator Center.

John Horvath, Richard K. Cooper, David Sutter, Eyvind Wichmann, David Judd, Robert Karplus, Alex Chao, Desmond Barber, Sateesh Mane, Peter Walstrom, C. Thomas Mottershead, Filippo Neri, David Douglas, Étienne Forest, Liam Healy, Marco Venturini, Robert Ryne, Dan Abell, Chad Mitchell.

I was like a boy playing on the sea-shore, and diverting myself now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me.

Isaac Newton

For in Him we live and move and have our being.

Acts 17:28

Assertion made by Saint Paul about the "unknown god" and attributed by Paul to an unnamed Greek poet, now thought to be Epimenides of Knossos because this line appears in his poem *Cretica*.

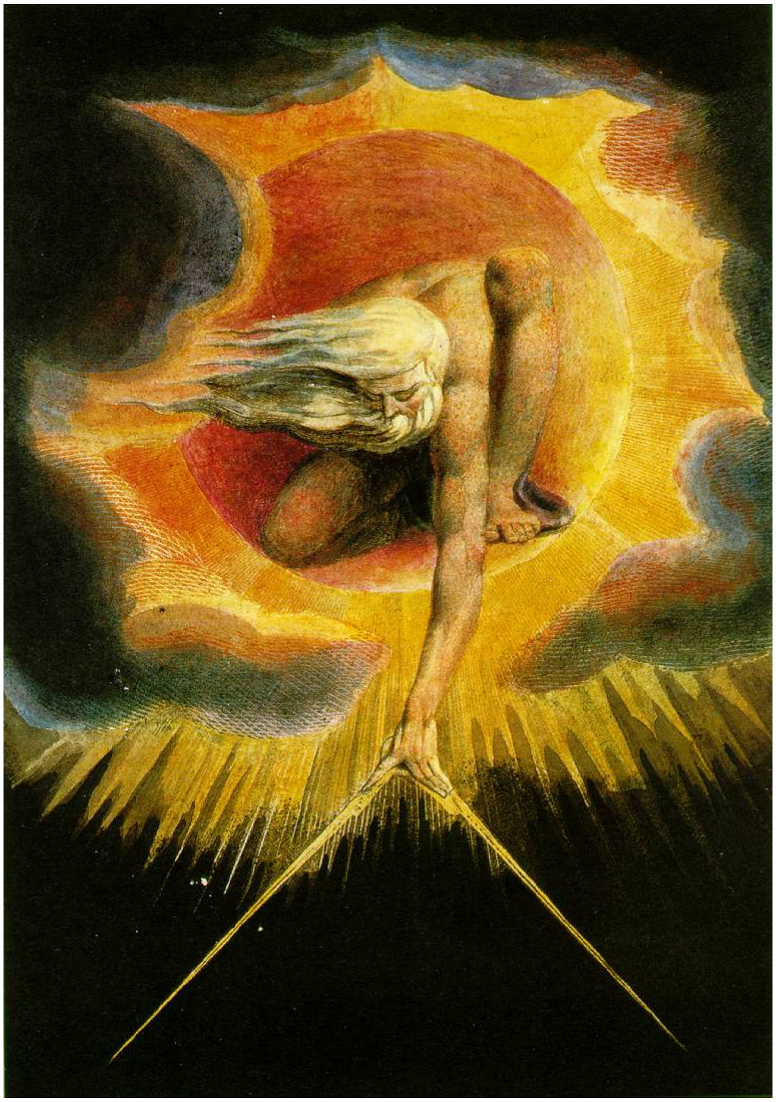


Figure 0.0.1: The Ancient of Days. “If the doors of perception were cleansed, everything would appear to man as it is: Infinite.” William Blake (1757-1827)

Chapter 1

Introductory Concepts

This book is devoted to the subject of Nonlinear Dynamics and the use of Lie Methods for the description and study of Nonlinear Dynamics. Where appropriate, special attention will be given to the application of these methods to the field of charged-particle electro-magnetic optics in general and Accelerator Physics in particular.[1] The purpose of this chapter is to provide introductory background material that will be needed throughout the book. The first four sections of this chapter provide an introduction to the history and use of *maps*, and their relation to differential equations, Hamiltonian dynamics, and Lie theory. Some terms in these sections may not be completely familiar. (If they are, perhaps you need not read further save as a cure for insomnia.) They will be explained and/or properly referenced subsequently. The remaining three sections treat some required aspects of Hamiltonian dynamics.

1.1 Transfer Maps

The use and analysis of maps now plays a major role in nonlinear dynamics and accelerator physics. Much of the material of this book will be dedicated to a map approach. The current use of maps arises from the confluence of two mathematical/physical streams of thought. The first of these streams originates in Geometry, and dates back to the ancient Greeks. The second is related to Dynamics, and originates largely in the discoveries of Isaac Newton (1642-1727).[2]

[1] Lie Methods can also be applied to Light Optics. See Appendix X.

[2] Newton published his first edition of *Philosophiæ Naturalis Principia Mathematica* in 1687, and subsequent editions in 1713 and 1726. Concerning Newton, Laplace said "There is but one law of the cosmos, and Newton has discovered it." Vladimir Arnold was asked: "Mathematics is a very old and important part of human culture. What is your opinion about the place of mathematics in cultural heritage?" Arnold replied: "The word 'mathematics' means science about truth. It seems to me that modern science (i.e., theoretical physics along with mathematics) is a new religion, a cult of truth, founded by Newton three hundred years ago."

1.1.1 Maps and Dynamics

> Prediction is very difficult, especially about the future.
>
> *Niels Bohr* (1885-1962), *Yogi Berra* (1925-2015)

> Nature and Nature's laws lay hid in night:
> God said, Let Newton be! and all was light.
>
> *Alexander Pope* (1688-1744)

> And from my pillow, looking forth by light
> Of moon or favoring stars, I could behold
> The antechapel where the statue stood
> Of Newton with his prism and silent face,
> The marble index of a mind forever
> Voyaging through strange seas of thought, alone.
>
> *William Wordsworth* (1770-1850)

> Then ye who now on heavenly nectar fare,
> Come celebrate with me in song the name
> Of Newton, to the Muses dear; for he
> Unlocked the hidden treasuries of Truth:
> So richly through his mind had Phoebus cast
> The radiance of his own divinity.
> Nearer the gods no mortal may approach.
>
> *Edmond Halley* (1656-1742)

> So few went to hear him, and fewer understood him, and ofttimes he did, for want of hearers, read to the walls. He usually stayed about half an hour; when he had no auditors he commonly returned in a quarter of that time.
>
> *Teaching Evaluation of Professor Newton* (circa 1690)

Let us begin with the second stream, the stream of Dynamics. Newton's basic and most remarkable discovery was that motion is governed by *mathematical laws*, and the nature of these laws is such that the *future* can be determined/predicted from a knowledge of the *present*![3] We illustrate this fact with the sketch in Figure 1.1. Suppose we think of the

[3] Roger Cotes, Newton's student, wrote the preface to the second edition of Newton's *Principia*. Much of this preface is devoted to defending the thesis that the ability to generate and respond to gravity in proportion to its mass is a *natural* property of every object (Cavendish did his experiment 70 years after Newton's death), and not an *occult* property as many critics complained, and to criticizing Descartes' rival theory of vortices. He also writes, with regard to what we call *natural laws*, "The business of true philosophy is to derive the nature of things from causes truly existent; and to enquire after those laws on which the Great Creator actually chose to found this most beautiful Frame of the World; not those by which he might have done the same, had he so pleased. $\cdots$ Without all doubt this World, so diversified with that variety

present as a set of *initial conditions*, and regard the future as a set of *final conditions*. Newton's laws, when appropriately formulated, can be regarded as a set of first-order ordinary differential equations. Indeed, Newton viewed differential equations and their applicability to describing nature as one of his fundamental discoveries, so important that he kept it secret initially by revealing it [in a 1676 letter (via Oldenburg) to his calculus rival Leibniz (1646-1716)] only in the form of an cypher/anagram/cryptogram:

6accdae13eff7i3l9n4o4qrr4s8t12ux

which Newton's friend Wallis years later (in his 1693 book *Algebra*, second edition) disclosed stood for

of forms and motions we find in it, could arise from nothing but the perfectly free will of God directing and presiding over all. From this fountain it is that those laws, which we call the laws of Nature, have flowed; in which there appear many traces indeed of the most wise contrivance, but not the least shadow of necessity. $\cdots$ He who thinks to find the true principles of physics and the laws of natural things by the force alone of his own mind, and the internal light of his reason must either suppose that the World exists by necessity, and by the same necessity follows the laws proposed; or if the order of Nature was established by the will of God, that himself, a miserable reptile, can tell what was fittest to be done. $\cdots$ He must be blind who from the most wise and excellent contrivances of things cannot see the infinite Wisdom and Goodness of their Almighty Creator, and he must be mad and senseless who refuses to acknowledge them." Newton himself wrote (in his book *Opticks*): "The main Business of natural Philosophy is to argue from Phenomena without feigning Hypotheses, and to deduce Causes from Effects, till we come to the very first Cause, which certainly is not mechanical."

With regard to the concept of *necessity*, it is interesting that centuries later Einstein (perhaps with his reptilian brain?) wrote: "What I am really interested in is whether God could have created the world in a different way; that is, whether the necessity of logical simplicity leaves any freedom at all?$\cdots$I would like to state a theorem which at present cannot be based on anything more than faith in the simplicity, i.e., intelligibility, of nature: there are no *arbitrary* constants $\cdots$ that is to say, nature is so constituted that it is possible logically to lay down such strongly determined laws that within these laws only rationally determined constants occur."

Newton himself wrote the preface to the first edition of the *Principia*, and laid out his goals as follows: "$\cdots$ for the whole burden of philosophy seems to consist of this - from the phenomena of motions to investigate the forces of nature, and then from these forces to demonstrate the other phenomena; and to this end the general propositions in the first and second Books are directed. In the third Book I give an example of this in the explication of the System of the World; for by the propositions mathematically demonstrated in the former books, in the third I derive from the celestial phenomena the forces of gravity with which bodies tend to the sun and the several planets. Then from these forces, by other propositions which are also mathematical, I deduce the motion of the planets, the comets, the moon, and the sea. I wish we could derive the rest of the phenomena of Nature by the same kind of reasoning from mechanical principles, for I am induced by many reasons to suspect that they may all depend upon certain forces by which the particles of bodies, by some causes hitherto unknown, are either mutually impelled towards one another, and cohere in regular figures, or are repelled and recede from one another. These forces being unknown, philosophers have hitherto attempted the search of Nature in vain; but I hope the principles here laid down will afford some light either to this or some truer method of philosophy."

With but a few editorial changes Newton's words could equally well serve today as justification for the support of contemporary basic research! If, for the sake of argument, we identify the the aims of "basic research" with those of High Energy Elementary Particle Physics, at the risk of offending a few colleagues, then we see that the goal remains the same, and even the subject matter has changed relatively little. Under the rubric of *bound states* and *scattering theory*, we still wonder about fundamental "forces" and "particles", and why they "cohere in regular figures, or are repelled and recede form one another." And as Newton hoped, his "principles have afforded some light on the truer methods" of Quantum Mechanics, Quantum Field Theory including the Standard Model of Particle Physics, General Relativity including the Standard Model of Cosmology, and the mysteries of Dark Matter and Dark Energy.

> *Data aequatione quotcunque fluentes quantitates involvente, fluxiones invenire; et vice versa*

and means

> Given an equation involving any number of fluent quantities, find the fluxions; and vice versa.

In effect, Newton said it is useful to formulate and solve differential equations.[4]

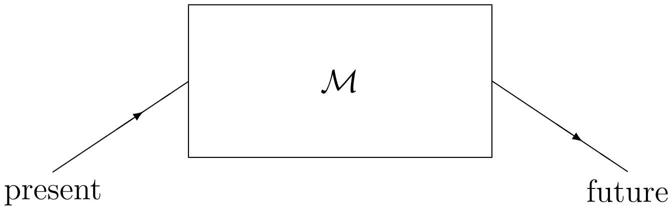


Figure 1.1.1: In Dynamics the future can be determined by performing a certain operation, called a mapping $\mathcal{M}$, on the present.

As will be described in Section 1.3, there are mathematical theorems about first-order ordinary differential equations to the effect that they generally have solutions. (That is, solutions *exist*.) Moreover, under quite general circumstances, these solutions are *unique* and are completely determined by the initial conditions. Thus there is a rule, or *mapping* $\mathcal{M}$, that sends the initial conditions (the present) into the final conditions (the future): one simply integrates Newton's equations in first-order form, perhaps numerically on a computer.[5]

In the same era, on the continent across the Channel from Newton, Leibniz wrote (in the context of a problem for which the future depends very sensitively on the present):

> That everything is brought forth through an established destiny is just as certain as that three times three is nine. $\cdots$ If, for example, one sphere meets another sphere in free space and if their sizes and their paths and directions before collision are known, we can then foretell and calculate how they will rebound and what course they will take after the impact. Very simple laws are followed

[4] For the Leibniz-Newton calculus controversy, see the Web link `https://en.wikipedia.org/wiki/Leibniz-Newton_calculus_controversy`. With regard to their rivalry, there is equity in the universe. Most modern calculus notation such as dy/dx and $\int ydx$ is due to Leibniz. He also coined the term *calculus*. Moreover, there might appear to be parity on the cookie front. There are the Fig Newton (1891) and the Leibniz Butterkeks (also 1891). But, alas for I. Newton, the Fig Newton is named for a Massachusetts town whose original name was Newtown.

[5] Given the final conditions (the future), one can equally well integrate backwards in time to find/retrodict the initial conditions (the present), or even farther back to find the past. Thus, we may equally well say that the future determines the present and even the past. The conditions at any instant determine the conditions at all other instants, both future and past. Mathematically, this means that the transfer map $\mathcal{M}$ associated with any set of first-order ordinary differential equations is *invertible*.

> which also apply, no matter how many spheres are taken or whether objects are taken other than spheres. From this one sees then that everything proceeds mathematically -that is, infallibly- in the whole wide world, so that if someone could have a sufficient insight into the inner parts of things, and in addition had remembrance and intelligence enough to consider all the circumstances and to take them into account, he would be a prophet and would see the future in the present as in a mirror.

That this concept (in the context of motion) was generally understood in scholarly circles a generation after Newton and Leibniz is evident from the work of the Serbian Jesuit scholar Boscovich (1711-1787). In 1763 he wrote:

> Any point of matter $\cdots$ must describe some continuous curved line, the determination of which can be reduced to the following general problem. Given a number of points of matter, and given, for each of them, the point of space that it occupies at any given instant of time; also given $\cdots$ the tangential velocity $\cdots$; and given the law of forces $\cdots$; it is required to find the path of each of the points, that is to say, the line along which each of them moves. How difficult this mechanical problem may become, how it may surpass all powers of the human mind, can be easily understood by anyone who is versed in Mechanics and is not quite unaware that the motion of even three bodies only, and those possessed of a perfectly simple law of force, have not yet been completely determined in general $\cdots$. Now although a problem of such a kind surpasses all the powers of the human intellect, yet any geometer can easily see thus far that the problem is determinate $\cdots$. Now, if the law of forces were known, and the position, velocity and direction of all the points at any given instant (were known), it would be possible for a mind of this type to foresee all the necessary subsequent motions and states, and to predict all the phenomena that necessarily followed from them.

Laplace (1749-1827) subsequently stated this concept equally explicitly in 1814 when he wrote:

> We ought then to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it—an intelligence sufficiently vast to submit these data to analysis—it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes. The human mind offers, in the perfection which it has been able to give to astronomy, a feeble idea of this intelligence. Its discoveries in mechanics and geometry, added to that of universal gravity, have enabled it to comprehend in the same analytical expressions the past and future states of the system of the world. Applying the same method to some other objects of its knowledge, it has succeeded in referring to general laws observed phenomena and in foreseeing those which given circumstances ought to produce. All these efforts

> in the search for truth tend to lead it back continually to the vast intelligence which we have just mentioned, but from which it will always remain infinitely removed. This tendency, peculiar to the human race, is that which renders it superior to animals; and their progress in this respect distinguishes nations and ages and constitutes their true glory.

Note the similarity in language![6] The same perception is echoed by Thomasina Coverly in Tom Stoppard's 1993 play *Arcadia*. In Act I she says:

> If you could stop every atom in its position and direction, and if your mind could comprehend all the actions thus suspended, then if you were really, really good at algebra you could write the formula for all the future; and although nobody can be so clever as to do it, the formula must exist just as if one could.

In modern terminology, Leibniz, Boscovich, Laplace, and Thomasina (Stoppard) were describing what we call a *transfer map*.[7]

All this would have pleased the ancient Greek Stoic philosophers both in buttressing their belief in determinism and in addressing their desire to divine the future. As Cicero explained in his 44 B.C. work *On Divination*,

> Besides, since everything happens by fate, as will be shown elsewhere, if there could be any mortal who could observe with his mind the interconnection of all causes, nothing indeed would escape him. For he who knows the causes of things that are to be necessarily knows all the things that are going to be. But since no one but God could do this, what is left for man is that he should be aware of future things in advance by certain signs which make clear what will follow. For the things which are going to be do not come into existence suddenly, but the passage of time is like the unwinding of a rope, producing nothing new but unfolding what was there at first.

Newton showed that what was needed to determine the future was a knowledge of the initial conditions and the universal force laws (the inverse square law for gravity in his case), followed by the integration of his equations of motion. And integration of the equations of motion, particularly when carried out time-step by time-step numerically (see Chapter 2), does resemble, in some ways, the unwinding of a rope. Whatever is produced is not "new", but rather already inherent in the initial conditions.

[6] Later commentators and philosophers of science sometimes refer to Laplace's *vast intelligence* by the (what might be understood as pejorative) term Laplace's *demon*, perhaps in analogy to Maxwell's demon. Laplace never used that term, and based on his usage above it could be argued that he was envisioning an admirable/exalted transcendent/divine being. Actually, Maxwell didn't use the term demon for his being either. It was first introduced by Kelvin in 1874, and he implied that he intended the mediating, rather than malevolent, connotation of the word.

[7] The use of the terminology *transfer map* in this context is not to be confused with the use of the same terminology in other contexts including computer graphics, statistical mechanics, various aspects of group theory, and the articulation of courses between different universities and colleges. Our usage is motivated by terminology in (light) ray optics. In ray optics the the linear (paraxial) approximation of what we call a transfer map is called a *transfer matrix*.

Let us continue our historical narrative: Lagrange (1736–1813) and others discovered that for many systems of physical interest all the differential equations of motion could be generated by (derived from) a *single* master function now called the Lagrangian L.[8] These equations of motion were second order. Building on this work, Hamilton (1805–1865) showed that it was possible to write a related set of first-order equations, and that all these equations could also be generated by a single master function now called the Hamiltonian H.[9]

Being well aware of the aforementioned properties of first-order differential equations, Hamilton made a detailed study of the nature of the relation between initial and final conditions (the transfer map $\mathcal{M}$) for Hamiltonian systems. In modern language, he showed that such maps must be *symplectic* (canonical). He also discovered mixed-variable generating functions, and showed that they can be used at will to produce symplectic maps. Finally, he and Jacobi (1804-1851) studied how symplectic maps could be used to transform Hamiltonians with the aim of simplifying them, and thus also the differential equations and flows they generate. In modern terminology, their work was the beginning of the Theory of Normal Forms for differential equations, Hamiltonians, and Symplectic Maps.

Poincaré (1854-1912) was the next person to champion the use of maps and explore their properties: He introduced what we now call stroboscopic maps and Poincaré surface-of-section maps. He showed that the existence of an infinite number of periodic orbits in the gravitational 3-body problem would follow from proving the existence of two fixed points for a certain symplectic map of an annulus (in the plane) into itself.[10] He discovered what we now call the Poincaré invariants and showed that they are preserved by symplectic maps. He studied normal forms for differential equations and showed that attempts to use symplectic maps to bring certain classes of Hamiltonians to a certain kind of normal form, which if successful would prove the existence of integrals of motion, seemed fraught with intractable difficulties due to the appearance of so-called *small denominators* that potentially spoil the convergence of the series designed to construct the desired normal form.[11] He also discovered what are now called *homoclinic tangles*, emphasized their generic existence, and demonstrated that their presence destroys integrability and leads to chaos.

Birkhoff (1884-1944), in addition to making other outstanding contributions to mathematics, extended the program of Poincaré. In a celebrated early paper he was able to prove

[8] Together Lagrange and Euler (1707–1783), and later Hamilton, also developed variational calculus and showed that Lagrangian and variational formulations are equivalent. Presently it is commonly assumed that any fundamental theory of nature will be Lagrangian in form. See Section 5.

[9] The function H, its relation to L by way of a Legendre (1752–1833) transformation, and the resulting equations of motion were actually introduced earlier by Lagrange when Hamilton was still a child. Lagrange used the letter H to honor Huygens (1629–1695). Hamilton wrote definitive papers on light optics and dynamics in which he introduced characteristic (generating) functions and also employed the H of Lagrange. See Appendix X. To his great fortune, after that H became known as the Hamiltonian. With regard to Lagrange, Hamilton wrote "Lagrange has perhaps done more than any other to give extent and harmony to such deductive researches by showing that the most varied consequences . . . may be derived from one radical formula, the beauty of the method so suiting the dignity of the results as to make his great work a kind of scientific poem."

[10] The conjecture that the symplectic map of the annulus into itself must have two fixed points is called Poincaré's last geometric theorem. He in fact knew that the existence of one fixed point already entailed the existence of two fixed points, and therefore it is only necessary to prove the existence of one fixed point.

[11] Poincaré was unable to prove either the convergence or divergence of the series in question, but inclined toward the opinion that such series were generally divergent.

what, despite considerable effort, had eluded Poincaré: the map for the 3-body problem developed by Poincaré did indeed have two fixed points. (He proved that the assumption that Poincaré's annulus map had no fixed point entailed a contradiction.) He also studied the possibility of using symplectic maps to bring certain classes of Hamiltonians to what is now called Birkhoff normal form. Again, he found that the appearance of small denominators potentially destroyed convergence, and left the convergence question unanswered. Finally, Birkhoff made other significant contributions to Dynamics including fundamental work on ergodic theory and the areas we now call bifurcation theory and symbolic dynamics.

Siegel (1896-1981) was the first to master the small denominator problem in the context of analytic maps of the complex plane into itself. Subsequently, Moser (1928-1999) overcame this problem for "twist" and area-preserving (symplectic) maps of the plane into itself under the assumption of only sufficiently high-order differentiability. Kolmogorov (1903-1987) and Arnold (1937-2010) handled the small denominator problem for the case of symplectic maps/Hamiltonian systems in any number of dimensions under the assumption of analyticity. Together their work proved, under suitable assumptions, the existence of KAM (Kolmogorov-Arnold-Moser) tori for symplectic/Hamiltonian systems. Major advances/extensions in KAM theory were made subsequently by Aubry, Mather, Nekhoroshev, Chirikov, and others.

Smale (1930-) greatly extended symbolic dynamics, invented his horseshoe construction which he described using symbolic dynamics, and showed that Poincaré's homoclinic tangle contained a horseshoe.

Recent advances in nonlinear dynamics include bifurcation and chaos theory, symplectic differential geometry and symplectic topology, and special numerical integration methods often referred to as *geometric/structure-preserving/symplectic* integration.

1.1.2 Maps and Accelerator Physics

Let us momentarily turn our attention to accelerator physics. Courant and Snyder pioneered the use of matrices to characterize transverse beam behavior in the linear (first-order or paraxial) approximation. These matrices were enlarged by Penner to include chromatic (energy dependent) effects.[12] In subsequent work Brown made the important step of extending the linear matrix formalism to include nonlinear effects through second order. From the perspective of maps, we may view the use of a matrix as making a linear approximation to the underlying transfer map $\mathcal{M}$, and the inclusion of second-order effects as introducing the first nonlinear terms that appear in a Taylor expansion of $\mathcal{M}$ about some design orbit. It is now relatively easy to compute the terms in a Taylor expansion of $\mathcal{M}$ to very high order. This computation is made possible by two tools. The first is the use of Lie methods. The second consists of *Truncated Power Series Algebra* (TPSA) and/or *Automatic Differentiation* (AD) computer programs that manipulate very high-order polynomials and various other familiar functions in several variables.[13] Both these topics will be described extensively in subsequent chapters.

[12] Time-dependent effects were first included in the Lie algebraic code *MaryLie*.

[13] Some authors refer to AD as *Differential Algebra* (DA).

1.1.3 Maps and Geometry

Euclid alone has looked on Beauty bare.
Let all who prate of Beauty hold their peace,
And lay them prone upon the earth and cease
To ponder on themselves, the while they stare
At nothing, intricately drawn nowhere
In shapes of shifting lineage; let geese
Gabble and hiss, but heroes seek release
From dusty bondage into luminous air.

O blinding hour, O holy, terrible day,
When first the shaft into his vision shone
Of light anatomized! Euclid alone
Has looked on Beauty bare. Fortunate they
Who, though once only and then but far away,
Have heard her massive sandal set on stone.

Edna St. Vincent Millay (1892-1950)

We now consider the first stream, the stream of Geometry. A fundamental notion in geometry as conceived by Euclid (c. 300 B.C.) is that of *congruence*. Roughly speaking, we regard two triangles as congruent if one can be placed over the other with a resulting perfect fit. From the perspective of maps, we have in mind the operations of translations and rotations which map Euclidean space into itself. Together these operations form a group, the *Euclidean group*. Thus, following Felix Klein (1849-1925), we may say that two triangles are congruent if one can be *transformed* into the other under the action of the Euclidean group. And two triangles are *similar* if one can be transformed into the other under the action of the Euclidean group augmented by scale transformations.

The concepts underlying the Euclidean group were subsequently broadened by Klein (as part of his Erlangen program) and others to include the idea of general transformation groups that map various kinds of spaces or various classes of objects into themselves.[14] Sophus Lie (1842-1899), and others both before and after him (including Poincaré), studied transformation groups for their applications to both geometry and function theory, and (in what amounts to a systematic procedure for transforming variables) the simplification and perhaps even solution of certain classes of differential equations.[15] Lie studied in particular the properties of what we now call Lie groups: groups that can be *generated* by near-identity operations. The generators of these near-identity operations form algebras which we now call Lie algebras. For example, in the case of the rotation group (a subgroup of the Euclidean group) there exist small (infinitesimal) rotations, and any group element

[14] The importance of groups was not always universally appreciated. In 1910 a board of experts including Oswald Veblen and Sir James Jeans, upon reviewing the mathematics curriculum at Princeton, concluded that group theory ought to be thrown out as useless. And, in the early days of Quantum Mechanics, the work of those physicists/mathematicians who sought to apply group theory to this new field was referred to as *Gruppenpest*.

[15] By developing a theory of continuous groups, Lie aspired to do for differential equations what Galois had done for algebraic (polynomial) equations using finite groups.

can be constructed (infinitesimally generated) from these near-identity operations. When a matrix representation is used (and assuming a three-dimensional space), the generators of the infinitesimal rotations are three matrices, call them L_x, L_y, and L_z, that obey the commutation (Lie algebraic multiplication) rules

$$\{L_x, L_y\} = L_x L_y - L_y L_x = L_z, \text{ etc.} \tag{1.1.1}$$

The elements of the set of all continuous and invertible maps of a space into itself are called *homeomorphisms*. Topology (another area pioneered largely by Poincaré) is the study of those properties of spaces, and objects in these spaces, that are invariant under homeomorphisms. Homeomorphisms that are differentiable are called *diffeomorphisms*. The set of all diffeomorphisms forms a group that is a Lie group. Differential geometry is the study of those properties of spaces, and objects in these spaces, that are invariant under diffeomorphisms.

The set of all symplectic maps (sometimes called *symplectomorphisms*) also forms a Lie group, and this Lie group is a subgroup of the Lie group of diffeomorphisms. In both the group of all diffeomorphisms and the group of all symplectic maps, *Lie transformations* are those group elements produced by a single generator. Hori (1932–) and Deprit (1926–2006) were the first (in the context of Dynamics) to use Lie transformations for the production of symplectic maps. They employed these maps to try to bring to normal form various Hamiltonians that arise in celestial mechanics, and showed that the use of Lie transformations is often much more convenient than the method of mixed-variable generating functions developed earlier by Hamilton and Jacobi. As will be described in subsequent chapters, Lie algebraic methods also have important applications to Accelerator Physics. In this case Lie transformations, and products of Lie transformations, can be used to represent symplectic transfer maps, and Lie algebraic formulas (the Baker-Campbell-Hausdorff and Zassenhaus formulas) can be used to multiply and factorize maps. Lie methods can also be used to bring transfer maps to normal form. Among other things, normal form theory generalizes Courant-Snyder theory to the nonlinear regime.

1.2 Map Iteration and Other Background Material

There are important situations where it is desirable to know the effect of a map when it is applied a large number of times. Consider, for example, the case of a charged-particle storage ring. Such rings can be characterized by a one-turn map; call this map $\mathcal{M}$. Since storage rings are intended to hold particles for long periods of time and correspondingly a large number of turns, we find we are interested in properties of $\mathcal{M}^n$ for values of n in the range 10^8 to 10^{10}.

We observe that the concept of map iteration, or equivalently the study of $\mathcal{M}^n$ for large n, introduces an *infinity* into the game. Consequently, we might anticipate that phenomena arising from the iteration of maps could be very complicated. This is indeed the case.

1.2.1 Logistic Map

Consider, as a simple example, the biological subject of insect population growth. Let P_n be the population in year n (of some insect species), and let P_{n+1} be the population the following year. Then we might imagine that there is some kind of rule (or map) $\mathcal{M}$ that relates the population in two successive years as shown schematically in Figure 2.1.

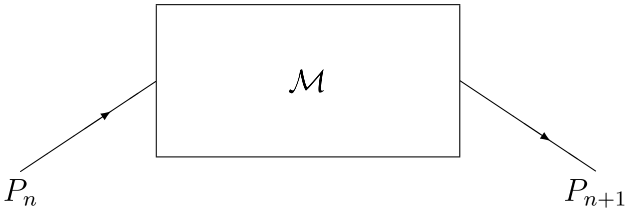


Figure 1.2.1: The insect populations in two successive years are related by a map $\mathcal{M}$.

The simplest form for the map $\mathcal{M}$ is a relation of the kind

$$P_{n+1} = \alpha P_n, \tag{1.2.1}$$

where α is viewed as some *fixed* growth rate. However, depending on the size of α, the recursion relation (2.1) has only exponentially damped or exponentially growing solutions; and both these possibilities are unphysical – the actual insect population is neither dropping to zero nor growing indefinitely.

An improved model would be to assume that the growth rate itself depends on the current population. For example, we might imagine that if the population were small, then food would be plentiful, and the growth rate should be high. Conversely, if the population were at some maximum value $P_{\max}$, then food might be in such short supply that there would be no reproduction at all. A simple form for α having this property is obtained by writing

$$\alpha(P) = \beta(P_{\max} - P). \tag{1.2.2}$$

With this improved model the map $\mathcal{M}$ takes the form

$$P_{n+1} = \beta(P_{\max} - P_n)P_n. \tag{1.2.3}$$

Finally, for mathematical convenience, let us introduce the *fractional* population x defined by the rule

$$x = P/P_{\max}. \tag{1.2.4}$$

In terms of this variable the relation (2.3) takes the form known as the *logistic map* or *Verhulst process*,

$$x_{n+1} = f(\lambda, x_n) = \lambda x_n(1 - x_n). \tag{1.2.5}$$

(Here $\lambda = \beta P_{\max}$.) Note that (2.5) has the physically desirable property that $x_{n+1} \in [0, 1]$ if $x_n \in [0, 1]$ provided $\lambda \in [0, 4]$.

Let us solve (2.5) for an *equilibrium* value (*fixed point*) x_e. By definition, and using map notation, this value must satisfy the relation

$$\mathcal{M}x_e = x_e, \tag{1.2.6}$$

from which we find the result

$$x_e = \lambda x_e(1 - x_e) \tag{1.2.7}$$

with the solutions

$$x_e = 0, \tag{1.2.8}$$

$$x_e = (\lambda - 1)/\lambda. \tag{1.2.9}$$

Suppose we select some value x_0 for an initial (fractional) population and apply the map $\mathcal{M}$ repeatedly for a total of m times to find the result

$$x_m = \mathcal{M}^m x_0. \tag{1.2.10}$$

That is, we carry out the operation (2.5) for a total of m times. Then we might wonder what happens in the limit of large m. (The set of all points x_m for all integer m is called the *orbit* of x_0 under the action of $\mathcal{M}$). For example, do the x_m approach x_e (in which case x_e is called an *attractor*), or does something else happen?

Figure 2.2 shows the values x_m as a function of m starting with $x_0 = 1/2$ for the case $\lambda = 2.8$. Other starting values of x_0 give similar similar results as m becomes large. Evidently the x_m converge to the value x_e given by (2.9) as $m \to \infty$, and x_e is an attractor. All points (starting values) x_0 such that the associated x_m converge to x_e are said to be in the *basin of attraction* of x_e. Let x_f be an attracting fixed point for some map $\mathcal{M}$,

$$\mathcal{M}x_f = x_f. \tag{1.2.11}$$

In set theoretic language, $B(x_f)$, the basin of x_f under the action of $\mathcal{M}$, is defined by the rule

$$B(x_f) = \{x \mid \lim_{n\to\infty} \mathcal{M}^n x = x_f\}. \tag{1.2.12}$$

By contrast, Figure 2.3 shows the values x_m as a function of m starting with $x_0 = 1/2$ for the case $\lambda = 3.01$. Again other starting values of x_0 give similar similar results as m becomes large. Now we see that x_e, while still a fixed point, is no longer an attractor. Instead, as m becomes large, the successive values of x_m settle down to *two alternating* values; and it now takes *two* years for each of these values to repeat itself. We say that *period doubling* has occurred so that for $\lambda = 3.01$ the map $\mathcal{M}^2$ has two attracting fixed points, and $\mathcal{M}$ itself sends each into the other. Insects living in this regime experience alternating fat and lean years! Since the map $\mathcal{M}^2$ has two attracting fixed points, there will be two basins of attraction for $\mathcal{M}^2$, one for each fixed point.

Figure 2.4 shows, as a function of λ, the limiting values, called x_∞, that occur as $m \to \infty$. Such a graphic is often called a *final-state* or *Feigenbaum diagram.* The calculations for this graphic were again made using $x_0 = 1/2$, but other choices in the interval (0,1) would have given the same result. We see that x_∞ is unique for $1 < \lambda < 3$, and can be verified to have

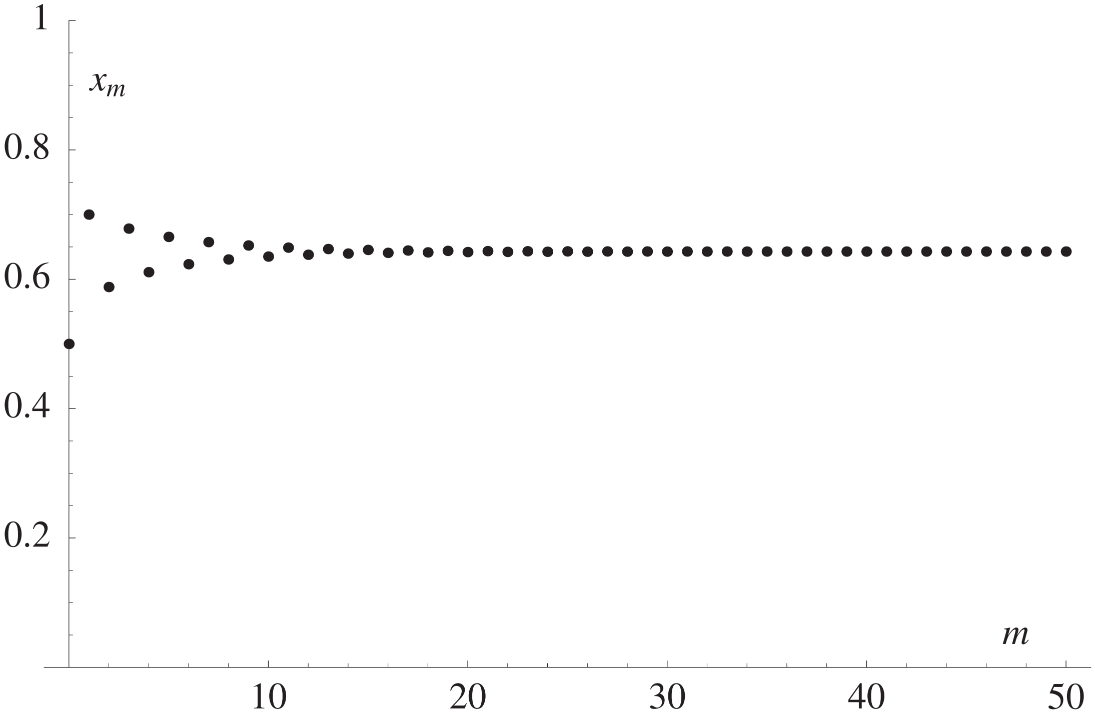


Figure 1.2.2: The values x_m as a function m for the case $\lambda = 2.8$.

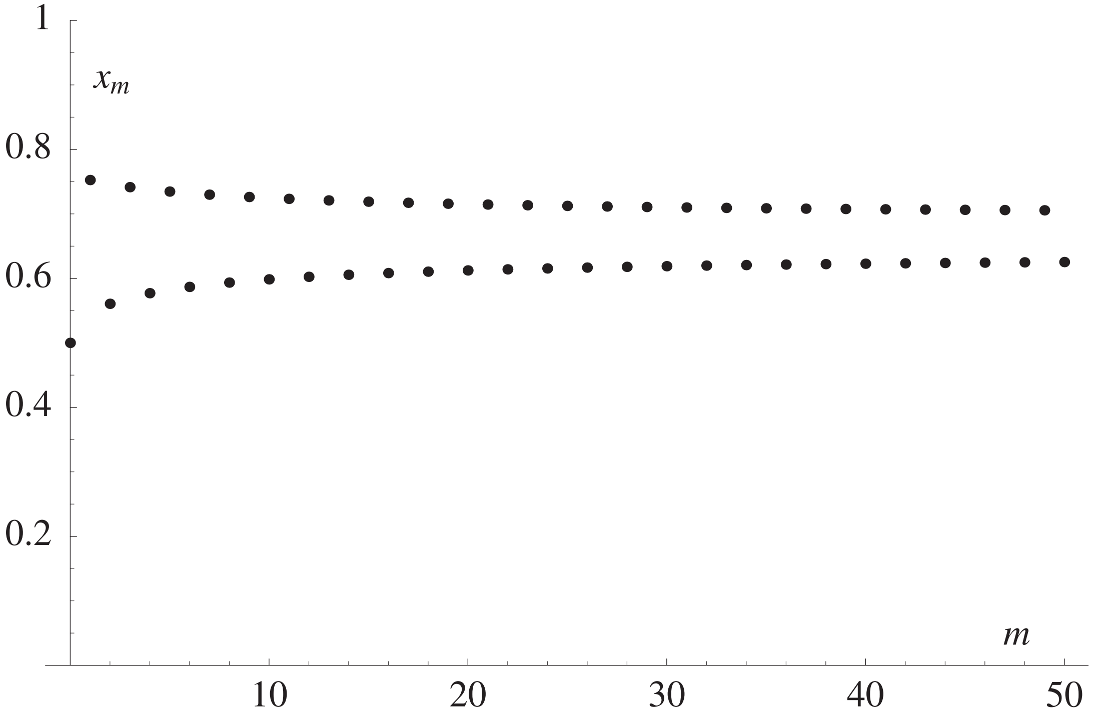


Figure 1.2.3: The values x_m as a function m for the case $\lambda = 3.01$.

the value x_e given by (2.9). That is, this fixed point x_e is attracting (stable) for $1 < \lambda < 3$. However, this x_e is *repelling* (unstable) for $\lambda > 3$ and, although it still is a fixed point, it no longer appears in the figure for these λ values.[16] (A fixed point is called a *repeller* if points near it move away under repeated application of $\mathcal{M}$.) Instead *bifurcation* (period doubling) occurs at $\lambda = 3$ so that, as seen in Figure 2.3, $\mathcal{M}^2$ has two stable fixed points for λ slightly larger than 3.

Inspection of Figure 2.4 shows that there is a cascade of period doublings as λ increases beyond 3. For example, for λ slightly larger than $3.449\cdots$, there are four fixed points of $\mathcal{M}^4$. Application of $\mathcal{M}$ cyclically permutes these points among themselves, and it takes four years for each of these points to repeat itself. Moreover, further inspection shows that an *infinite* number of doublings have occurred by the time λ reaches the *critical* value $\lambda_{\rm cr} \simeq 3.569$. Let $\lambda_1,\ \lambda_2,\ \ldots$ denote the λ values at which successive period doublings occur. The first few values are given by the relations

$$\lambda_1 = 3,\ \ \lambda_2 = 1+\sqrt{6} = 3.449\cdots,\ \ \lambda_3 = 3.544\cdots. \tag{1.2.13}$$

Let us also write $\lambda_\infty = \lambda_{\rm cr} \simeq 3.569$. Then it can be shown that (for sufficiently large j) the λ_j converge to λ_∞ as $j \to \infty$ in the fashion

$$\lambda_j = \lambda_\infty + \gamma\delta^{-j} + \text{higher-order terms}, \tag{1.2.14}$$

with

$$\begin{aligned}\lambda_\infty &= 3.569\cdots,\\ \gamma &= -2.66\cdots,\\ \delta &= 4.6692016\cdots.\end{aligned} \tag{1.2.15}$$

The values of λ_∞ and γ are specific to the logistic map. However, the quantity δ, called the *Feigenbaum constant*, is *universal*. Examination of a graph of the right side of (2.5) shows that the logistic map is produced by a function with one hump (an inverted parabola in this case), and the second derivative of the function does not vanish at the top of the hump. It can be shown that all maps with this property undergo an infinite cascade of period doublings as some appropriate parameter is varied, and there is a relation of the form (2.14) with the *same* (Feigenbaum's) value of δ. [Strictly speaking, what is required is that the *Schwarzian* derivative of the function be negative. If f is any function, its Schwarzian derivative, denoted by Sf, is defined by the rule

$$Sf = \frac{f'''}{f'} - \frac{3}{2}\left(\frac{f''}{f'}\right)^2. \tag{1.2.16}$$

The condition $Sf < 0$ is true for the logistic map, for example, since in this case $f''' = 0$.]

Many systems in nature exhibit a cascade of period doublings, and it is often found experimentally that these cascades behave according to (2.14), again with Feigenbaum's

[16] Sometimes Feigenbaum diagrams are called *bifurcation* diagrams. However. strictly speaking, bifurcation diagrams should also display the unstable fixed points, and Feigenbaum diagrams generally do not. The use of the term *bifurcation* in the context of Dynamics is due to Poincaré.

value. See, for illustration, the case of the Duffing equation treated in Chapter 23. Finally, we remark that there are maps for which the Feigenbaum period-doubling cascade begins as some parameter is varied, but does not complete. Rather, as the parameter is further increased after some finite number of period doublings have occurred, the cascade undoes itself. See Appendix J. There are also systems of physical interest that exhibit this kind of behavior. Again see Chapter 23.

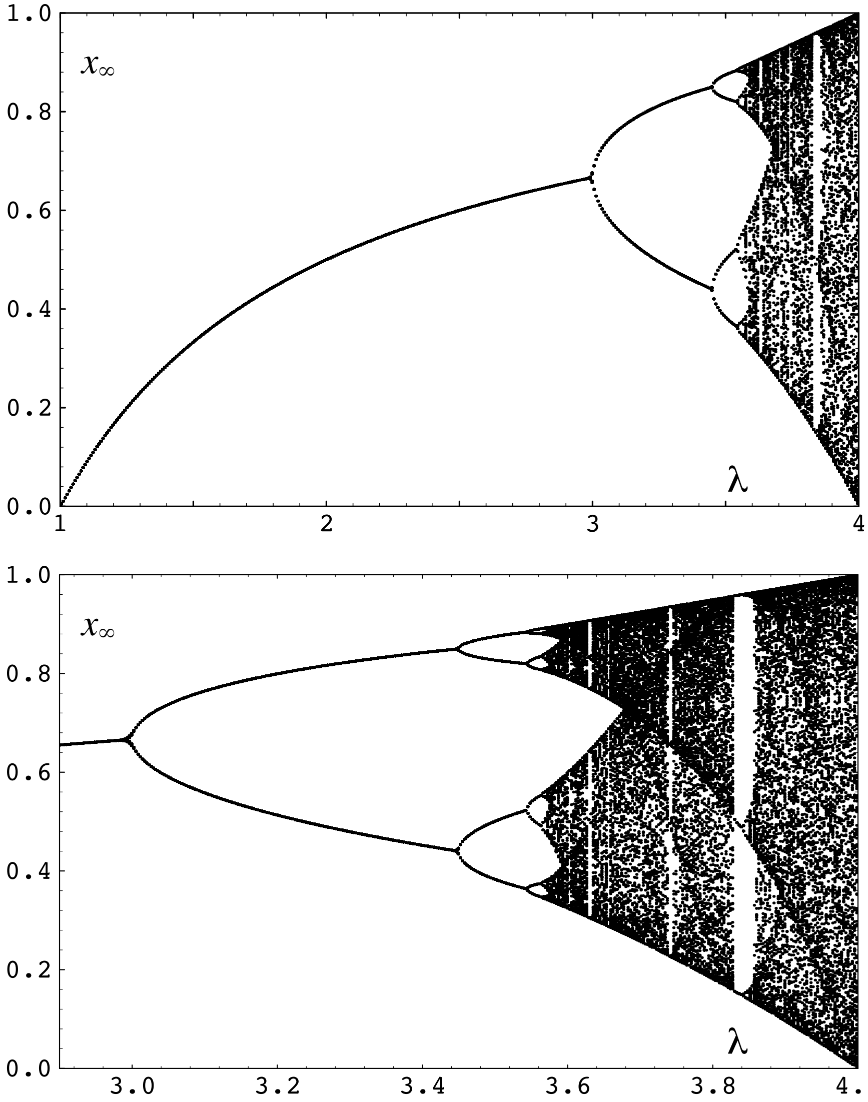


Figure 1.2.4: Feigenbaum diagram showing limiting values x_∞ as a function of λ for the logistic map.

Yet more can be said. Figure 2.5 shows an enlargement of the bifurcation cascade for the logistic map. Suppose d is the distance between two forks just as they themselves are about to bifurcate ($d = 0.409\cdots$ for the first fork in the logistic map, see Exercise 2.2). Then (to ever better approximation the farther one proceeds in the cascade), the distances between the next two forks when they are about to bifurcate are d/α and d/α^2 where

$$\alpha = 2.5029\,0787\,5 \cdots . \tag{1.2.17}$$

Moreover, there is an explicit splitting rule for determining which distance will be d/α and which will be d/α^2. For example, consider the *upper* fork after the first bifurcation, and let d^U be the distance between the two new forks produced when this fork bifurcates. Then, see Figure 2.5, one has the relation $d^U \approx d/\alpha^2$. Similarly, let d^L be the corresponding distance when the *lower* fork bifurcates. Then one has the relation $d^L \approx d/\alpha$. Next, let d^{UL} be the distance for the lower fork of the preceding upper fork. Then one has the relation $d^{UL} \approx d^U/\alpha$, etc. Again consult Figure 2.5.

The splitting rule and the scaling factor α are also universal for all one-hump maps (with negative Schwarzian derivative), and α is sometimes called the second Feigenbaum constant.

How does this universality arise? Feigenbaum found an explanation that involves a study of certain maps acting on *function* space. The explanation is deep, and we will only be able to sketch part of it. Inspired by the observation of scaling, let $\mathcal{R}$ be a map that acts on functions $\psi(x)$ according to the rule

$$\mathcal{R} : \psi \to \bar{\psi} \tag{1.2.18}$$

with

$$\bar{\psi}(x) = -a\psi(\psi(-x/a)). \tag{1.2.19}$$

In words, $\mathcal{R}$ scales the argument x, lets ψ act twice on this scaled argument, and then rescales the result. Operations of this kind occur elsewhere in physics, and are called *renormalization*. It can be shown that the map $\mathcal{R}$ has a "fixed point" in the space of *analytic* functions *if and only if* a has the Feigenbaum scaling value α,

$$a = \alpha, \tag{1.2.20}$$

and this fixed point (function) is unique up to a normalization. Specifically, for $a = \alpha$, there is a unique analytic function $g(x)$ such that

$$g(x) = -ag(g(-x/a)) \tag{1.2.21}$$

provided g is normalized so that

$$g(0) = 1; \tag{1.2.22}$$

and there is no analytic function satisfying (2.21) for $a \neq \alpha$. Indeed, it can be shown that g has a convergent Taylor expansion of the form

$$g(x) = 1-(1.5276329\cdots)x^2+(0.1048151\cdots)x^4+(0.0267056\cdots)x^6-(0.0035274\cdots)x^8+\cdots . \tag{1.2.23}$$

We have been informed that the second Feigenbaum constant α is a property of $\mathcal{R}$. We will next learn that the first Feigenbaum constant δ is also a property of $\mathcal{R}$. Let $\mathcal{L}$ be the linear part of $\mathcal{R}$ about the fixed point g. It is defined by the relation

$$\mathcal{R}[g(x) + \epsilon h(x)] = g(x) + \epsilon\mathcal{L}[h(x)] + O(\epsilon^2) \tag{1.2.24}$$

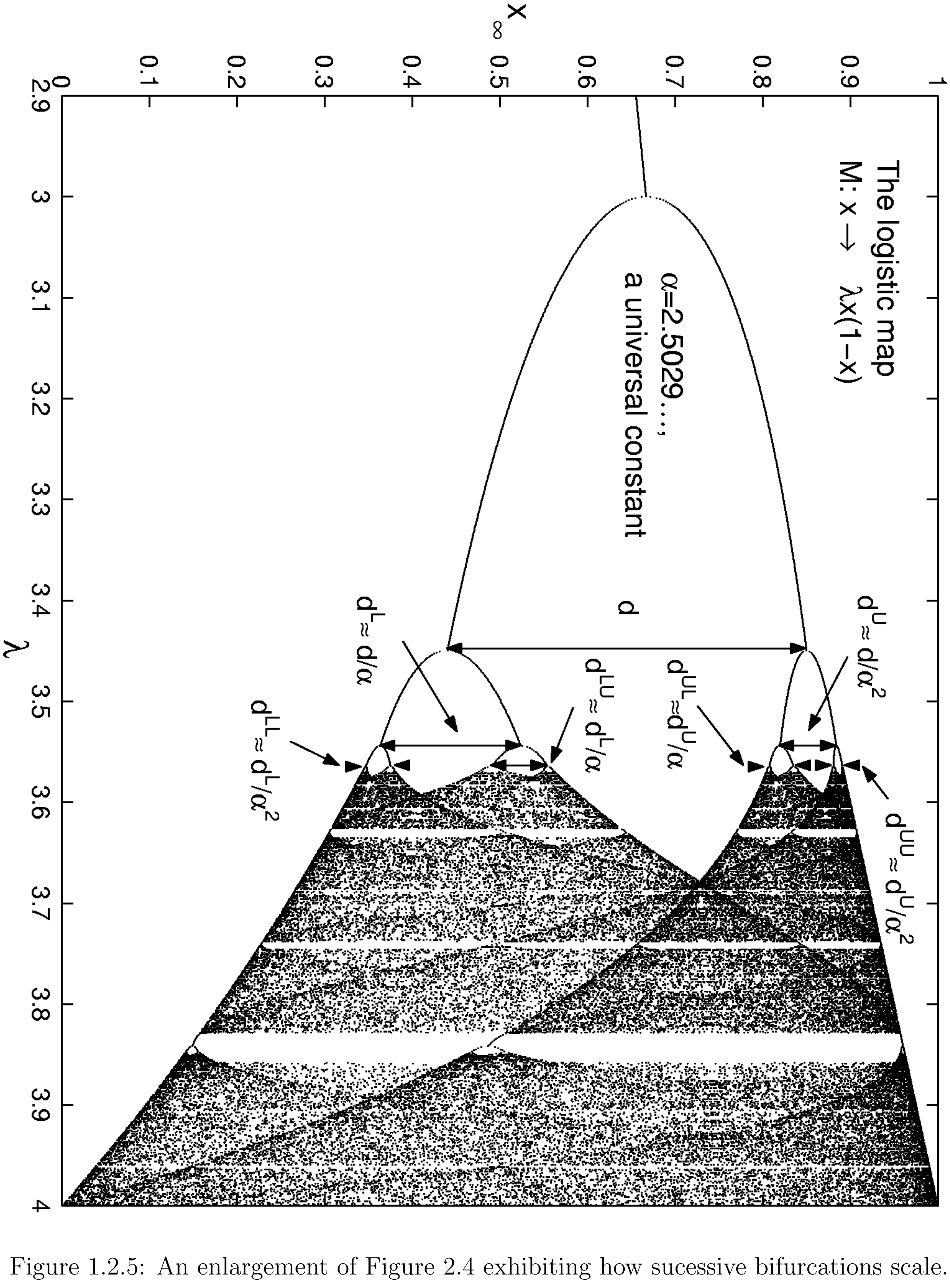


Figure 1.2.5: An enlargement of Figure 2.4 exhibiting how sucessive bifurcations scale.

for ϵ small and h any function. It follows from (2.19) and (2.24) that $\mathcal{L}$ is given explicitly by the relation

$$\mathcal{L}[h(x)] = -\alpha h(g(-x/\alpha)) - \alpha[g'(g(-x/\alpha))]h(-x/\alpha). \tag{1.2.25}$$

It can be shown that $\mathcal{L}$, which evidently and as expected is a linear operator, has eigenfunctions and eigenvalues. Moreover, there is an eigenfunction, call it h_δ, that has the Feigenbaum constant δ as its eigenvalue,

$$\mathcal{L}h_\delta = \delta h_\delta. \tag{1.2.26}$$

All other eigenvalues of $\mathcal{L}$ (there are an infinite number of them) lie inside the unit circle of the complex plane. Thus, $\mathcal{L}$ has a unique eigenvalue that lies outside the unit circle, and this eigenvalue is δ. (Note that $\delta > 1$.) Put another way, in language that will become clearer later, $\mathcal{L}$ has *one* repelling "direction" (eigenfunction) in function space associated with the eigenvalue δ and *all* other directions are attracting.

We have learned that both α and δ are properties of $\mathcal{R}$, and have told the part of the story that is easy to relate, if not to prove. What remains to be shown is that there is a connection between the set of maps that exhibit infinite period doubling cascades as some parameter is varied and the operator $\mathcal{R}$. For example, if $f(\lambda, x)$ is a function that produces any such map by the rule

$$\bar{x} = f(\lambda, x), \tag{1.2.27}$$

and the parameter λ has the critical value λ_∞ for which an infinite period doubling cascade has just occurred, then it can be shown that (with $a = \alpha$)

$$\lim_{n\to\infty} \mathcal{R}^n[f(\lambda_\infty, x)] = g(x). \tag{1.2.28}$$

For the whole story, the reader is referred to the references at the end of this chapter.

Let us, having made this pleasant detour through function space, return to a further discussion of the logistic map. We have sketched the behavior of $\mathcal{M}$ as λ approaches $\lambda_{\rm cr}$. For λ slightly beyond $\lambda_{\rm cr}$ the set of x_∞ points is infinite, and the action of $\mathcal{M}$ on these points is chaotic. Then, remarkably, as λ is increased still further, there are occasional *windows of stability* again followed by period doublings and subsequent chaotic regimes. For example, there is a period-three window (a regime having three values for x_∞) beginning at $\lambda = 1 + \sqrt{8} = 3.828\cdots$. Note that, by construction, only *stable* periodic orbits are displayed in Figures 2.4 and 2.5. Thus, as mentioned earlier, the x_e given by (2.9) no longer is shown for $\lambda > 3$. It can be demonstrated that, while there are only a finite number of stable periodic orbits in the windows of stability (as Figures 2.4 and 2.5 indicate), there are an infinite number of unstable periodic orbits. (By the way, all this behavior is also universal.)

1.2.2 Complex Logistic Map and the Mandelbrot Set

According to Paul Painlevé (1863-1933) and popularized by Jacques Hadamard (1865-1963),

> The shortest path between two truths in the real domain passes though the complex domain.

In a similar vein, Gaston Julia (1893-1978) frequently instructed the students in his class, one of whom was Benoit Mandelbrot (1924-2010),

> To simplify, you should 'complexify'. That is, when you have a complicated problem and wish to simplify it, it is a good idea to replace all reals by complex numbers.

For example, the behavior of power series is understood more simply using complex variables rather than real variables.

With this lesson in mind, and following Mandelbrot, suppose we extend both x and λ in (2.5) to complex values. Then the map $\mathcal{M}$ takes the form

$$z_{n+1} = \mathcal{M} z_n = f(\gamma, z_n) = \gamma z_n(1 - z_n) \tag{1.2.29}$$

where z is the complex extension of x, and γ is the complex extension of λ. (See Exercise 2.5.) Associated with the map (2.29) are two complex planes. One of these, the z plane, will be called the *mapping* plane since the map sends this plane into itself. The other, the γ plane, will be called the *control* plane.

The nature of what happens in the mapping plane under repeated iteration depends sensitively on where γ is in the control plane. For example, Figure 2.6 shows the nature of the map for $\gamma = 2.55268 - 0.959456i$. Points in the black area of the mapping plane remain there indefinitely under repeated application of (2.29). By contrast, any point launched in the white area eventually iterates away to infinity. (We may view the point $z = \infty$ as an attractor for $\mathcal{M}$. See Exercise 2.6.) In Accelerator Physics language, we would call the black area the *dynamic aperture*. (Mathematicians call it the *filled Julia* set.[17]) It can be shown that the boundary of the dynamic aperture (the Julia set) is fractal. Remarkably, it is nevertheless possible to name in a precise way every point on the boundary.

If γ is changed, the dynamic aperture also is changed. Figure 2.6 shows what is called *Douady's* rabbit; for some other values of γ the dynamic aperture disintegrates into a cloud of isolated points called *Fatou* dust.[18] Since the nature of what happens under repeated iteration in the mapping plane depends sensitively on the location of γ in the control plane, we may turn the matter around. That is, we may characterize points in the γ plane by the behavior (under repeated iteration) of points in the mapping plane. Suppose we consider those points M in the control plane for which the dynamic aperture in the mapping plane is a *connected* set. This set M in the control plane is called the *Mandelbrot* set.[19] It is shown in Figure 2.7.

There is another definition of the Mandelbrot set that is more computationally tractable, and which can be shown to be equivalent to that just given. The function $f(\gamma, z)$ has a *critical* point (a point where $\partial f/\partial z = 0$) at $z = 1/2$. Now consider the points $\mathcal{M}^n(1/2)$. They form the orbit of (1/2) under the action of $\mathcal{M}$. If, for a particular value of γ, this orbit goes to

[17] Gaston Julia (1893-1978) and Pierre Fatou (1878-1929) began the study of complex dynamics during the early 20th century.

[18] Adrien Douady (1935-2006) made significant contributions to the fields of analytic geometry and dynamical systems.

[19] Elsewhere in this book the symbol M will commonly be used to denote the linear part of a map $\mathcal{M}$. But here it is used to honor Mandelbrot.

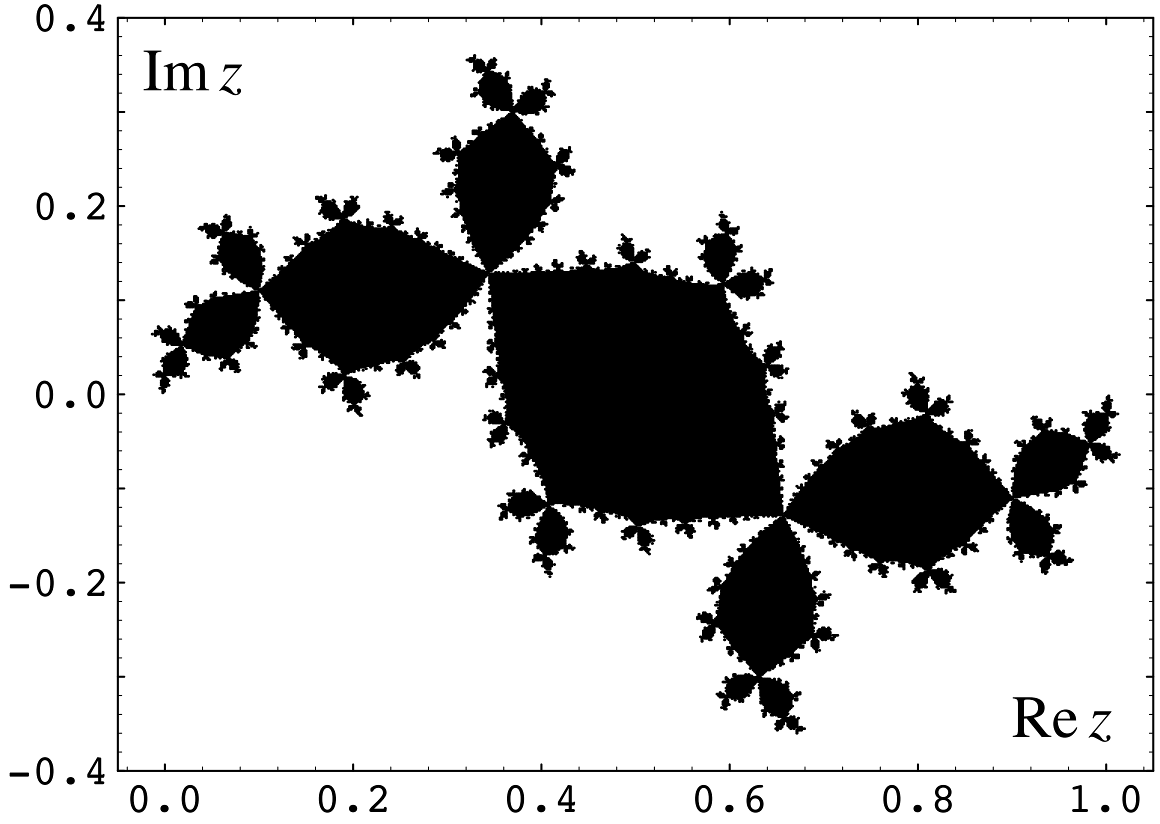


Figure 1.2.6: Douady's rabbit, the dynamic aperture in the mapping plane z for the case $\gamma = 2.55268 - 0.959456i$.

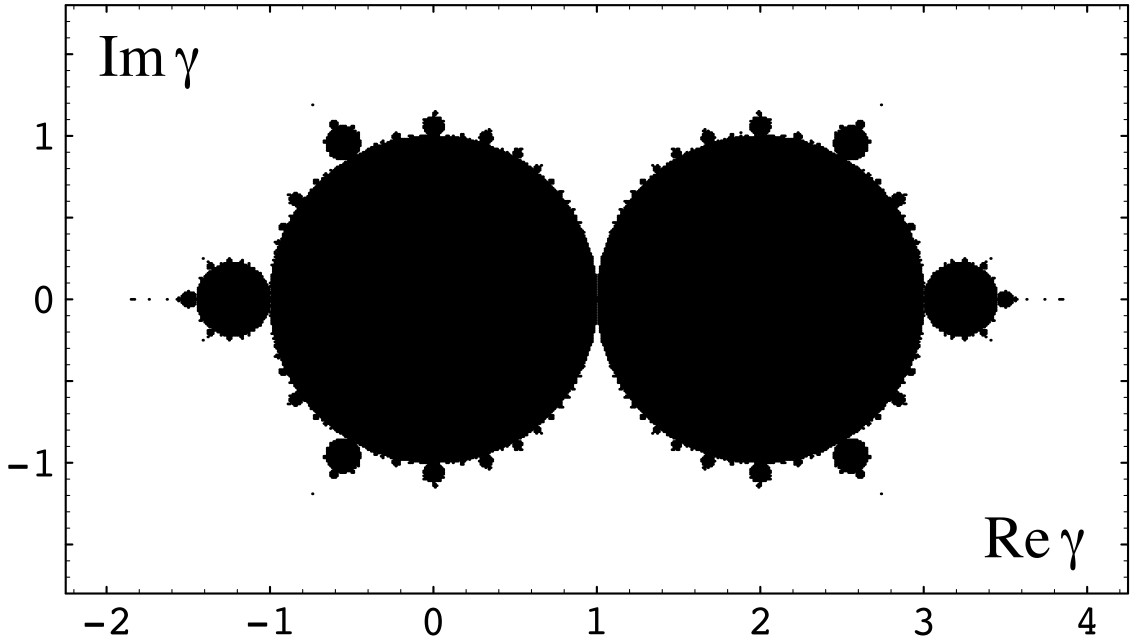


Figure 1.2.7: The Mandelbrot set M in the control plane γ.

infinity, then γ is *not* in the Mandelbrot set M. If the orbit of (1/2) does *not* go to infinity for a particular value of γ, then this value of γ *is* in the Mandelbrot set. Technically, we say that (1/2) is in the basin of attraction for the attractor $z = \infty$ if its orbit goes to infinity. Thus, γ is in the Mandelbrot set if (1/2) is not in the basin of $z = \infty$; and γ is not in the Mandelbrot set if (1/2) is in the basin of $z = \infty$. Finally, we remark that it is not necessary to follow an orbit to infinity by iterating infinitely often. See Exercise 2.6 to learn that a point z is in the basin of infinity, i.e. will go to infinity under infinite iteration, if z lies outside the disk specified by $|z| = 1 + 1/|\gamma]$. See (2.109). Therefore, if any point on the orbit of (1/2) falls outside this disk, it is not necessary to iterate further to determine the ultimate fate of points on the orbit.

When viewed from a distance, the Mandelbrot set M appears to be a mainland consisting of two back-to-back discs with sprouts. The discs are tangent at the point $\gamma = (1, 0)$, and M has reflection symmetry about both the lines $\mathrm{Re}\,\gamma = 1$ and $\mathrm{Im}\,\gamma = 0$. Closer examination reveals the presence of what appear to be very small islands around the mainland. (In fact these islands, when suitably magnified, resemble the mainland, and the whole structure of the Mandelbrot set is fractal.) Since γ is the complexification of λ, one can see that λ values in the range $(1, \lambda_{\rm cr})$ correspond to *real* γ values lying in the right disc and its sprouts and its subsprouts. In addition, it can be shown that λ values for the windows of stability seen in Figure 2.4 correspond to real γ values lying in small islands on the real γ axis to the right of the mainland. Finally, contrary to superficial appearances, it can be shown that the Mandelbrot set is *connected* (and, indeed, *simply connected*). There are thin filiments, too small to be seen in Figure 2.7, that connect the visible apparent islands to the mainland. Thus, there is really only a mainland (and this mainland has no holes)!

Consider the value of γ for Douady's rabbit. It lies in the sprout located at the five-o'clock position of the right disc in Figure 2.7. For this value of γ the complexified version of (2.8) and (2.9) yields for $\mathcal{M}$ the fixed points $z_f = 0$ and $z_f = .656747 - .129015i$. These fixed points are both repellers. Also, there is a fixed point at ∞, and it is attracting. See Exercises 2.6 and 2.11. See also Exercise 5.5 of Chapter 22 where the machinery is developed to deal with the nature of fixed points in 2-dimensional maps.

Moreover, it can be shown that for this γ value the map (2.29) has three *attracting* complex period-three fixed points. Indeed, Douady's rabbit turns out to be the basins of attraction for these fixed points. The three attracting fixed points of $\mathcal{M}^3$ have the locations

$$z^1 = 0.499997032420304 - (1.221880225696050\text{E}{-}006)i \quad \text{(red)}, \tag{1.2.30}$$

$$z^2 = 0.638169999974373 - (0.239864000011495)i \quad \text{(green)}, \tag{1.2.31}$$

$$z^3 = 0.799901291393262 - (0.107547238170383)i \quad \text{(yellow)}. \tag{1.2.32}$$

The action of $\mathcal{M}$ on these fixed points is given by the relations

$$\mathcal{M}z^1 = z^2, \tag{1.2.33}$$

$$\mathcal{M}z^2 = z^3, \tag{1.2.34}$$

$$\mathcal{M}z^3 = z^1. \tag{1.2.35}$$

Figure 2.8 shows Douady's rabbit again, this time in color. The red, green, and yellow points lie in the basins $B(z^1)$, $B(z^2)$, and $B(z^3)$ of $\mathcal{M}^3$, respectively. The white points lie in the basin $B(\infty)$ of $\mathcal{M}$. Corresponding to the relations (2.33) through (2.35) there are the results

$$\mathcal{M}B(z^1) = B(z^2) \quad \text{or} \quad \mathcal{M} \text{ red} \subseteq \text{green}, \tag{1.2.36}$$

$$\mathcal{M}B(z^2) = B(z^3) \quad \text{or} \quad \mathcal{M} \text{ green} \subseteq \text{yellow}, \tag{1.2.37}$$

$$\mathcal{M}B(z^3) = B(z^1) \quad \text{or} \quad \mathcal{M} \text{ yellow} \subseteq \text{red}. \tag{1.2.38}$$

Note the marvelous fractal structure at the basin boundaries.

In addition to the attracting fixed points of $\mathcal{M}^3$, there exist another three *repelling* complex period-three fixed points that lie on the boundary of the rabbit. Now continuously vary the value of γ until it enters the island for the period-three window, and eventually takes on a real value corresponding to a λ value lying in the period-three window of Figure 2.4. As γ varies, the period-three fixed points move. They may change their nature, (*e.g.* they all become repellers when γ leaves the sprout), but they *cannot* disappear. (See, for example, Exercise 2.2.) It can be shown that in this case, as γ changes from the Douady-rabbit value in the sprout to a real value in the period-three window, all the associated period-three fixed points of Douady's rabbit move from their original complex values to the real line. Furthermore, the three period-three fixed points that begin as repellers when γ lies in the sprout become the three attractors x_∞ when γ reaches the island. The other three period-three fixed points, which begin as attractors when γ lies in the sprout and become repellers when γ leaves the sprout, remain repellers when γ reaches the island. Thus, by extending the logistic map to the complex domain, we have learned that seemingly isolated phenomena are in fact related.

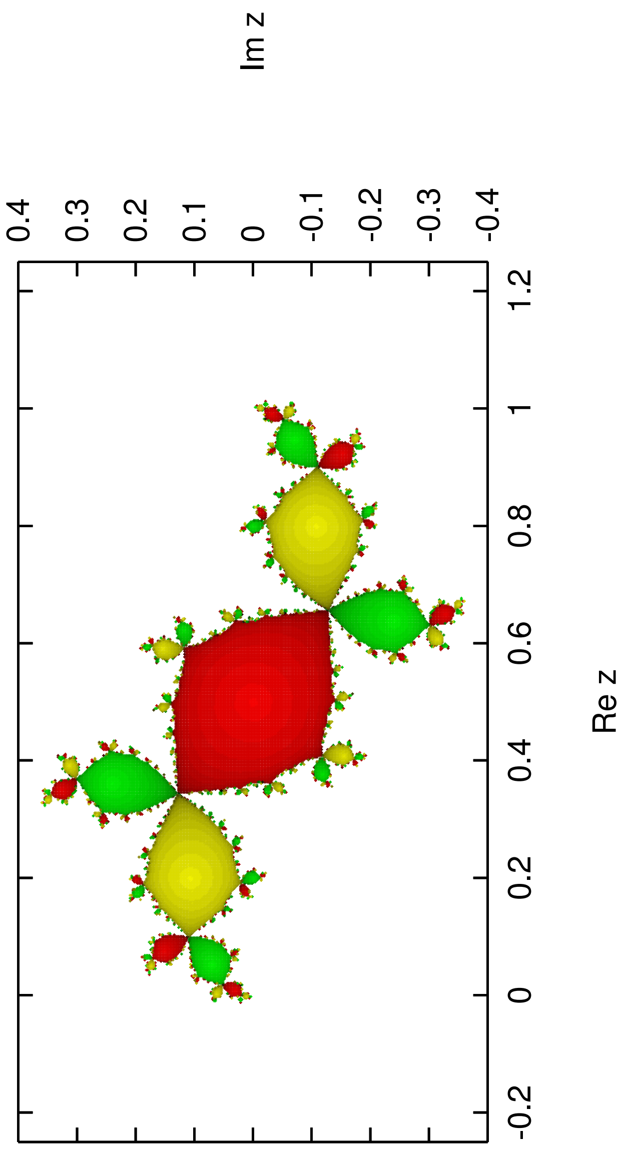


Figure 1.2.8: Douady's rabbit in color. The white points lie in the basin of ∞ under the action of $\mathcal{M}$. The origin is a repelling fixed point of $\mathcal{M}$. The other repelling fixed point has the location $z_f = .656747 - .129015i$. Under the action of $\mathcal{M}^3$, red points lie in the basin of z^1, green points lie in the basin of z^2, and yellow points lie in the basin of z^3.

1.2.3 Simplest Nonlinear Symplectic Map

The complex logistic map (2.29) may be viewed as the simplest nonlinear two-dimensional *analytic* map. In the same spirit, the simplest nonlinear two-dimensional *symplectic* map is the *Hénon* map.[20] It, too, is a quadratic map. We take the opportunity here to describe it in Lie algebraic terms. To do this, we will need to introduce some Lie algebraic tools. These tools will be described briefly below. Their full exposition is given in subsequent chapters.

We begin by redefining the symbol z; it will now stand for a canonically conjugate pair of position and momentum variables q and p,

$$z = (q, p). \tag{1.2.39}$$

Next, let $f(z)$ denote any function of q, p. We will associate with each such function a *differential* operator, called a *Lie operator* and denoted by the symbol $:f:$, by making the definition

$$:f: \overset{\text{def}}{=} (\partial f/\partial q)(\partial/\partial p) - (\partial f/\partial p)(\partial/\partial q). \tag{1.2.40}$$

Then if g is any other function of the phase-space variables z, we have the result

$$:f: g = (\partial f/\partial q)(\partial g/\partial p) - (\partial f/\partial p)(\partial g/\partial q) = [f, g], \tag{1.2.41}$$

where $[*, *]$ denotes the familiar Poisson bracket. (See Section 1.7.) Powers of $:f:$ are defined by repeated application of (2.40) or (2.41),

$$\begin{aligned} :f:^2 g &= [f, [f, g]], \\ :f:^3 g &= [f, [f, [f, g]]], \text{ etc.} \end{aligned} \tag{1.2.42}$$

Finally, we define $:f:^0$ to be the identity operator,

$$:f:^0 = \mathcal{I} \Leftrightarrow :f:^0 g = g. \tag{1.2.43}$$

Now that powers of Lie operators have been defined, we can also define power series. Of particular interest is the power series for the exponential function,

$$\exp(:f:) = \sum_{k=0}^{\infty} :f:^k /k!. \tag{1.2.44}$$

This object is referred to as a *Lie transformation*, and $:f:$ (or f) is called its *generator*. Specifically, if g is any function, we have the result/action

$$\exp(:f:)g = g + [f, g] + [f, [f, g]]/2! + \cdots . \tag{1.2.45}$$

With regard to its action on the phase-space coordinates q and p, it can be shown that any Lie transformation produces a symplectic map. See Section 7.1.

[20] Michel Hénon (1931-2013), a French mathematician and astronomer, invented this map to model a Poincaré map.

At this point the reader should verify the results

$$\exp(:q^3:)q = q, \tag{1.2.46}$$
$$\exp(:q^3:)p = p + 3q^2. \tag{1.2.47}$$

In Accelerator Physics terminology, the Lie transformation $\exp(:q^3:)$ produces the phase-space mapping associated with a *thin sextupole kick*. See Section 13.10.

Similarly, the reader should verify the results

$$\exp\bigl(-(\phi/2):p^2 + q^2:\bigr)q = q\cos\phi + p\sin\phi, \tag{1.2.48}$$
$$\exp\bigl(-(\phi/2):p^2 + q^2:\bigr)p = -q\sin\phi + p\cos\phi. \tag{1.2.49}$$

This verification requires the summation of an infinite series. In Accelerator Physics terminology, the Lie transformation $\exp[-(\phi/2):p^2 + q^2:]$ produces the phase-space mapping for a *simple phase advance* (rotation in phase space) of amount ϕ.

With this background in mind, let us consider the map $\mathcal{M}$ given by the product

$$\mathcal{M}(\theta) = \exp\bigl(-(\theta/4):p^2 + q^2:\bigr)\exp(:q^3:)\exp\bigl(-(\theta/4):p^2 + q^2:\bigr). \tag{1.2.50}$$

The map consists of a $\theta/2$ phase advance, followed by a sextupole kick, followed again by a $\theta/2$ phase advance. Figure 2.9 illustrates this map schematically. In Accelerator Physics terminology, it may be viewed as describing horizontal betatron motion in an idealized storage ring with a single thin *sextupole* insertion S, and an *observation* point O (Poincaré surface of section) located diametrically across the ring from the sextupole insertion. As seen from (2.46) through (2.49), the map (2.50) does indeed consist of linear and quadratic terms, as advertised. Since Lie transformations produce symplectic maps when acting on phase-space coordinates, and symplectic maps form a group, it follows that (2.50) is a symplectic map. Finally, it can be shown that this map is a variant of the usual Hénon map, and differs from it only by a linear change of variables. See Chapter 29 for a study of general quadratic maps in two dimensions. We also remark that, unlike the logistics map (real or complex), the Hénon map, like all symplectic maps, is invertible.

The Hénon map has been studied in detail. As simple as it appears, it is known to have very complicated properties: these include homoclinic points, chaotic behavior, and period bifurcations. Figure 2.10 shows the dynamic aperture for our variant of the Hénon map for the case $\theta/2\pi = 0.22$. Points in the black area of the q, p (mapping) plane remain there under repeated application of the map. [Actually, the points shown remain there for at least 10,000 iterations ($\mathcal{M}^n$ with $n \leq 10,000$).][21] By contrast, any point launched in the white area eventually iterates away to infinity. Inspection of the figure suggests, and it can in fact be proved, that the dynamic aperture for our variant of the Hénon map is symmetrical about the q axis.

Figure 2.11 illustrates how the size and shape of the dynamic aperture for our variant of the Hénon map depend on the total phase advance θ. As is evident from examination

[21] We remark that the dynamic aperture is not known for the Hénon map, or any other nontrivial symplectic map for that matter, when $n = \infty$. See Section 20.10.

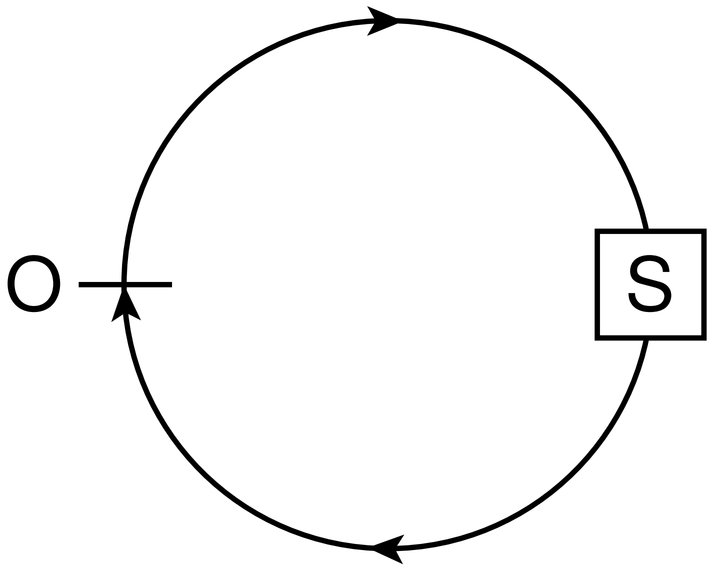


Figure 1.2.9: Schematic representation of the map (2.50).

of (2.46) through (2.50) and Figure 2.11, the dynamic aperture shrinks to the phase-space origin as θ goes to zero. By contrast, when $\theta = \pi$, one has the results

$$\mathcal{M}(\pi)q = -q + 3p^2, \tag{1.2.51}$$

$$\mathcal{M}(\pi)p = -p, \tag{1.2.52}$$

$$\mathcal{M}^2(\pi) = \mathcal{I}. \tag{1.2.53}$$

Correspondingly, the dynamic aperture in this case is *all* of phase space. For general θ it can be shown that the dynamic aperture for the map $\mathcal{M}(-\theta)$ is the same as that for the map $\mathcal{M}(\theta)$ save for a 180° rotation about the phase-space origin. Moreover, the dynamic aperture for the map $\mathcal{M}(\pi + \phi)$ is the same as that for $\mathcal{M}(\pi - \phi)$. Finally, the dynamic aperture for $\mathcal{M}(\theta)$ is periodic in θ with period 4π. It follows that the information presented in the figure is sufficient to deduce the dynamic aperture for all (real) values of θ.

The study of phenomena arising from the iteration of symplectic maps is still in its infancy, and much remains to be done in even the very simplest of cases. For example, in analogy with what has been learned in the case of the logistic map, one might wonder if further insight could be gained by complexifying the Hénon map, i.e. by making both q and p complex. Then (2.50) would become a mapping of $\mathbb{C}^2$ (the space of two complex variables) into itself. Also, the control parameter θ could be made complex. By such a study one might hope, for example, to better understand the boundary of the dynamic aperture. *Hubbard* and *Oberste-Vorth* have begun this exploration, and results to date indicate that the complex Hénon map is a remarkably complicated object. This should be a sobering thought to accelerator physicists, because they are interested in knowing the behavior of far more complicated symplectic maps in more (four and six) dimensions. When complexified, four- and six-dimensional phase spaces become $\mathbb{C}^4$ and $\mathbb{C}^6$. Thus it is no wonder that questions of dynamic aperture for realistic accelerators are so complicated. Nor, in analogy to the properties of the Mandelbrot set, should we be surprised that the

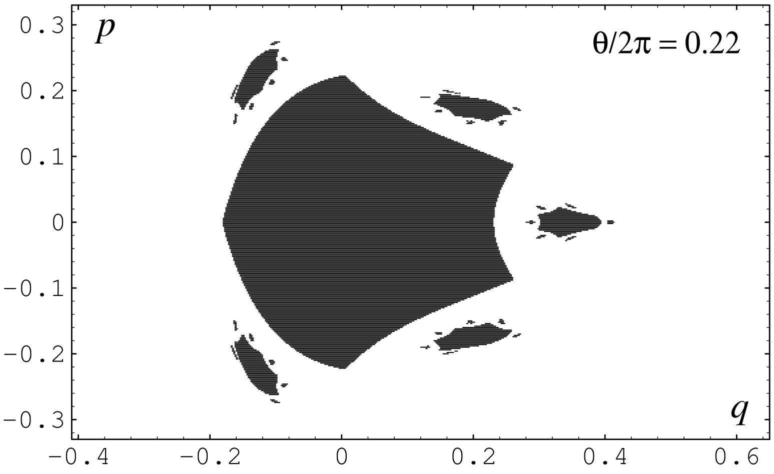


Figure 1.2.10: The dynamic aperture of the Hénon map for the case $\theta/2\pi = 0.22$.

dynamic aperture depends sensitively on the choice of accelerator parameters such as tunes, local phase advances, multipole strengths, etc. What we are observing in all these instances is that complicated properties can arise as a result of an infinite process, namely that of indefinite iteration.

1.2.4 Goal for Use of Maps in Accelerator Physics

In some areas of nonlinear dynamics, e.g. celestial/galactic dynamics, the Hamiltonian is dictated by Nature and the goal is to understand/predict the dynamics arising from this Hamiltonian. In Accelerator Physics, the Hamiltonian can, more or less, be engineered; and the goal is to engineer the Hamiltonian in such a way that particles will be accelerated, stored, and directed to achieve various desired ends. In particular, in the context of Accelerator Physics, the long-term goal of map methods is to be able to describe, predict, and control nonlinear properties with the same facility with which we now handle linear properties. Much has been accomplished in this direction, particularly with regard to single-pass systems and short-to-moderate-term behavior in circulating systems.

It is known that once-differentiable symplectic maps (and probably even analytic symplectic maps) *generically* have simultaneously hyperbolic fixed points, elliptic fixed points, and homoclinic points that are all *everywhere dense* in phase space. (The meaning of the terms *hyperbolic*, *elliptic*, and *homoclinic* will be defined subsequently.) Consequently, the detailed long-term behavior of most symplectic maps under repeated iteration must be complicated beyond comprehension.[22] However, there is still the hope that it may be possible to

[22] Thus, the properties of the Hénon map are vastly more complicated than those of the already very complicated complex logistic map. For example, apart from the behavior of the Julia set which is sent into itself in a complicated way, the behavior at most points in the mapping plane for the complex logistic map is governed by a few attractors. By contrast, we will see in Chapter 3 that symplectic maps have no attractors (and also no repellers). Therefore the orbits produced by a symplectic map never settle down, and something new should always be expected. But this "newness" would not be surprising to the vast intelligence described by Laplace. It is surprising only to those who have not done enough computation to see how results depend on the initial conditions and the number of iterations.

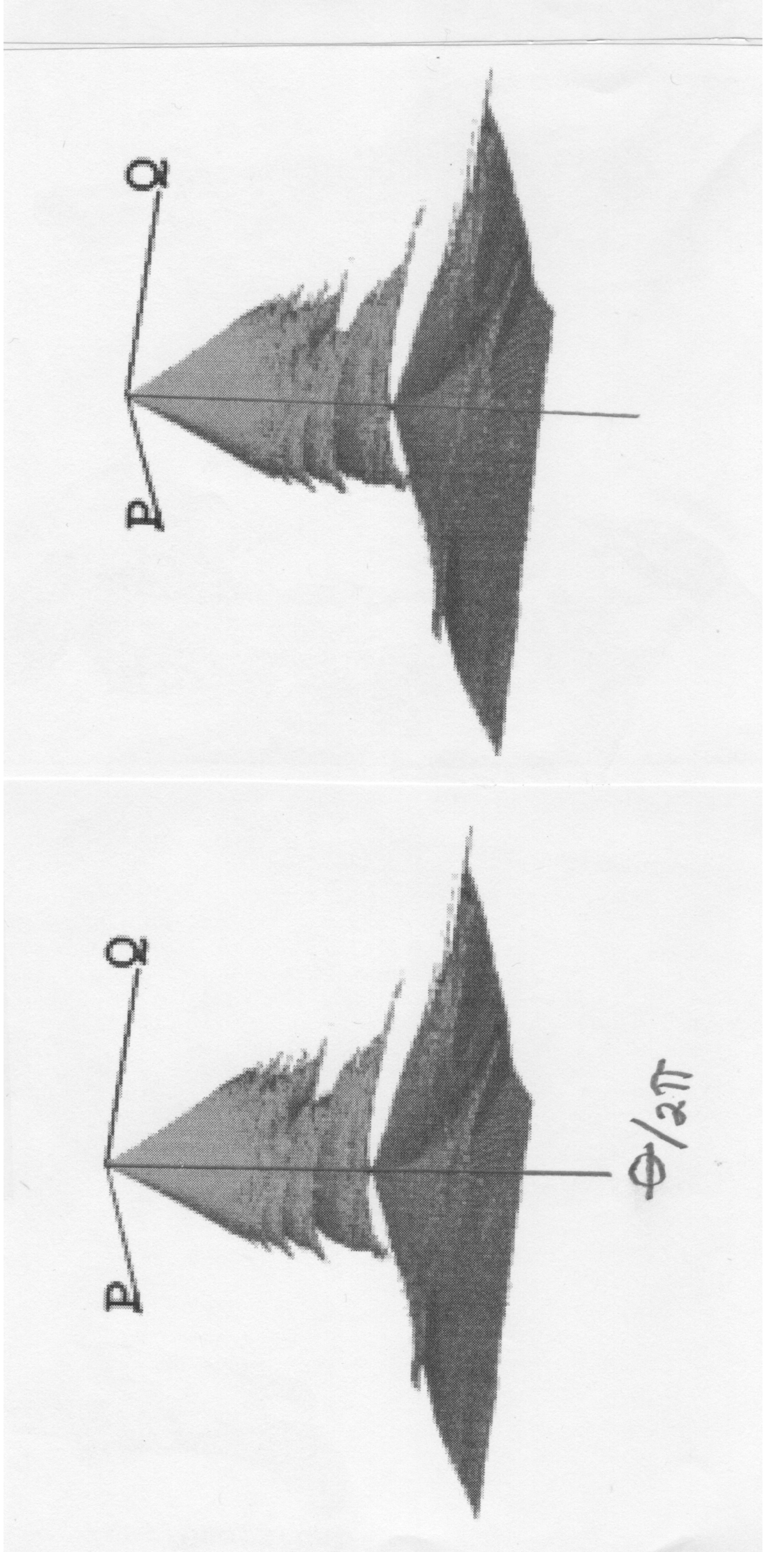


Figure 1.2.11: Stereographic view of the dynamic aperture of the Hénon map as a function of the parameter θ. The region shown is $q \in [-.8, .8]$, $p \in [-.7 .7]$, $\theta/2\pi \in [0, .5]$.

compute gross long-term properties: the rough size of the dynamic aperture, approximate (but useful) lower bounds on the life time for some sizable fraction of a circulating beam, etc.

We now know that generically Hamiltonian motion is *chaotic* in the sense that final conditions (in the long-term) generally depend very sensitively on initial conditions. (And, we know that final conditions can also depend very sensitively on parameter values.)[23] This possibility was already envisioned by Maxwell, and subsequently by Poincaré. In 1873 Maxwell wrote:

> When the state of things is such that an infinitely small variation of the present state will alter only by an infinitely small quantity the state at some future time, the condition of the system $\cdots$ is said to be stable; but when an infinitely small variation in the present state may bring about a finite difference in the state of the system in a finite time, the condition of the system is said to be unstable. It is manifest that the existence of unstable conditions renders impossible the prediction of future events, if our knowledge of the present state is only approximate, and not accurate $\cdots$ It is a metaphysical doctrine that from the same antecedents follow the same consequences. No one can gainsay this.[24] But it is not of much use in a world like this, in which the same antecedents never occur, and nothing ever happens twice.

Strictly speaking, if continuity holds as we know it does for solutions of differential equations under quite general circumstances, Maxwell was not correct in the assertion that infinitesimal changes in initial conditions could produce (in finite time) a finite change in final conditions. But his ideas were correct in spirit. In 1903, in the same spirit and with more precision, Poincaré wrote:

> If we knew exactly the laws of nature and the situation of the universe at some initial moment, we could predict exactly the situation of that same universe at a succeeding moment. But even if it were the case that the natural laws had no longer any secret for us, we could still only know the initial situation approximately. If that enabled us to predict the succeeding situation with the *same approximation*, that is all we require, and we should say that the phenomenon had been predicted, that it is governed by laws. But it is not always so: it may happen that small differences in the initial conditions produce very great ones in the final phenomena. A small error in the former will produce an enormous error in the latter. Prediction then becomes impossible, and we have the fortuitous phenomenon.

[23] The word *chaotic* can have a variety of meanings. The least stringent is sensitive dependence on initial conditions. A more stringent definition is to require in addition that for a map $\mathcal{M}$ to exhibit chaotic behavior in some domain $\mathcal{D}$ it must be *transitive* in $\mathcal{D}$ in the sense that if $\mathcal{E}$ and $\mathcal{F}$ are any two subdomains in $\mathcal{D}$, then there is some point in $\mathcal{E}$ such that applying $\mathcal{M}$ enough times to this point yields some point in in $\mathcal{F}$. Finally, we require that the set of periodic points of $\mathcal{M}$ and its powers be dense in some subdomain $\mathcal{G}$ of $\mathcal{D}$. According to Exercise 2.9 the logistic map is chaotic, following this more stringent definition, when $\lambda = 4$ and for some $\lambda < 4$.

[24] Note that quantum mechanics does gainsay this.

One of the goals of accelerator design is to minimize chaotic behavior and its effects, and to minimize sensitive dependence on parameter values.[25]

For the most part we will restrict our attention to single-particle dynamics. To the extent that multiparticle dynamics is considered, we will generally assume that interactions between individual particles can be neglected, or that we are interested only in single-particle dynamics occurring in the presence of an already specified multiparticle background. That is, we will not attempt a *self-consistent* treatment of many-particle effects such as wake fields, space-charge forces, and strong-strong beam-beam interactions. As Newton already realized, the self-consistent inclusion of even relatively *few*-particle effects raises a whole new set of complications:

> The orbit of any one planet depends on the combined motion of all the planets, not to mention the actions of all these on each other. To consider simultaneously all these causes of motion and to define these motions by exact laws allowing of convenient calculation exceeds, unless I am mistaken, the forces of the entire human intellect.

In the case of the solar system, the “forces that exceed those of the entire human intellect” have recently been provided by special-purpose super computers running special-purpose integration algorithms (based, as it turns out, on map methods). And, by following orbits for sufficiently long times, it has been found that solar-system dynamics is chaotic.[26] Routine

[25] In the context of chaotic behavior,“sensitive dependence on initial conditions” is now generally taken to mean that, to achieve a given accuracy in the final conditions after a given time or, in the case of maps, a given number of map iterations, the required accuracy in the initial conditions ultimately grows *exponentially* in time or the number of map iterations. Since parameter values may also be viewed as dynamical variables and therefore as initial conditions, see Section 10.12, the same is also possibly true of parameter values. See Exercise 2.9 for an example of how sensitive dependence on initial conditions can occur in the case of the logistic map.

[26] Under the assumption of an inverse square gravitational force law for point masses, Newton was able to show that the gravitational forces between rigid extended (macroscopic) spherical distributions of point masses (assuming they do not collide) are the same as as if they were point masses with the mass of each distribution (body) concentrated at its center. Next, Newton was able to show for two point masses that, under their mutual gravitation, their center of mass would move with constant velocity, and the motion of each about their center of mass would be an ellipse (more generally a conic section). This conclusion of Newton [about what is now called the Kepler (1571-1630) problem] had to be extracted from him by Edmond Halley (of cometary fame) after Robert Hooke (of spring-force law fame) had failed to deliver on a promised proof that an inverse square force law led to Kepler’s laws of planetary motion. When subsequently asked by Halley, Newton claimed that he had proved it four years earlier, but then, because he had apparently lost his notes, was able to produce a new and improved proof only after three months delay. Although Newton had invented the basics of calculus, his actual armamentarium of mathematical concepts and tools was quite limited by modern standards. After this, and at the urging of Halley, Christopher Wren (of architectural fame), and others, he began to work in earnest on writing his *Principia.* When completed, it was edited by Halley and published at Halley’s expense.

In the approximation that all the planets have very small masses compared to that of the sun, and with neglect of mutual interactions among the planets, the orbits of all the planets would be ellipses. Correspondingly, in this approximation, the solar system would be *stable* for *all* time. But what happens if the very small mass approximation is not made for the planets and if mutual planetary interactions are included? This so-called *gravitational N-body problem* is difficult for two reasons: First, the consideration of arbitrarily long times introduces an infinity into the problem. Second, the idealization of treating extended detailed treatment of long-term *many*-particle effects awaits the advent of readily accessible super computers routinely operating at or exceeding petaFLOPS speed.

macroscopic bodies as point masses means that bodies can become arbitrarily close with their associated gravitational potential energies possibly supplying an unbounded amount of energy to other bodies that could lead to their ejection from the system. (In fact, in some cases even relatively close encounters might provide enough energy for the ejection of others.) Thus, the singular nature of the $1/r^2$ gravitational force introduces additional possible infinities.

We remark in passing that there are exotic "solutions" to the gravitational N-body problem for which some body escapes to *infinity* in *finite* time with *infinite* velocity. Such "solutions", constructed with great ingenuity, exploit the singular nature at $r = 0$ of the $1/r^2$ idealized model for the gravitational force in that they require arbitrarily close encounters and thereby entail arbitrarily large forces. But in so doing they violate the "finiteness" conditions to be presented in Theorem 3.1. For example they do not occur if $1/r^2$ in the gravitational force law is replaced by $1/(r^2 + a^2)$ for any nonzero but arbitrarily small value of a, for then there are no infinite forces and the conditions of Theorem 3.1 are met. These exotic "solutions" are sometimes cited as evidence for instances in which *determinism* in classical mechanics is violated. This is a misunderstanding. Their true nature is that they are instances where singularities arising from idealizations are allowed to play a hidden but nonetheless decisive role. They have no deep philosophical significance. They are, however, of great mathematical interest because they clarify/prove some long-open conjectures about the nature of singular "solutions" in the gravitational N-body problem. Moreover, they have heuristic value, for they suggest that there may be nearby true solutions for which no infinities arise (forces remain bounded) but for which large (but finite) excursions may occur. Thus, for example, there are instances in which it is possible to employ relatively close encounters to achieve deep-space satellite missions with a minimum expenditure of fuel. Finally we remark that even in the two-body problem and in the case of elliptic orbits so that all body coordinates are well defined for all *real* time, the "virtual possibility" of a two-body collision (thus bringing the $1/r^2$ singularity into evidence) appears in the form of singularities in the *complex* time plane. If the orbits are highly elliptic/eccentric so that very close encounters are possible, these complex singularities lie very close to the real time axis thereby making numerical integration very difficult near times of close encounters. This problem is generally treated by *regularization* of the equations of motion prior to numerical integration.

Let us return to the main discussion. As indicated by the quotation above, Newton apparently viewed the N-body gravitational problem as humanly intractable. Nevertheless he attempted to estimate the effects of mutual interactions and concluded that they would rapidly become noticeable and detrimental to stability. Since he believed that the solar system had and should continue to exhibit regular motion for a long period of time (based on his Biblical studies, to which he devoted more time than to physics, he believed that the world would last at least until 2060), he concluded that *divine intervention/reformation* was required from time to time to correct the effect of these mutual interactions: "$\cdots$. By the help of these principles, all material things seem to have been composed of the hard and solid particles above-mentioned, variously associated in the first creation by the counsel of an intelligent agent: for it became Him who created them to set them in order. And if He did so, it is unphilosophical to seek for any other origin of the world, or to pretend that it might arise out of chaos by the mere laws of Nature; though, being once formed, it may continue by these laws for many ages. For while comets move in very eccentric orbs in all manner of positions, blind fate could never make all the planets move one and the same way in orbs concentric, some inconsiderable irregularities excepted, which may have arisen from the mutual actions of comets and planets on one another, and which will be apt to increase, till this system wants a reformation. Such a wonderful uniformity in the planetary system must be allowed the effect of choice; $\cdots$"

With regard to the solar system itself and God, Newton (in the General Scholium that appears as an appendix to the second edition of the Principia) wrote: "This most beautiful system of the sun, planets, and comets, could only proceed from the counsel and dominion of an intelligent and powerful Being. And if the fixed Stars are the centers of other like systems, these, being form'd by the like wise counsel, must be all subject to the dominion of One; especially since the light of the fixed Stars is of the same nature with the

1.2.5 Maps from Hamiltonian Differential Equations

There is one last set of motivational remarks to be made. Often, as already described and to be illustrated subsequently in Section 1.4 and later, we are interested in maps produced by integrating differential equations. In the case that these differential equations arise from a

light of the Sun, and from every system light passes into all the other systems. And lest the systems of the fixed Stars should, by their gravity, fall into each other mutually, he hath placed these Systems at immense distances from one another.$\cdots$ This Being governs all things, not as the soul of the world, but as Lord over all; and on account of his dominion he is wont to be called Lord God $\pi\alpha\nu\tau o\kappa\rho\alpha\tau\omega\rho$, or Universal Ruler $\cdots .\cdots$He is eternal and infinite, omnipotent and omniscient; that is, his duration reaches from eternity to eternity; his presence from infinity to infinity; he governs all things, and knows all things that are or can be done."

In the terminology of the philosophy of religion or natural theology, Newton's invoking divine action to "reform" (adjust) from time to time the solar system is an early example of *God of the gaps*: When something is not understood or a theory appears to fail, direct action by God is invoked as an explanation. For further discussion of Newton's Biblical and historical studies see the book of Jed Z. Buchwald and Mordechai Feingold and the book of Rob Iliffe cited in the Bibliography at the end of this chapter.

Kepler's discoveries of elliptical planetary orbits also posed unanswered questions. Like his contemporaries he initially believed, based on philosophical grounds dating back to Greek/Platonic ideas, that circular motion was the most perfect of all motions, and therefore the planets might naturally be expected to move in circular orbits. What then is the explanation for the small transverse deviations from circular motion associated with elliptical motion? Rather then invoking supernatural agents or unphysical powers, he eventually came to the physical hypothesis that both the underlying circular motions and the deviations from it arose from magnetic effects associated with a rotating sun. Of course, with Newton's discovery that the orbits associated with an inverse-square force law must be conic sections, the need for further explanation vanished, and one need not think that circular motion is the most perfect of all motions.

We also digress to note that Kepler made other scientific/mathematical contributions in addition to his laws of planetary motion. He discovered that the eye has a lens, and that the action of this lens forms an (inverted) image on the back of the eye. He also studied sphere packing, and calculated the packing fraction for a particular configuration that has since been conjectured to be optimal (the so called *Kepler conjecture*). In 2014 Thomas Hales, leader of the Flyspeck Team, announced that this conjecture was finally proved. The proof involved 300 pages of text, about 3 gigabytes of computer programs and data, and about 5000 processor-hours.

With much improved mathematical tools and a century later Laplace, in his books *Exposition du Système du Monde* and *Mécaniqué Céleste*, claimed to show that the effects of mutual interactions of the planets and the sun essentially average to zero over large times, and therefore no "reformation" is required. He also studied solar system formation mechanisms for which the planets would be expected to orbit in essentially the same plane and in the same direction. Laplace's claims might actually have pleased Newton because Newton also maintained that, "No more causes of natural things should be admitted than are both true, and sufficient to explain their phenomena."

There is a story, perhaps apocryphal/embellished, to the effect that Napoleon met Laplace and said, "I understand you have written a large book on the system of the universe and have not mentioned its creator." To this comment Laplace replied, "I had no need of that hypothesis." Napoleon, greatly amused by this response, later related this interchange to Lagrange. Lagrange reportedly replied, "Ah, it is a beautiful hypothesis; it explains many things." Subsequent versions of the Laplace-Napoleon event claim that Laplace was not denying the existence of God or his ability to intervene should he so desire, but only denying that it was necessary for God to intervene from time to time to set the planets back on a regular course. [In Exposition du Système du Monde, Laplace quotes Newton's assertion that "This most beautiful system of the sun, planets, and comets, could only proceed from the counsel and dominion of an intelligent and powerful Being." This, says Laplace, is a "thought in which he (Newton) would be even more confirmed, if he had known what we have shown, namely that the conditions of the arrangement of the planets and their satellites are precisely those which ensure its stability". Laplace originally trained for the priesthood before taking up mathematics, and received last rites at his death. But there are also indications that Laplace was *time-independent* Hamiltonian H, the associated map $\mathcal{M}(t^f, t^i)$ that takes initial conditions q^i, p^i at time t^i to final conditions q^f, p^f at time t^f can formally be written as the Lie transformation

$$\mathcal{M}(t^i, t^f) = \exp\big(-(t^f - t^i) :H(q^i, p^i):\big). \tag{1.2.54}$$

This result is proved in Section 7.4. How to capitalize on this result, and what to do in the time-dependent case, are discussed in subsequent sections. There are related results for non-Hamiltonian differential equations. One can then work with exponentials of what are called non-Hamiltonian vector fields.

Exercises

1.2.1. The purpose of this exercise is to examine the stability of the fixed points x_e given by (2.8) and (2.9). Re-express the logistic map (2.5) by using the notation

$$\bar{x} = \mathcal{M}x = f(\lambda, x) = \lambda x(1 - x). \tag{1.2.55}$$

very skeptical about the occurrence of miracles in general and transubstantiation in particular.] In effect, there was no gap that needed special filling.

Leibniz had thought, from the beginning and on philosophical grounds, that Newton's view was ill conceived because surely God could and therefore necessarily would create a universe that did not constantly require maintenance. In fact, Leibniz held, this world (universe) is the *best* of all *possible* worlds: "In whatever manner God created the world, it would always have been regular and in a certain general order. God, however, has chosen the most perfect, that is to say, the one which is at the same time the simplest in hypothesis and the richest in phenomena."

It is now known that Laplace's stability calculations are inconclusive for long-term stability (although presumably satisfactory to show stability through the year 2060) because in his perturbative method he neglected some important high-order terms. Moreover, he did not consider the possibility, now known to be generically likely, that the perturbative series he was generating would ultimately be divergent and therefore useless for determining stability.

Early in his career Poincaré also crossed swords with the N-body gravitational problem in the form of determining stability in the restricted 3-body approximation. His work won the King Oscar II of Sweden prize. But when it came time for publication a year later, Poincaré found he had made a major error, stopped the presses, paid for the printing costs himself, and wrote a corrected manuscript that was published yet a year later. See the book by June Barrow-Green cited in the Bibliography at the end of this chapter. The question of 3-body stability and solar-system stability remained unresolved.

It is now believed possible that one or more planetary ejections from the solar system may have indeed occurred in the distant past. (Numerical and analytical studies of the gravitational N-body problem indicate that there are indeed solutions for which one or more bodies escape to infinity. Moreover, numerical simulations of stellar globular clusters indicate that they routinely "boil off" individual stars.) Thus, in its early history, the solar system may have been unstable. However, long-term numerical integrations indicate that the solar system we now observe should survive far into the future. (It takes approximately 50 million years, when integrating forward or backward in time, for the uncertainties in orbital positions to grow to substantial values due to chaotic sensitivity to initial condition and parameter value uncertainties.) Of course, such calculations do not rule out collisions with small unknown objects such as asteroids. But, as locally damaging as such collisions might be to various planets and moons, they would not seriously perturb the solar system as a whole. Google *solar system stability* or *stability of the solar system* and look for, among others, the Web sites http://www.scholarpedia.org/article/Stability_of_the_solar_system and https://www.ias.edu/about/publications/ias-letter/articles/2011-summer/solar-system-tremaine. See also the book of Dumas on The KAM Story cited in the Classical/Celestial $\cdots$ section of the bibliography at the end of this chapter.

Introduce *deviation* variables δ and $\bar{\delta}$ about the fixed point x_e by the relations

$$x = x_e + \delta, \quad \bar{x} = x_e + \bar{\delta}. \tag{1.2.56}$$

Show that in terms of these deviation variables the logistic map (2.44) takes the form

$$\bar{\delta} = \mu\delta - \lambda\delta^2 \tag{1.2.57}$$

where

$$\mu = \lambda(1 - 2x_e). \tag{1.2.58}$$

The first term on the right side of (2.57) is called the *linear* part of $\mathcal{M}$ about x_e, and μ is called the eigenvalue of the linear part. Evidently, unless $\mu = 0$, the behavior of (2.57) under repeated iteration, and for δ sufficiently small, is governed by the linear part, which in turn is described by μ. That is, we may neglect the δ^2 term in (2.57). Show that if $|\mu| < 1$, then x_e is stable; and if $|\mu| > 1$, then x_e is unstable. In particular, suppose (2.57) is rewritten in the form

$$\delta_{n+1} = \mu\delta_n - \lambda(\delta_n)^2 \tag{1.2.59}$$

and assume $|\mu| < 1$ but $\mu \neq 0$. Show that, for sufficiently small δ_0, (2.59) yields the asymptotic behavior

$$\delta_n \simeq \mu^n \delta_0. \tag{1.2.60}$$

Show that if x_e is given by (2.8), then μ is given by the relation

$$\mu = \lambda. \tag{1.2.61}$$

Show that if x_e is given by (2.9), then μ is given by the relation

$$\mu = 2 - \lambda. \tag{1.2.62}$$

For $\lambda \in (0, 1)$, verify that the fixed point given by (2.8) is stable, and that given by (2.9) is unstable. Show that their stability roles are reversed for $\lambda \in (1, 3)$. Show that when $\lambda = 1$, $\mu = 1$ for both values of x_e, and show that the two fixed points then also coincide. Show that the x_e given by (2.9) is especially attractive when $\lambda = 2$. You will have to retain the δ^2 terms in (2.57) because now $\mu = 0$. In particular, show that (2.59) now yields the asymptotic behavior

$$\delta_n \simeq -(1/\lambda)(-\lambda\delta_0)^{2^n}. \tag{1.2.63}$$

When $\mu = 0$, the associated fixed point x_e is called *super attractive* or *super stable*. For $\lambda > 2$ show that μ as given by (2.62) is negative, $\mu < 0$. Use this fact to explain the behavior of the x_m in Figure 2.2. Show that μ as given by (2.62) has the value $\mu = -1$ when $\lambda = 3$, and that the fixed point given by (2.9) is unstable for $\lambda > 3$. Verify from Figures 2.4 and 2.5 that period doubling occurs when $\lambda = 3$. See also Exercise 2.2. That is, period doubling for a fixed point occurs when the associated value of μ passes through the value $\mu = -1$.

1.2.2. For $\lambda \geq 3$ the maps $\mathcal{M}$ and hence $\mathcal{M}^2$ continue to have the x_e given by (2.8) and (2.9) as fixed points. Show that, for $\lambda > 3$, $\mathcal{M}^2$ also has the two additional fixed points ${}^2x_e^{\pm}$ given by

$${}^2x_e^{\pm} = [(\lambda + 1)/(2\lambda)] \pm [(\lambda - 3)(\lambda + 1)]^{1/2}/(2\lambda), \tag{1.2.64}$$

and that these points are mapped into each other under the action of $\mathcal{M}$. (In point of fact, $\mathcal{M}^2$ also has these fixed points for $\lambda \leq 3$, but then they are complex. For an analytic map fixed points cannot be created or destroyed.) Verify that x_e as given by (2.9) and ${}^2x_e^{\pm}$ agree when $\lambda = 3$. Verify also that

$$\partial({}^2x_e^{\pm})/\partial\lambda = \pm\infty \text{ at } \lambda = 3. \tag{1.2.65}$$

Thus the curves ${}^2x_e^{\pm}(\lambda)$ have infinite slope at $\lambda = 3$. See Figure 2.5. Finally, verify that

$$d = ({}^2x_e^{+} - {}^2x_e^{-})|_{\lambda=3.449\cdots} = 0.409\cdots . \tag{1.2.66}$$

Again see Figure 2.5.

1.2.3. It has already been mentioned, and in Section 1.4 we will see in more detail, that differential equations produce maps. Moreover, in Chapter 10 and Section 24.12 we will learn how to compute these maps, how to find their fixed points, and how to expand them in deviation variables (see Exercise 2.1). Suppose a map has been expanded up to some order in deviation variables about a fixed point. Can this expansion be used to predict period doubling and other bifurcation phenomena? If so, to what order must the map be expanded to make such predictions? The purpose of this exercise is to explore these questions for the simplest case of one-dimensional maps.

Let $\mathcal{M}$ be a one-dimensional map and suppose [in analogy to (2.57)] that it has an expansion, in deviation variables about a fixed point, of the form

$$\bar{\delta} = a\delta + b\delta^2 + c\delta^3 + d\delta^4 + e\delta^5 + \cdots . \tag{1.2.67}$$

Suppose we employ the notation

$$\bar{\delta} = \mathcal{M}\delta \tag{1.2.68}$$

and

$$\bar{\bar{\delta}} = \mathcal{M}\bar{\delta}. \tag{1.2.69}$$

Show that $\mathcal{M}^2$, the square of $\mathcal{M}$, then has an expansion about the same fixed point of the form

$$\bar{\bar{\delta}} = \alpha\delta + \beta\delta^2 + \gamma\delta^3 + \sigma\delta^4 + \tau\delta^5 + \cdots , \tag{1.2.70}$$

where

$$\alpha = a^2, \tag{1.2.71}$$

$$\beta = ab + a^2b = ab(1 + a), \tag{1.2.72}$$

$$\gamma = 2ab^2 + ac + a^3c, \tag{1.2.73}$$

$$\sigma = b^3 + 2abc + 3a^2bc + ad + a^4d, \tag{1.2.74}$$

$$\tau = 2b^2c + 3ab^2c + 3a^2c^2 + 2abd + 4a^3bd + ae + a^5e. \tag{1.2.75}$$

Evaluate α, β, γ, σ, and τ for the logistic map, see (2.57), and show that in this case the terms beyond order 4 in (2.70) vanish. Now let δ_e be a fixed point of $\mathcal{M}^2$. According to (2.70) it must satisfy the equation

$$\delta_e = \alpha\delta_e + \beta\delta_e^2 + \gamma\delta_e^3 + \sigma\delta_e^4 + \tau\delta_e^5 + \cdots . \tag{1.2.76}$$

One solution to (2.76), which we already know about because it is also a fixed point of $\mathcal{M}$, is $\delta_e = 0$. Upon dividing both sides of (2.76) by δ_e, we see that any *nonvanishing* solution must satisfy the relation

$$1 - \alpha = \beta\delta_e + \gamma\delta_e^2 + \sigma\delta_e^3 + \tau\delta_e^4 + \cdots . \tag{1.2.77}$$

Show for the logistic map that the terms beyond order 3 in (2.77) vanish. Show in fact that, for the logistic map, (2.77) becomes the relation

$$Q(\delta_e) \overset{\text{def}}{=} \delta_e^3 - (2\mu/\lambda)\delta_e^2 + [\mu(\mu+1)/(\lambda^2)]\delta_e - [(\mu^2-1)/(\lambda^3)] = 0. \tag{1.2.78}$$

For the logistic map we also know from (2.57) that

$$\delta_e = (\mu - 1)/\lambda \tag{1.2.79}$$

is a second fixed point of $\mathcal{M}$. Verify this assertion. The quantity δ_e given by (2.79) is therefore also a fixed point of $\mathcal{M}^2$, and consequently is also a solution of (2.78). Indeed, verify that

$$P(\delta_e) \overset{\text{def}}{=} Q(\delta_e)/[\delta_e - (\mu-1)/\lambda] = \delta_e^2 - [(1+\mu)/\lambda]\delta_e + (1+\mu)/(\lambda^2). \tag{1.2.80}$$

Solve the equation $P(\delta_e) = 0$ and use (2.62) to find the results

$$\delta_e^{\pm} = (3-\lambda)/(2\lambda) \pm (1/2\lambda)[(\lambda-3)(\lambda+1)]^{1/2}. \tag{1.2.81}$$

Check that these results agree with (2.64).

At this point it is convenient to introduce the quantity ϵ defined by the relation

$$\epsilon = -(\mu + 1). \tag{1.2.82}$$

Evidently ϵ will be small when $\mu \simeq -1$, namely when μ is near the bifurcation value $\mu = -1$. Show that in terms of the quantity ϵ, see (2.62), the relation (2.81) has the expansion

$$\delta_e^{\pm} = \pm\,(1/3)(\epsilon)^{1/2} - (1/6)(\epsilon) \mp (5/72)(\epsilon)^{3/2} + (1/18)(\epsilon)^2 + \cdots . \tag{1.2.83}$$

For the general one-dimensional map (2.67), we do not have at our disposal a second fixed point besides the first fixed point $\delta_e = 0$. Therefore we cannot solve (2.77) directly by factorization. However, we may still proceed as follows: We see from (2.83) that for the logistic map the δ_e of interest are small when ϵ is small. We might therefore try to solve (2.77) perturbatively under the assumption that in the general case the desired δ_e are small near a bifurcation, and consequently sufficiently high powers of δ_e may be neglected. Suppose we neglect all powers of δ_e in (2.77) beyond the first. Then (2.77) has the tentative solution

$$\delta_e \overset{?}{=} (1-\alpha)/\beta = (1-a^2)/[ab(1+a)] = (1-a)/ab. \tag{1.2.84}$$

Here we have used (2.71) and (2.72). However, since the parameter a in (2.67) plays the role of μ in (2.57), near a bifurcation we expect that $a \simeq -1$. Therefore (2.84) does not produce a solution near 0, and our assumption about being able to neglect terms in (2.77) beyond the first is unjustified. The quantity $(1-\alpha=1-a^2)$ is small, which is expected and desirable, but the quantity $[\beta=ab(1+a)]$ that multiplies δ_e is also small. Therefore the product $\beta\delta_e$ is not large compared to higher powers of δ_e.

We need to make a careful expansion in small quantities. To do so, in analogy with the case of the logistic map, now define ϵ by the relation

$$\epsilon=-(a+1). \tag{1.2.85}$$

Presumably the quantities a, b, $\cdots$ in (2.67), and correspondingly the quantities α, β, $\cdots$ in (2.70), depend analytically on some common parameter, and it is the change in this parameter that causes bifurcation. Without loss of generality, we may replace this parameter with the quantity ϵ using (2.85). The quantities α, β, $\cdots$ may then be expanded in terms of ϵ to yield relations of the form

$$\alpha-1=a^2-1=2\epsilon+\epsilon^2, \tag{1.2.86}$$

$$\beta=ab(1+a)=\beta_1\epsilon+\beta_2\epsilon^2+\cdots, \tag{1.2.87}$$

$$\gamma=\gamma_0+\gamma_1\epsilon+\cdots, \tag{1.2.88}$$

$$\sigma=\sigma_0+\sigma_1\epsilon+\cdots, \tag{1.2.89}$$

$$\tau=\tau_0+\tau_1\epsilon+\cdots,\ \text{etc.} \tag{1.2.90}$$

Here we have made explicit use of (2.71) and (2.72).

Now we are ready to proceed. Write (2.77) in the form

$$\delta_e^2=(1-\alpha)/\gamma-(\beta/\gamma)\delta_e-(\sigma/\gamma)\delta_e^3-(\tau/\gamma)\delta_e^4+\cdots. \tag{1.2.91}$$

Suppose now we neglect all powers of δ_e in (2.91) beyond the second. Then (2.91) has the tentative solution

$$\delta_e^{\pm}\overset{?}{=}-\beta/(2\gamma)\pm(1/2)[(\beta/\gamma)^2-4(\alpha-1)/\gamma]^{1/2}. \tag{1.2.92}$$

Verify (2.92) and show that inserting (2.86) through (2.88) into it yields the expansion

$$\delta_e^{\pm}\overset{?}{=}\pm[2/(-\gamma_0)]^{1/2}(\epsilon)^{1/2}+[\beta_1/(-2\gamma_0)](\epsilon)+\cdots. \tag{1.2.93}$$

According to (2.93), δ_e is now of order $(\epsilon)^{1/2}$. Assuming this to be true, let us examine the orders of the various terms on the right side of (2.91): The term $(1-\alpha)/\gamma$ is of order ϵ. See (2.86) and (2.88). The term $(\beta/\gamma)\delta_e$ is of order $(\epsilon)^{3/2}$. See (2.87) and (2.88). Moreover, the term $(\sigma/\gamma)\delta_e^3$ is also of order $(\epsilon)^{3/2}$. See (2.88) and (2.89). Finally, the terms $(\tau/\gamma)\delta_e^4$ etc. are of order ϵ^2 and higher.

With these estimates in mind, we will now seek to solve (2.91) by iteration. For the zeroth iteration we will first write

$$(\delta_e^{(0)})^2=(1-\alpha)/\gamma \tag{1.2.94}$$

with the solution

$$\delta_e^0=[(1-\alpha)/\gamma]^{1/2}=\ \pm[2/(-\gamma_0)]^{1/2}(\epsilon)^{1/2}\pm(1/4)[2/(-\gamma_0)]^{1/2}[1-2(\gamma_1/\gamma_0)](\epsilon)^{3/2}+\cdots. \tag{1.2.95}$$

More simply, for our purposes, it suffices to start with the approximation

$$\delta_e^0 = \pm [2/(-\gamma_0)]^{1/2}(\epsilon)^{1/2}. \tag{1.2.96}$$

For subsequent iterations we will rewrite (2.91) in the form

$$(\delta_e^{(n+1)})^2 = (1-\alpha)/\gamma - (\beta/\gamma)\delta_e^{(n)} - (\sigma/\gamma)(\delta_e^{(n)})^3 - (\tau/\gamma)(\delta_e^{(n)})^4 + \cdots . \tag{1.2.97}$$

Verify that carrying out this iterative solution yields the expansion

$$\delta_e^{\pm} = \pm [2/(-\gamma_0)]^{1/2}(\epsilon)^{1/2} + [(\sigma_0)/(\gamma_0^2) - (\beta_1)/(2\gamma_0)]\epsilon \pm (**)(\epsilon)^{3/2} + \cdots . \tag{1.2.98}$$

As a sanity check on this procedure, verify that (2.98) yields (2.83) for the case of the logistic map.

We conclude that finding the leading behavior of $\delta_e^{\pm}$, the coefficient of $(\epsilon)^{1/2}$ in (2.98), requires a knowledge of γ_0. This knowledge in turn, according to (2.73), requires a knowledge of the quantities a through c in (2.67). We see that $\mathcal{M}$ must be known through *third* order, that is through terms of order δ^3, to find the leading bifurcation behavior. And finding subsequent terms in the expansion of $\delta_e^{\pm}$ requires knowing $\mathcal{M}$ to successively higher orders. For example, finding the order ϵ term in (2.98) requires a knowledge of σ_0, which in turn according to (2.74) requires a knowledge of the *fourth*-order coefficient d.

This is the result for the case of one-dimensional maps. Since one-dimensional maps can be parts of many-dimensional maps, we conclude that a *necessary* condition to find the leading bifurcation behavior of a many-dimensional map is also that we know its expansion in deviation variables (about a fixed point) through *third* order. We speculate that this information is also *sufficient*. See Section 24.12.

1.2.4. Assuming that (2.14) is asymptotically correct, show that δ can be determined by the limiting process

$$\lim_{j\to\infty}[(\lambda_j - \lambda_{j-1})/(\lambda_{j+1} - \lambda_j)] = \delta. \tag{1.2.99}$$

Suppose a map is reparameterized by introducing the parameter $\mu = g(\lambda)$, where g is any invertible differentiable function. Show that the $\mu_j = g(\lambda_j)$ also satisfy (2.99).

1.2.5. For the complex logistic map in the form (2.29), write

$$z = x + iy, \tag{1.2.100}$$

$$\gamma = \alpha + i\beta. \tag{1.2.101}$$

Show that in terms of these quantities the complex logistic map in the form (2.29) is equivalent to the two-dimensional real quadratic map given by the relations

$$x_{n+1} = \alpha x_n - \beta y_n - \alpha(x_n^2 - y_n^2) + 2\beta x_n y_n, \tag{1.2.102}$$

$$y_{n+1} = \beta x_n + \alpha y_n - \beta(x_n^2 - y_n^2) - 2\alpha x_n y_n. \tag{1.2.103}$$

1.2.6. Consider the transformation

$$z = 1/w \text{ or } w = 1/z, \tag{1.2.104}$$

which interchanges the origin and the point at infinity. Show that under this change of variables the logistic map (2.29) takes the form

$$w_{n+1} = -(1/\gamma)(w_n)^2/(1 - w_n). \tag{1.2.105}$$

Evidently $w = 0$ is a fixed point. Suppose that w_n is sufficiently close to the origin so that

$$|w_n| = \tau|\gamma|/(1 + |\gamma|) \text{ with } \tau < 1. \tag{1.2.106}$$

Show that then there is the inequality

$$|w_{n+1}| \leq \tau|w_n|. \tag{1.2.107}$$

Thus, $w = 0$ is an attractor, and its basin, at the very least, contains the open disk

$$|w| < |\gamma|/(1 + |\gamma|). \tag{1.2.108}$$

Show, in fact, that $w = 0$ is super attractive. See Exercise 2.1. Show that all points z that satisfy

$$|z| > 1 + 1/|\gamma| \tag{1.2.109}$$

iterate to ∞ under the action of $\mathcal{M}$ as given by (2.29). We remark that this exercise shows that the complex logistic map can better be viewed as a mapping into itself of the Riemann sphere rather than the complex plane. We also remark that the Julia set may be viewed as the *boundary* of the basin of attraction for the attractor $z = \infty$. That the Julia set is fractal is an instance of the theorem that basin boundaries are generally fractal.

1.2.7. Show that under the change of variables

$$z = -(w/\gamma) + (1/2) \tag{1.2.110}$$

and the parameter change

$$\mu = (\gamma^2/4) - (\gamma/2) = (\gamma - 1)^2/4 - (1/4), \tag{1.2.111}$$

the logistic map (2.29) takes the form

$$w_{n+1} = w_n^2 - \mu. \tag{1.2.112}$$

Show that the logistic map is two-to-one, and therefore not globally invertible. Show that it is, however, locally invertible in the neighborhood of each fixed point. Verify the symmetry claimed for the Mandelbrot set shown in Figure 2.7. Figure 2.12 shows the Mandelbrot set in the complex μ plane. Verify that μ is unchanged under the substitution $\gamma \to 2-\gamma$. Verify that (2.111) maps the two disks in Figure 2.7 into a cardioid. See Figure 2.12. Verify that the point $\gamma = 1$ in Figure 2.7 corresponds to the point $\mu = -(1/4)$ in Figure 2.12, and that this point is at the cusp of the cardioid. Verify that the points $\gamma = 2$, $\gamma = 3$, $\gamma = \lambda_{\text{cr}}$, and

$\gamma = 4$ correspond to the points $\mu = 0$, $\mu = (3/4)$, $\mu = \mu_{\rm cr} \simeq 1.40$, and $\mu = 2$. Find μ for Douady's rabbit, and describe the location of this μ value in Figure 2.12. Show that the map (2.112) has the equilibrium (fixed) points

$$w_e^{\pm} = (1/2) \pm [\mu + (1/4)]^{1/2}, \tag{1.2.113}$$

and relate these points to the x_e given by (2.8) and (2.9). Show that w_e^- is stable for μ real and in the interval $(-1/4, 3/4)$, and w_e^+ is unstable.

Figure 2.13 is the analog of Figure 2.4 for μ real and in terms of the variable w. Only the trail of w_e^-, as μ is varied, is shown because w_e^+ is unstable. However, if both were shown and according to (2.113), verify that the trails $w_e^{\pm}(\mu)$ would together comprise a parabola lying on its side and extending to the right with vertex $\mu = -(1/4)$, $w = (1/2)$. Note that since $w_e^{\pm}(\mu)$ are complex for $\mu < -(1/4)$, these fixed points do not appear in Figure 2.13. Thus, from the perspective of one living in the real world, two fixed points have appeared "out of the blue" at $\mu = -(1/4)$. There are no (real) fixed points for $\mu < -(1/4)$ and there are two, one stable and one unstable, for $\mu > -(1/4)$. This appearance of two fixed points out of nowhere is an example of a *blue sky* or *saddle-node* bifurcation. Finally, verify that this behavior is not manifest when x is used as a variable and λ is used as a parameter, see (2.5), because according to (2.111) γ and hence λ is complex when $\mu < -(1/4)$.

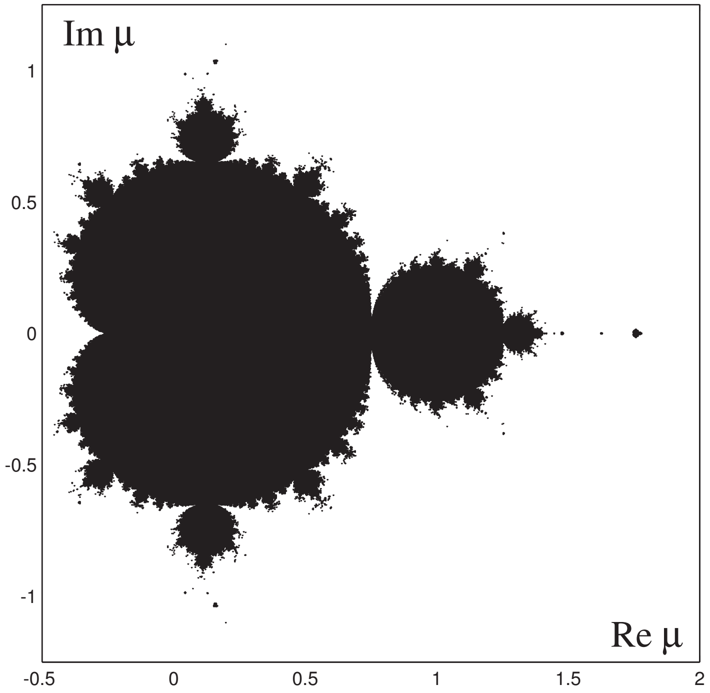


Figure 1.2.12: The Mandelbrot set in the μ plane. The "plate" has been somewhat "overexposed" compared to Figure 2.7 to bring out the island chains.

1.2.8. The general one-variable *analytic* quadratic map is of the form

$$z_{n+1} = a + bz_n + cz_n^2 \tag{1.2.114}$$

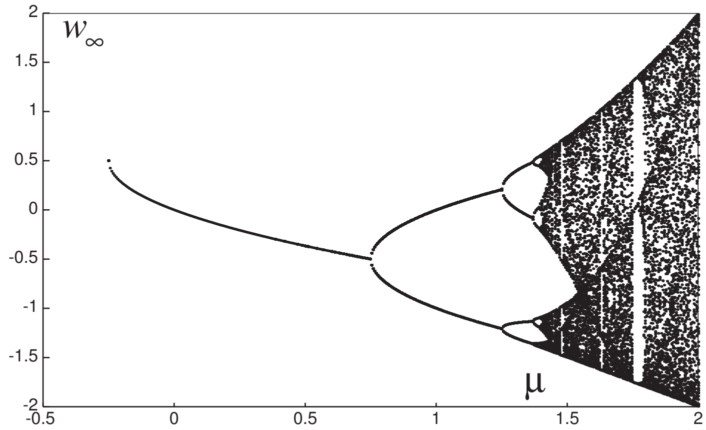


Figure 1.2.13: The analog of Figure 2.4 for μ real and the variable w.

with $c \neq 0$. Show that, under the change of variables

$$z = w/c - b/(2c), \tag{1.2.115}$$

this map takes the form (2.112) with

$$\mu = b^2/4 - b/2 - ac. \tag{1.2.116}$$

1.2.9. The behavior of the real logistic map (2.5) can be analyzed fully in the case $\lambda = 4$. This analysis also provides a simple example of *symbolic dynamics.*

Suppose x_0 is some number in the interval $[0, 1]$,

$$x_0 \in [0, 1]. \tag{1.2.117}$$

Define a related angle ϕ_0 by the rule

$$x_0 = (1/2) - (1/2)\cos\phi_0. \tag{1.2.118}$$

Show that (2.118) has a unique solution satisfying

$$\phi_0 \in [0, \pi]. \tag{1.2.119}$$

Now define a sequence $\{x_0, x_1, x_2, \cdots\}$ by the rule

$$x_n = (1/2) - (1/2)\cos(2^n\phi_0). \tag{1.2.120}$$

Show that these points satisfy the recursion relation (2.5). Define α_0 by the rule

$$\alpha_0 = \phi_0/\pi \tag{1.2.121}$$

and verify that

$$\alpha_0 \in [0, 1]. \tag{1.2.122}$$

Next define a map $\mathcal{B}$, called the *Bernoulli shift*, that acts on a sequence $\{\alpha_0, \alpha_1, \alpha_2, \cdots\}$ by the rule

$$\alpha_{n+1} = \mathcal{B}\alpha_n \overset{\text{def}}{=} 2\alpha_n \mod 2. \tag{1.2.123}$$

Show that this recursion relation, with the intial condition α_0, has the solution

$$\alpha_n = 2^n \alpha_0 \mod 2. \tag{1.2.124}$$

Verify that, because of the periodicity of the cosine function, we may rewrite (2.120) in the form

$$x_n = (1/2) - (1/2)\cos(\pi \alpha_n). \tag{1.2.125}$$

If we call the points α_n the *orbit* of α_0 under the action of the Bernoulli shift, and call the points x_n the orbit of x_0 under the action of the logistic map, then we see that the logistic orbit is the image of the Bernoulli orbit under the relation (2.125). Show, by drawing a suitable graph, that the relation (2.125) is two to one.

Suppose α_n for some n is written in *binary* form. Then we get an expression of the form

$$\alpha_n = a_1.a_2a_3a_4 \cdots \tag{1.2.126}$$

where the entries a_i are 0 or 1. For example, there are the relations

$$0 = 0.000\cdots ,$$

$$3/2 = 1.100\cdots ,$$

$$1 = 1.000\cdots = 0.11111\cdots ,$$

$$1/2 = 0.1000\cdots ,$$

$$1/4 = 0.01000\cdots ,$$

$$2/5 = 0.0110011001100\cdots . \tag{1.2.127}$$

Show that α_{n+1} then has the binary expansion

$$\alpha_{n+1} = a_2.a_3a_4 \cdots . \tag{1.2.128}$$

That is, the binary sequence for α_{n+1} is gotten by *shifting* the binary sequence for α_n one entry to the left and then discarding the first term. In the language of symbolic dynamics, the quantities 0 and 1 are called *symbols* (or letters from an alphabet if letters are used in place of digits) and the sequences (2.126) are called *words.* The Bernoulli map is an example of a dynamical operation on symbols.

By using (2.126) and (2.128) one can show that the Bernoulli map has many more or less evident properties that are reflected, in turn, in the behavior of the logistic map (when $\lambda = 4$). As a simple example, suppose α_n' is a number whose binary expansion is the same as that given for α_n in (2.127) save that the first entry, the one before the binary point, is different from a_1. Then, according to (12.128), the result of $\mathcal{B}$ acting on α_n' is the same as the result of $\mathcal{B}$ acting on α_n. We immediately see that $\mathcal{B}$, and hence $\mathcal{M}$, is two to one.

Next consider some more complicated examples. To begin, suppose α_0 has a repeating binary expansion. Then $\mathcal{B}$ acting repeatedly on α_0 produces a periodic orbit, and so will $\mathcal{M}$ acting repeatedly on x_0. Verify that, when α_0 has the value

$$\alpha_0 = .0100100100\cdots = 2/7, \tag{1.2.129}$$

the map $\mathcal{B}$ has a 3-cycle (period three orbit) consisting of the values 2/7, 4/7, 8/7. Correspondingly, the map $\mathcal{M}$ has a 3-cycle when acting on the associated x_0 given by the relation

$$x_0 = (1/2) - (1/2)\cos(2\pi/7) = .188255099\cdots . \qquad (1.2.130)$$

Verify that when α_0 has the value 2/5, the map $\mathcal{B}$ has the 4-cycle 2/5, 4/5, 8/5, 16/5 = 6/5 mod 2. See the expansion given for 2/5 in (2.127). Correspondingly, one might expect that $\mathcal{M}$ has a 4-cycle when acting on the associated $x_0 = .345491503\cdots$. However, it actually has a 2-cycle because the relation (2.125) is two to one.

Conversely, if α_0 does not have a binary expansion that eventually repeats, then the α_n will never repeat and the corresponding x_n given by (2.125) will never repeat. As a special case of this circumstance, suppose the successive a_j in the binary expansion for α_0 are determined by tossing a coin with $a_j = 1$ if the jth toss gives a head, and $a_j = 0$ if the jth toss is tails. Then we may say that α_0 is a random number, and the successive α_n and their corresponding x_n will also relect this randomness. Thus, in this sense we can say that the long-term behavior of some orbits of $\mathcal{M}$ is as random as a coin toss.

Next, suppose x^a and x^b are any two points in $[0, 1]$. Let the asociated α^a and α^b have the binary expansions $a_1.a_2a_3\cdots$ and $b_1.b_2b_3\cdots$. Define a number α^ξ by the rule

$$\alpha^\xi = a_1.a_2a_3\cdots a_N b_1b_2b_3\cdots . \qquad (1.2.131)$$

Note that in (2.131) the sequence for α^a has been truncated after N terms and the full sequence for α^b has been appended at the end. Let x^ξ be the point associated with α^ξ using (2.125). Show that x^ξ can be made arbitrarily near x^a by making N large enough. That is, study how $|x^a - x^\xi|$ goes to 0 for large N. Next show that

$$\mathcal{M}^N x^\xi = x^b. \qquad (1.2.132)$$

Thus, in any vicinity of an arbitary point x^a there are points x^ξ, and these points can be sent to any other point x^b by a sufficiently high power of $\mathcal{M}$.

This construction also illustrates that the long-term behavior of an orbit generated by $\mathcal{M}$ depends very sensitively on the initial condition x_0. Indeed, we see that to determine the effect of $\mathcal{M}^N$ on x_0 we must know at least the first few digits beyond the first N digits of the binary expansion of α_0. Thus, to achieve a given accuracy in the final condition $\mathcal{M}^N x_0$, the required accuracy in α_0, and hence also in x_0, grows exponentially in N. Verify this claim. Moreover, this construction reveals that chaotic behavior in the orbit x_n, if any, arises from random behavior, if any, in the binary expansion of α_0.

Extend the construction just given to an arbitrary sequence of points x^a, x^b, x^c, $\cdots$ and show that there are points arbitrarily near x^a which, when taken as initial conditions, have orbits that pass arbitrarily near (and in sequence) the remaining points $x^b, x^c\cdots$. You have demonstrated that there are orbits of $\mathcal{M}$ that are *ergodic*.

As one last observation, suppose α^d is the number having the binary expansion

$$\alpha^d = .\{[0][1]\}\,\{[00][01][10][11]\}\,\{[000][001][010][011][100][101][110][111]\}\,\{[\cdots . \qquad (1.2.133)$$

Here the curly and square brackets {} and [ ] are to be removed. They simply guide the eye to indicate that α^d consists first of all one-letter words, then all two-letter words, then all three-letter words, etc., with the words for each fixed length listed, when viewed as binary numbers, in ascending order.[27] Evidently, under sufficiently many Bernoulli shifts acting on α^d, it will happen that any finite string will eventually occur as the leading string in the shifted α^d. Let x^d be the point associated with α^d using (2.125). Show that the orbit of x^d under the action of $\mathcal{M}$ is *dense* on the interval $[0,1]$. That is, it comes arbitrarily close to any point in the interval. In fact, show that it does so infinitely often. Show that If one wishes to minimize (to any finite degree) the effect, on a word, of nearby words, one can separate adjacent words by strings of 0's of any desired (but finite) lengths so that α^d is of the general form (2.133) except for strings of 0's inserted between the words.

Remark: When $\lambda = 4$ you have shown that the logistic orbit is the image of the Bernoulli orbit. Let (2.125) define a map $\mathcal{T}$ so that we may write

$$x_n = \mathcal{T}\alpha_n. \tag{1.2.134}$$

Then the relation between the two orbits is equivalent to the equation

$$\mathcal{M}\mathcal{T}\alpha_n = \mathcal{T}\mathcal{B}\alpha_n \tag{1.2.135}$$

or, more abstractly,

$$\mathcal{M}\mathcal{T} = \mathcal{T}\mathcal{B}. \tag{1.2.136}$$

We say that $\mathcal{M}$ is *conjugate* to $\mathcal{B}$ under the action of $\mathcal{T}$. (See Section 19.2.) Thus, you have shown that the logistic map is conjugate to the Bernoulli map when $\lambda = 4$. The same can be proved (although with considerable more difficulty) for some λ values less than 4. Of course, when $\lambda \neq 4$, the conjugating map $\mathcal{T}$ is no longer given by (2.125).

1.2.10. Show that it follows from the fixed-point property (2.21) and the normalization condition (2.22) that

$$g(1) = -1/\alpha. \tag{1.2.137}$$

Evaluate the series (2.23) at $x = 1$ and compare your result with (2.137).

1.2.11. Verify (2.25) using (2.19) through (2.21) and (2.24).

1.2.12. This exercise studies the complex logistic map (2.29). The complexified version of (2.8) gives

$$z_f = 0 \tag{1.2.138}$$

as a fixed point of $\mathcal{M}$. Locate this point in Figure 2.8. Let z' be the point

$$z' = 1. \tag{1.2.139}$$

Locate it in Figure 2.8. Show analytically that

$$\mathcal{M}z' = z_f. \tag{1.2.140}$$

Find points z'' such that

$$\mathcal{M}z'' = z' \tag{1.2.141}$$

[27] Constructions of this kind were first made by *D. G. Champernowne.*

and hence

$$\mathcal{M}^2 z'' = z_f. \tag{1.2.142}$$

Can you find points z''' such that

$$\mathcal{M} z''' = z'', \tag{1.2.143}$$

and hence

$$\mathcal{M}^3 z''' = z_f, \text{ etc.?} \tag{1.2.144}$$

Verify that the complexified version of (2.9) gives (for Douady's γ value) the second fixed point

$$z_f = 1 - 1/\gamma = .656747 - .129015i. \tag{1.2.145}$$

To an uninformed botanist, Douady's rabbit, particularly in color, might look more like a *cactus*.[28] Again see Figure 2.8. Adopting this terminology, verify, by examining Figure 2.8, that this fixed point z_f is located at the point where the three lobes containing the period-three fixed points z^1, z^2, and z^3 meet. Define a point z' by the relation

$$z' = 1/\gamma = .343253 + .129015i. \tag{1.2.146}$$

Verify, again by examination, that three lobes also meet at this point. Show analytically that

$$\mathcal{M} z' = z_f. \tag{1.2.147}$$

Can you again find points z'', z''', etc., such that (2.141) through (2.144), etc. hold for z_f given by (2.145)?

Next consider the yellow lobe containing the point $z^{\text{in}} = .2 + .1i$. View z^{in} as an *initial* condition. Find the successive lobes that the orbit of z^{in} belongs to under successive applications of $\mathcal{M}$, and list their colors. Carry out the same exercise for the green point $z^{\text{in}} = .05 + .08i$ and the red point $z^{\text{in}} = .08 + .15i$. Suggestion: Study Exercise 2.5, and write and execute a suitable computer program.

1.2.13. Verify (2.46) and (2.47).

1.2.14. Verify (2.48) and (2.49).

1.2.15. Verify (2.51) through (2.53).

1.2.16. Show that the dynamic aperture for the map (2.50) is periodic in θ with period 4π.

[28] In fact, it is sometimes called a cactus fractal.

1.3 Essential Theorems for Differential Equations

> Among all the disciplines of mathematics, the *theory of differential equations* is the most important one. All areas of physics pose problems which lead to the integration of differential equations. In fact, it is the theory of differential equations which shows the way to understanding all time-dependent phenomena. If, on the one hand, the theory of differential equations has extreme *practical* significance, then, on the other hand, it attains a corresponding *theoretical* importance because it leads in a rational way to the study of new functions or classes of functions.
>
> *Sophus Lie (1894)*

In this book we shall be concerned primarily with processes and maps that are described by or arise from differential equations. When all is said and done, the Laws of Motion for a Newtonian Dynamical System, however formulated, reduce to a set of second-order ordinary differential equations of the form

$$\begin{aligned}\ddot{q}_1 =& h_1(q_1, q_2, \dots ; \dot{q}_1, \dot{q}_2, \dots ; t),\\ \ddot{q}_2 =& h_2(q_1, q_2, \dots ; \dot{q}_1, \dot{q}_2, \dots ; t),\\ &\text{etc.}\end{aligned} \tag{1.3.1}$$

where the quantities $q_j(t)$ refer directly or indirectly to the instantaneous coordinates of various particles, and (following William Jones' and Newton's convention) a dot above a letter denotes differentiation with respect to time.[29] Do differential equations such as (3.1) actually contain information about trajectories? If so, how much? To these questions mathematicians have given answers in the form, as is their custom, of theorems. Actually, their theorems apply to sets of first-order differential equations. But that is no problem. We can easily convert a set of n second-order equations such as (3.1) into a set of $2n$ first-order equations. We define $2n$ variables $y_j(t)$ by the rule

$$\begin{aligned} y_1(t) &= q_1(t)\\ &\vdots\\ y_n(t) &= q_n(t),\\ y_{n+1}(t) &= \dot{q}_1(t)\\ &\vdots\\ y_{2n}(t) &= \dot{q}_n(t). \end{aligned} \tag{1.3.2}$$

The equations (3.1) are then equivalent to the first-order set

$$\dot{y}_j = y_{n+j}, \quad j \ \leq \ n$$

[29] Surprisingly, nowhere in Newton's *Principia* does Newton's second law of motion appear in the familiar equation forms $F = ma$ or $a = F/m$, not to mention in the then unavailable concise vector notation form $\boldsymbol{a} = \boldsymbol{F}/m$. He writes no equation, but employs only the words "A change in motion is proportional to the motive force impressed and takes place along the straight line in which that force is impressed".

$$\dot{y}_j = h_{j-n}(y_1, \cdots y_{2n}, t) \quad n < j \leq 2n. \tag{1.3.3}$$

Alternatively, if the original equations (3.1) arose from a Lagrangian, they can also be converted into a first-order set by passing to a Hamiltonian formulation. See Sections 1.5 and 1.6.

Now hear the pronouncements of mathematicians. They provide definitive results for what is called the *Cauchy* (or initial value) problem:

Theorem 1.3.1. *Consider any set of m first-order differential equations of the form*

$$\dot{y}_j = f_j(y_1, \ldots, y_m; t), \quad j = 1, \ldots, m. \tag{1.3.4}$$

Here m may be even or odd. Assume that the right sides of (3.4), which define the set of differential equations, are sufficiently well behaved. In particular, assume that the f_j and the partial derivatives $\partial f_j/\partial y_k$ exist and are continuous in the y_k and in t within some region R of the m-dimensional space $y_1, \ldots, y_m$ and for t in some interval T about a fixed value t^0. Let $(y_1^0, \cdots y_m^0)$ be a point in R. Then there exists a unique *solution*

$$y_j(t) = g_j(y_1^0 \cdots y_m^0; t^0; t), \quad j = 1, \ldots, m \tag{1.3.5}$$

of (3.4) with the property

$$y_j(t^0) = g_j(y_1^0 \cdots y_m^0; t^0; t^0) = y_j^0, \quad j = 1, \ldots, m. \tag{1.3.6}$$

This solution is guaranteed to exist for a finite interval of time about the point t^0, and can be extended forward or backward in time as long as the f_j are continuous in the y_k and t, and the $y_j(t)$ remain within a region R' where the $\partial f_j/\partial y_k$ exist and are continuous in the y_k and t. Furthermore, the solution (3.5) is continuous *(and* bounded*) in all the variables y_j^0, t^0, and t. See Figure 3.1. The quantities y_j^0 are called* initial conditions *and t^0 is called the* initial time. *To put the matter naively, we may think of first-order differential equations as a set of "marching orders" instructing us how to move at each instant of time. Once the initial starting time t^0 and the initial starting point (the initial conditions y_j^0) for the march are specified, the whole march is completely determined.*

Theorem 1.3.2. *Suppose the f_j also depend on a set of parameters $\lambda_1 \cdots \lambda_n$. Assume that all $\partial f_j/\partial \lambda_k$ are continuous. Then the solution (3.5) will also be continuous in the parameters λ_k.*

Theorem 1.3.3. *Suppose the f_j are* analytic *in the variables y_j, λ_k, and t. (A function is analytic in some variable if it has a convergent Taylor series expansion in that variable when all other variables are held fixed. For more detail, see Sections 38.1 and 38.2.) Then the solution (3.5) will also be* analytic *in the variables y_j^0, λ_k, t^0, and t.*

The proofs of these theorems may be found in most reasonably complete books on differential equations. Including possible parameter dependence, the m differential equations to be solved are of the form

$$\dot{y}_j = f_j(y_1 \cdots y_m; \lambda_1 \cdots \lambda_n; t), \quad j = 1, \ldots, m, \tag{1.3.7}$$

and are equivalent to the integral equations,

$$y_j(t) = y_j^0 + \int_{t^0}^{t} f_j[y_1(\tau) \cdots y_m^{p-1}(\tau); \lambda_1 \cdots \lambda_n; \tau] d\tau. \tag{1.3.8}$$

(Note that these integral equations automatically incorporate the initial conditions y_j^0.) In turn, these integral equations are usually analyzed by showing that successive *Picard* iterations y_j^p of (3.8) defined by

$$y_j^p(t) = y_j^0 + \int_{t^0}^{t} f_j[y_1^{p-1}(\tau) \cdots y_m^{p-1}(\tau); \lambda_1 \cdots \lambda_n; \tau] d\tau, \ p \geq 1, \tag{1.3.9}$$

converge to g_j as $p \to \infty$, and that the limit has the stated properties.[30]

We should also mention that Theorem 3.1 can be proved under weaker conditions than the existence of various partial derivatives. For example, *Peano* proved existence under the assumption of simple continuity of the f_j in t and the y_j (however, as illustrated in the Exercises for this section, in this case there are examples for which uniqueness fails); simple continuity in t and *Lipschitz* continuity in the y_j are sufficient for both existence and uniqueness.[31] Usually, however, the results we have stated are adequate.

Next a few words about the content of the theorems themselves. Theorem 3.1, when applied to the second-order equations (3.1), says that these equations have a unique solution providing we specify the initial coordinates

$$q_j(t^0) = q_j^0$$

and the initial "velocities"

$$\dot{q}_j(t^0) = \dot{q}_j^0.$$

[Alternatively, in a Hamiltonian formulation, these equations have a unique solution providing we specify the initial coordinates q_j^0 (as before) and the initial momenta p_j^0.] Again we call these quantities, when taken together, a set of initial conditions. Thus, in general there is a unique trajectory for each set of initial conditions, and each trajectory varies continuously with the initial conditions, their time of imposition t^0, and the time t. Needless to say, this continuity is in accord with our physical intuition of motion. However, the fact that initial coordinates and velocities *alone* are enough to completely specify a trajectory, i.e. that the physical equations of motion (3.1) are of second order, is not at all obvious. Or, put another way, it is not obvious that all effects of past history are in fact subsumed in a knowledge of present positions and velocities. Rather, this fact should be regarded as one of the greater discoveries of our ancestors.

Theorem 3.2, and particularly Theorem 3.3, are often of practical computational use. First, parameters often occur either quite naturally or can be introduced into problems of physical interest. Consider the motion, for example, of the sun-earth-moon system. There the mass ratios $\lambda_1 = M_{\text{moon}}/M_{\text{sun}}$ and $\lambda_2 = M_{\text{earth}}/M_{\text{sun}}$ appear in a natural way. Their

[30] Picard was a son-in-law of Hermite.

[31] *Hadamard*, a student of Picard, defined a problem to be *well posed* if a solution exists, is unique, and depends continuously on initial conditions and parameters. Thus, the assumptions of Theorems 3.1 through 3.3 assure that the problem of computing trajectories is well posed.

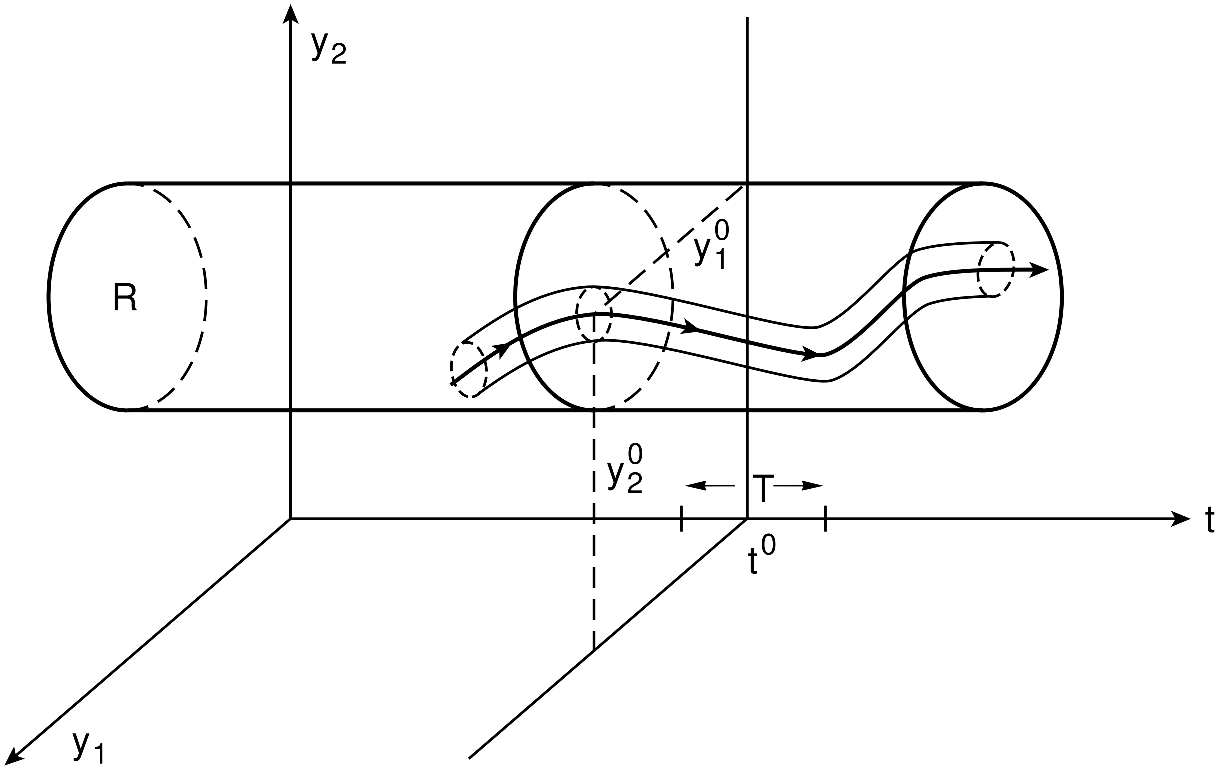


Figure 1.3.1: An illustration of Theorem 3.1 in the case that "$\boldsymbol{y}$" space is two dimensional. The solution $\boldsymbol{y}$ exists, is unique, and is continuous in t as long as it remains within the large cylinder of base R where $\boldsymbol{f}$ is continuous and the $\partial \boldsymbol{f}/\partial y_j$ are continuous. If the point $\boldsymbol{y}^0$ is varied slightly, the solution also changes only slightly so that nearby solutions form a bundle.

smallness suggests the possibility of making a power series expansion of the equations of motion in terms of λ_1 and λ_2, and then solving the resulting equations term by term. The success of such a perturbation technique is intimately related to the contents of Theorem 3.3. The use of perturbative power series was first systematically studied by Poincaré. In fact, Theorem 3.3 is often called *Poincaré's holomorphic lemma* or its results are referred to as *Poincaré analyticity*.[32] Second, as will be seen later, it is often useful to expand a solution as a power series in the initial conditions. Finally, analyticity in t, or at least the existence of several derivatives in t, is supposed in carrying out numerical integration. See Chapter 2.

We also note that the conditions for Theorem 3.3 can be relaxed. Suppose the f_j are analytic in the y_j and the λ_k, but only have n derivatives in t. Remarkably, the final conditions will still be analytic functions of the initial conditions and the parameters, and will have $n+1$ derivatives in t. If the f_j are analytic in the y_j and the λ_k, but are only continuous in t, then the final conditions will still be analytic functions of the initial conditions and the parameters, and will have first derivatives in t. If the f_j are analytic in the y_j and the λ_k, but are only piece-wise continuous in t, then the final conditions will still be analytic functions of the initial conditions and the parameters, and will be piece-wise (first) differentiable in t. Finally, as it stands, the notation (3.6) indicates that the initial conditions are assumed to be independent of any parameters. All conclusions concerning analyticity continue to hold if the initial conditions are allowed to depend on parameters providing this dependence is analytic.

As an application of these relaxed conditions, suppose the time axis is broken up into a finite number of intervals and that the f_j are analytic in the y_j, the λ_k, and at least continuous in t for each interval. Then the final conditions will be piece-wise differentiable in t and will still be analytic functions of the initial conditions and the parameters. In the context of Accelerator Physics, where some coordinate related to path length plays the role of time, this situation arises in the idealization that an accelerator is treated as a sequence of discrete beam-line elements with a separate Hamiltonian, and therefore a separate transfer map, for each element. See Subsection 2.4 and Sections 4 and 6. Each such transfer map will be analytic in the initial conditions and parameters, and their product will then also be analytic in these quantities.

Finally, we remark that Poincaré's holomorphic lemma has important applications outside of Classical Mechanics. It is used in advanced Quantum Mechanics, for example, to show that solutions to the Schrödinger equation are analytic in energy, angular momentum, and coupling constant. This analyticity is in turn used to suggest that various processes involving elementary particles at high energies obey certain integral conditions called dispersion relations.

[32] The terms *analytic* and *holomorphic* are commonly used interchangeably, particularly in the context of several complex variables. (The definitions of analytic and holomorphic are different, but can be proven to be mathematically equivalent. See Sections 38.1 and 38.2.) Poincaré derived his analyticity results on a case-by-case basis as needed using *Cauchy's* method of *majorants*.

Exercises

1.3.1. Consider the differential equation

$$t^3 \dot{y} = 2y$$

with the initial condition $y(0) = 0$. Show that it has *two* solutions: $y(t) = e^{-(1/t)^2}$, $y(0) = 0$; and $y(t) = 0$ for $t \leq 0$, $y(t) = e^{-(1/t)^2}$ for $t > 0$. Does this lack of uniqueness violate Theorem 3.1? Are there even more solutions?

1.3.2. Consider the differential equation

$$\dot{y} = -(1-y)^{1/2}$$

with the initial condition $y(0) = 1$. Show that it has the *two* solutions

$$y(t) \equiv 1 \text{ and } y(t) = 1 - t^2/4.$$

What causes this lack of uniqueness?

1.3.3. Consider the differential equation

$$\dot{y} = (1-y)^{-1}$$

with the initial condition $y(0) = 0$. Find the solution and show that it cannot be extended arbitrarily far forward in time. In view of Theorem 3.1, what went wrong?

1.3.4. Consider the growth of a crystal in a supersaturated solution. Let V be the volume of the crystal and A its surface area. We assume the growth rate is proportional to the surface area, that is,

$$\dot{V} = k_1 A$$

where k_1 is some constant. But for a regular geometric figure there is a definite relation between A and V of the form

$$A = k_2 V^{2/3}.$$

For example, $k_2 = (36\pi)^{1/3}$ for a sphere and $k_2 = 6$ for a cube. Thus, for a regular figure we have a growth law of the form

$$\dot{V} = kV^{2/3}.$$

Show that with the initial condition $V(0) = 0$, one has the *family* of solutions

$$\begin{aligned} V(t) &= 0 \quad , \quad 0 \leq t \leq \tau \\ &= [(k/3)(t-\tau)]^3 \quad , \quad t \geq \tau \end{aligned}$$

for any positive τ. What causes this lack of uniqueness mathematically? Physically, τ is the time that elapses before random fluctuations form a "seed" which initiates crystal growth.

1.3.5. Consider one-dimensional motion with position coordinate x. Let $f(x)$ be a position dependent but time independent force defined by the rule

$$f(x) = 0 \text{ for } x \leq 0;\ f(x) = +12x^{1/2} \text{ for } x \geq 0. \tag{1.3.10}$$

Note that $f(x)$ is continuous and satisfies $f(x) \geq 0$. Consider the equation of motion

$$\ddot{x} = f(x) \tag{1.3.11}$$

with the initial conditions

$$x(0) = \dot{x}(0) = 0. \tag{1.3.12}$$

Let c be any constant satisfying $c \geq 0$. Verify that (3.11) with the initial conditions (3.12) has the solution

$$x(t) = 0 \text{ for } t \leq c \text{ and } x(t) = (t - c)^4 \text{ for } t \geq c. \tag{1.3.13}$$

That is, verify that both (3.11) and (3.12) are satisfied. Note that $x(t)$ is continuous. How many continuous derivatives does it have? Why is the solution not unique? Are there still more solutions? What are the solutions to (3.11) for other initial conditions?

1.3.6. In computing and managing the trajectory of a space craft, one is obliged to use tracking data that inevitably contain at least some small errors. Also various parameters, such as anomalies in the gravitational field, the mass of the space ship, and the impulses provided by various rockets and thrusters, are not exactly known. Comment on the effect of these errors in view of Theorems 3.1 through 3.3.

1.3.7. Consider a set of differential equations of the form (3.4), and assume that the existence and uniqueness conditions of Theorem 3.1 are met. Show that no two different trajectories in $(\boldsymbol{y}, t)$ space can ever join or intersect in finite time. Suppose the quantities f_j are independent of the time t. Then the set of differential equations is called *autonomous*. Show that in this case no trajectory in $\boldsymbol{y}$ space can *cross* itself in finite time. (We say that a trajectory crosses itself if the two tangent lines to the two portions of the trajectory at the point of intersection have a finite angle between them.) Show that if a trajectory does intersect itself in finite time, it must join itself smoothly to form a periodic trajectory.

1.4 Transfer Maps Produced by Differential Equations

Suppose we rewrite the set of first-order differential equations (3.4) in the more compact vector form

$$\dot{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{y}; t). \tag{1.4.1}$$

Then, according to Theorem 3.1 and again using vector notation, their solution can be written in the form

$$\boldsymbol{y}(t) = \boldsymbol{g}(\boldsymbol{y}^0; t^0; t). \tag{1.4.2}$$

That is, the quantities $\boldsymbol{y}(t)$ at any time t are uniquely specified by the initial quantities $\boldsymbol{y}^0$ given at the initial time t^0.

We capitalize on this fact by introducing a slightly different notation. First, use t^i instead of t^0 to denote the *initial* time, and similarly use $\boldsymbol{y}^i$ to denote initial conditions by writing

$$\boldsymbol{y}^i = \boldsymbol{y}^0 = \boldsymbol{y}(t^i). \tag{1.4.3}$$

Next, let t^f be some *final* time, and define final conditions $\boldsymbol{y}^f$ by writing

$$\boldsymbol{y}^f = \boldsymbol{y}(t^f). \tag{1.4.4}$$

Then, with this notation, (4.2) can be rewritten in the form

$$\boldsymbol{y}^f = \boldsymbol{g}(\boldsymbol{y}^i; t^i; t^f). \tag{1.4.5}$$

We now view (4.5) as a map that sends the initial conditions $\boldsymbol{y}^i$ to the final conditions $\boldsymbol{y}^f$. This map will be called the *transfer map* between the times t^i and t^f, and will often be denoted by the symbol $\mathcal{M}$. What we have learned is that a set of first-order differential equations of the form (4.1) can be integrated to produce a transfer map $\mathcal{M}$. We express the fact that $\mathcal{M}$ sends $\boldsymbol{y}^i$ to $\boldsymbol{y}^f$ in symbols by writing the equation

$$\boldsymbol{y}^f = \mathcal{M}\boldsymbol{y}^i, \tag{1.4.6}$$

and illustrate this relation by the picture shown in Figure 4.1. Finally, as noted earlier, $\mathcal{M}$ is always invertible: Given $\boldsymbol{y}^f$, t^f, and t^i, we can always march (integrate) backward in time to the moment t^i and thereby find the initial conditions $\boldsymbol{y}^i$.

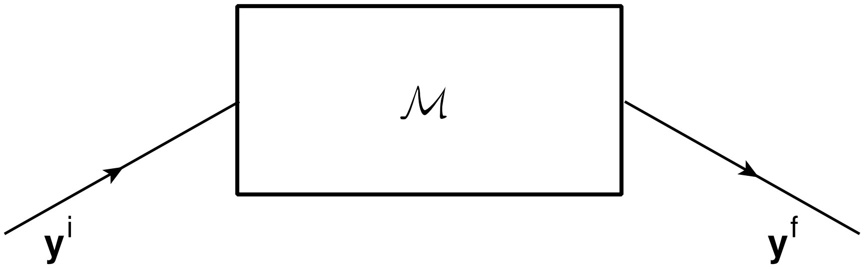


Figure 1.4.1: The transfer map $\mathcal{M}$ sends the initial conditions $\boldsymbol{y}^i$ to the final conditions $\boldsymbol{y}^f$.

1.4.1 Map for Simple Harmonic Oscillator

To fix these ideas more clearly in the mind, we consider three examples. The first is a one-dimensional harmonic oscilator described by the Hamiltonian

$$H = p^2/(2m) + (k/2)q^2. \tag{1.4.7}$$

In this case the equations of motion are

$$\dot{q} = \partial H/\partial p = p/m,$$

$$\dot{p} = -\partial H/\partial q = -kq. \tag{1.4.8}$$

(See Section 1.5 for a review of Hamilton's equations of motion.) These equations can be solved easily enough. However, for future use, it is convenient to make the (canonical) change of variables

$$Q = (km)^{1/4} q,$$
$$P = (km)^{-1/4} p. \tag{1.4.9}$$

In these new variables the equations of motion become

$$\dot{Q} = \omega P,$$
$$\dot{P} = -\omega Q, \tag{1.4.10}$$

where

$$\omega = \sqrt{(k/m)}. \tag{1.4.11}$$

It is easily verified that the equations of motion (4.10) are produced by the new Hamiltonian K given by the relation

$$K = (\omega/2)(P^2 + Q^2). \tag{1.4.12}$$

The equations (4.10) are easily integrated to give the transfer map $\mathcal{M}$ described by the relations

$$Q^f = Q^i \cos[\omega(t^f - t^i)] + P^i \sin[\omega(t^f - t^i)],$$
$$P^f = -Q^i \sin[\omega(t^f - t^i)] + P^i \cos[\omega(t^f - t^i)]. \tag{1.4.13}$$

We see that for this example the transfer map is a linear relation between the initial and final conditions and (in the Q, P variables) simply consists of a (clockwise) rotation in phase space by the angle $[\omega(t^f - t^i)]$.

In view of the assertion (2.43), the map described by (4.13) can also be written formally as

$$\mathcal{M} = \exp\{-(t^f - t^i) : (\omega/2)[(P^i)^2 + (Q^i)^2] :\}. \tag{1.4.14}$$

Note that this claim is consistent with (2.37) and (2.38).

1.4.2 Maps for Monomial Hamiltonians

The second example of a transfer map is somewhat more complicated, and leads to a nonlinear relation between initial and final conditions. It too will be useful in the future. Consider, for the case of a two-dimensional phase space, the monomial Hamiltonian

$$H = \lambda q^r p^s. \tag{1.4.15}$$

Here λ is a parameter, and r and s are integers. The Hamiltonian (4.15) produces the equations of motion

$$\dot{q} = \lambda s q^r p^{s-1}, \tag{1.4.16}$$
$$\dot{p} = -\lambda r q^{r-1} p^s. \tag{1.4.17}$$

Since H has no explicit time dependence, we conclude that H must be a constant of motion. If you doubt this, see (5.14) in the next section. Let us solve (4.15) for p. Doing so gives the result

$$p = (H/\lambda)^{\frac{1}{s}} q^{-\frac{r}{s}}. \tag{1.4.18}$$

Next substitute (4.18) into (4.16) to get the relation

$$\dot{q} = \lambda s (H/\lambda)^{\frac{s-1}{s}} q^{\frac{r}{s}}. \tag{1.4.19}$$

Assume for the moment that $r \neq s$. Then (4.19) can be integrated immediately to give the result

$$(q^f)^{\frac{s-r}{s}} - (q^i)^{\frac{s-r}{s}} = \lambda(s-r)(t^f - t^i)(H/\lambda)^{\frac{s-1}{s}}. \tag{1.4.20}$$

Also, since H is a constant of motion, we may write

$$H = \lambda (q^i)^r (p^i)^s. \tag{1.4.21}$$

Equations (4.20) and (4.21) can now be combined and solved for q^f in terms of q^i and p^i. Finally, (4.17) can be integrated in a similar manner. The net result is the transfer map relations

$$q^f = q^i[1 + \lambda(s-r)(t^f - t^i)(q^i)^{r-1}(p^i)^{s-1}]^{\frac{s}{s-r}}, \tag{1.4.22}$$

$$p^f = p^i[1 + \lambda(s-r)(t^f - t^i)(q^i)^{r-1}(p^i)^{s-1}]^{\frac{r}{r-s}}, \tag{1.4.23}$$

when $r \neq s$.

The equations of motion for the case $r = s$ can also be solved. In this case (4.18) can be integrated in terms of logarithms. Also, (4.17) can be integrated similarly. The net result is the transfer map relations

$$q^f = q^i \exp[\lambda r (t^f - t^i)(q^i p^i)^{r-1}], \tag{1.4.24}$$

$$p^f = p^i \exp[-\lambda r (t^f - t^i)(q^i p^i)^{r-1}], \tag{1.4.25}$$

when $r = s$.

Note that the relations (4.22) through (4.25) are indeed nonlinear. The transfer maps for monomial Hamiltonians in higher dimensional phase spaces can also be found exactly. See Exercise 4.3. Also, we remark that the relations (4.22) and (4.23) can become *singular* in finite time. That is, the solutions to the equations of motion (4.16) and (4.17) cannot always be extended arbitrarily far forward and backward in time. See Exercise 4.4.

Again because of (2.43), the maps described by (4.22) through (4.25) can formally be written as

$$\mathcal{M} = \exp\{-(t^f - t^i) : \lambda (q^i)^r (p^i)^s :\}. \tag{1.4.26}$$

And summation of the exponential series (4.26), when acting on the initial conditions, will produce the maps (4.22) through (4.25).

1.4.3 Stroboscopic Maps and Duffing Equation Example

For a last example of a transfer map produced by a differential equation, we will begin a study of the behavior of a periodically driven damped *nonlinear* oscillator described by the equation of motion

$$\ddot{x} + a\dot{x} + bx + cx^3 = d\cos(\Omega t + \psi). \tag{1.4.27}$$

This equation, or sometimes a variant with x^3 replaced by x^2, is commonly called *Duffing's equation*. Here ψ is an arbitrary phase factor that is often set to zero. For our purposes it is more convenient to set

$$\psi = \pi/2. \tag{1.4.28}$$

Evidently any particular choice of ψ simply results in a shift of the origin in time, and this shift has no physical consequence since the left side of (4.27) is independent of time.

We assume $b, c > 0$, which is the case of a positive hard spring restoring force.[33] We make these assumptions because we want the Duffing oscillator to behave like an ordinary harmonic oscillator when the amplitude is small, and we want the motion to be bounded away from infinity when the amplitude is large. Then, by a suitable choice of time and length scales that introduces new variables q and τ, the equation of motion can be brought to the form

$$\ddot{q} + 2\beta\dot{q} + q + q^3 = -\epsilon \sin \omega\tau, \tag{1.4.29}$$

where now a dot denotes $d/d\tau$ and we have made use of (4.28). See Exercise 4.10. In this form it is evident that there are 3 free parameters: β, ϵ, and ω.

Unlike the previous examples, this problem is dissipative (assuming $\beta > 0$) and time dependent. There is, however, the simplifying feature that the driving force is *periodic* with period

$$T = 2\pi/\omega. \tag{1.4.30}$$

Let us convert (4.29) into a pair of first-order equations by making the definition

$$p = \dot{q}, \tag{1.4.31}$$

with the result

$$\dot{q} = p,$$
$$\dot{p} = -2\beta p - q - q^3 - \epsilon \sin \omega\tau. \tag{1.4.32}$$

Let q_0, p_0 denote initial conditions at $\tau = 0$, and let q_1, p_1 be the final conditions resulting from integrating the pair (4.32) one full period to the time $\tau = T$. Let $\mathcal{M}$ denote the transfer map that relates q_1, p_1 to q_0, p_0. Then, using the definition (2.39) and the notation (4.6), we may write

$$z_1 = \mathcal{M} z_0. \tag{1.4.33}$$

Suppose we now integrate for a second full period to find q_2, p_2. Since the right side of (4.32) is periodic, the rules for integrating from $\tau = T$ to $\tau = 2T$ are the same as the rules for integrating from $\tau = 0$ to $\tau = T$. Therefore we may write

$$z_2 = \mathcal{M} z_1 = \mathcal{M}^2 z_0, \tag{1.4.34}$$

and in general

$$z_{n+1} = \mathcal{M} z_n = \mathcal{M}^{n+1} z_0. \tag{1.4.35}$$

[33] Other authors consider other cases, particularly the 'double well' case $b < 0$ and $c > 0$.

We may regard the quantities z_n as the result of viewing the motion of the Duffing oscillator by the light provided by a stroboscope that flashes at the times[34]

$$\tau^n = nT. \tag{1.4.36}$$

Because of the periodicity of the right side of the equations of motion, the rule for sending z_n to z_{n+1} over the intervals between successive flashes is always the same, namely $\mathcal{M}$. For these reasons $\mathcal{M}$ is called a *stroboscopic map.* Despite the explicit time dependence in the equations of motion, because of periodicity we have been able to describe the long-term motion by the repeated application of a single fixed map. A moment's reflection shows that what we have done here for the Duffing oscillator is quite general. The behavior of any periodically (not necessarily sinusoidally) driven system can be described by a stroboscopic map.[35]

It follows from (4.35) that the long-term behavior of the driven Duffing oscillator is equivalent to the behavior of the Duffing stroboscopic map $\mathcal{M}$ under repeated iteration. As we have seen from the examples of Section 1.2, the iteration of maps generally leads to enormous complications. Correspondingly, the driven Duffing oscillator displays an enormously rich behavior that varies widely with the parameter values β, ϵ, ω. This richness is typical of the long-term behavior of damped driven nonlinear systems. Indeed, without editorial restraint, the detailed study of any one of them could fill this entire book. However, rich as it is, the behavior of the driven Duffing oscillator, since it is governed by relatively few *attracting* (due to the presence of damping) fixed points, is trivial compared to that of most nonlinear *Hamiltonian* systems where fixed points are numerous and none are attracting.

Because even providing an overview of what can happen under repeated iteration of the stroboscopic Duffing map requires considerable work, at least an entire chapter is required for this purpose. Such an overview is provided in Chapter 28 where the subject is studied numerically and Section 29.12 where the behavior of polynomial approximations to the stroboscopic Duffing map is explored. See also Sections 10.12.7 and 10.12.8 and Appendix S.4.

Exercises

1.4.1. Verify equations (4.8) through (4.13) and all assertions made about them.

[34] Note that, with the choice (4.28) for ψ, the driving term described by the right side of (4.29) vanishes at the stroboscopic times τ^n.

[35] Consider a set of n second-order differential equations of the form (3.1) with the further assumption that the h_j do not depend on the time. We will say that such a set of equations (which is equivalent to a set of $2n$ autonomous first-order differential equations) describes a system having n *autonomous* degrees of freedom. Suppose next that the h_j do depend on the time, and in an *arbitrary* way. By choosing a new independent variable, it is possible to convert such a set of second-order differential equations into $(2n + 2)$ first-order autonomous differential equations. (See Exercise 6.5 for a discussion of how this can be done in the Hamiltonian case.) Thus, when t is present in the equations (3.1), we may say that these $2n$ nonautonomous equations describe a system having $(n + 1)$ autonomous degrees of freedom. As the discussion of this section shows, the case where the h_j depend on the time in a *periodic* way lies somewhere in between. Such systems are sometimes said to have $(n + 1/2)$ autonomous degrees of freedom. Thus, the Duffing oscillator may be said to have $3/2 = 1$ and $\frac{1}{2}$ degrees of freedom.

1.4.2. Verify equations (4.16) through (4.25). Suppose $s = 0$ and $r \neq 0$. Show that in this case

$$q^f = q^i,$$
$$p^f = p^i - \lambda r(t^f - t^i)(q^i)^{r-1}. \tag{1.4.37}$$

Suppose $s \neq 0$ and $r = 0$. Show that in this case

$$q^f = q^i + \lambda s(t^f - t^i)(p^i)^{s-1},$$
$$p^f = p^i. \tag{1.4.38}$$

1.4.3. Consider, for the case of a four-dimensional phase space, the monomial Hamiltonian

$$H = \lambda q_1^{r_1} p_1^{s_1} q_2^{r_2} p_2^{s_2}. \tag{1.4.39}$$

Define "sub" Hamiltonians H_1 and H_2 by the relations

$$H_j = q_j^{r_j} p_j^{s_j}, \quad j = 1 \text{ and } 2. \tag{1.4.40}$$

With these definitions (4.43) can be rewritten in the form

$$H = \lambda H_1 H_2. \tag{1.4.41}$$

Show that both H_1 and H_2 are constants (in fact, integrals) of motion. [Hint: If you are having trouble, use (7.4) and (7.7) of Section 1.7.] Show that the equations of motion generated by H can be integrated and (when $r_j \neq s_j$) have solutions of the form

$$q_1^f = q_1^i[1 + \lambda(s_1 - r_1)(t^f - t^i)(q_2^i)^{r_2}(p_2^i)^{s_2}(q_1^i)^{r_1-1}(p_1^i)^{s_1-1}]^{\frac{s_1}{s_1-r_1}}, \text{ etc.} \tag{1.4.42}$$

Find complete results for all q_j^f, p_j^f and for all cases of the exponents r_j, s_j.

1.4.4. Consider the solution to (4.16) and (4.17) as given by (4.22) and (4.23) for the case $r = 1$ and $s = 4$. Show that the solution has a branch point in t at a finite time. Find other integer values of r, s for which the solution (4.22) and (4.23) has singularities for finite time. Conversely, given any neighborhood of the origin in the initial conditions q^i and p^i, show that (for suitable r, s values) the solution (4.22) and (4.23), when viewed as a function of the initial conditions, has singularities in this neighborhood for t finite (and real) providing t is sufficiently large. In view of Theorems 3.1 and 3.3, what is going wrong?

1.4.5. From the general discussion of transfer maps it is clear that non-Hamiltonian systems also can be described in terms of maps. All that is required is that the set of differential equations be written in the first-order form (4.1). Consider the one-dimensional motion of an object moving vertically and subject to gravity and viscous drag. Newton's equation of motion for such an object can be written in the form

$$m\ddot{z} = -mg - \gamma\dot{z}. \tag{1.4.43}$$

Here m is the mass of the particle, g is the acceleration due to gravity, and γ (with $\gamma > 0$) is some measure of the viscous drag. Convert (4.47) into a first-order set of differential equations, and find the associated transfer map.

1.4.6. Let $\mathcal{M}$ be a map of an m-dimensional space into itself as in (4.6). What happens to the final conditions when small changes are made in the initial conditions? From calculus we have the differential relation

$$dy_j^f = \sum_k (\partial y_j^f / \partial y_k^i) dy_k^i, \tag{1.4.44}$$

which can be written in the form

$$dy_j^f = \sum_k M_{jk}(\boldsymbol{y}^i) dy_k^i \tag{1.4.45}$$

where $M(\boldsymbol{y}^i)$ is the $m \times m$ matrix

$$M_{jk}(\boldsymbol{y}^i) = \partial y_j^f / \partial y_k^i. \tag{1.4.46}$$

This matrix is called the *Jacobian* matrix of $\mathcal{M}$. According to (4.81) it describes how small changes in the initial conditions $\boldsymbol{y}^i$ produce small changes in the final conditions $\boldsymbol{y}^f$. Note that generally the Jacobian matrix depends on the initial conditions, and therefore we write $M(\boldsymbol{y}^i)$.

In the case that $\mathcal{M}$ is a transfer map arising from a differential equation as in (4.1), the associated Jacobian matrix can be found by integrating the *variational* equations derived from (4.1). Here, as before, we assume $\boldsymbol{y}$ has m components. Let $\boldsymbol{y}^i$ be a set of initial conditions and let $\boldsymbol{y}^d(t)$ be the trajectory (sometimes called the *design* trajectory) that has these initial conditions,

$$\boldsymbol{y}^d(t^i) = \boldsymbol{y}^i. \tag{1.4.47}$$

Because it is a trajectory, it satisfies the differential equation

$$\dot{\boldsymbol{y}}^d = \boldsymbol{f}(\boldsymbol{y}^d; t). \tag{1.4.48}$$

Next consider nearby trajectories of the form

$$\boldsymbol{y}(t) = \boldsymbol{y}^d(t) + \epsilon \boldsymbol{\eta}(t) \tag{1.4.49}$$

where ϵ is small. Insertion of (4.49) into (4.1) gives the equation

$$\dot{\boldsymbol{y}}^d + \epsilon \dot{\boldsymbol{\eta}} = \boldsymbol{f}(\boldsymbol{y}^d + \epsilon \boldsymbol{\eta}; t). \tag{1.4.50}$$

Now take components of both sides of (4.50) and expand in powers of ϵ to find the relation

$$\dot{\boldsymbol{y}}_j^d + \epsilon \dot{\boldsymbol{\eta}}_j = f_j(\boldsymbol{y}^d; t) + \sum_k [(\partial f_j / \partial y_k) \Big|_{\boldsymbol{y}=\boldsymbol{y}^d}] \epsilon \eta_k + O(\epsilon^2) \ . \tag{1.4.51}$$

Define the $m \times m$ matrix $A(t)$ by the rule

$$A_{jk}(t) = (\partial f_j / \partial y_k) \Big|_{\boldsymbol{y}=\boldsymbol{y}^d} \ . \tag{1.4.52}$$

Use (4.48), (4.51), and (4.52) and equate powers of ϵ to show that $\boldsymbol{\eta}$ satisfies the set of equations

$$\dot{\boldsymbol{\eta}} = A(t) \boldsymbol{\eta}. \tag{1.4.53}$$

These are the *variational equations* associated with (4.1) around the trajectory $\boldsymbol{y}^d$.[36] Note that there are m such (usually coupled) equations because $\boldsymbol{\eta}$ is m dimensional, and that they are *linear* even if (4.1) is nonlinear.

Let $L(t)$ be the $m \times m$ matrix defined by the *matrix* differential equation (a collection of m^2 ordinary differential equations)

$$\dot{L} = A(t)L \tag{1.4.54}$$

with the initial condition

$$L(t^i) = I \tag{1.4.55}$$

where I denotes the $m \times m$ identity matrix. Show that the solution to (4.53) with the initial condition $\boldsymbol{\eta}^i$ is given by the prescription

$$\boldsymbol{\eta}(t) = L(t)\boldsymbol{\eta}^i. \tag{1.4.56}$$

Show that the desired Jacobian matrix is given in terms of $L(t)$ by the relation

$$M = L(t^f). \tag{1.4.57}$$

The solution of the differential equations (4.48) for the design trajectory, which is required to determine A using (4.52), generally requires numerical integraton. Solution of the variational equations (4.53), or their matrix counterpart (4.54), even though they are linear, also generally requires numerical integration because they are coupled and A is usually time dependent. However, assuming A is known, it is possible to calculate the *determinant* of M analytically. Use (4.54) to write the Taylor expansion

$$\begin{aligned} L(t+dt) &= L(t) + \dot{L}(t)dt + O[(dt)^2] \\ &= L(t) + dtA(t)L(t) + O[(dt)^2] \\ &= [I + dtA(t)]L(t) + O[(dt)^2]. \end{aligned} \tag{1.4.58}$$

Take determinants of both sides of (4.58) to get the result

$$\begin{aligned} \det[L(t+dt)] &= \det\{[I + dtA(t)][L(t)]\} + O[(dt)^2] \\ &= \{\det[I + dtA(t)]\}\{\det[L(t)]\} + O[(dt)^2] \\ &= \{1 + dt\,\mathrm{tr}[A(t)]\}\{\det[L(t)]\} + O[(dt)^2]. \end{aligned} \tag{1.4.59}$$

Here use has been made of (3.7.132). Show that (4.59) produces the differential equation

$$(d/dt)\det[L(t)] = \{\mathrm{tr}[A(t)]\}\{\det[L(t)]\} \tag{1.4.60}$$

and, in view of (4.55), that this equation has the explicit solution

$$\det[L(t)] = \exp\{\int_{t^i}^{t} dt'\,\mathrm{tr}[A(t')]\}. \tag{1.4.61}$$

[36] We could more accurately call them the first-variation equations or lowest-order variational equations. For what we call the *complete* variational equations, see Section 10.12.

In particular, there is the result

$$\det(M) = \exp\{\int_{t^i}^{t^f} dt\, \mathrm{tr}[A(t)]\}. \tag{1.4.62}$$

Subsequently, this result will be related, in the context of Hamiltonian dynamics, to what is called *Liouville's theorem.* The result itself, in the context of linear differential equations, which is what the variational equations are, is sometimes called the *Abel-Liouville-Jacobi-Ostrogradski formula.*

From this result show that the determinant of the Jacobian matrix associated with any transfer map arising from a *real* differential equation must satisfy the condition

$$\det(M) > 0. \tag{1.4.63}$$

Geometrically, this condition means that $\mathcal{M}$ preserves *orientation.* For example, in the case $m = 3$, the $\mathcal{M}$ arising from any real differential equation cannot send a right-handed triad into a left-handed triad. Comment: There is also a simpler but more subtle topological argument that leads to the result (4.63). Since a transfer map arising from integrating a differential equation evolves in a continuous way starting from the identity map, it can be written as a product of several transfer maps, all of which are near the identity map. See Section 6.4.1. Since each of these maps is near the identity, by continuity the determinant of the Jacobian matrix of each must be positive. But, by the chain rule, the Jacobian matrix of a product of maps must be the product of the Jacobian matrices of the individual factors. Finally, the determinant of a product of matrices is the product of the determinants of the individual factors.

The determinant of the Jacobian matrix also has further geometrical significance. For the purpose of this exercise, let us refer to the m-dimensional space we have been considering as *variable* space. This variable space need not be phase space because the dimension may be odd, and even if m is even the equations of motion need not be Hamiltonian in form and the coordinates may not necessarily come in canonically conjugate pairs. The equations of motion and the coordinates can be completely general.

Consider a particular trajectory with initial conditions given by (4.47) and also all other trajectories whose initial conditions lie within a small volume dV^i about the initial conditions for the particular trajectory. Then at some final time t^f the final conditions for these trajectories will lie within a small volume dV^f about the final conditions for the particular trajectory. From standard advanced calculus lore the initial and final volumes are related by the equation

$$dV^f = \{\det[M(\boldsymbol{y}^i)]\}dV^i. \tag{1.4.64}$$

Thus, the determinant of M specifies the evolution of volume elements in variable space.

1.4.7. For the case of the complex logistic map in the form (2.112), write

$$w = u + iv \tag{1.4.65}$$

and show that the Jacobian matrix is given by the relation

$$M(w_n) = 2\begin{pmatrix} u_n & -v_n \\ v_n & u_n \end{pmatrix}. \tag{1.4.66}$$

See Exercise 4.6. Thus for this map

$$\det[M(w_n)] = 4(u_n^2 + v_n^2) = 4|w_n|^2, \tag{1.4.67}$$

and the map preserves orientation except at the origin. Verify that the map is not invertible in the neighborhood of the origin. Consider any map of the form

$$w_{n+1} = f(w_n) \tag{1.4.68}$$

where f is an analytic function. Show that

$$\det[M(w_n)] = |f'(w_n)|^2. \tag{1.4.69}$$

Thus, all analytic maps are orientation preserving.

1.4.8. Consider the Hénon map in the product form (2.23). Compute the Jacobian matrix for each factor. See Exercise 4.6. Verify that the Jacobian matrix for each factor has determinant one and therefore, by the chain rule, the determinant of the Jacobian matrix for the full map also has determinant one. It follows, as will be described in detail later, that the Hénon map is *area preserving.*

1.4.9. Let $\delta_{per}(t)$ denote the 2π *periodic* delta function defined by the relation

$$\delta_{per}(t) = \sum_{n=-\infty}^{\infty} \delta(t + 2n\pi). \tag{1.4.70}$$

Show that the map $\mathcal{M}(\theta)$ given by (2.39) is the stroboscopic map resulting from integrating from $t^i = 0$ to $t^f = 2\pi$ the motion arising from the 2π periodic Hamiltonian

$$H = [\theta/(4\pi)](p^2 + q^2) - \delta_{per}(t - \pi)q^3. \tag{1.4.71}$$

1.4.10. Choose appropriate time and length scales by writing $x = \lambda q$ and $t = \sigma\tau$ to convert (4.27) into (4.29).

1.5 Lagrangian and Hamiltonian Equations

It is a remarkable discovery that all the known fundamental dynamical laws of Nature are expressible in Lagrangian or Hamiltonian form, and therefore also in variational form. Indeed, as Euler wrote in his (and the first by any author) publication on variational calculus,

> Because the shape of the whole universe is most perfect and, in fact, designed by the wisest Creator, nothing in all of the world will occur in which no maximum or minimum rule is somehow shining forth.
>
> Since the construction of the entire universe is absolutely perfect and is due to a Creator with infinite knowledge, nothing exists in the world which does not exhibit some property of maximum or minimum. Therefore, there cannot be any doubt whatsoever about the possibility that all the effects are determined by their final aims with the help of the maxima method, in the same way in which they are also determined by the initial causes.

The last three sections of this chapter are devoted to needed aspects of Lagrangian and Hamiltonian dynamics.

Since much of this book concerns the motion of charged particles in electromagnetic fields, we recall that the relativistic Lagrangian for the motion of a particle of mass m and charge q in an electromagnetic field is given by the expression

$$L(\boldsymbol{r},\boldsymbol{v},t) = -mc^2(1-v^2/c^2)^{1/2} - q\psi(\boldsymbol{r},t) + q\boldsymbol{v}\cdot\boldsymbol{A}(\boldsymbol{r},t). \tag{1.5.1}$$

Here ψ and $\boldsymbol{A}$ are the scalar and vector potentials defined in such a way that the electromagnetic fields $\boldsymbol{E}$ and $\boldsymbol{B}$ are given by the standard relations

$$\boldsymbol{B} = \nabla\times\boldsymbol{A},$$

$$\boldsymbol{E} = -\nabla\psi - \partial\boldsymbol{A}/\partial t. \tag{1.5.2}$$

We note that this formulation ignores spin, radiation reaction (synchrotron radiation), and quantum effects.[37]

1.5.1 The Nonsingular Case

Lagrange's equations of motion for a system having n degrees of freedom are

$$\frac{d}{dt}\frac{\partial L}{\partial\dot{q}_i} - \frac{\partial L}{\partial q_i} = 0, \tag{1.5.3}$$

where $(q_1\cdots q_n)$ is any set of generalized coordinates. [Note that in general L is a function of the q_i, $\dot{q}_i$, and t; $L = L(q,\dot{q},t)$.] According to Section 1.3, what we ultimately need are equations of the form (3.1). By the chain rule there are the relations

$$\frac{d}{dt}\frac{\partial L}{\partial\dot{q}_j} = \frac{\partial^2 L}{\partial t\partial\dot{q}_j} + \sum_i\left[\frac{\partial^2 L}{\partial\dot{q}_i\partial\dot{q}_j}\ddot{q}_i + \frac{\partial^2 L}{\partial q_i\partial\dot{q}_j}\dot{q}_i\right] \tag{1.5.4}$$

so that Lagrange's equations can also be written in the form

$$\sum_i\frac{\partial^2 L}{\partial\dot{q}_i\partial\dot{q}_j}\ddot{q}_i = \frac{\partial L}{\partial q_j} - \frac{\partial^2 L}{\partial t\partial\dot{q}_j} - \sum_i\frac{\partial^2 L}{\partial q_i\partial\dot{q}_j}\dot{q}_i. \tag{1.5.5}$$

[37] We also note that the Lagrangian L given by (5.1), while relativistically correct, it is not manifestly Lorentz invariant. The connection between this Lagrangian and a manifestly Lorentz invariant formulation is explored in Exercises 6.7 and 6.8. It is assumed that the reader already has some background in Special Relativity. For further discussion of the Lorentz group and related material, see Exercises 6.17, 6.18, 6.2.6, 6.2.7, 6.2.12, 6.2.13, 7.3.26 through 7.3.36, and 8.2.14 through 8.2.21.

Poincaré, in a 1905 paper, coined the terms *Lorentz transformation* and *Lorentz group*. Hendrik Lorentz (1853-1928) was a Dutch physicist who made many contributions to Physics including the discovery and theoretical explanation of the Zeeman effect for which they jointly shared the 1902 Nobel Prize in Physics. For a video of Lorentz's funeral procession, which included Einstein, see `https://www.youtube.com/watch?v=H2VtrJD0xJk`. Pieter Zeeman (1865-1943) was a student and subsequent colleague of Lorentz. Hendrik Lorentz is not to be confused with the Danish physicist Ludvig Lorenz (1829-1891) for whom the Lorenz gauge/condition is named or with the meteorologist Edward Norton Lorenz (1917-2008) who was a pioneer in chaos theory.

The quantity $[\partial^2 L/\partial\dot{q}_i\partial\dot{q}_j]$ is called the *Hessian* (matrix) of L. In order to solve the relations (5.5) for the $\ddot{q}_i$ to obtain equations of the form (3.1), the Hessian of L must be invertible/nonsingular and therefore must satisfy the condition

$$\det(\partial^2 L/\partial\dot{q}_i\partial\dot{q}_j) \neq 0. \tag{1.5.6}$$

We call this the *nonsingular* or *regular* case; and, when (5.6) fails to hold, we call this the *singular* case.[38]

The momentum p_i canonically conjugate to the variable q_i is defined by the relation

$$p_i = p_i(q, \dot{q}, t) = \partial L/\partial\dot{q}_i, \tag{1.5.7}$$

and the Hamiltonian H associated with the Lagrangian L is defined by the *Legendre* transformation

$$H(q,p,t) = \sum_i p_i\dot{q}_i - L(q,\dot{q},t). \tag{1.5.8}$$

(For a study of Legendre transformations, see Exercise 6.2.9.) Note that as it stands, and in view of (5.7), the right side of (5.8) is a function of the variables $q, \dot{q}, t$. However the left side describes H as a function of the variables q, p, t. That is, the variables $\dot{q}$ are to be eliminated in terms of the p's. According to the *inverse function theorem*, this is possible if and only if the determinant of the associated *Jacobian* matrix is nonzero,

$$\det(\partial p_i/\partial\dot{q}_j) \neq 0. \tag{1.5.9}$$

From (5.7) there is the relation

$$\partial p_i/\partial\dot{q}_j = \partial^2 L/\partial\dot{q}_j\partial\dot{q}_i = \partial^2 L/\partial\dot{q}_i\partial\dot{q}_j. \tag{1.5.10}$$

Here we have used the equality of mixed partial derivatives.[39] Thus, the conditions (5.6) and (5.9) are the same.

Hamilton's equations of motion for the $2n$ canonical variables $(q_1 \cdots q_n)$ and $(p_1 \cdots p_n)$ are given in terms of the Hamiltonian $H(q,p,t)$ by the rules

$$\dot{q}_i = \partial H/\partial p_i \quad , \quad \dot{p}_i = -\partial H/\partial q_i. \tag{1.5.11}$$

There is also the additional relation

$$\partial H/\partial t = -\partial L/\partial t. \tag{1.5.12}$$

For later use, it is convenient to append yet one more equation to the set (5.11) and (5.12). Consider the total time rate of change of the Hamiltonian H along a trajectory in q, p space. Using the chain rule, one finds the result

$$dH/dt = \partial H/\partial t + \sum_i [(\partial H/\partial q_i)\dot{q}_i + (\partial H/\partial p_i)\dot{p}_i]. \tag{1.5.13}$$

[38] Abraham and Marsden call the nonsingular case *hyperregular* if the the map $q, \dot{q}, t \leftrightarrow q, p, t$ is a diffeomorphism; that is, it is a differentiable map with a differentiable inverse. See (5.6) through (5.10). (For our purposes we are happy to assume differentiability, or even analyticity.) They call the singular case *degenerate*. Some other authors call the singular case *irregular*.

[39] The Clairaut-Schwarz theorem.

However, the quantity under the summation sign vanishes because of (5.11). It follows that the Hamiltonian has the special property

$$dH/dt = \partial H/\partial t = -\partial L/\partial t. \tag{1.5.14}$$

Suppose that H (or L) does not depend explicitly on the time ($\partial H/\partial t = 0$ or $\partial L/\partial t = 0$). A system that does not explicitly depend on the independent variable (the time) is called *autonomous*. We see from (5.14) that if a Hamiltonian system is autonomous, then the Hamiltonian H must be a *constant* of motion, and conversely. Moreover, because it has no explicit time dependence, such an H is also an *integral* of motion. For a discussion of constants and integrals of motion see Section 5.2.

1.5.2 A Common Singular Case

We end this section by noting that there is a fairly frequently encountered case in which (5.6) and (5.9) fail to hold, namely when L is *homogeneous* of degree *one* in the $\dot{q}_i$,

$$L(q, \lambda\dot{q}, t) = \lambda L(q, \dot{q}, t). \tag{1.5.15}$$

See, for examples, Exercises 5.15, 6.5, 6.9, and 6.16. In this case the p_i are homogeneous of degree *zero* in the $\dot{q}_i$ and, according to Euler's relation, there will be the result

$$\sum_j (\partial p_i/\partial \dot{q}_j)\dot{q}_j = 0. \tag{1.5.16}$$

See Exercise 5.12. The quantities $(\partial p_i/\partial \dot{q}_j)$ may be viewed as the entries in a matrix, and the quantities $\dot{q}_j$ may be viewed as the entries in a vector. Since (5.16) must hold for any value of the $\dot{q}_j$, we conclude that the matrix $(\partial p_i/\partial \dot{q}_j)$ has a nonzero vector as an eigenvector with eigenvalue 0. It follows that in this case

$$\det(\partial p_i/\partial \dot{q}_j) = \det(\partial^2 L/\partial \dot{q}_i \partial \dot{q}_j) = 0. \tag{1.5.17}$$

Moreover, since L is assumed homogeneous of degree 1 in the $\dot{q}_i$, Euler's relation also gives the result

$$\sum_i p_i \dot{q}_i = \sum_i (\partial L/\partial \dot{q}_i)\dot{q}_i = L, \tag{1.5.18}$$

and hence, according to (5.8), the Hamiltonian associated with L *vanishes identically*.

Finally, suppose $\mathcal{A}$ is the action functional associated with L and that L does not explicitly depend on the time,

$$\mathcal{A}[q(t)] = \int_{t^1}^{t^2} L(q, dq/dt)dt. \tag{1.5.19}$$

Let $\tau(t)$ be any monotonic function of t so that we may also write $t = t(\tau)$. Here we view τ as a parameter. Given a path $q(t)$, we will define a related path $Q(\tau)$ by the rule

$$Q_i(\tau) = q_i(t(\tau)). \tag{1.5.20}$$

Assign an action to any such path using the same functional (5.19),

$$\mathcal{A}[Q(\tau)] = \int_{\tau^1}^{\tau^2} L(Q, dQ/d\tau)d\tau. \tag{1.5.21}$$

By the chain rule we have the relations

$$d\tau = (d\tau/dt)dt, \tag{1.5.22}$$

$$dQ_i/d\tau = (dq_i/dt)(dt/d\tau). \tag{1.5.23}$$

Therefore, upon changing integration variables, there is the result

$$\mathcal{A}[Q(\tau)] = \int_{t^1}^{t^2} L\{q, (dq/dt)(dt/d\tau)\}(d\tau/dt)dt. \tag{1.5.24}$$

Under the assumption that L is homogeneous of degree one in the $\dot{q}_i$, there is also the relation

$$L\{q, (dq/dt)(dt/d\tau)\} = L(q, dq/dt)(dt/d\tau). \tag{1.5.25}$$

See (5.15). Inserting (5.25) into (5.24) gives the final result

$$\mathcal{A}[Q(\tau)] = \int_{t^1}^{t^2} L(q, dq/dt)(dt/d\tau)(d\tau/dt)dt = \int_{t^1}^{t^2} L(q, dq/dt)dt = \mathcal{A}[q(t)]. \tag{1.5.26}$$

We have learned that in this case $\mathcal{A}[Q(\tau)]$ is *independent* of the parameterization employed. That is, there are an infinite number of paths $Q(\tau)$, corresponding to the infinite number of parameterizations $t(\tau)$, all of which have the same action. This independence implies that we should *not* expect to find a unique solution that extremizes $\mathcal{A}$ since any reparameterization also gives a solution.

Is all lost when L is homogeneous of degree one in the $\dot{q}_i$? The answer is *no.* What we may expect in this case is that additional information beyond Hamilton's principle (or Lagrange's equations) will be required to specify a trajectory. Some further information has to be provided about the parameterization. Again see, for examples, Exercises 5.15, 6.5, 6.9, and 6.16.

Exercises

1.5.1. For the Lagrangian (5.1), show that the *canonical* momenta in Cartesian coordinates are given by the equation

$$\boldsymbol{p}^{\text{can}} = m\boldsymbol{v}/(1 - v^2/c^2)^{1/2} + q\boldsymbol{A}. \tag{1.5.27}$$

Here we have used the superscript *can* to emphasize that we are deriving the *canonical* momenta. Note that the first term in (5.27) is just the relativistic *mechanical* momentum,

$$\boldsymbol{p}^{\text{mech}} = m\boldsymbol{v}/(1 - v^2/c^2)^{1/2} = \gamma m\boldsymbol{v} \tag{1.5.28}$$

where γ is the standard relativistic factor

$$\gamma = 1/(1 - v^2/c^2)^{1/2}. \tag{1.5.29}$$

Consequently, the relation (5.27) can also be written in the forms

$$\boldsymbol{p}^{\text{can}} = \boldsymbol{p}^{\text{mech}} + q\boldsymbol{A} \quad \text{and} \quad \boldsymbol{p}^{\text{mech}} = \boldsymbol{p}^{\text{can}} - q\boldsymbol{A}. \tag{1.5.30}$$

1.5.2. The purpose of this exercise is to derive and study the equations of motion associated with the Langrangian L given by (5.1). Begin by reviewing Exercise 5.1. Let $\boldsymbol{p}^{\text{mech}}$ denote the *mechanical* momentum given by (5.28). In the case that the generalized coordinates are taken to be the usual Cartesian coordinates, verify that Lagrange's equations for the Lagrangian (5.1) produce for the mechanical momentum the equation of motion

$$\dot{\boldsymbol{p}}^{\text{mech}} = d\boldsymbol{p}^{\text{mech}}/dt = \boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}). \tag{1.5.31}$$

Here $\boldsymbol{F}$ is the Lorentz force.

For reasons that will become clear shortly, let us calculate the quantity $(d\gamma/dt)$. Rewrite (5.28) in the form

$$\boldsymbol{v} = \boldsymbol{p}^{\text{mech}}/(\gamma m). \tag{1.5.32}$$

Verify that squaring and inverting both sides of (5.29), and use of (5.32), produce the chain of relations

$$1/\gamma^2 = 1 - v^2/c^2 = 1 - (\boldsymbol{p}^{\text{mech}} \cdot \boldsymbol{p}^{\text{mech}})/(\gamma mc)^2, \tag{1.5.33}$$

from which it follows that

$$\gamma^2 = 1 + (\boldsymbol{p}^{\text{mech}} \cdot \boldsymbol{p}^{\text{mech}})/(mc)^2. \tag{1.5.34}$$

Next differentiate both sides of (5.34) to find that

$$\gamma(d\gamma/dt) = [1/(mc)^2](\boldsymbol{p}^{\text{mech}} \cdot \dot{\boldsymbol{p}}^{\text{mech}}) = [\gamma/(mc^2)](\boldsymbol{v} \cdot \dot{\boldsymbol{p}}^{\text{mech}}), \tag{1.5.35}$$

from which it follows that

$$d\gamma/dt = [1/(mc^2)](\boldsymbol{v} \cdot \dot{\boldsymbol{p}}^{\text{mech}}). \tag{1.5.36}$$

Now use (5.31) to show that

$$\boldsymbol{v} \cdot \dot{\boldsymbol{p}}^{\text{mech}} = \boldsymbol{v} \cdot \boldsymbol{F} = q\boldsymbol{v} \cdot \boldsymbol{E}, \tag{1.5.37}$$

and thereby verify that

$$d\gamma/dt = [1/(mc^2)](\boldsymbol{v} \cdot \boldsymbol{F}) = [q/(mc^2)](\boldsymbol{v} \cdot \boldsymbol{E}). \tag{1.5.38}$$

Define the relativistic energy $\mathcal{E}$ by the rule

$$\mathcal{E} = \gamma mc^2. \tag{1.5.39}$$

Show from (5.36) that it obeys the equation of motion

$$d\mathcal{E}/dt = \boldsymbol{v} \cdot \boldsymbol{F} = q(\boldsymbol{v} \cdot \boldsymbol{E}). \tag{1.5.40}$$

Note that $\boldsymbol{v}\cdot\boldsymbol{F}$ is simply the rate at which work is being done by the Lorentz force. Indeed, verify that (5.40) is equivalent to the differential relation

$$d\mathcal{E} = \boldsymbol{F}\cdot d\boldsymbol{r} = q(\boldsymbol{E}\cdot d\boldsymbol{r}). \tag{1.5.41}$$

As a particle moves, its change in energy equals the work done by the Lorentz force, more specifically by the *electric* part of the Lorentz force.

Verify from (5.34) and (5.39) that there is the relation

$$\mathcal{E}^2 = m^2c^4 + (\boldsymbol{p}^{\rm mech}\cdot\boldsymbol{p}^{\rm mech})c^2. \tag{1.5.42}$$

Show that (5.40) also follows from (5.31) and (5.42).

Solve (5.32) and (5.34) for $\boldsymbol{v}$ to find the relation

$$\dot{\boldsymbol{r}} = \boldsymbol{v} = \boldsymbol{p}^{\rm mech}/(m^2 + \boldsymbol{p}^{\rm mech}\cdot\boldsymbol{p}^{\rm mech}/c^2)^{1/2}. \tag{1.5.43}$$

Show that (5.31) can be rewritten in the form

$$\dot{\boldsymbol{p}}^{\rm mech} = q(\boldsymbol{E}+\boldsymbol{v}\times\boldsymbol{B}) = q\{\boldsymbol{E} + [\boldsymbol{p}^{\rm mech}/(m^2 + \boldsymbol{p}^{\rm mech}\cdot\boldsymbol{p}^{\rm mech}/c^2)^{1/2}]\times\boldsymbol{B}\}. \tag{1.5.44}$$

Taken together, (5.43) and (5.44) provide equations of motion for the quantities $\boldsymbol{r}$ and $\boldsymbol{p}^{\rm mech}$ in terms of $\boldsymbol{r}$, $\boldsymbol{p}^{\rm mech}$, and t. Note that these equations only involve the fields $\boldsymbol{E}$ and $\boldsymbol{B}$, and not the vector and scalar potentials $\boldsymbol{A}$ and ψ. They are therefore gauge independent.

Suppose we seek equations of motion for the quantities $(\boldsymbol{r};\boldsymbol{v})$ with t taken to be the independent variable. That is, what are desired are equations for the quantities $\ddot{\boldsymbol{r}}$ in terms of the variables $\boldsymbol{r}$, $\boldsymbol{v}$, and t. Verify that differentiating (5.32) yields the result

$$\ddot{\boldsymbol{r}} = \dot{\boldsymbol{v}} = \dot{\boldsymbol{p}}^{\rm mech}/(m\gamma) - \boldsymbol{p}^{\rm mech}[1/(m\gamma^2)](d\gamma/dt) = \dot{\boldsymbol{p}}^{\rm mech}/(m\gamma) - (\boldsymbol{v}/\gamma)(d\gamma/dt). \tag{1.5.45}$$

For the first term on the far right side of (5.45), namely the term involving $\dot{\boldsymbol{p}}^{\rm mech}$, we will use (5.31). For the second term involving the $d\gamma/dt$ factor we will use (5.38). Verify that use of (5.31) and (5.38) in (5.45) yields, in the form desired, the result

$$\ddot{\boldsymbol{r}} = [q/(\gamma m)](\boldsymbol{E}+\boldsymbol{v}\times\boldsymbol{B}) - [q/(\gamma mc^2)]\boldsymbol{v}(\boldsymbol{v}\cdot\boldsymbol{E}). \tag{1.5.46}$$

Equivalently, there is the coupled pair of first-order equations

$$\dot{\boldsymbol{r}} = \boldsymbol{v}, \tag{1.5.47}$$

$$\dot{\boldsymbol{v}} = \ddot{\boldsymbol{r}} = [q/(\gamma m)](\boldsymbol{E}+\boldsymbol{v}\times\boldsymbol{B}) - [q/(\gamma mc^2)]\boldsymbol{v}(\boldsymbol{v}\cdot\boldsymbol{E}). \tag{1.5.48}$$

1.5.3. Show that the Hamiltonian associated with the Lagrangian (5.1) is given in Cartesian coordinates by the expression

$$H = [m^2c^4 + c^2(\boldsymbol{p}^{\rm can} - q\boldsymbol{A})\cdot(\boldsymbol{p}^{\rm can} - q\boldsymbol{A})]^{1/2} + q\psi = [m^2c^4 + c^2(\boldsymbol{p}^{\rm can} - q\boldsymbol{A})^2]^{1/2} + q\psi. \tag{1.5.49}$$

Verify, using (5.27) through (5.30), that there is the result

$$H = [m^2c^4 + c^2(\boldsymbol{p}^{\rm mech}\cdot\boldsymbol{p}^{\rm mech})]^{1/2} + q\psi = \gamma mc^2 + q\psi. \tag{1.5.50}$$

Here we have used the superscripts *can* and *mech* to emphasize the distinction between *canonical* and *mechanical* momenta.

1.5.4. Let x, y, z denote the usual Cartesian coordinates. In the x, z plane introduce polar coordinates ρ, ϕ by the relations

$$x = \rho \ \cos \ \phi, \tag{1.5.51}$$
$$z = \rho \ \sin \ \phi.$$

View the triplet ρ, y, ϕ as a cylindrical coordinate system, and let $\boldsymbol{e}_\rho, \boldsymbol{e}_y, \boldsymbol{e}_\phi$ be the associated right-handed orthonormal triad. See Figure 5.1. (Note that this choice of cylindrical coordinates differs from the usual choice ρ, ϕ, z.) The purpose of this exercise is to find the canonical momenta and the Hamiltonian associated with the Lagrangian (5.1) when the cylindrical coordinates ρ, y, ϕ are used as generalized coordinates.

Verify that there are the relations

$$\boldsymbol{r} = x\boldsymbol{e}_x + y\boldsymbol{e}_y + z\boldsymbol{e}_z = \rho \ \cos \ \phi \ \boldsymbol{e}_x + y\boldsymbol{e}_y + \rho \ \sin \ \phi \ \boldsymbol{e}_z, \tag{1.5.52}$$

and

$$\boldsymbol{e}_\rho = \cos \ \phi \ \boldsymbol{e}_x + \sin \ \phi \ \boldsymbol{e}_z, \tag{1.5.53}$$
$$\boldsymbol{e}_\phi = -\sin \ \phi \ \boldsymbol{e}_x + \cos \ \phi \ \boldsymbol{e}_z,$$

so that there is also the relation

$$\boldsymbol{r} = \rho\boldsymbol{e}_\rho + y\boldsymbol{e}_y. \tag{1.5.54}$$

Note that the directions of $\boldsymbol{e}_\rho$ and $\boldsymbol{e}_\phi$ depend on ϕ, and hence on x and z. For example, the pair $\boldsymbol{e}_\rho$ and $\boldsymbol{e}_\phi$ appearing in Figure 5.1 are shown pointing in the direction they would have at the x, z location where they are displayed. Verify that $\boldsymbol{e}_\rho, \boldsymbol{e}_y, \boldsymbol{e}_\phi$ do indeed form a right-handed orthonormal triad.

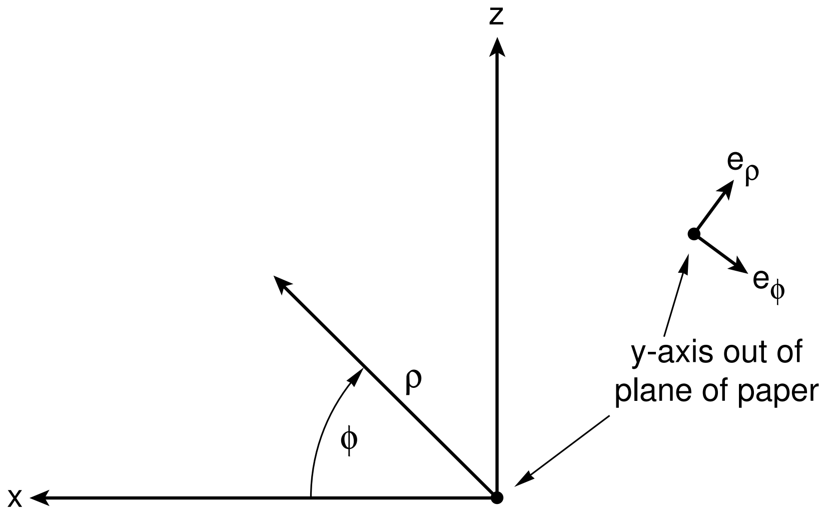


Figure 1.5.1: Illustration of the ρ, y, ϕ cylindrical coordinate system and a sample unit-vector pair $\boldsymbol{e}_\rho$ and $\boldsymbol{e}_\phi$.

Answer: It is easily verified from (5.52) and (5.53) that $\boldsymbol{e}_\rho$ and $\boldsymbol{e}_\phi$ satisfy the equations

$$\boldsymbol{e}_\rho = \frac{\partial \boldsymbol{r}}{\partial \rho} / \parallel \frac{\partial \boldsymbol{r}}{\partial \rho} \parallel, \tag{1.5.55}$$

$$\boldsymbol{e}_\phi = \frac{\partial \boldsymbol{r}}{\partial \phi} / \parallel \frac{\partial \boldsymbol{r}}{\partial \phi} \parallel,$$

and therefore are properly defined. Moreover, it is easily checked that $\boldsymbol{e}_\rho, \boldsymbol{e}_y$, and $\boldsymbol{e}_\phi$ are orthonormal and satisfy the relation

$$\boldsymbol{e}_\rho \times \boldsymbol{e}_y = \boldsymbol{e}_\phi. \tag{1.5.56}$$

They therefore form a right-handed triad.

Verify that it also follows from (5.54) or the second part of (5.52) that

$$d\boldsymbol{r} \cdot d\boldsymbol{r} = (d\rho)^2 + (dy)^2 + \rho^2 (d\phi)^2. \tag{1.5.57}$$

Consequently, the line element squared can be written in the standard form

$$d\boldsymbol{r} \cdot d\boldsymbol{r} = h_1^2 (dq_1)^2 + h_2^2 (dq_2)^2 + h_3^2 (dq_3)^2, \tag{1.5.58}$$

where

$$h_1 = 1 \quad , \quad h_2 = 1 \quad , \quad h_3 = \rho, \tag{1.5.59}$$

and

$$q_1 = \rho \quad , \quad q_2 = y \quad , \quad q_3 = \phi. \tag{1.5.60}$$

Correspondingly, the unit vectors are numbered in the order

$$\boldsymbol{e}_1 = \boldsymbol{e}_\rho \quad , \quad \boldsymbol{e}_2 = \boldsymbol{e}_y \quad , \quad \boldsymbol{e}_3 = \boldsymbol{e}_\phi. \tag{1.5.61}$$

With the above prescription, the curl of an arbitrary vector $\boldsymbol{A}$ is given by the relation

$$(\nabla \times \boldsymbol{A})_1 = \frac{1}{h_2 h_3} [\frac{\partial (h_3 A_3)}{\partial q_2} - \frac{\partial (h_2 A_2)}{\partial q_3}], \tag{1.5.62}$$

and the relations obtained from it by cyclic permutations of the coordinate indices. Here the components A_i of $\boldsymbol{A}$ are defined by the relations

$$A_i = \boldsymbol{e}_i \cdot \boldsymbol{A}. \tag{1.5.63}$$

Verify that in terms of the coordinates (5.51 there are the relations

$$\dot{x} = \dot{\rho} \ \cos\ \phi - \rho \dot{\phi} \ \sin\ \phi, \tag{1.5.64}$$

$$\dot{z} = \dot{\rho} \ \sin\ \phi + \rho \dot{\phi} \ \cos\ \phi.$$

Show from (5.51) and (5.64) that consequently there are the relations

$$\begin{aligned} \boldsymbol{v} &= d\boldsymbol{r}/dt = \dot{x}\boldsymbol{e}_x + \dot{y}\boldsymbol{e}_y + \dot{z}\boldsymbol{e}_z \\ &= \dot{\rho}(\cos\phi\ \boldsymbol{e}_x + \sin\phi\ \boldsymbol{e}_z) + \dot{y}\boldsymbol{e}_y + \rho\dot{\phi}(\cos\phi\ \boldsymbol{e}_z - \sin\phi\ \boldsymbol{e}_x) \\ &= \dot{\rho}\boldsymbol{e}_\rho + \dot{y}\boldsymbol{e}_y + \rho\dot{\phi}\boldsymbol{e}_\phi, \end{aligned} \tag{1.5.65}$$

$$v^2 = \dot{x}^2 + \dot{y}^2 + \dot{z}^2 = \dot{\rho}^2 + \dot{y}^2 + \rho^2 \dot{\phi}^2, \tag{1.5.66}$$

$$\boldsymbol{v} \cdot \boldsymbol{A} = \dot{x} A_x + \dot{y} A_y + \dot{z} A_z = \dot{\rho} A_\rho + \dot{y} A_y + \rho \dot{\phi} A_\phi, \tag{1.5.67}$$

where

$$A_\rho = \cos\ \phi\ A_x + \sin\ \phi\ A_z = \boldsymbol{e}_\rho \cdot \boldsymbol{A}, \tag{1.5.68}$$

$$A_\phi = -\sin\ \phi\ A_x + \cos\ \phi\ A_z = \boldsymbol{e}_\phi \cdot \boldsymbol{A}.$$

Using these results, show that the Lagrangian (5.1) can also be written in the form

$$L = -mc^2[1 - (\dot\rho^2 + \dot y^2 + \rho^2\dot\phi^2)/c^2]^{1/2} - q\psi + q(\dot\rho A_\rho + \dot y A_y + \rho\dot\phi A_\phi). \tag{1.5.69}$$

The Hamiltonian H corresponding to the Lagrangian L given by (5.69) can now be found by the usual procedure. Show that for the conjugate momenta there are the results

$$p_\rho = \frac{\partial L}{\partial \dot\rho} = \frac{m\dot\rho}{\sqrt{1 - (\dot\rho^2 + \dot y^2 + \rho^2\dot\phi^2)/c^2}} + qA_\rho,$$

$$p_y = \frac{\partial L}{\partial \dot y} = \frac{m\dot y}{\sqrt{1 - (\dot\rho^2 + \dot y^2 + \rho^2\dot\phi^2)/c^2}} + qA_y,$$

$$p_\phi = \frac{\partial L}{\partial \dot\phi} = \frac{m\rho^2\dot\phi}{\sqrt{1 - (\dot\rho^2 + \dot y^2 + \rho^2\dot\phi^2)/c^2}} + q\rho A_\phi. \tag{1.5.70}$$

Finally, verify that H is given by the relation

$$\begin{aligned} H &= \dot\rho p_\rho + \dot y p_y + \dot\phi p_\phi - L \qquad (1.5.71)\\ &= \{m^2c^4 + c^2[(p_\rho - qA_\rho)^2 + (p_y - qA_y)^2 + (p_\phi/\rho - qA_\phi)^2]\}^{1/2} + q\psi. \end{aligned}$$

Here is a cautionary note: Let $\boldsymbol{p}$ be the momentum vector as defined by (5.27). Then, from (5.65) and (5.70), verify that there are the results

$$p_\rho = \boldsymbol{p} \cdot \boldsymbol{e}_\rho, \tag{1.5.72}$$

$$p_y = \boldsymbol{p} \cdot \boldsymbol{e}_y, \tag{1.5.73}$$

but

$$p_\phi = \rho\boldsymbol{p} \cdot \boldsymbol{e}_\phi \neq \boldsymbol{p} \cdot \boldsymbol{e}_\phi. \tag{1.5.74}$$

1.5.5. Show that a uniform electric field in the z direction can be derived from the scalar and vector potentials

$$\psi = 0, \tag{1.5.75}$$

$$\boldsymbol{A} = -Et\boldsymbol{e}_z.$$

1.5.6. Show that a uniform electric field in the z direction can be derived from the scalar and vector potentials

$$\psi = -Ez, \tag{1.5.76}$$

$$\boldsymbol{A} = 0.$$

1.5.7. Show that a uniform vertical magnetic field $\boldsymbol{B} = B\boldsymbol{e}_y$, such as that produced by an idealized (normal) dipole bending magnet, can be derived from the scalar and vector potentials

$$\psi = 0, \tag{1.5.77}$$

$$\boldsymbol{A} = -Bx\boldsymbol{e}_z.$$

Assuming the magnet has iron pole faces, sketch the pole faces and windings required to produce such a field, and label the pole faces N and S. Also sketch the magnetic field lines and the directions the current must flow in the windings.

1.5.8. Show that when cylindrical coordinates ρ, y, ϕ are used, a uniform magnetic field in the y direction can be derived from the scalar and vector potentials

$$\psi = 0, \tag{1.5.78}$$

$$\boldsymbol{A} = -(\rho/2)B\boldsymbol{e}_\phi.$$

Answer: See Figure 5.1. From (5.62) one has the results

$$B_\rho = (\nabla \times \boldsymbol{A})_\rho = \frac{\partial A_\phi}{\partial y} - \frac{1}{\rho}\frac{\partial A_y}{\partial \phi} = 0, \tag{1.5.79}$$

$$B_y = (\nabla \times \boldsymbol{A})_y = \frac{1}{\rho}\frac{\partial A_\rho}{\partial \phi} - \frac{1}{\rho}\frac{\partial}{\partial \rho}(\rho A_\phi) = \frac{1}{\rho}\frac{\partial}{\partial \rho}(\frac{\rho^2 B}{2}) = B,$$

$$B_\phi = (\nabla \times \boldsymbol{A})_\phi = \frac{\partial A_y}{\partial \rho} - \frac{\partial A_\rho}{\partial y} = 0.$$

1.5.9. Review Exercises 5.1 and 5.2. Suppose there is a *uniform* static magnetic field given by

$$\boldsymbol{B} = B\boldsymbol{e}_y \text{ with } B > 0, \tag{1.5.80}$$

and no electric field. Consider charged-particle motion in this simple case. Show that a possible trajectory is uniform motion on a circle of radius ρ in the $y = 0$ plane, and show that $p = ||\boldsymbol{p}^{\text{mech}}|| = ||\gamma m\boldsymbol{v}||$, the magnitude of the mechanical momentum given by (5.28), is related to B and ρ by the equation

$$B\rho = p/|q|. \tag{1.5.81}$$

The product $B\rho$ is called the *magnetic rigidity.* In Accelerator Physics it is common to characterize the mechanical momentum of a particle by its equivalent magnetic rigidity.

Show that, in the case of uniform circular motion, the circle is traced out with angular velocity ω given by the relation

$$\omega = |q|B/(\gamma m). \tag{1.5.82}$$

The quantity ω, particularly in the nonrelativistic limit $\gamma \simeq 1$, is called the *cyclotron frequency.*

1.5.10. Show that a magnetic quadrupole field with midplane ($\pm y$) symmetry can be derived from the scalar and vector potentials

$$\psi = 0, \tag{1.5.83}$$

$$\boldsymbol{A} = -(Q/2)(x^2 - y^2)\boldsymbol{e}_z.$$

Answer:

$$B_x = Qy, \tag{1.5.84}$$

$$B_y = Qx, \tag{1.5.85}$$

$$B_z = 0. \tag{1.5.86}$$

Assuming the quadrupole magnet has iron pole faces, sketch the pole faces and windings required to produce such a field, and label the pole faces N and S. Also sketch the magnetic field lines and the directions the current must flow in the windings.

1.5.11. Show that a magnetic sextupole field with midplane ($\pm y$) symmetry can be derived from the scalar and vector potentials

$$\psi = 0, \tag{1.5.87}$$

$$\boldsymbol{A} = -(S/3)(x^3 - 3xy^2)\boldsymbol{e}_z.$$

Assuming the sextupole magnet has iron pole faces, sketch the pole faces and windings required to produce such a field, and label the pole faces N and S. Also sketch the magnetic field lines and the directions the current must flow in the windings.

1.5.12. Let f be a function of the ℓ variables $z_1, \cdots z_\ell$. The function f is said to be *homogeneous* of degree m if it satisfies the relation

$$f(\lambda z) = \lambda^m f(z) \text{ (for } \lambda > 0). \tag{1.5.88}$$

Evidently homogeneous polynomials provide examples of homogeneous functions. However, a function need not be polynomial to be homogeneous. Verify, for example, that the function

$$f = (ax^2 + bxy + cy^2)^{1/2} \tag{1.5.89}$$

is homogeneous of degree 1. Show that if f is homogeneous of degree m, then the functions $(\partial f/\partial z_j)$ are homogeneous of degree $(m-1)$. Show that if f is homogeneous of degree m, then it satisfies *Euler's* relation

$$\sum_a z_a(\partial f/\partial z_a) = mf, \tag{1.5.90}$$

and conversely.

1.5.13. Given a Lagrangian L, one can find the associated Hamiltonian H by a Legendre transformation provided (5.6) is satisfied. Consider the inverse question. Given H, show that one can find an associated L using the *inverse* Legendre transformation provided by rewriting (5.8) in the form

$$L(q, \dot{q}, t) = \sum_i p_i \dot{q}_i - H(q, p, t) \tag{1.5.91}$$

with the proviso that

$$\dot{q}_i = \partial H/\partial p_i. \tag{1.5.92}$$

Show, as required, that the variables p can be eliminated in terms of the $\dot{q}$'s, provided

$$\det(\partial \dot{q}_i/\partial p_j) \neq 0. \tag{1.5.93}$$

Verify that

$$\partial \dot{q}_i/\partial p_j = \partial^2 H/\partial p_i \partial p_j \tag{1.5.94}$$

so that (5.93) is equivalent to the condition

$$\det(\partial^2 H/\partial p_i \partial p_j) \neq 0. \tag{1.5.95}$$

In analogy to (5.6), when (5.95) holds we will call this the nonsingular Hamiltonian case.

Suppose, as is true in these kinds of calculations, that the variables q are held fixed. Show that then, by the chain rule, there is the differential relation

$$d\dot{q}_i = \sum_j (\partial \dot{q}_i/\partial p_j) dp_j \tag{1.5.96}$$

which can be written in the matrix-vector form

$$d\dot{q} = T dp \tag{1.5.97}$$

where T is the matrix

$$T_{ij} = (\partial \dot{q}_i/\partial p_j). \tag{1.5.98}$$

Show, in view of (5.93), that (5.97) may be solved for the dp to yield the relation

$$dp = T^{-1} d\dot{q}. \tag{1.5.99}$$

Argue, on the other hand, that there is the relation

$$dp = U d\dot{q} \tag{1.5.100}$$

where U is the matrix

$$U_{ij} = (\partial p_i/\partial \dot{q}_j). \tag{1.5.101}$$

Verify that comparison of (5.99) and (5.100) gives the result

$$U = T^{-1}. \tag{1.5.102}$$

Finally, show that there is the two-directional logical implication

$$\det(\partial^2 L/\partial \dot{q}_i \partial \dot{q}_j) \neq 0 \;\Leftrightarrow\; \det(\partial^2 H/\partial p_i \partial p_j) \neq 0. \tag{1.5.103}$$

Thus, if a Legendre transformation can be made in one direction, it can also be made in the reverse direction. The nonsingular Lagrangian case leads to the nonsingular Hamiltonian case, and vice versa.

1.5.14. Review Subsection 5.2 and Exercise 5.13. Show that if the Hamiltonian $H(q,p,t)$ is homogeneous of degree one in the p_i, then (5.95) fails to hold, and we are dealing with the singular Hamiltonian case. Make an analysis of this case similar to that which was done for the Lagrangian case of Subsection 5.2. Show, in particular, that the Lagrangian associated with H vanishes identically.

1.5.15. Review Exercises 5.3 and 5.13. Show, by making an inverse Legendre transformation, that the Lagrangian associated with the Hamiltonian (5.49) is the Lagrangian (5.1).

1.5.16. Let $x(\tau), y(\tau)$ be a parameterized path in two-dimensional space. Let $\mathcal{A}$ be the *distance* functional defined by

$$\mathcal{A} = \int ds = \int (ds/d\tau)d\tau = \int (\dot{x}^2 + \dot{y}^2)^{1/2} d\tau \tag{1.5.104}$$

where

$$(ds)^2 = (dx)^2 + (dy)^2 \tag{1.5.105}$$

and a dot denotes $d/d\tau$. Specifically, consider all paths for $\tau \in [0,1]$ with the end points

$$x(0) = y(0) = 0, \tag{1.5.106}$$

$$x(1) = y(1) = 1. \tag{1.5.107}$$

Then we may write

$$\mathcal{A} = \int_0^1 L(\dot{x}, \dot{y}) d\tau \tag{1.5.108}$$

with

$$L = (\dot{x}^2 + \dot{y}^2)^{1/2}. \tag{1.5.109}$$

Verify that L is homogeneous of degree one in the quantities $\dot{x}, \dot{y}$, and verify by direct calculation that (5.6) fails. Visualize the paths $x(\tau), y(\tau)$ as curves in the three-dimensional x, y, τ space. Show that there are an *infinity* of curves (corresponding to different parameterizations) that extremize $\mathcal{A}$. Show that all of these curves, when projected onto the x, y plane, fall on the straight line joining $(0,0)$ to $(1,1)$. Specifically, consider all curves of the form

$$x(\tau) = \tau + f(\tau), \tag{1.5.110}$$

$$y(\tau) = \tau + f(\tau), \tag{1.5.111}$$

where f is any function satisfying

$$f(0) = f(1) = 0, \tag{1.5.112}$$

$$|f'(\tau)| \leq 1. \tag{1.5.113}$$

Show that all these curves extremize $\mathcal{A}$. Show that for all these curves $\mathcal{A}$ has the value

$$\mathcal{A} = \sqrt{2}. \tag{1.5.114}$$

1.5.17. Review Exercise 5.16. Instead of using the parameterization $x(\tau), y(\tau)$, simply write

$$y = y(x), \tag{1.5.115}$$

which is equivalent to taking the coordinate x to be the parameter,

$$\tau = x, \tag{1.5.116}$$

and thereby providing information about the parameterization.

Verify that in this case the distance functional takes the form

$$\mathcal{A} = \int ds = \int (ds/dx)dx = \int [1 + (y')^2]^{1/2} dx \tag{1.5.117}$$

where a prime denotes d/dx. Specifically, consider all paths $y(x)$ with the end points

$$y(0) = 0, \tag{1.5.118}$$

$$y(1) = 1. \tag{1.5.119}$$

Show that now we may write

$$\mathcal{A} = \int_0^1 L(y')dx \tag{1.5.120}$$

with

$$L = [1 + (y')^2]^{1/2}. \tag{1.5.121}$$

Verify that this L is *not* homogeneous of degree one in the quantity y'. Show that

$$p_y = \partial L/\partial y' = y'/[1 + (y')^2]^{1/2}, \tag{1.5.122}$$

and verify that this relation can be solved for y' in terms of p_y to give the result

$$y' = p_y/(1 - p_y^2)^{1/2}. \tag{1.5.123}$$

Therefore, we are dealing with the nonsingular case. Verify that, in fact,

$$\partial^2 L/(\partial y')^2 \neq 0. \tag{1.5.124}$$

so that (5.6) holds. Show that the Hamiltonian associated with the Lagrangian (5.121) is given by the relation

$$H = -(1 - p_y^2)^{1/2}. \tag{1.5.125}$$

Show that the solution to Lagrange's (or Hamilton's) equations in this case takes the form

$$y(x) = ax + b \tag{1.5.126}$$

where a and b are constants to be determined by the end-point conditions (5.118) and (5.119). Show that imposition of the end-point conditions yields the *unique* solution

$$y(x) = x, \tag{1.5.127}$$

the straight line between the end points. Verify that for this path $\mathcal{A}$ has the value (5.114).

1.5.18. Exercises 5.16 and 5.17 treated a simple example of finding *geodesics*, the shortest paths between two points, in terms of the distance functional. It involved the Lagrangians (5.109) and (5.121). Consider as before the parameterized path $x(\tau), y(\tau)$, but now employ instead the Lagrangian

$$\hat{L} = (1/2)(\dot{x}^2 + \dot{y}^2) \tag{1.5.128}$$

and seek to extremize what is called the *energy* functional $\hat{\mathcal{A}}$ defined by

$$\hat{\mathcal{A}} = \int \hat{L} d\tau. \tag{1.5.129}$$

The solution to this goal is an example of an *affine* geodesic. For a further description of geodesics and affine geodesics, see Exercise 6.16.

Verify that the Hessian of $\hat{L}$ is invertible so that we are dealing with the nonsingular case. Show that the Lagrange equations associated with $\hat{L}$ have, for the end-point conditions (5.106) and (5.107), the *unique* solution

$$x(\tau) = \tau, \tag{1.5.130}$$

$$y(\tau) = \tau. \tag{1.5.131}$$

Note that the extremizing path is again the straight line between the end points. Show that for this path $\hat{\mathcal{A}} = 1$.

1.5.19. This problem concerns fluid flow in two dimensions and its relation to Hamiltonian dynamics. Consider a fluid flowing in two dimensions and let $v_x(x, y, t)$ and $v_y(x, y, t)$ be the components of the velocity $\boldsymbol{v}$ of a small portion of the fluid at the point with coordinates x and y. (It is assumed that there is no motion/velocity in the z direction and that $\boldsymbol{v}$ does not depend on z.) We are interested in the solutions to the coupled pair of differential equations

$$\dot{x} = v_x(x, y, t), \tag{1.5.132}$$

$$\dot{y} = v_y(x, y, t). \tag{1.5.133}$$

Moreover assume that the flow is divergence free (which follows from the assumption that the fluid density remains constant, i.e., the flow is incompressible) so that

$$\nabla \cdot \boldsymbol{v} = \partial_x v_x + \partial_y v_y = 0. \tag{1.5.134}$$

Define an associated two-dimensional vector field $\boldsymbol{u}(x, y, t)$ by the rule

$$u_x = -v_y, \tag{1.5.135}$$

$$u_y = v_x. \tag{1.5.136}$$

Verify that(5.134) through (5.136) imply the relation

$$\partial_x u_y = \partial_x v_x = -\partial_y v_y = \partial_y u_x. \tag{1.5.137}$$

That is, the differential form associated with the vector field $\boldsymbol{u}(x, y, t)$ is closed. Consequently there is a function $\psi(x, y, t)$ defined by

$$\psi(x, y, t) = \int^{x,y} [u_x(x', y', t)dx' + u_y(x', y', t)dy'] \tag{1.5.138}$$

such that

$$u_x = \partial_x \psi, \tag{1.5.139}$$

$$u_y = \partial_y \psi. \tag{1.5.140}$$

See Exercise 6.1.1.

Verify that the results obtained so far can be combined to yield the differential equation pair

$$\dot{x} = \partial_y \psi, \tag{1.5.141}$$

$$\dot{y} = -\partial_x \psi. \tag{1.5.142}$$

Evidently these are Hamilton's equations with ψ playing the role of the Hamiltonian and x and y playing the roles of q and p.

In the case that $\boldsymbol{v}$ is time independent, $\boldsymbol{u}$ and therefore ψ will have no explicit time dependence. Then, because ψ is a Hamiltonian, there will be the relation

$$\psi\{x(t), y(t)\} = \text{constant} \tag{1.5.143}$$

on any solution of the pair (5.132) and (5.133). Call the pair $x(t)$ and $y(t)$ a *flow line.* According to (5.143), lines of constant ψ (*level lines* of ψ) are flow lines. For this reason (and the fact that Lagrange first arrived at this result in 1781) ψ is called a (Lagrange) *stream function.*

It is also possible to set up a steam function in three dimensions in the case of axial symmetry. The result is called a *Stokes* (1819-1903) stream function. Let ρ,ϕ,z be the usual choice of cylindrical coordinates with associated unit vectors $\boldsymbol{e}_\rho$,$\boldsymbol{e}_\phi$,$\boldsymbol{e}_z$. See (15.2.12) through (15.2.14), (15.2.20) through (15.2.25), and Exercise 15.2.2. Suppose the fluid velocity has only $\boldsymbol{e}_\rho$ and $\boldsymbol{e}_z$ components and does not depend on ϕ,

$$\boldsymbol{v}(\rho, z, t) = v_\rho(\rho, z, t)\boldsymbol{e}_\rho + v_z(\rho, z, t)\boldsymbol{e}_z. \tag{1.5.144}$$

Recall that in general there is the relation

$$\boldsymbol{v} = d\boldsymbol{r}/dt = \dot{\rho}\boldsymbol{e}_\rho + \rho\dot{\phi}\boldsymbol{e}_\phi + \dot{z}\boldsymbol{e}_z \tag{1.5.145}$$

so that we are then interested in the solutions to the coupled pair

$$\dot{\rho} = v_\rho(\rho, z, t), \tag{1.5.146}$$

$$\dot{z} = v_z(\rho, z, t). \tag{1.5.147}$$

Again assume the flow is divergence free so that

$$\nabla \cdot \boldsymbol{v} = (1/\rho)\partial_\rho(\rho v_\rho) + \partial_z v_z = 0. \tag{1.5.148}$$

Multiply the last two pieces of (5.148) by ρ to get the result

$$\partial_\rho(\rho v_\rho) + \partial_z(\rho v_z) = 0. \tag{1.5.149}$$

Define a vector $\boldsymbol{u}(\rho, z, t)$ by the rule

$$u_\rho = -(\rho v_z), \tag{1.5.150}$$

$$u_z = (\rho v_\rho). \tag{1.5.151}$$

Verify from (5.149) through (5.151) that there is the relation

$$\partial_\rho u_z = \partial_\rho(\rho v_\rho) = -\partial_z(\rho v_z) = \partial_z u_\rho. \tag{1.5.152}$$

That is, the differential form associated with the vector field $\boldsymbol{u}(\rho, z, t)$ is closed. Consequently there is a function $\psi_S(\rho, z, t)$ defined by

$$\psi_S(\rho, z, t) = \int^{\rho,z} [u_\rho(\rho', z', t)d\rho' + u_z(\rho', z', t)dz'] \tag{1.5.153}$$

such that

$$u_\rho = \partial_\rho \psi_S, \tag{1.5.154}$$

$$u_z = \partial_z \psi_S. \tag{1.5.155}$$

Here we have added a subscript S to ψ to distinguish it from Lagrange's stream function and to honor Stokes.

Verify that the results obtained so far for the case of axial symmetry can be combined to yield the differential equation pair

$$\dot{\rho} = (1/\rho)\partial_z \psi_S, \tag{1.5.156}$$

$$\dot{z} = -(1/\rho)\partial_\rho \psi_S. \tag{1.5.157}$$

Because of the $(1/\rho)$ factor these are not Hamilton's equations. Nevertheless we will be able to draw from them similar conclusions.

In the case that $\boldsymbol{v}$ is time independent, $\boldsymbol{u}$ and therefore ψ_S will have no explicit time dependence. For this case let us compute the change in ψ_S along a flow line. Verify that so doing yields, with the aid of (5.156) and (5.157), the result

$$d\psi_S/dt = \partial_\rho \psi_S \dot{\rho} + \partial_z \psi_S \dot{z} = -\rho \dot{z} \dot{\rho} + \rho \dot{\rho} \dot{z} = 0. \tag{1.5.158}$$

Thus ψ_S, which is called the Stokes stream function, has the property that lines of constant ψ_S (level lines of ψ_S) are flow lines.

1.6 Hamilton's Equations with a Coordinate as an Independent Variable

In the usual Hamiltonian formulation (as in the usual Lagrangian formulation) the time t plays the distinguished role of an *independent* variable, and all the q's and p's are *dependent* variables. That is, the canonical variables are viewed as functions $q(t), p(t)$ of the independent variable t.

In some cases, it is more convenient to take some coordinate to be the independent variable rather than the time. So doing may facilitate the use of maps. For example, consider the passage of a collection of particles through a rectangular magnet such as is shown in Figures 6.1 and 6.2. In such a situation, particles with different initial conditions will require different times to pass through the magnet. If the quantities of interest are primarily the locations and momenta of the particles as they leave the exit face of the magnet, then it would clearly be more convenient to use some coordinate that measures the progress of a particle through the magnet as an independent variable. With such a choice, the relation between entering coordinates and momenta and exiting coordinates and momenta could be treated as a transfer map.

In the case of a magnet with parallel faces as shown in Figures 6.1 and 6.2, a convenient independent variable would be the z coordinate. In the case of a wedge magnet as shown in Figure 6.3, a convenient independent variable would be the angle ϕ of a cylindrical coordinate triad ρ, y, ϕ. See Exercise 5.4.

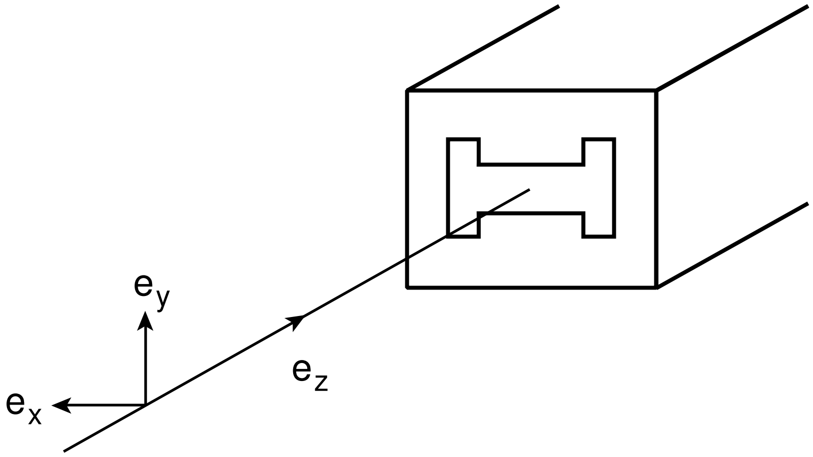


Figure 1.6.1: Typical choice of a Cartesian coordinate system for the description of charged-particle trajectories in a magnet.

Suppose some coordinate is indeed chosen to be the independent variable. Is it then still possible to have a Hamiltonian (or Lagrangian) formulation of the equations of motion? The answer in general is *yes* as is shown by the following theorem:

Theorem 1.6.1. *Suppose $H(q, p, t)$ is a Hamiltonian for a system having n degrees of freedom. Suppose further that $\dot{q}_1 = \partial H/\partial p_1 \neq 0$ for some interval of time T in some region R of the phase space described by the $2n$ variables $(q_1, \ldots, q_n)$ and $(p_1, \ldots, p_n)$. Then in*

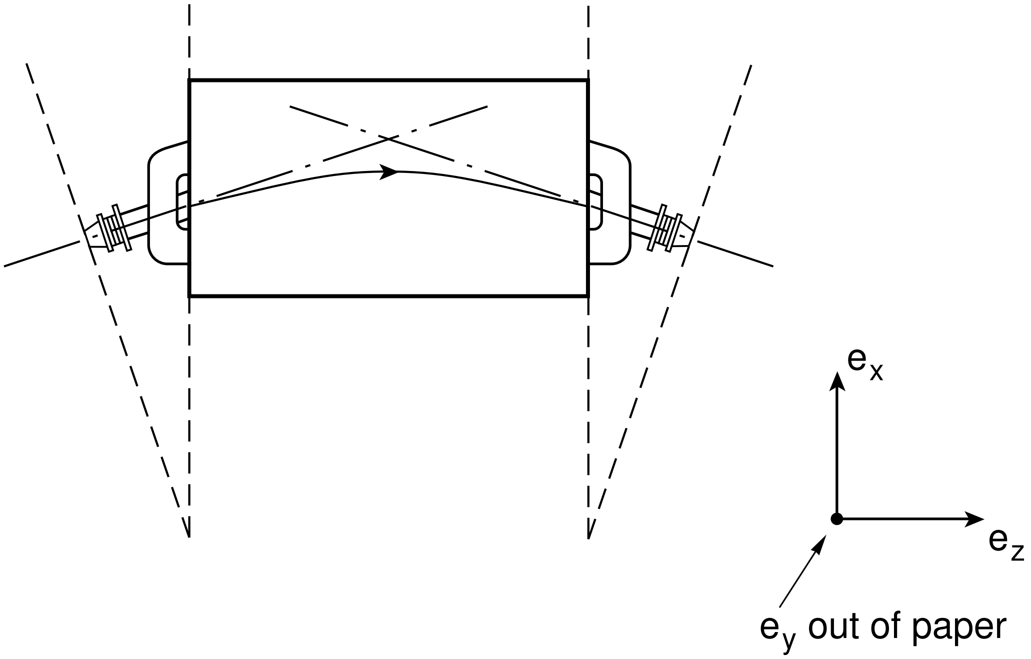


Figure 1.6.2: Top view of a particle trajectory in a rectangular magnet.

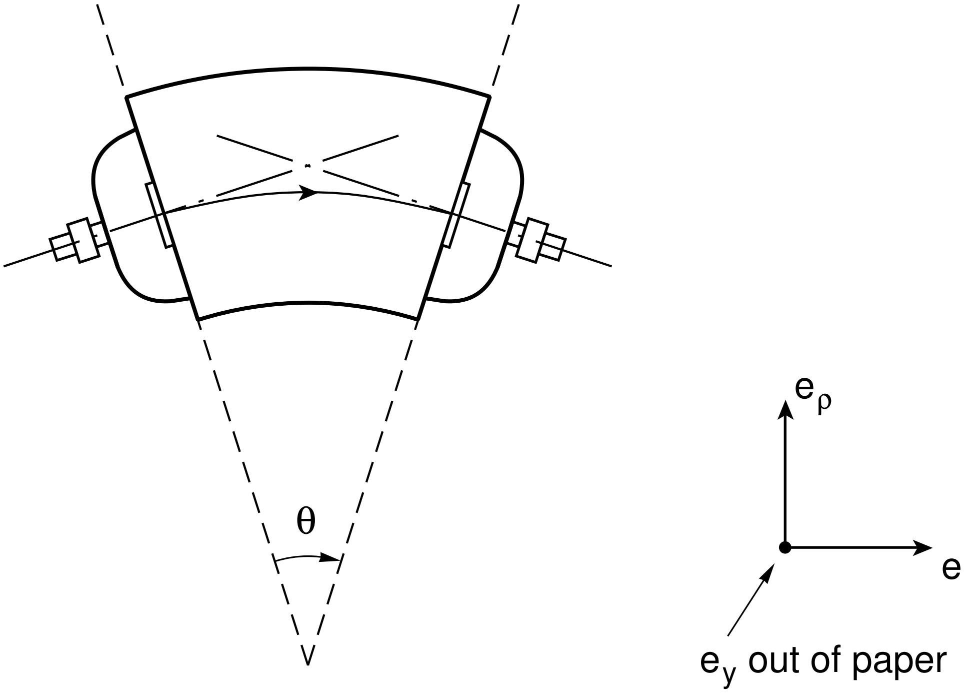


Figure 1.6.3: Top view of a particle trajectory in a wedge magnet. The trajectory is conveniently described using the cylindrical coordinates ρ, y, ϕ. See Figure 5.1.

this region and time interval, q_1 can be introduced as an independent variable in place of the time t. Moreover, the equations of motion with q_1 as an independent variable can be obtained from a Hamiltonian that will be called K.

Proof. Consider the $2n-2$ quantities $(q_2, \ldots, q_n)$ and $(p_2, \ldots, p_n)$. They obey Hamilton's equations of motion

$$\begin{aligned} \dot{q}_i &= \partial H/\partial p_i, \quad & i = 2, \ldots, n, \\ \dot{p}_i &= -\partial H/\partial q_i, \quad & i = 2, \ldots, n. \end{aligned} \tag{1.6.1}$$

Denote total derivates with respect to q_1 by a prime. Then, applying the chain rule to equations (6.1), one finds the relations

$$\begin{aligned} q_i' &= dq_i/dq_1 = (dq_i/dt)(dt/dq_1) = (\partial H/\partial p_i)(\partial H/\partial p_1)^{-1}, \\ p_i' &= dp_i/dq_1 = (dp_i/dt)(dt/dq_1) = -(\partial H/\partial q_i)(\partial H/\partial p_1)^{-1}. \end{aligned} \tag{1.6.2}$$

To these $2n-2$ relations it is convenient to add two more. First, suppose the time t is added to the list of *coordinates* as a *dependent* variable. Then one immediately has the relation

$$t' = dt/dq_1 = (dq_1/dt)^{-1} = (\partial H/\partial p_1)^{-1}. \tag{1.6.3}$$

Second, suppose the quantity p_t defined by writing $p_t = -H$ is formally added to the list of momenta. Then, using (5.11) and (5.14), one finds the relation

$$p_t' = dp_t/dq_1 = (dp_t/dt)(dt/dq_1) = -(\partial H/\partial t)(\partial H/\partial p_1)^{-1}. \tag{1.6.4}$$

Equations (6.2) through (6.4) are the desired equations of motion for the $2n$ variables $(t, q_2, \ldots, q_n)$ and $(p_t, p_2, \ldots, p_n)$ with q_1 as an independent variable. What remains to be shown is that the quantities on the right sides of these equations can be obtained by applying the standard rules to some Hamiltonian K.

Look once again at the defining relation for p_t,

$$p_t = -H(q, p, t). \tag{1.6.5}$$

Suppose that this relation is solved for p_1 to give a relation of the form

$$p_1 = -K(t, q_2, \ldots, q_n; p_t, p_2, \ldots, p_n; q_1). \tag{1.6.6}$$

Such an inversion is possible according to the inverse function theorem because $\partial H/\partial p_1 \neq 0$ by assumption. Then, as the notation is intended to suggest, K is the desired new Hamiltonian.

To see that this is so, compute the total differential of (6.5) to find the relation

$$dp_t = -(\partial H/\partial t)dt - \sum_i (\partial H/\partial q_i)dq_i - \sum_i (\partial H/\partial p_i)dp_i. \tag{1.6.7}$$

Now solve (6.7) for dp_1 to get the relation

$$dp_1 = \left(\frac{\partial H}{\partial p_1}\right)^{-1} \left[-dp_t - (\partial H/\partial t)dt - \sum_i (\partial H/\partial q_i)dq_i - \sum_{i \neq 1} (\partial H/\partial p_i)dp_i\right]. \tag{1.6.8}$$

Also, compute the total differential of (6.6) to find the relation

$$dp_1 = -(\partial K/\partial p_t)dp_t - (\partial K/\partial t)dt - \sum_i (\partial K/\partial q_i)dq_i - \sum_{i\neq 1}(\partial K/\partial p_i)dp_i. \tag{1.6.9}$$

Upon comparing (6.8) and (6.9), and looking at equations (6.1–6.4), one obtains the advertised result:

$$\begin{aligned}
\partial K/\partial p_t &= (\partial H/\partial p_1)^{-1} = t', \\
\partial K/\partial p_i &= (\partial H/\partial p_i)(\partial H/\partial p_1)^{-1} = q_i', \quad i = 2, \ldots, n, \\
\partial K/\partial t &= (\partial H/\partial t)(\partial H/\partial p_1)^{-1} = -p_t', \\
\partial K/\partial q_i &= (\partial H/\partial q_i)(\partial H/\partial p_1)^{-1} = -p_i', \quad i = 2, \ldots, n.
\end{aligned} \tag{1.6.10}$$

That is, the indicated partial derivates of K do indeed produce the required right sides of equations (6.2) through (6.4). Note that according to equations (6.10), the quantity p_t may be viewed as the momentum canonically conjugate to the time t. □

How might one have guessed that (6.6) gives the desired Hamiltonian? One way is to employ (a modified) Hamilton's principle. According to this principle, the *action* $\mathcal{A}$ associated with a path in phase space should be defined by the relation

$$\mathcal{A} = \int dt(\sum_{i=1}^{n} p_i\dot{q}_i - H) = \int (\sum_{i=1}^{n} p_i dq_i - Hdt); \tag{1.6.11}$$

and the equations of motion (5.11) through (5.14) follow from requiring that $\mathcal{A}$ be an extremum,

$$\delta\mathcal{A} = 0, \tag{1.6.12}$$

and use of the calculus of variations. Now introduce the notation

$$q_{n+1} = t \ , \quad p_{n+1} = -H = p_t. \tag{1.6.13}$$

With this notation the action (6.11) takes the symmetrical form

$$\mathcal{A} = \int \sum_{i=1}^{n+1} p_i dq_i. \tag{1.6.14}$$

In this form it is evident that we may regard any of the p_i as being related to some Hamiltonian. Suppose we choose p_1, and then write (6.6). When this is done, $\mathcal{A}$ takes the form

$$\begin{aligned}
\mathcal{A} &= \int \sum_{i=2}^{n+1} p_i dq_i - K dq_1 = \int dq_1[\sum_{i=2}^{n+1} p_i(dq_i/dq_1) - K] \\
&= \int dq_1(\sum_{i=2}^{n+1} p_i q_i' - K).
\end{aligned} \tag{1.6.15}$$

Since the requirement (6.12) is intrinsic in nature and therefore coordinate independent, it must also hold when $\mathcal{A}$ is written in the form (6.15). (An extremum is an extremum independent of parameterization.) But then use of (6.12), and application of the calculus of variations to (6.15), give (6.10).

Exercises

1.6.1. Find the Hamiltonian K corresponding to the Hamiltonian H given by (5.49) when the z coordinate is taken to be the independent variable. Assume that $\dot{z} > 0$ for the trajectories in question. Answer:

$$K = -[(p_t + q\psi)^2/c^2 - m^2c^2 - (p_x - qA_x)^2 - (p_y - qA_y)^2]^{1/2} - qA_z. \tag{1.6.16}$$

Here the quantities p_x and p_y denote *canonical* momenta. Note that according to (6.5), p_t is usually negative. Show, using (5.50), that

$$p_t = -[m^2c^4 + c^2(\boldsymbol{p}^{\rm mech} \cdot \boldsymbol{p}^{\rm mech})]^{1/2} - q\psi = -\gamma mc^2 - q\psi. \tag{1.6.17}$$

1.6.2. Find the Hamiltonian K corresponding to the Hamiltonian H given by (5.71) when the coordinate ϕ is taken to be the independent variable. Assume that $\dot{\phi} > 0$ for trajectories of interest. Answer:

$$K = -\rho[(p_t + q\psi)^2/c^2 - m^2c^2 - (p_\rho - qA_\rho)^2 - (p_y - qA_y)^2]^{1/2} - q\rho A_\phi. \tag{1.6.18}$$

Here the quantities p_ρ and p_y denote *canonical* momenta. Verify that (6.17) continues to hold.

1.6.3. The derivation of (6.10) based on the modified Hamilton's principle, Equations (6.11) through (6.15), is a bit heuristic. Make the derivation more precise by indicating exactly what changes of variables are being made; what the limits of integration are in (6.11), (6.14), and (6.15); what (6.12) means; etc. Begin your discussion by reviewing exactly how (5.11) and (5.14) follow from (6.11) and (6.12). Hint: To derive (5.14) from Hamilton's principle, consider variations in t as well as those in the q_i and p_i. That is, introduce a new independent variable τ such that the dependent variables are parameterized in the form $q_i(\tau)$, $p_i(\tau)$, $t(\tau)$.

1.6.4. How might one have guessed that p_t should be defined as in (6.5)? According to Hamilton's principle stated in Lagrangian terms, the action $\mathcal{A}$ associated with a path in configuration space is given by the relation

$$\mathcal{A} = \int_{t^1}^{t^2} L(q, \dot{q}, t)dt. \tag{1.6.19}$$

Suppose we introduce a new independent variable τ such that the time t and the other dependent variables are parameterized in the form $t(\tau)$, $q_i(\tau)$. Then, using a prime to denote differentiation with respect to τ, we have the relation

$$dt = (dt/d\tau)d\tau = t'd\tau, \tag{1.6.20}$$

$$\dot{q}_i = dq_i/dt = (dq_i/d\tau)(d\tau/dt) = q_i'/t'. \tag{1.6.21}$$

Correspondingly, the action (6.18) takes the form

$$\mathcal{A} = \int Ldt = \int Lt'd\tau = \int [L(q, q'/t', t)t']d\tau, \tag{1.6.22}$$

and we see that in terms of τ there is an *effective* Lagrangian L^{eff} given by the expression

$$L^{\text{eff}}(q,t;q',t') = L(q,q'/t',t)t'. \tag{1.6.23}$$

Justify this assertion by treating all the necessary details. (See the analogous case of Exercise 6.3.) Following the standard procedure (5.7), the momentum p_t canonically conjugate to the variable t is defined by the relation

$$p_t = \partial L^{\text{eff}}/\partial t'. \tag{1.6.24}$$

By using (6.23) and (6.24) and the chain rule show that

$$p_t = L - \sum_i p_i \dot{q}_i = -H. \tag{1.6.25}$$

The Lagrange equation for the t coordinate is

$$\frac{d}{d\tau}\frac{\partial L^{\text{eff}}}{\partial t'} - \frac{\partial L^{\text{eff}}}{\partial t} = 0. \tag{1.6.26}$$

Show that p_t is conserved if L (and therefore L^{eff}) does not explicitly contain the time t. In view of (6.25), we may say that energy (the Hamiltonian) is conserved if the time is an *ignorable* coordinate. Use (5.14) to obtain the same result.

1.6.5. Review Exercise 6.4. Suppose we wish to find the Hamiltonian H^{eff} associated with L^{eff}. To do so we must first compute all the conjugate momenta p_i^{eff}. Using (6.23), show that

$$p_i^{\text{eff}} = \partial L^{\text{eff}}/\partial q_i' = \partial L/\partial \dot{q}_i = p_i. \tag{1.6.27}$$

Next, following the rule (5.8), find the result

$$\begin{aligned} H^{\text{eff}} &= p_t t' + \sum_i p_i^{\text{eff}} q_i' - L^{\text{eff}} = p_t t' + \sum_i p_i \dot{q}_i t' - Lt' \\ &= t'(p_t + H). \end{aligned} \tag{1.6.28}$$

At this stage, two complications arise: First, in view of (6.25) and (6.27), it is evident that p_t does not depend on t', and therefore the Jacobian determinant (5.9) vanishes. Verify this assertion. Second, because of (6.25), we see from (6.28) that H^{eff} vanishes identically.

These complications should not surprise us. Review Subsection 5.2. Show that L^{eff} as given by (6.23) is homogeneous of degree one in the velocities and does not explicitly depend on τ. Therefore, these complications must occur.

What to do? Some further information has to be provided about the parameterization. Suppose we make the dependence of t on τ a bit more explicit by writing a relation of the form

$$d\tau = f(q,p,t)dt \tag{1.6.29}$$

where f is a function to be specified. Then, by the chain rule, we have the relations

$$q_j' = (dq_j/dt)(dt/d\tau) = (1/f)(\partial H/\partial p_j),$$

$$t' = (1/f),$$
$$p'_j = (dp_j/dt)(dt/d\tau) = -(1/f)(\partial H/\partial q_j),$$
$$H' = (dH/dt)(dt/d\tau) = (1/f)(\partial H/\partial t). \tag{1.6.30}$$

In the last of these equations use has been made of (5.14). Is there a Hamiltonian that will produce these equations?

There is. Inspired by (6.28), *define* an effective Hamiltonian $\bar{H}^{\text{eff}}$ [on the *extended* $(2n+2)$-dimensional phase space consisting of the variables $q_1, q_2, \cdots q_n, t$ and $p_1, p_2 \cdots p_n, p_t$] by writing

$$\bar{H}^{\text{eff}}(q, t; p, p_t) = (1/f)(p_t + H) \tag{1.6.31}$$

where the relation (6.25) is to be ignored (but soon recovered as a special case).[40] Then, taking partial derivatives, we find the results

$$\begin{aligned}\partial\bar{H}^{\text{eff}}/\partial p_j &= (1/f)(\partial H/\partial p_j) + (p_t + H)[\partial(1/f)/\partial p_j] \\ &= q'_j + (p_t + H)[\partial(1/f)/\partial p_j],\end{aligned}$$

$$\partial\bar{H}^{\text{eff}}/\partial p_t = (1/f) = t',$$

$$\begin{aligned}\partial\bar{H}^{\text{eff}}/\partial q_j &= (1/f)(\partial H/\partial q_j) + (p_t + H)[\partial(1/f)/\partial q_j] \\ &= -p'_j + (p_t + H)[\partial(1/f)/\partial q_j],\end{aligned}$$

$$\begin{aligned}\partial\bar{H}^{\text{eff}}/\partial t &= (1/f)(\partial H/\partial t) + (p_t + H)[\partial(1/f)/\partial t] \\ &= H' + (p_t + H)[\partial(1/f)/\partial t].\end{aligned} \tag{1.6.32}$$

Here, we have also used (6.30). Next observe that $\bar{H}^{\text{eff}}$ does not depend on the independent variable τ, and therefore must be *constant* on each trajectory it generates. Consider those trajectories on which $\bar{H}^{\text{eff}} = 0$. Then for those trajectories (6.25) holds and the relations (6.32) take the form

$$q'_j = \partial\bar{H}^{\text{eff}}/\partial p_j,$$
$$t' = \partial\bar{H}^{\text{eff}}/\partial p_t,$$
$$p'_j = -\partial\bar{H}^{\text{eff}}/\partial q_j,$$
$$p'_t = -\partial\bar{H}^{\text{eff}}/\partial t. \tag{1.6.33}$$

Thus, a *special* class of trajectories generated by $\bar{H}^{\text{eff}}$, namely those for which $\bar{H}^{\text{eff}} = 0$, gives $q(\tau)$, $t(\tau)$, $p(\tau)$, and $p_t(\tau)$.

A particularly simple case is to set $f = 1$ so that

$$t' = (1/f) = 1. \tag{1.6.34}$$

In this case find the result

$$\bar{H}^{\text{eff}}(q, t; p, p_t) = p_t + H(q, p, t). \tag{1.6.35}$$

[40] The transformation (6.31) is sometimes called a *Poincaré transformation*, is useful for *regularization*, but should not be confused with the Poincaré transformations of Relativity Theory. See Section 2.7.4.

Show, by one of Hamilton's equations, that there is now the relation

$$t' = \partial \bar{H}^{\text{eff}}/\partial p_t = 1, \tag{1.6.36}$$

which is consistent with the requirement (6.34).

Suppose, for any choice of f, we consider trajectories in q, t, p, p_t space generated by $\bar{H}^{\text{eff}}$ for which the initial conditions happen to satisfy the relation

$$p_t = -H \tag{1.6.37}$$

at some initial value of τ. Then $\bar{H}^{\text{eff}} = 0$ at this value of τ. But since $\bar{H}^{\text{eff}}$ is constant on trajectories, (6.37) must then hold all along such trajectories.

One moral of this exercise is that a nonautonomous Hamiltonian system can always be converted into an autonomous one in an extended phase space (with the two additional phase-space variables t, p_t) by use of (6.35) or, more generally, (6.31). Another is that the time t can be replaced by a new independent variable τ, while remaining within a Hamiltonian framework, such that a relation of the form (6.29) holds. Such a replacement may be useful for *regularization.* See Section 2.7 and the regularization references at the end of Chapter 2.

1.6.6. Read Exercise 6.4. Let K be the Hamiltonian defined by (6.6). By reversing the Legendre transformation that relates a Lagrangian and a Hamiltonian, see (5.8), show that the Lagrangian L_K associated with K is given by the relation

$$L_K = p_t t' + \sum_{i=2}^{n} p_i q_i' - K = (\dot{q}_1)^{-1} L. \tag{1.6.38}$$

Suppose we rewrite (6.19) in the form

$$\mathcal{A} = \int L dt = \int L(dt/dq_1) dq_1 = \int L^{\text{eff}} dq_1, \tag{1.6.39}$$

with

$$L^{\text{eff}} = L(dt/dq_1). \tag{1.6.40}$$

Verify the relation

$$L_K = L^{\text{eff}}. \tag{1.6.41}$$

Conversely show that, starting with the Lagrangian L_K defined by (6.40), one arrives at the Hamiltonian K defined by (6.6).

1.6.7. Review Exercises 5.2 and 5.3. The expression (5.1) for the Lagrangian L and the expression (5.49) for the Hamiltonian H do not seem particularly aesthetically pleasing because they contain a square root and because they are not manifestly Lorentz invariant. The purpose of this exercise and the next is to explore another possible Lagrangian, and to show that the particular forms of the Lagrangian (5.1) and the Hamiltonian (5.49) come about because of a decision to treat time as an independent variable, and the coordinates as dependent variables. We will also find other interesting results along the way. Finally, in this exercise and what follows we will take special care to deal properly with "down" (covariant) and "up" (contravariant) indices.[41]

In the spirit of relativity, and following the insight of Hermann Minkowski (1864-1909), it is reasonable to try to treat space and time on a similar footing. Suppose the world line of a particle through space-time is parameterized in terms of some parameter τ by specifying four functions $x^\mu(\tau)$ that, taken together, form a vector with four contravariant components x^μ. We adopt the convention that the first three components of x^μ are the spatial coordinates of the particle, and the fourth (with a factor of c) is its temporal coordinate. Specifically, we write $\mu = 1, 2, 3, 4$ with $x^4 = ct$. That is, we write

$$x^\mu = (x, y, z, ct) = (\boldsymbol{r}, ct). \tag{1.6.42}$$

Also, let $(x')^\mu$ denote the four quantities defined by the equations

$$(x')^\mu = dx^\mu/d\tau. \tag{1.6.43}$$

Under the assumption that the parameterization is unchanged by a Lorentz transformation, $(x')^\mu$ is evidently also a 4-vector, which will be called the 4-velocity. The 3-velocity $\boldsymbol{v}$ of a particle is given by the ratio $\boldsymbol{v} = (d\boldsymbol{r}/d\tau)/(dt/d\tau)$. Since the speed of a massive particle must be less than c, $||\boldsymbol{v}|| < c$, verify that the 4-velocity must satisfy the condition

$$x' \cdot x' = (x')^\mu (x')^\nu g_{\mu\nu} > 0. \tag{1.6.44}$$

Here $g_{\mu\nu}$ denotes the metric tensor, and we have employed the usual Einstein convention that repeated indices are to be summed over. In Cartesian coordinates and for flat space-time, only the diagonal entries of g are nonzero, and we take them to have the values

$$g_{11} = g_{22} = g_{33} = -1,$$

$$g_{44} = 1. \tag{1.6.45}$$

That is, the space-time interval ds is taken to be given by the relation

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu = c^2 dt^2 - (d\boldsymbol{r})^2. \tag{1.6.46}$$

We remark that the notation ds^2 appearing in (6.46), although universally employed, can be misleading since, depending on circumstances, ds^2 can be negative, zero, or positive, and therefore is not necessarily the square of anything. But note that $ds^2 > 0$ for time-like displacements. Space-time endowed with the metric (6.45) is sometimes called *Minkowski* space.[42]

[41] Is there an easy way to remember the association between down/up and covariant/contravariant? Here is one way: The third letter from the left of the word covariant is a v, which may be viewed as the tip of a *downward* pointing arrow. Correspondingly, covariant components have down indices. And the third letter from the left in the word contravariant is an n, which, with somewhat more imagination, may be viewed as the tip of a blunted *upward* pointing arrow. Correspondingly, contravariant components have up indices.

[42] Many authors adopt the convention $\mu = 0, 1, 2, 3$ with $x^0 = ct$ and the remaining x^μ being the spatial coordinates. Accordingly, they would write $x^\mu = (ct, x, y, z) = (ct, \boldsymbol{r})$ and $g = diag(1, -1, -1, -1)$. The definition (6.46) holds in either case.

For future use we also define quantities $g^{\mu\nu}$ by the rule

$$g^{\mu\nu} = (g^{-1})_{\mu\nu}. \tag{1.6.47}$$

For the choice (6.45) there is the immediate relation

$$g^{\mu\nu} = g_{\mu\nu}. \tag{1.6.48}$$

The metric tensor can be used to raise and lower indices. For example there are the relations

$$x_\mu = g_{\mu\nu} x^\nu, \tag{1.6.49}$$

$$x^\mu = g^{\mu\nu} x_\nu, \tag{1.6.50}$$

$$g_\mu{}^\sigma = g_{\mu\nu} g^{\nu\sigma} = \delta_\mu^\sigma. \tag{1.6.51}$$

In particular, x_μ has the entries

$$x_\mu = (-x, -y, -z, ct) = (-\boldsymbol{r}, ct). \tag{1.6.52}$$

We will also define a 4-potential A^μ with entries

$$A^\mu = (A_x, A_y, A_z, \psi/c) = (\boldsymbol{A}, \psi/c). \tag{1.6.53}$$

We will soon need the antisymmetric tensor $F^{\mu\nu}$ and its lowered counterpart $F_{\mu\nu}$ defined by the relations

$$\begin{aligned} F^{\mu\nu} &= \partial^\mu A^\nu - \partial^\nu A^\mu, \\ F_{\mu\nu} &= \partial_\mu A_\nu - \partial_\nu A_\mu. \end{aligned} \tag{1.6.54}$$

Here we have used the notation

$$\begin{aligned} &\partial^\mu = \partial/\partial x_\mu = (-\partial/\partial x, -\partial/\partial y, -\partial/\partial z, c^{-1}\partial/\partial t) = (-\nabla, c^{-1}\partial/\partial t), \\ &\partial_\mu = \partial/\partial x^\mu = (\partial/\partial x, \partial/\partial y, \partial/\partial z, c^{-1}\partial/\partial t) = (\nabla, c^{-1}\partial/\partial t), \end{aligned} \tag{1.6.55}$$

which reminds us, for example, that the derivative of a scalar with respect to a covariant variable yields a contravariant result, and vice versa. (See Exercise 6.18.) Verify that the entries of $F^{\mu\nu}$ are the components of $\boldsymbol{B}$ and $\boldsymbol{E}/c$ arranged in the form

$$F^{\mu\nu} = \begin{pmatrix} 0 & -B_z & B_y & E_x/c \\ B_z & 0 & -B_x & E_y/c \\ -B_y & B_x & 0 & E_z/c \\ -E_x/c & -E_y/c & -E_z/c & 0 \end{pmatrix}, \tag{1.6.56}$$

so that $F^{12} = -B_z$, etc. Recall (5.2).

Consider the *relativistic* Lagrangian L_R defined by the relation

$$L_R = (1/2)mc(x')^\mu (x')^\nu g_{\mu\nu} + q(x')^\mu A^\nu g_{\mu\nu}. \tag{1.6.57}$$

It has the pleasing property that it is algebraically simple and treats space and time on a similar footing. In particular, L_R is evidently a scalar. That is, it is invariant under Lorentz transformations.[43]

We will now find the equations of motion that L_R produces and will also find the associated Hamiltonian H_R.

a) Show that the *canonical* momenta p_μ are given by the relation

$$p_\mu = \partial L_R / \partial (x')^\mu = mc(x')_\mu + qA_\mu, \tag{1.6.58}$$

which can also be written in the form

$$p_\mu = p_\mu^{\rm mech} + qA_\mu \tag{1.6.59}$$

where the *mechanical* momenta are given by

$$p_\mu^{\rm mech} = mc(x')_\mu. \tag{1.6.60}$$

Note again that the derivative of a scalar (in this case a Lagrangian) with respect to a contravariant ("up" index) variable yields a covariant ("down" index) result. [44] Verify that there are the results

$$p^\mu = mc(x')^\mu + qA^\mu = (p^{\rm mech})^\mu + qA^\mu \tag{1.6.61}$$

where

$$(p^{\rm mech})^\mu = mc(x')^\mu. \tag{1.6.62}$$

[43] The quantity $(x')^\mu A^\nu g_{\mu\nu}$ is a scalar under Lorentz transformations providing the 4-potential A^ν actually transforms as a 4-vector. This can be shown to be the case if the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields described by A^ν arise from an external current $j^\nu_{\rm ext}$ that vanishes sufficiently rapidly at infinity. But in some cases, such as that of an electromagnetic plane wave or wave packet, the associated 4-potential A^ν is sourceless and does not transform like a 4-vector under Lorentz transformations. Instead the new 4-potential is the Lorentz transformation of the old (as if it were a 4-vector) plus a gauge transformation term. However, the additional gauge transformation term, when combined with the term arising from $(x')^\mu g_{\mu\nu}$, forms a total τ derivative. As discussed in standard Classical Mechanics texts, such total derivatives, when added to the Lagrangian, have no effect on the equations of motion. Moreover, they do not contribute to the variation of the action $\mathcal{A}$ associated with the Lagrangian when the path is varied with fixed end points. They may therefore be dropped. Thus, Lorentz invariance is again restored, even in those cases in which the 4-potential A^ν does not transform as a 4-vector.

A similar discussion of Lorentz invariance is required in the case of the Lagrangian (classical or quantal) for the combined system of electromagnetic fields and charged particles. In this case, the charge conservation relation $\partial_\nu j^\nu = 0$ again allows conversion of possibly non Lorentz invariant terms into total derivatives that may be dropped. Note that in the single particle case, the quantities $(x')^\nu$ may be viewed as being proportional to the single-particle current 4-vector. Thus, the single particle case is a special instance of the general case.

[44] We also remark that, according to (6.62), the mechanical momentum transforms like a 4-vector under Lorentz transformations because $(x')^\mu$ transforms like a 4-vector. From (6.61) we see that the canonical momentum also transforms like a 4-vector to the extent that the 4-potential does so. If a gauge transformation is also involved in the transformation of the 4-potential, then this same additional term appears in the transformation of the canonical momentum. According to Exercise 6.2.8 this additional term may be viewed as the result of a symplectic map. Finally, we remark that a Lorentz transformation is itself a symplectic map. See Exercise 6.2.6.

b) Verify that

$$\partial^2 L_R/\partial(x')^\mu \partial(x')^\nu = mcg_{\mu\nu}, \tag{1.6.63}$$

and therefore (5.6) is satisfied if $m \neq 0$.

c) Show that differentiating and rearranging both sides of (6.59) produces the relation

$$dp_\mu^{\text{mech}}/d\tau = dp_\mu/d\tau - q(dA_\mu/d\tau), \tag{1.6.64}$$

and verify by the chain rule that

$$q(dA_\mu/d\tau) = q\sum_\nu (\partial A_\mu/\partial x^\nu)(dx^\nu/d\tau). \tag{1.6.65}$$

Show from Lagrange's equations that the canonical momenta obey the equations of motion

$$p'_\mu = dp_\mu/d\tau = \partial L_R/\partial x^\mu = q\sum_\nu (x')^\nu(\partial A_\nu/\partial x^\mu). \tag{1.6.66}$$

d) Work out Lagrange's equations of motion, with τ as an independent variable, for the Lagrangian L_R. Show, in view of (6.54), (6.60), and (6.64) through (6.66), that they yield for the mechanical momenta and coordinates the equations of motion

$$d(p^{\text{mech}})^\mu/d\tau = qF^{\mu\nu}g_{\nu\sigma}(dx^\sigma/d\tau) = qF^{\mu\nu}(dx_\nu/d\tau), \tag{1.6.67}$$

$$d^2x^\mu/d\tau^2 = [q/(mc)]F^{\mu\nu}g_{\nu\sigma}(dx^\sigma/d\tau) = [q/(mc)]F^{\mu\nu}(dx_\nu/d\tau). \tag{1.6.68}$$

The equations of motion, when written in the forms (6.67) and (6.68), are manifestly Lorentz invariant.[45] Indeed, this is an ideal opportunity to reiterate the meaning of Lorentz invariance: Lorentz invariance, as embodied by (6.68), states that if the world line $x^\mu(\tau)$ is a solution of the equations of motion, then so is its Lorentz transformed world line $\bar{x}^\mu(\tau)$ provided $F^{\mu\nu}$ is replaced by $\bar{F}^{\mu\nu}$ where $\bar{F}^{\mu\nu}$ is the tensor composed of $\bar{\boldsymbol{E}}$ and $\bar{\boldsymbol{B}}$, the Lorentz transformed electric and magnetic fields. We note that this happy circumstance comes about because, as we have already seen, the variation of the action $\mathcal{A}$ associated with L_R is *unchanged* by a Lorentz transformation. Therefore if $x^\mu(\tau)$ with its specified endpoints extremizes $\mathcal{A}$, so will $\bar{x}^\mu(\tau)$ with its transformed end points. Finally, we observe that the equations of motion (6.67) and (6.68) do not involve the vector potential, but only the fields $\boldsymbol{E}$ and $\boldsymbol{B}$. They are therefore gauge independent.

[45] Some authors would say instead that the equations of motion (6.67) and (6.68) are *covariant*. We prefer not to use such terminology because we wish to reserve the use of the term *covariant*, and the complementary term *contravariant*, to refer to the "down" and "up" index components of vectors and tensors. Perhaps even better would be to say that the equations of motion (6.67) and (6.68) are *form* invariant; they have the same form in every inertial frame.

e) Suppose we wish to use (6.67) to compute a world line (trajectory). Verify that inverting (6.62) yields the relations

$$dx^\mu/d\tau = [1/(mc)](p^{\rm mech})^\mu. \tag{1.6.69}$$

Use (6.69) to rewrite (6.67) in the form

$$d(p^{\rm mech})^\mu/d\tau = [q/(mc)]F^{\mu\nu}g_{\nu\sigma}(p^{\rm mech})^\sigma. \tag{1.6.70}$$

Verify that taken together the relations (6.69) and (6.70) constitute a (coupled) set of first-order differential equations for the variables x^μ and $(p^{\rm mech})^\mu$.

f) suppose we wish to use (6.68) to compute a world line. Introduce auxiliary variables u^μ by the rule

$$u^\mu = dx^\mu/d\tau. \tag{1.6.71}$$

Verify that (6.68) and (6.71) can be rewritten in the form

$$dx^\mu/d\tau = u^\mu, \tag{1.6.72}$$

$$du^\mu/d\tau = [q/(mc)]F^{\mu\nu}g_{\nu\sigma}u^\sigma \tag{1.6.73}$$

to yield a (coupled) set of first-order differential equations for the variables x^μ and u^μ.

g) Show that the equation of motion (6.67) has the constant and integral of motion

$$(p^{\rm mech})^\mu p^{\rm mech}_\mu = \text{const}, \tag{1.6.74}$$

and the equation of motion (6.68) has the constant and integral of motion

$$(x')^\mu(x')_\mu = \text{const}. \tag{1.6.75}$$

Whatever values these quantities have for some initial value of τ, they retain these same values for all values of τ.

h) Define the associated relativistic Hamiltonian H_R by the rule

$$\begin{aligned} H_R &= \{\sum_\mu [\partial L_R/\partial(x')^\mu](x')^\mu\} - L_R \\ &= \{\sum_\mu p_\mu(x')^\mu\} - L_R. \end{aligned} \tag{1.6.76}$$

Show that H_R is given by the relation

$$\begin{aligned} H_R &= (1/2)mc(x')^\mu(x')^\nu g_{\mu\nu} = [1/(2mc)](p^\mu - qA^\mu)(p^\nu - qA^\nu)g_{\mu\nu} \\ &= [1/(2mc)](p_\mu - qA_\mu)(p_\nu - qA_\nu)g^{\mu\nu} \\ &= [1/(2mc)](p^\mu - qA^\mu)(p_\mu - qA_\mu). \end{aligned} \tag{1.6.77}$$

Note that H_R, like L_R, is Lorentz invariant.

i) If H_R is viewed as a function of the variables x^μ, p_μ, and τ, it has the total differential

$$dH_R = \{\sum_\mu (\partial H_R/\partial x^\mu)dx^\mu + (\partial H_R/\partial p_\mu)dp_\mu\} + (\partial H_R/\partial \tau)d\tau. \tag{1.6.78}$$

On the other hand, if it is viewed as a function of the variables x^μ, $(x')^\mu$, and τ, H_R has [using (6.76)] the total differential

$$\begin{aligned} dH_R &= \{\sum_\mu [p_\mu - \partial L_R/\partial(x')^\mu]d(x')^\mu + (x')^\mu dp_\mu - (\partial L_R/\partial x^\mu)dx^\mu\} - (\partial L_R/\partial\tau)d\tau \\ &= \{\sum_\mu (x')^\mu dp_\mu - (p')_\mu dx^\mu\} - (\partial L_R/\partial\tau)d\tau. \end{aligned} \tag{1.6.79}$$

Here we have also used (6.58) and (6.66). By comparing (6.77) and (6.78), deduce the equations of motion.

$$(x')^\mu = \partial H_R/\partial p_\mu, \tag{1.6.80}$$

$$(p')_\mu = -\partial H_R/\partial x^\mu, \tag{1.6.81}$$

$$\partial H_R/\partial\tau = -\partial L_R/\partial\tau. \tag{1.6.82}$$

j) Let us check that use of the Lorentz invariant Hamiltonian H_R given by (6.77), and the associated equations of motion (6.80) through (6.82), reproduces some previous results. Verify that use of (6.80) yields (6.62). Also work out the consequences of (6.81) and compare your results with those produced by use of (6.67). Show that (6.68) is a consequence of the Hamiltonian equations (6.80) and (6.81).

k) Verify that L_R as given by (6.57) does not depend explicitly on τ,

$$\partial L_R/\partial\tau = 0. \tag{1.6.83}$$

It follows, see (5.14), that

$$dH_R/d\tau = \partial H_R/\partial\tau = -\partial L_R/\partial\tau = 0. \tag{1.6.84}$$

That is, H_R is a constant and integral of motion and therefore the quantity $[ds^2/(d\tau)^2]$ defined by

$$ds^2/(d\tau)^2 = g_{\mu\nu}(x')^\mu(x')^\nu = (x')^\mu(x')_\mu = x'\cdot x' \tag{1.6.85}$$

is a constant and an integral of motion,

$$ds^2/(d\tau)^2 = \text{ const.} \tag{1.6.86}$$

Note that this result agrees with (6.75).

l) Suppose we restrict our attention to those solutions that satisfy the relation

$$x'\cdot x' = \lambda \tag{1.6.87}$$

where λ is a constant that can have any value including negative and zero values as well as positive values. Show that for these solutions $(p^{\mathrm{mech}})^\mu$, as given by (6.62), satisfies the mass-shell condition

$$p^{\mathrm{mech}}_\mu (p^{\mathrm{mech}})^\mu = \lambda m^2 c^2. \tag{1.6.88}$$

Thus there are solutions for which the quantity $p^{\mathrm{mech}}_\mu (p^{\mathrm{mech}})^\mu$ can have any value including negative and zero values as well as positive values. Show that for these solutions H_R has the values

$$H_R = \lambda(mc/2). \tag{1.6.89}$$

Thus H_R can also have any value including negative and zero values as well as positive values.

m) Suppose we restrict our attention to those solutions that satisfy the relation

$$x' \cdot x' = \lambda = 1. \tag{1.6.90}$$

Show that for these solutions the particle has mass m,

$$p^{\mathrm{mech}}_\mu (p^{\mathrm{mech}})^\mu = m^2 c^2, \tag{1.6.91}$$

and H_R has the value

$$H_R = mc/2. \tag{1.6.92}$$

n) For those solutions that satisfy (6.90) verify that $ds^2 > 0$ and therefore we may select, in accord with (6.46), (6.85), and (6.90), a parameterization such that

$$ds/d\tau = 1. \tag{1.6.93}$$

Show that these solutions satisfy the equations

$$d(p^{\mathrm{mech}})^\mu/ds = qF^{\mu\nu}(dx_\nu/ds), \tag{1.6.94}$$

$$d^2x^\mu/ds^2 = [q/(mc)]F^{\mu\nu}(dx_\nu/ds). \tag{1.6.95}$$

o) Again restrict attention to those solutions that satisfy (6.90). Show that for these solutions there is the result

$$x' = dx/d\tau = (dx/d\tau)(d\tau/ds) = dx/ds = (dx/dt)(dt/ds) = \dot{x}(dt/ds). \tag{1.6.96}$$

Verify from (6.42) that

$$\dot{x}^\mu = dx^\mu/dt = (d\boldsymbol{r}/dt, c) = (\boldsymbol{v}, c). \tag{1.6.97}$$

Also verify, starting with (6.46), that there is the relation

$$ds/dt = c(1 - v^2/c^2)^{1/2} = c/\gamma, \tag{1.6.98}$$

and therefore

$$dt/ds = \gamma/c. \tag{1.6.99}$$

Recall the definition (5.29). Conclude that

$$(x')^\mu = (\gamma/c)(\boldsymbol{v}, c), \tag{1.6.100}$$

and therefore, by the definition (6.62), there is the relation

$$(p^{\text{mech}})^\mu = mc(x')^\mu = (m\gamma)(\boldsymbol{v}, c) = (\boldsymbol{p}^{\text{mech}}, \mathcal{E}/c) \tag{1.6.101}$$

with

$$\boldsymbol{p}^{\text{mech}} = \gamma m\boldsymbol{v} = (p_x^{\text{mech}}, p_y^{\text{mech}}, p_z^{\text{mech}}) \tag{1.6.102}$$

and

$$\mathcal{E} = \gamma mc^2. \tag{1.6.103}$$

Recall (5.28) and (5.39). Verify that combining (6.91) and (6.101) yields the relation

$$\mathcal{E}^2 = m^2c^4 + (\boldsymbol{p}^{\text{mech}} \cdot \boldsymbol{p}^{\text{mech}})c^2. \tag{1.6.104}$$

Verify also that

$$p^4 = (p^{\text{mech}})^4 + qA^4 = \mathcal{E}/c + q\psi/c = (1/c)(\gamma mc^2 + q\psi) = -(1/c)p_t. \tag{1.6.105}$$

Recall Exercise 6.1.

p) Show for the solutions of the equations of motion that satisfy (6.90) there is the relation

$$p^\mu = \{\boldsymbol{p}^{\text{can}}, -(1/c)p_t\} \tag{1.6.106}$$

with

$$p_t = -q\psi - \gamma mc^2 = -q\psi - \mathcal{E} \tag{1.6.107}$$

and

$$\boldsymbol{p}^{\text{can}} = \boldsymbol{p}^{\text{mech}} + q\boldsymbol{A} = \gamma m\boldsymbol{v} + q\boldsymbol{A}. \tag{1.6.108}$$

q) Multiply both sides of (6.94) by ds/dt to find the intermediate result

$$[d(p^{\text{mech}})^\mu/ds](ds/dt) = qF^{\mu\nu}(dx_\nu/ds)(ds/dt). \tag{1.6.109}$$

Verify the relations

$$[d(p^{\text{mech}})^\mu/ds](ds/dt) = d(p^{\text{mech}})^\mu/dt, \tag{1.6.110}$$

$$(dx_\nu/ds)(ds/dt) = dx_\nu/dt, \tag{1.6.111}$$

and conclude that (6.94) can be rewritten in the form

$$d(p^{\text{mech}})^\mu/dt = qF^{\mu\nu}(dx_\nu/dt). \tag{1.6.112}$$

Verify using (6.52) that

$$dx_\nu/dt = (-\boldsymbol{v}, c). \tag{1.6.113}$$

Use this result to show that (6.109) yields and is equivalent to the relations

$$d\boldsymbol{p}^{\text{mech}}/dt = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}), \tag{1.6.114}$$

$$d\mathcal{E}/dt = q\boldsymbol{v} \cdot \boldsymbol{E}. \tag{1.6.115}$$

Recall (5.31) and (5.40).

1.6.8. With the preparation provided by Exercise 6.7, the purpose of this exercise is to relate the Lagrangian L_R given by (6.57) and the Hamiltonian H_R given by (6.77) to the Lagrangian L given by (5.1) and the Hamiltonian H given by (5.49).

a) Review the machinery of Section 1.6. Start with the Hamiltonian H_R given by (6.77), for which τ is the independent variable, and make the definition

$$p_\tau = -H_R. \tag{1.6.116}$$

Verify that (6.116) can be rewritten in the form

$$\begin{aligned}(p_4 - qA_4)^2 &= (p^4 - qA^4)^2 = [-2mcp_\tau + (\boldsymbol{p}^{\rm can} - q\boldsymbol{A})\cdot(\boldsymbol{p}^{\rm can} - q\boldsymbol{A})] \\ &= [-2mcp_\tau + (\boldsymbol{p}^{\rm can} - q\boldsymbol{A})^2],\end{aligned} \tag{1.6.117}$$

from which it follows that

$$p_4 - qA_4 = \pm[-2mcp_\tau + (\boldsymbol{p}^{\rm can} - q\boldsymbol{A})^2]^{1/2}. \tag{1.6.118}$$

Here we have made the definition

$$\boldsymbol{p}^{\rm can} = \boldsymbol{p}^{\rm mech} + q\boldsymbol{A}. \tag{1.6.119}$$

Observe that (6.59) and (6.60) can be combined and rewritten rewritten in the form

$$p_4 - qA_4 = p_4^{\rm mech} = mc(x')_4 = mc^2(dt/d\tau). \tag{1.6.120}$$

Require that the parameterization of the world line be such that $dt/d\tau > 0$. Verify that, upon taking into account this requirement, (6.118) can be rewritten in the form

$$p_4 = qA_4 + [-2mcp_\tau + (\boldsymbol{p}^{\rm can} - q\boldsymbol{A})^2]^{1/2}. \tag{1.6.121}$$

b) Let K be the new Hamiltonian for which x^4 is the independent variable. Recall that x^4 and p_4 are canonically conjugate. See also Exercise 7.6 for further discussion of this point. Verify that there is the result

$$K(\boldsymbol{r}, \tau, \boldsymbol{p}^{\rm can}, p_\tau; x^4) = -p_4 = -qA_4 - [-2mcp_\tau + (\boldsymbol{p}^{\rm can} - q\boldsymbol{A})^2]^{1/2}. \tag{1.6.122}$$

c) Note that H_R and hence K are, in fact, independent of τ. Therefore p_τ is a constant of motion. Relate this constant to equation (6.89). That is, verify the relation

$$p_\tau = -\lambda(mc/2). \tag{1.6.123}$$

d) Since K is independent of τ, and p_τ is a constant, suppose attention is restricted to the remaining variables in K. Moreover, let us assign to p_τ the value it has for trajectories of interest, namely those with $\lambda = 1$. That is, we restrict our attention to the case where

$$p_\tau = -mc/2. \tag{1.6.124}$$

Verify that there is then the result

$$K(\boldsymbol{r}, \tau, \boldsymbol{p}^{\rm can}, -mc/2; x^4) = -qA_4 - [m^2c^2 + (\boldsymbol{p}^{\rm can} - q\boldsymbol{A})^2]^{1/2}. \tag{1.6.125}$$

Upon comparing (5.49) and (6.125), verify that there must be the relation

$$K(\boldsymbol{r}, \tau, \boldsymbol{p}^{\rm can}, -mc/2; x^4) = -(1/c)H. \tag{1.6.126}$$

e) Does (6.126) agree with what we already know? Suppose K, as given by (6.125), is used to produce equations of motion. Then, in view of (6.79) and (6.80), show that we expect the results

$$(1/c)(dx^\mu/dt) = dx^\mu/dx^4 = \partial K/\partial p_\mu \text{ for } \mu = 1, 2, 3; \tag{1.6.127}$$

$$(1/c)(dp_\mu/dt) = dp_\mu/dx^4 = -\partial K/\partial x^\mu \text{ for } \mu = 1, 2, 3. \tag{1.6.128}$$

But, there are the relations

$$p_\mu = -p^\mu \text{ for } \mu = 1, 2, 3. \tag{1.6.129}$$

Verify that, consequently, (6.127) and (6.128) can be rewritten in the form

$$(1/c)(dx^\mu/dt) = -\partial K/\partial p^\mu \text{ for } \mu = 1, 2, 3; \tag{1.6.130}$$

$$(1/c)(dp^\mu/dt) = +\partial K/\partial x^\mu \text{ for } \mu = 1, 2, 3. \tag{1.6.131}$$

But, as we already know, we wish to have the relations

$$(1/c)(dx^\mu/dt) = (1/c)(\partial H/\partial p^\mu) \text{ for } \mu = 1, 2, 3; \tag{1.6.132}$$

$$(1/c)(dp^\mu/dt) = -(1/c)(\partial H/\partial x^\mu) \text{ for } \mu = 1, 2, 3. \tag{1.6.133}$$

Verify that (6.130) through (6.133) are consistent with (6.126).

Let us summarize our results. In Exercise 6.7 you showed that use of the manifestly Lorentz invariant Lagrangian L_R given by (6.57) leads to the manifestly Lorentz invariant Hamiltonian H_R given by (6.77). Subsequently, in this exercise you showed that deciding to treat the time as the independent variable, and restricting attention to the variables x^μ and p^μ (with $\mu = 1, 2, 3$), leads from H_R to the Hamiltonian K given by (6.122) and then to the Hamiltonian H given by (5.49). Finally, see Exercise 5.13, by an inverse Legendre transformation the Hamiltonian H yields the Lagrangian L given by (5.1).

1.6.9. Review Exercise 6.7. Some texts claim that the equations of motion for relativistic charged-particle motion can also be derived from the action functional $\mathcal{A}[x]$ given by

$$\mathcal{A}[x] = \int mcds + \int qg_{\mu\nu}A^\mu dx^\nu \tag{1.6.134}$$

with ds^2 given by (6.46). Since this $\mathcal{A}$ can also be written in the form

$$\mathcal{A}[x] = \int [mc(ds/d\tau) + qg_{\mu\nu}A^\mu (x')^\nu]d\tau \tag{1.6.135}$$

where τ parameterizes the world line $x^\mu(\tau)$, show that the use of this action is equivalent to using the Lagrangian L given by the rule

$$L = mc[g_{\mu\nu}(x')^\mu(x')^\nu]^{1/2} + q(x')^\mu A^\nu g_{\mu\nu}. \tag{1.6.136}$$

Evidently this Lagrangian, like (6.57), is also invariant under Lorentz transformations provided the parameterization is Lorentz invariant.

a) Show that L is homogeneous of degree one in the velocities and does not explicitly depend on τ. Verify directly that $\mathcal{A}[x]$ is *independent* of the parameterization employed. This independence implies that we should not expect to find a unique solution that extremizes $\mathcal{A}$ since any reparametrization also gives a solution. See the discussion at the end of Subsection 5.2. Consequently, as expected, additional information will be required. By contrast, show that the action $\mathcal{A}_R[x]$ associated with the Lagrangian L_R given by (6.57) is *not* parameterization independent.

b) Show that for the Lagrangian (6.136) the *canonical* momenta $p_\mu^{\rm can}$ are given by the relations

$$p_\mu^{\rm can} = \partial L/\partial (x')^\mu = p_\mu^{\rm mech} + qA_\mu \tag{1.6.137}$$

where

$$p_\mu^{\rm mech} = mc(x')_\mu/(x' \cdot x')^{1/2}. \tag{1.6.138}$$

Here, consistent with (6.44), the parameterization and the sign of the square root are selected in such a way that both $(x')^4$ and $(p^{\rm mech})^4$ are positive. Show that both $p^{\rm mech}$ and $p^{\rm can}$ are independent of the choice of parameterization τ. Verify that the quantities $p_\mu^{\rm mech}$ comprise a 4-vector, and that there is the Lorentz invariant relation

$$p_\mu^{\rm mech}(p^{\rm mech})^\mu = m^2c^2. \tag{1.6.139}$$

c) Show that Lagrange's equations of motion yield the result

$$d(p^{\rm mech})^\mu/d\tau = qF^{\mu\nu}(dx_\nu/d\tau). \tag{1.6.140}$$

The equations of motion, when written in the form (6.140), are manifestly Lorentz invariant. However note that, while superficially similar, (6.140) is not the same as (6.67) because the definitions of $p^{\rm mech}$ in (6.62) and (6.138) are not the same.

Show that the form of the equations of motion (6.140) is unchanged if the world-line parameterization is changed. Show that the equations of motion (6.140) preserve the relation (6.139).

d) Verify that, as it stands, L as given by (6.136) is not a very promising Lagrangian because it has the property

$$\det[\partial^2 L/\partial (x')^\mu \partial (x')^\nu] = 0. \tag{1.6.141}$$

That is, the requirement (5.6) is violated. [Compare (6.141) with the analogous result in Exercise 6.7.] Also, because L is homogeneous of degree one in the variables $(x')^\mu$, it satisfies the relation

$$\sum_\mu [\partial L/\partial (x')^\mu](x')^\mu = L. \tag{1.6.142}$$

See Exercise 5.12. Consequently, verify that the Hamiltonian associated with L vanishes identically! By contrast, the Lagrangian L_R given by (6.79) satisfies (5.6), has a well-defined Hamiltonian counterpart H_R, and also automatically provides the supplementary condition (6.86).

e) In point of fact the equations (6.140), in the absence of further information, do *not* provide equations of motion in the form (3.1) as is required in order to specify trajectories. To see this, compute $d(p^{\text{mech}})^\mu/d\tau$ using the chain rule,

$$d(p^{\text{mech}})^\mu/d\tau = \sum_\nu [\partial(p^{\text{mech}})^\mu/\partial(x')^\nu](x'')^\nu. \tag{1.6.143}$$

Show that

$$\det[\partial(p^{\text{mech}})^\mu/\partial(x')^\nu] = 0 \tag{1.6.144}$$

so that (6.140) and (6.143) *cannot* be solved for the $(x'')^\nu$ to produce equations of motion of the form (3.1). Hint: Either verify (6.144) directly by brute force using (6.138) or, more elegantly and following the discussion in Subsection 5.2, show that each $(p^{\text{mech}})^\mu$ is a homogeneous function of degree zero. It then follows from Euler's relation, see Exercise 5.12, that there is the result

$$\sum_\nu [(x')^\nu][\partial(p^{\text{mech}})^\mu/\partial(x')^\nu] = 0. \tag{1.6.145}$$

This result shows that the matrix $[\partial(p^{\text{mech}})^\mu/\partial(x')^\nu]$ has the (generally nonzero) vector x' as an eigenvector with eigenvalue zero. Therefore (6.144) must hold.

f) Nevertheless, as we will see, the equations of motion provided by L give satisfactory results when supplemented by additional information. As might be expected, what is required is some information about how the parameter τ is to be selected. Suppose, for example, that τ is selected in such a way that

$$x^4 = c\tau \text{ or } t = \tau. \tag{1.6.146}$$

(Alternatively, we may proceed as in Exercise 6.10.) Note that this parameterization is not Lorentz invariant. However, since both p^{mech} and p^{can} do not depend on the choice of parameter, they continue to be 4-vectors. With the parameter choice (6.146) there is the additional information

$$(x')^4 = c, \tag{1.6.147}$$

and the equations of motion (6.140) take the form

$$d(p^{\text{mech}})^\mu/dt = qF^{\mu\nu}(dx_\nu/dt). \tag{1.6.148}$$

Show that if (6.147) holds, then there is the relation

$$x' \cdot x' = c^2(1 - v^2/c^2) = c^2/\gamma^2, \tag{1.6.149}$$

and therefore there is the relation

$$(x' \cdot x')^{1/2} = c/\gamma. \tag{1.6.150}$$

Consequently, verify using (6.138) and (6.150) that $(p^{\text{mech}})^\mu$ now takes the form

$$(p^{\text{mech}})^\mu = (p_x, p_y, p_z, \mathcal{E}/c) = (\boldsymbol{p}, \mathcal{E}/c) \tag{1.6.151}$$

with the relativistic momentum $\boldsymbol{p}$ given the by the relation

$$\boldsymbol{p} = \gamma m \boldsymbol{v} \tag{1.6.152}$$

and the relativistic energy $\mathcal{E}$ given by the relation

$$\mathcal{E} = \gamma m c^2. \tag{1.6.153}$$

Show that, when written out in component form, the equations of motion (6.145) become

$$d\boldsymbol{p}/dt = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}), \tag{1.6.154}$$

$$d\mathcal{E}/dt = q\boldsymbol{v} \cdot \boldsymbol{E}. \tag{1.6.155}$$

Recall (5.31) and (5.40).

Let us summarize our results. We have seen that, because it is homogeneous of degree 1 in the variables $(x')^\mu$, the Lagrangian (6.136) has no Hamiltonian counterpart. It is therefore of limited interest if we wish, as we do in this book, to exploit the symplectic symmetries associated with Hamiltonian formulations. However, with the aid of additional information specifying the parameterization, it is possible to obtain the equations of motion (6.154) and (6.155).

1.6.10. This exercise is a continuation of Exercise 6.9. We have explored some consequences of using the parameterization (6.146). Another attractive possibility (which is Lorentz invariant) is to select τ in such a way that there is the relation

$$d\tau = ds \tag{1.6.156}$$

with $(ds)^2$ given by (6.46). That is, the world line is parameterized by the *space-time* path length. Show that in this case there is the additional (Lorentz invariant) information

$$(x') \cdot (x') = 1 \text{ for all } \tau \tag{1.6.157}$$

so that now

$$(p^{\text{mech}})^\mu = mc(x')^\mu, \tag{1.6.158}$$

and the equations of motion (6.140) take the (Lorentz invariant) form

$$d(p^{\text{mech}})^\mu/ds = qF^{\mu\nu}(dx_\nu/ds),$$

$$d^2x^\mu/ds^2 = [q/(mc)]F^{\mu\nu}(dx_\nu/ds) = [q/(mc)]F^{\mu\nu}g_{\nu\sigma}(dx^\sigma/ds). \tag{1.6.159}$$

Verify that these equations of motion preserve the conditions (6.139) and (6.157). Moreover, observe that they are of the desired form (3.1).

Since the equations of motion (6.159) agree with those given by (6.94) and (6.95), verify that one can use them to derive the remaining results in items o through q in Exercise 6.7

1.6.11. Review Exercises 6.7 and 6.8. Starting with the Hamiltonian H_R, as given by (6.77) and for which τ is the independent variable, find a new Hamiltonian (call it K) for which $x^3 = z$ is the independent variable. Use (6.105) and show that it is correct to make the identification $p_t = -p^4c = -p_4c$. Compare your result with (6.16).

1.6.12. Exercise 5.2 determined the equations of motion for the *mechanical* variables $\boldsymbol{r}$ and $\boldsymbol{p}$ with the time t as the independent variable. See (5.43) and (5.44). The purpose of this exercise is to determine the equations of motion for mechanical variables when some coordinate is taken to be the independent variable. Specifically, suppose that the coordinate z is taken to be the independent variable. Introduce the notation

$$D = [(p_t + q\psi)^2/c^2 - m^2c^2 - (p_x - qA_x)^2 - (p_y - qA_y)^2]^{1/2}. \tag{1.6.160}$$

From (6.16) derive the equations of motion

$$x' = \partial K/\partial p_x = (p_x - qA_x)/D, \tag{1.6.161}$$

$$y' = \partial K/\partial p_y = (p_y - qA_y)/D, \tag{1.6.162}$$

$$t' = \partial K/\partial p_t = -(1/c^2)(p_t + q\psi)/D, \tag{1.6.163}$$

$$\begin{aligned} p_x' &= -\partial K/\partial x \\ &= q[(p_x - qA_x)(\partial A_x/\partial x) + (p_y - qA_y)(\partial A_y/\partial x) + (1/c^2)(p_t + q\psi)(\partial\psi/\partial x)]/D \\ &\quad +q\partial A_z/\partial x, \end{aligned} \tag{1.6.164}$$

$$\begin{aligned} p_y' &= -\partial K/\partial y \\ &= q[(p_x - qA_x)(\partial A_x/\partial y) + (p_y - qA_y)(\partial A_y/\partial y) + (1/c^2)(p_t + q\psi)(\partial\psi/\partial y)]/D \\ &\quad +q\partial A_z/\partial y, \end{aligned} \tag{1.6.165}$$

$$\begin{aligned} p_t' &= -\partial K/\partial t \\ &= q[(p_x - qA_x)(\partial A_x/\partial t) + (p_y - qA_y)(\partial A_y/\partial t) + (1/c^2)(p_t + q\psi)(\partial\psi/\partial t)]/D \\ &\quad +q\partial A_z/\partial t, \end{aligned} \tag{1.6.166}$$

where a prime denotes d/dz. Next employ (6.161) through (6.163) in (6.164) through (6.166) to find the results

$$p_x' = q[x'(\partial A_x/\partial x) + y'(\partial A_y/\partial x) - t'(\partial\psi/\partial x)] + q\partial A_z/\partial x, \tag{1.6.167}$$

$$p_y' = q[x'(\partial A_x/\partial y) + y'(\partial A_y/\partial y) - t'(\partial\psi/\partial y)] + q\partial A_z/\partial y, \tag{1.6.168}$$

$$p_t' = q[x'(\partial A_x/\partial t) + y'(\partial A_y/\partial t) - t'(\partial\psi/\partial t)] + q\partial A_z/\partial t. \tag{1.6.169}$$

The relations (6.160) through (6.169) involve canonical momenta. Since we are interested in employing mechanical variables, introduce the notation

$$\tilde{p}_x = p_x - qA_x, \tag{1.6.170}$$

$$\tilde{p}_y = p_y - qA_y, \tag{1.6.171}$$

$$\tilde{p}_t = p_t + q\psi. \tag{1.6.172}$$

From (5.30) we see that $\tilde{p}_x$ and $\tilde{p}_y$ are mechanical momenta, and from (6.107) we conclude that

$$\tilde{p}_t = p_t + q\psi = -\gamma mc^2 = -\mathcal{E} = -E^{\rm mech}. \tag{1.6.173}$$

(See also Exercise 7.10.) In terms of these variables the equations of motion (6.161) through (6.163) for the coordinates take the form

$$x' = \tilde{p}_x/\tilde{D}, \tag{1.6.174}$$

$$y' = \tilde{p}_y/\tilde{D}, \tag{1.6.175}$$

$$t' = -(1/c^2)\tilde{p}_t/\tilde{D}, \tag{1.6.176}$$

where

$$\tilde{D} = [\tilde{p}_t^2/c^2 - m^2c^2 - \tilde{p}_x^2 - \tilde{p}_y^2]^{1/2}. \tag{1.6.177}$$

The remaining task is to find the equations of motion for the mechanical momenta. Differentiate and apply the chain rule to (6.170) through (6.172) to find the results

$$\begin{aligned}\tilde{p}_x' &= p_x' - qA_x' \\ &= p_x' - q[(\partial A_x/\partial x)x' + (\partial A_x/\partial y)y' + (\partial A_x/\partial z) + (\partial A_x/\partial t)t'],\end{aligned} \tag{1.6.178}$$

$$\begin{aligned}\tilde{p}_y' &= p_y' - qA_y' \\ &= p_y' - q[(\partial A_y/\partial x)x' + (\partial A_y/\partial y)y' + (\partial A_y/\partial z) + (\partial A_y/\partial t)t'],\end{aligned} \tag{1.6.179}$$

$$\begin{aligned}\tilde{p}_t' &= p_t' + q\psi' \\ &= p_t' + q[(\partial \psi/\partial x)x' + (\partial \psi/\partial y)y' + (\partial \psi/\partial z) + (\partial \psi/\partial t)t'].\end{aligned} \tag{1.6.180}$$

Now combine (6.167) and (6.178) to obtain the result

$$\begin{aligned}\tilde{p}_x' &= p_x' - qA_x' \\ &= q[x'(\partial A_x/\partial x) + y'(\partial A_y/\partial x) - t'(\partial \psi/\partial x)] + q\partial A_z/\partial x \\ &\quad -q[(\partial A_x/\partial x)x' + (\partial A_x/\partial y)y' + (\partial A_x/\partial z) + (\partial A_x/\partial t)t'] \\ &= q[y'(\partial A_y/\partial x - \partial A_x/\partial y) + (\partial A_z/\partial x - \partial A_x/\partial z)] \\ &\quad -qt'[(\partial \psi/\partial x) + (\partial A_x/\partial t)] \\ &= q[y'B_z - B_y] + qt'E_x.\end{aligned} \tag{1.6.181}$$

Here we have used (5.2). Similarly, verify that

$$\tilde{p}_y' = q[B_x - x'B_z] + qt'E_y. \tag{1.6.182}$$

Next, combine (6.169) and (6.180) to find the result

$$\begin{aligned}\tilde{p}_t' &= p_t' + q\psi' \\ &= q[x'(\partial A_x/\partial t) + y'(\partial A_y/\partial t) - t'(\partial \psi/\partial t)] + q\partial A_z/\partial t \\ &\quad +q[(\partial \psi/\partial x)x' + (\partial \psi/\partial y)y' + (\partial \psi/\partial z) + (\partial \psi/\partial t)t'] \\ &= q[x'(\partial \psi/\partial x + \partial A_x/\partial t) + y'(\partial \psi/\partial x + \partial A_x/\partial t) + (\partial \psi/\partial z + \partial A_z/\partial t)] \\ &= -q[x'E_x + y'E_y + E_z].\end{aligned} \tag{1.6.183}$$

Verify that the relations (6.181) through (6.183) are what one would expect in view of (6.114) and (6.115).

There is one final step. We would like the right sides of (6.181) through (6.183) to involve only the coordinates and mechanical momenta, and not the quantities x', x', and t'. This can be accomplished with the aid of (6.174) through (6.177). Show that the net results are equations of motion for the mechanical momenta in the form

$$\tilde{p}_x' = q[(\tilde{p}_y/\tilde{D})B_z - B_y] - q[(1/c^2)\tilde{p}_t/\tilde{D}]E_x, \tag{1.6.184}$$

$$\tilde{p}_y' = q[B_x - (\tilde{p}_x/\tilde{D})B_z] - q[(1/c^2)\tilde{p}_t/\tilde{D}]E_y, \tag{1.6.185}$$

$$\tilde{p}_t' = -q[(\tilde{p}_x E_x + \tilde{p}_y E_y)/\tilde{D} + E_z]. \tag{1.6.186}$$

Taken together, the relations (6.174) through (6.177) and (6.184) through (6.186) provide equations of motion in mechanical variables when z is taken to be the independent variable. That is, the dependent variables are $(x, y, t; \tilde{p}_x, \tilde{p}_y, \tilde{p}_t)$, and z is the independent variable. Note that these equations of motion, like their similar counterparts in Exercises 5.2, 6.7, 6.9, and 6.10, involve only the fields $\boldsymbol{E}$ and $\boldsymbol{B}$ and make no reference to the vector and scalar potentials $\boldsymbol{A}$ and ψ.

1.6.13. Review Exercise 6.12. It formulated equations of motion for the dependent variables $(x, y, t; \tilde{p}_x, \tilde{p}_y, \tilde{p}_t)$ with z taken to be the independent variable. Your task for this exercise is to formulate equations of motion for the dependent variables $(x, y, t; x', y', t')$ with z taken to be the independent variable. What are desired are equations for the quantities (x'', y'', t'') in terms of the variables $(x, y, t; x', y', t')$ and z. Here a prime denotes (d/dz).

1.6.14. Consider charged-particle motion in the case of a *static* magnetic field $\boldsymbol{B}(\boldsymbol{r})$ and no electric field. (Note that, according to Maxwell's equations, there must be an electric field if $\boldsymbol{B}$ is not static.) Show from (5.40) that in this case the energy $\mathcal{E}$ is constant and, by (5.39), γ is constant. Next show from (5.48) that the equations of motion take the form

$$m^* d^2\boldsymbol{r}/dt^2 = q(\boldsymbol{v} \times \boldsymbol{B}) \tag{1.6.187}$$

where

$$m^* = \gamma m. \tag{1.6.188}$$

Thus, in the case of a static magnetic field and no electric field, the only difference between relativistic and nonrelativstic motion is that m must be replaced by m^*. To be more precise, suppose m^* (with $m^* \geq m$) and hence γ are specified numbers. The equations of motion (6.187) have solutions for any set $(\boldsymbol{r}^{\text{in}}, \boldsymbol{v}^{\text{in}})$ of initial conditions. Those solutions for which $\boldsymbol{v}^{\text{in}}$ satisfies

$$[1 - (\boldsymbol{v}^{\text{in}}/c) \cdot (\boldsymbol{v}^{\text{in}}/c)]^{-1/2} = m^*/m = \gamma \tag{1.6.189}$$

will also be solutions of the relativistic equations of motion.

Review Exercise 5.9. Show that the results in that exercise are consistent with the results of this exercise.

1.6.15. Review Exercise 6.14. Again view m^* as a specified number. Show that the equations of motion (6.187) follow from the "nonrelativistic" Lagrangian

$$L = (m^*/2)\boldsymbol{v} \cdot \boldsymbol{v} + q\boldsymbol{v} \cdot \boldsymbol{A}(\boldsymbol{r}). \tag{1.6.190}$$

Show that the canonical momentum $\boldsymbol{p}$ is given by the relation

$$\boldsymbol{p} = m^*\boldsymbol{v} + q\boldsymbol{A}, \tag{1.6.191}$$

and that the Hamiltonian H associated with L is given by

$$H = (\boldsymbol{p} - q\boldsymbol{A}) \cdot (\boldsymbol{p} - q\boldsymbol{A})/(2m^*). \tag{1.6.192}$$

Finally show that for trajectories of physical interest, namely those that satisfy (6.189), H has the constant value

$$H = (1/2)mc^2(\gamma^2 - 1)/\gamma = (1/2)mc^2[(m^*/m) - (m/m^*)] = (1/2)m^*v^2. \tag{1.6.193}$$

1.6.16. This exercise describes geodesics and affine geodesics. As background, review Exercises 5.16 through 5.18. They treat the problem of finding shortest paths in two-dimensional Euclidean space. Two-dimensional Euclidean space is a simple example of a Riemannian manifold for which the metric tensor is constant and equal to the identity matrix. Roughly speaking, an n-dimensional *manifold* is a set that locally at each point looks like an n-dimensional space with local coordinates $x^1, \cdots, x^n$. When equipped with a (possibly position dependent) metric tensor $g(x)$, it becomes a *Riemannian* manifold.

Now consider a general Riemannian manifold with local coordinates x^i and metric tensor $g(x)$. We assume that g is invertible. This manifold is called *proper* Riemannian if g is positive definite, and *pseudo* Riemannian if g is not positive (or negative) definite. [Thus for example, according to (6.45), space-time in the theory of special relativity (Minkowski space) is a pseudo Riemannian manifold.] Let y and z be any two nearby points in the manifold, and consider all paths $x(\tau)$ joining y and z such that

$$x(0) = y,$$

$$x(1) = z. \tag{1.6.194}$$

Let a dot denote $(d/d\tau)$. If the manifold is proper Riemannian, we may define a *distance* functional $D[x]$ by the rule

$$D[x] = \int_0^1 d\tau \left[\sum_{ij} g_{ij}(x)\dot{x}^i\dot{x}^j\right]^{1/2}. \tag{1.6.195}$$

If the manifold is either proper or pseudo Riemannian, we may define an *energy* functional $E[x]$ by the rule

$$E[x] = (1/2)\int_0^1 d\tau \sum_{ij} g_{ij}(x)\dot{x}^i\dot{x}^j. \tag{1.6.196}$$

A path that extremizes D is called a *geodesic*, and a path that extremizes E is called an *affine geodesic*. Note that the functional $D[x]$ may not be defined for all paths in a pseudo-Riemannian space because in that case the argument of the square root appearing in (6.195) may be negative. Correspondingly, geodesics do not necessarily exist between all y, z pairs in a pseudo-Riemannian space. By contrast, the functional $E[x]$ is well defined for all paths in both the proper and pseudo-Riemannian cases. (Note that in this simplified discussion we have assumed that the topology of the manifold is that of Euclidean space since we have assumed global coordinates in defining D and/or E. A more general discussion would involve the use of overlapping local coordinate patches.)

Is there a relation between geodesics and affine geodesics?

a) Let us begin with the geodesic case. Show that the functional $D[x]$ does not depend on the parameterization of x. That is, one may replace $x(\tau)$ by $x(\sigma(\tau))$ where $\sigma(\tau)$ is any function satisfying

$$\sigma(0) = 0,$$

$$\sigma(1) = 1. \tag{1.6.197}$$

Therefore, as described in Subsection 5.2 and illustrated in Exercises 5.16, 5.17, 6.5, and 6.9, there will eventually be a need for further information.

b) The condition for a geodesic is $\delta D = 0$. Verify, by the standard calculus of variations, that this condition is equivalent to Lagrange's equations for the Lagrangian L_D given by

$$L_D = (g_{ij}\dot{x}^i\dot{x}^j)^{1/2} = ds/d\tau. \tag{1.6.198}$$

Here, and in what follows, we again employ the Einstein summation convention. Show that L_D has the (unpromising) property

$$\det(\partial^2 L_D/\partial\dot{x}^i\partial\dot{x}^j) = 0, \tag{1.6.199}$$

and that this property arises from the fact that L_D is homogeneous of degree one in the $\dot{x}^i$, which is why $D[x]$ is parameterization independent. Also show that the Hamiltonian H_D associated with L_D vanishes identically.

c) Nevertheless, let us push on. As a first step, show that

$$\partial L_D/\partial\dot{x}^i = g_{ij}\dot{x}^j/(g_{k\ell}\dot{x}^k\dot{x}^\ell)^{1/2} = g_{ij}\dot{x}^j/(ds/d\tau). \tag{1.6.200}$$

Next verify the relations

$$\frac{d}{d\tau}\left(\frac{\partial L_D}{\partial\dot{x}^i}\right) = \left[\frac{d(g_{ij}\dot{x}^j)}{d\tau}\right]\left(\frac{ds}{d\tau}\right)^{-1} + (g_{ij}\dot{x}^j)\frac{d(ds/d\tau)^{-1}}{d\tau}, \tag{1.6.201}$$

$$d(g_{ij}\dot{x}^j)/d\tau = g_{ij}\ddot{x}^j + (\partial g_{ij}/\partial x^k)\dot{x}^j\dot{x}^k, \tag{1.6.202}$$

$$\frac{d(ds/d\tau)^{-1}}{d\tau} = -\left(\frac{ds}{d\tau}\right)^{-2}\frac{d^2s}{d\tau^2}, \tag{1.6.203}$$

$$\partial L_D/\partial x^i = (1/2)(g_{jk}\dot{x}^j\dot{x}^k)^{-1/2}(\partial g_{jk}/\partial x^i)\dot{x}^j\dot{x}^k = (1/2)(ds/d\tau)^{-1}(\partial g_{jk}/\partial x^i)\dot{x}^j\dot{x}^k. \quad (1.6.204)$$

Also verify the identity

$$\partial g_{ij}/\partial x^k = (1/2)(\partial g_{ij}/\partial x^k + \partial g_{ik}/\partial x^j) + (1/2)(\partial g_{ij}/\partial x^k - \partial g_{ik}/\partial x^j), \quad (1.6.205)$$

which decomposes $(\partial g_{ij}/\partial x^k)$ into symmetric and antisymmetric parts under the interchange of j and k. Note that only the symmetric part contributes to the sum $(\partial g_{ij}/\partial x^k)\dot{x}^j\dot{x}^k$ that occurs in (6.202) and (6.204). Thus, show that Lagrange's equations (5.3) for L_D produce the relations

$$\begin{aligned}&g_{ij}\ddot{x}^j(ds/d\tau)^{-1} + (ds/d\tau)^{-1}(1/2)(\partial g_{ij}/\partial x^k + \partial g_{ik}/\partial x^j)\dot{x}^j\dot{x}^k\\ &-(g_{ij}\dot{x}_j)(ds/d\tau)^{-2}(ds^2/d\tau^2) = (1/2)(ds/d\tau)^{-1}(\partial g_{jk}/\partial x^i)\dot{x}^j\dot{x}^k. \quad (1.6.206)\end{aligned}$$

d) Next multiply through by $(ds/d\tau)$ and group terms to get the result

$$\begin{aligned}g_{ij}\ddot{x}^j &= g_{ij}\dot{x}^j(ds/d\tau)^{-1}(ds^2/d\tau^2) + (1/2)\{(\partial g_{jk}/\partial x^i)\\ &- [(\partial g_{ij}/\partial x^k) + (\partial g_{ik}/\partial x^j)]\}\dot{x}^j\dot{x}^k. \quad (1.6.207)\end{aligned}$$

Since g is invertible, it appears that we may solve (6.207) for the $\ddot{x}^j$. Indeed, multiply both sides of (6.207) by $g^{\ell i}$, where $g^{\ell i}$ is defined by the rule

$$g^{\ell i} = (g^{-1})_{\ell i}, \quad (1.6.208)$$

and sum over i to get the intermediate results

$$g^{\ell i}g_{ij}\ddot{x}^j = g^{\ell i}g_{ij}\dot{x}^j(ds/d\tau)^{-1}(d^2s/d\tau^2) - \Gamma^{\ell}_{jk}\dot{x}^j\dot{x}^k. \quad (1.6.209)$$

Here the Γ^{ℓ}_{jk} are the *Christoffel* symbols/coefficients defined by the rule

$$\Gamma^{\ell}_{jk} = (1/2)g^{\ell i}\{[(\partial g_{ij}/\partial x^k) + (\partial g_{ik}/\partial x^j)] - (\partial g_{jk}/\partial x^i)\}. \quad (1.6.210)$$

Note that they are symmetric under the interchange of the two lower indices. Show that carrying out the indicated sums in (6.209) and using (6.208) yield the final results

$$\ddot{x}^{\ell} = \dot{x}^{\ell}(ds/d\tau)^{-1}(d^2s/d\tau^2) - \Gamma^{\ell}_{jk}\dot{x}^j\dot{x}^k. \quad (1.6.211)$$

e) Have we, contrary to (6.199), succeeded in solving for the $\ddot{x}^j$? The answer is *no*, because in general the quantity $(d^2s/d\tau^2)$ also involves the $\ddot{x}^j$. What is needed is some information about the parameterization. One possibility, also discussed in Exercise 5.17, is to take one of the x^j as the parameter. Another, more democratic, approach is to select the parameterization in such a way that

$$d^2s/d\tau^2 = 0. \quad (1.6.212)$$

Verify that (6.212) implies relations of the form

$$ds/d\tau = \text{const} = a,$$

$$s = a\tau + b, \tag{1.6.213}$$

where a and b are constants. Moreover, it is natural to set $b = 0$ so that $s = a\tau$ and $s = 0$ when $\tau = 0$. Finally, since both sides of (6.211) are homogeneous of degree 2 in a when (6.213) holds, verify that we may as well set $a = 1$ so that

$$s = \tau. \tag{1.6.214}$$

When this is done, verify that the equations (6.211) for a geodesic become

$$d^2x^\ell/ds^2 + \Gamma^\ell_{jk}(dx^j/ds)(dx^k/ds) = 0, \tag{1.6.215}$$

and on this geodesic, according to (6.214), there is the relation

$$(ds/d\tau)^2 = g_{ij}\dot{x}^i\dot{x}^j = 1. \tag{1.6.216}$$

f) There is a consistency check that may dispel any lingering doubts about the correctness of what we have done. Suppose we solve the equations

$$d^2x^\ell/d\tau^2 + \Gamma^\ell_{jk}(dx^j/d\tau)(dx^k/d\tau) = 0. \tag{1.6.217}$$

What can be said about $L_D = ds/d\tau$ for such solutions? Show, by undoing some of the previous steps, that (6.217) is equivalent to the relation

$$g_{ij}\ddot{x}^j + [(\partial g_{ij}/\partial x^k) - (1/2)(\partial g_{jk}/\partial x^i)]\dot{x}^j\dot{x}^k = 0. \tag{1.6.218}$$

Next, show that multiplying (6.218) by $\dot{x}^i$ and summing over i yields the result

$$g_{ij}\dot{x}^i\ddot{x}^j + (1/2)(\partial g_{ij}/\partial x_k)]\dot{x}^i\dot{x}^j\dot{x}^k = 0. \tag{1.6.219}$$

According to (6.198) there is the relation

$$L_D^2 = g_{ij}\dot{x}^i\dot{x}^j. \tag{1.6.220}$$

Show that (6.220) implies the relation

$$L_D(dL_D/d\tau) = g_{ij}\dot{x}^i\ddot{x}^j + (1/2)(\partial g_{ij}/\partial x^k)\dot{x}^i\dot{x}^j\dot{x}^k, \tag{1.6.221}$$

and that (6.219) and (6.221, when combined, yield the relation

$$dL_D/d\tau = 0. \tag{1.6.222}$$

Therefore, the relation

$$ds/d\tau = \text{const} \tag{1.6.223}$$

is a consequence of (6.217), and hence is consistent with (6.217).

g) Having discussed geodesics at some length, let us now turn to affine geodesics. Show that, unlike $D[x]$, the functional $E[x]$ does depend on parameterization. Evidently the Lagrangian L_E for an affine geodesic is given by

$$L_E = (1/2) g_{ij} \dot{x}^i \dot{x}^j. \tag{1.6.224}$$

Show that in this case

$$\det(\partial^2 L_E / \partial \dot{x}^i \partial \dot{x}^j) \neq 0. \tag{1.6.225}$$

Show that

$$p_i = \partial L_E / \partial \dot{x}^i = g_{ij} \dot{x}^j = \dot{x}_i, \tag{1.6.226}$$

and that the Hamiltonian H_E associated with L_E is given by

$$H_E = p_i \dot{x}^i - L_E = (1/2) g_{ij} \dot{x}^i \dot{x}^j = (1/2) g^{ij} p_i p_j. \tag{1.6.227}$$

Show that H_E is a constant of motion, and hence

$$H_E = (1/2) g_{ij} \dot{x}^i \dot{x}^j = (1/2)(ds/d\tau)^2 = \text{const}. \tag{1.6.228}$$

Show that

$$dp_i/d\tau = g_{ij} \ddot{x}^j + (\partial g_{ij}/\partial x^k) \dot{x}^j \dot{x}^k, \tag{1.6.229}$$

$$\partial L_E / \partial x^i = (1/2)(\partial g_{jk}/\partial x^i) \dot{x}^j \dot{x}^k, \tag{1.6.230}$$

and hence Lagrange's equations of motion yield the relations

$$g_{ij} \ddot{x}^j + [(\partial g_{ij}/\partial x^k) - (1/2)(\partial g_{jk}/\partial x^i)] \dot{x}^j \dot{x}^k = 0. \tag{1.6.231}$$

Show that these relations can be solved for the $\ddot{x}^j$ to yield the results

$$\ddot{x}^\ell + \Gamma^\ell_{jk} \dot{x}^j \dot{x}^k = 0. \tag{1.6.232}$$

You have demonstrated that an affine geodesic satisfies (6.228) and (6.232). Comparison of (6.217) and (6.232) shows that a geodesic, when it exists and is parameterized to satisfy $\tau = s/a$, is also an affine geodesic. Conversely, an affine geodesic always exists, is always automatically parameterized to satisfy (6.228), and yields a geodesic parameterized to satisfy $\tau = s/a$ when such exists. Thus, there is no loss of generality in working with affine geodesics, and they have the advantage of being defined even when the metric is not positive definite.

There is yet one more set of remarks of interest. Let x^0 be some point and consider some affine geodesic through x^0 parameterized in such a way that $x(\tau) = x^0$ when $\tau = 0$. Let us see what can be said about the quantity $(1/2) g_{ij} \dot{x}^i \dot{x}^j$ at this point. Since the metric tensor $g_{ij}(x^0)$ at this point, when regarded as a matrix, is a real symmetric matrix, it can be diagonalized by a similarity transformation employing a real orthogonal matrix. Moreover, all its eigenvalues will be real. Next, by proper scaling of the coordinates, g_{ij} at this point can be brought to a diagonal form where each of its eigenvalues are either $+1$, 0, or -1; and the numbers of each kind are invariants.[46] Since we have assumed that g_{ij} is invertible, we will exclude from our discussion the case where any of the eigenvalues vanish. Then, in the case that g_{ij} is positive definite, all the eigenvalues (after diagonalization and suitable coordinate scaling) may be taken to be $+1$. Correspondingly, the value of $(1/2) g_{ij} \dot{x}^i \dot{x}^j$ at this point will be positive, and (after suitable rescaling of the parameter τ) we may confine our attention to the case for which it has the value $1/2$.[47] Similarly, if g_{ij} is negative definite,

[46] This result is called *Sylvestor's law of inertia* for quadratic forms.

[47] Note that (6.232) is invariant under rescaling of τ.

all the eigenvalues may be taken to be -1. In this case the value occurring $(1/2)g_{ij}\dot{x}^i\dot{x}^j$ at this point will be negative, and we may confine our attention to the case for which it has the value $-1/2$. Finally, in the pseudo-Riemannian case, some of the eigenvalues will be positive and some will be negative.[48] In this circumstance we may confine our attention to three classes of cases: those for which $(1/2)g_{ij}\dot{x}^i\dot{x}^j$ has the value $+1/2$, those for which it has the value 0, and those for which it has the value $-1/2$.

All these considerations apply to the possible values of $(1/2)g_{ij}\dot{x}^i\dot{x}^j$ at the point x^0 for affine geodesics through the point x^0. But now, according to (6.228), the value of $(1/2)g_{ij}\dot{x}^i\dot{x}^j$ remains constant *all along* an affine geodesic. Therefore in the positive definite case we may restrict our attention to affine geodesics for which the constant appearing on the right side of (6.228) has the value 1/2; and in the negative definite case we may restrict our attention to affine geodesics for which the constant has the value $-1/2$. Finally, in the pseudo Riemannian case, we may restrict our attention to those affine geodesics for which the constant has the values 1/2, 0, and $-1/2$.

1.6.17. This exercise examines how a Lorentz transformation acts on the electromagnetic field tensor $F^{\alpha\beta}$ as given by (6.56), which we repeat below:

$$F^{\mu\nu} = \begin{pmatrix} 0 & -B_z & B_y & E_x/c \\ B_z & 0 & -B_x & E_y/c \\ -B_y & B_x & 0 & E_z/c \\ -E_x/c & -E_y/c & -E_z/c & 0 \end{pmatrix}, \tag{1.6.233}$$

so that $F^{12} = -B_z$, etc. Let us review some background information: A Lorentz transformation, when acting on space-time, is a linear transformation described by a matrix which we will call Λ. See (6.2.49). Its action on the four-vector (6.42) is given by the relation

$$\bar{x}^\alpha = \sum_\mu \Lambda^{\alpha\mu} x^\mu. \tag{1.6.234}$$

Its action on a tensor $F^{\alpha\beta}$ is given by the relation

$$\bar{F}^{\alpha\beta} = \sum_{\mu\nu} \Lambda^{\alpha\mu}\Lambda^{\beta\nu}F^{\mu\nu}. \tag{1.6.235}$$

See Exercise 6.2.6. The matrix Λ satisfies the relation

$$\Lambda g \Lambda^T = g. \tag{1.6.236}$$

See (6.2.51). Verify, if we view F as a matrix, then (6.235) can be written in the form

$$\bar{F} = \Lambda F \Lambda^T. \tag{1.6.237}$$

[Note that the left side of (6.236) and the right side of (6.237) have an identical structure. That is because both g and F are rank two tensors, and therefore are acted upon by a

[48] For example, in the case (6.45), one of the eigenvalues is +1 and three are -1.

Lorentz transformation Λ in the same fashion. But the difference between g and F is that g is invariant under this action, and F is not.]

Watch as we now perform some trickery: Define a matrix G by the rule

$$G = F - \lambda g \tag{1.6.238}$$

where λ is a parameter. Define a transformed matrix $\bar{G}$ by the rule

$$\bar{G} = \Lambda G \Lambda^T. \tag{1.6.239}$$

Show that there is the result

$$\bar{G} = \bar{F} - \lambda g. \tag{1.6.240}$$

Define polynomials $P(F, \lambda)$ and $P(\bar{F}, \lambda)$ by the rules

$$P(F, \lambda) = \det(G), \tag{1.6.241}$$

$$P(\bar{F}, \lambda) = \det(\bar{G}). \tag{1.6.242}$$

For the identity component of the Lorentz group the Lorentz transformation matrix Λ has the property

$$\det(\Lambda) = 1. \tag{1.6.243}$$

See Exercise 7.3.27. Show, using (6.239), that

$$\det(\bar{G}) = \det(G) \tag{1.6.244}$$

and consequently

$$P(\bar{F}, \lambda) = P(F, \lambda). \tag{1.6.245}$$

It therefore behooves us to compute $P(F, \lambda)$.

Show, because F is antisymmetric, only even powers of λ can occur in $P(F, \lambda)$. Indeed, show that

$$P(F, \lambda) = -\lambda^4 + \lambda^2[\boldsymbol{B} \cdot \boldsymbol{B} - (1/c^2)\boldsymbol{E} \cdot \boldsymbol{E}] + (1/c^2)(\boldsymbol{E} \cdot \boldsymbol{B})^2. \tag{1.6.246}$$

Define functions I_1 and I_2 of F (the electromagnetic fields $\boldsymbol{E}$ and $\boldsymbol{B}$) by the rules

$$I_1(F) = \boldsymbol{B} \cdot \boldsymbol{B} - (1/c^2)\boldsymbol{E} \cdot \boldsymbol{E}, \tag{1.6.247}$$

$$I_2(F) = (1/c)(\boldsymbol{E} \cdot \boldsymbol{B}). \tag{1.6.248}$$

Then, by (6.245) and upon equating powers of λ, we find using (6.246) that

$$I_1(\bar{F}) = I_1(F), \tag{1.6.249}$$

and

$$[I_2(\bar{F})]^2 = [I_2(F)]^2 \tag{1.6.250}$$

where $\bar{F}$ is the field tensor composed of $\bar{\boldsymbol{E}}$ and $\bar{\boldsymbol{B}}$, the fields resulting from applying the Lorentz transformation Λ to the fields $\boldsymbol{E}$ and $\boldsymbol{B}$. That is, the field functions I_1 and $[I_2]^2$ are Lorentz invariant.

In fact, the function I_2 itself, and not just its square, is Lorentz invariant,

$$I_2(\bar{F}) = I_2(F). \tag{1.6.251}$$

Evidently, (6.250) follows from (6..251). But can (6.251) be proved from (6.250)? It can, by *continuity*: First, suppose F is such that $I_2(F) = 0$. Then we have the chain of reasoning

$$I_2(F) = 0 \Rightarrow [I_2(F)]^2 = 0 \Rightarrow [I_2(\bar{F})]^2 = 0 \Rightarrow I_2(\bar{F}) = 0, \tag{1.6.252}$$

which establishes (6.251) in this case. But what about the case $I_2(F) \neq 0$? In this case, without further reasoning, we may only conclude from (6.250) that

$$I_2(\bar{F}) = I_2(F) \text{ or } I_2(\bar{F}) = -I_2(F). \tag{1.6.253}$$

We wish to rule out the second possibility.

Let $\hat{\Lambda}(\tau)$ be a continuous *path* in the identity component of the Lorentz group that connects the identity element I to the element Λ. Such a path is easily specified using, in Lie form, a *polar decomposition* for elements in the identity component of the Lorentz group. For details, see Exercise 7.3.27. There it is shown that Lorentz group elements Λ in the identity component can be written in the form

$$\Lambda(\lambda, \boldsymbol{m}; \theta, \boldsymbol{n}) = \exp(\lambda \boldsymbol{m} \cdot \boldsymbol{N}) \exp(\theta \boldsymbol{n} \cdot \boldsymbol{L}) \tag{1.6.254}$$

where $\boldsymbol{N}$ and $\boldsymbol{L}$ are Lie generators for *boosts* and *rotations*, respectively.[49] We now define a path $\hat{\Lambda}(\tau)$ with the desired properties by the rule

$$\hat{\Lambda}(\tau) = \exp(\tau\lambda \boldsymbol{m} \cdot \boldsymbol{N}) \exp(\tau\theta \boldsymbol{n} \cdot \boldsymbol{L}). \tag{1.6.255}$$

By construction, this path satisfies the relations

$$\hat{\Lambda}(0) = I \text{ and } \hat{\Lambda}(1) = \Lambda. \tag{1.6.256}$$

Next define a sequence of field tensors $\hat{F}(\tau)$ by the rule

$$\hat{F}(\tau) = \hat{\Lambda}(\tau) F \hat{\Lambda}^T(\tau) \tag{1.6.257}$$

with the results that

$$\hat{F}(0) = F \text{ and } \hat{F}(1) = \bar{F}. \tag{1.6.258}$$

Finally, define a function $\hat{I}_2(\tau)$ by the rule

$$\hat{I}_2(\tau) = I_2[\hat{F}(\tau)]. \tag{1.6.259}$$

Verify that $\hat{I}_2(\tau)$ is a continuous function, and has the properties

$$\hat{I}_2(0) = I_2(F) \text{ and } \hat{I}_2(1) = I_2(\bar{F}). \tag{1.6.260}$$

[49] Note that the parameters λ appearing in (6.238) and (6.254) are not the same.

We now have all the required ingredients to complete our argument. We have already assumed $I_2(F) \neq 0$. Next assume that the second option in (6.253) holds,

$$I_2(\bar{F}) = -I_2(F). \tag{1.6.261}$$

Show that then

$$\hat{I}_2(1) = I_2(\bar{F}) = -I_2(F) = -\hat{I}_2(0). \tag{1.6.262}$$

That is, the function $\hat{I}_2(\tau)$ *changes sign* as τ goes from $\tau = 0$ to $\tau = 1$. At this point *Bernard Bolzano* (1781-1848) would exclaim, based on his *intermediate value theorem*, that $\hat{I}_2(\tau)$ must vanish for some τ value somewhere in the interval $\tau \in (0,1)$. Let $\tau_0 \in (0,1)$ be a/the τ value for which $\hat{I}_2(\tau)$ takes on the intermediate value 0,

$$\hat{I}_2(\tau_0) = 0. \tag{1.6.263}$$

Now we may make the reasoning chain

$$\hat{I}_2(\tau_0) = 0 \Rightarrow [\hat{I}_2(\tau_0)]^2 = 0 \Rightarrow [\hat{I}_2(\tau_0)]^2 \neq [I_2(F)]^2, \tag{1.6.264}$$

which shows that $[I_2]^2$ is not Lorentz invariant. We have reached a contradiction, and therefore the first option in (6.253) must hold. That is, I_2 itself must be Lorentz invariant, as claimed.

There is also an alternate proof that I_1 and I_2 are Lorentz invariant and, in fact are manifestly Lorentz invariant. Begin by recalling some standard definitions: The tensor $F_{\mu\nu}$, the lower-index/covariant version of $F^{\alpha\beta}$, is defined by the rule

$$F_{\mu\nu} = \sum_{\alpha\beta} g_{\mu\alpha} g_{\nu\beta} F^{\alpha\beta}. \tag{1.6.265}$$

Verify that

$$F_{\mu\nu} = \begin{pmatrix} 0 & -B_z & B_y & -E_x/c \\ B_z & 0 & -B_x & -E_y/c \\ -B_y & B_x & 0 & -E_z/c \\ E_x/c & E_y/c & E_z/c & 0 \end{pmatrix}, \tag{1.6.266}$$

so that $F_{12} = -B_z$, etc. Evidently the elements of $F_{\mu\nu}$ are obtained from those of $F^{\mu\nu}$ by making the substitution $\boldsymbol{E} \rightarrow -\boldsymbol{E}$. The tensor ${}^*F^{\mu\nu}$, the tensor *dual* to $F^{\alpha\beta}$, is defined by the rule

$${}^*F^{\mu\nu} = (1/2) \sum_{\alpha\beta} \epsilon^{\mu\nu\alpha\beta} F_{\alpha\beta}. \tag{1.6.267}$$

Here $\epsilon^{\alpha\beta\gamma\delta}$ is the completely antisymmetric tensor with $\epsilon^{1234} = 1$. Verify that a particular application of the rule (6.267) yields the relation ${}^*F^{12} = F_{34}$. Verify the general result

$${}^*F^{\mu\nu} = \begin{pmatrix} 0 & -E_z/c & E_y/c & -B_x \\ E_z/c & 0 & -E_x/c & -B_y \\ -E_y/c & E_x/c & 0 & -B_z \\ B_x & B_y & B_z & 0 \end{pmatrix}, \tag{1.6.268}$$

so that ${}^*F^{12} = -E_z/c$, etc. Evidently the elements of ${}^*F^{\mu\nu}$ are obtained from those of $F^{\mu\nu}$ by making the substitutions $\boldsymbol{E}/c \to -\boldsymbol{B}$ and $\boldsymbol{B} \to \boldsymbol{E}/c$.

With these definitions in hand, employ them to show that there are the relations

$$I_1(F) = (1/2)\sum_{\alpha\beta} F^{\alpha\beta}F_{\alpha\beta}, \tag{1.6.269}$$

$$I_2(F) = (1/8)\sum_{\alpha\beta\gamma\delta} \epsilon^{\alpha\beta\gamma\delta}F_{\alpha\beta}F_{\gamma\delta} = (1/4)\sum_{\alpha\beta} {}^*F^{\alpha\beta}F_{\alpha\beta}. \tag{1.6.270}$$

Finally, for future use in Exercise 8.2.12, note from (6.246) that there is the result

$$\det(F) = P(F,0) = (1/c^2)(\boldsymbol{E}\cdot\boldsymbol{B})^2 = [I_2(F)]^2. \tag{1.6.271}$$

You have shown that (6.249) and (6.251) comprise a *necessary* condition for a field pair $\bar{\boldsymbol{E}}$,$\bar{\boldsymbol{B}}$ and $\boldsymbol{E}$,$\boldsymbol{B}$ to be related by a Lorentz transformation. It can be shown that (6.249) and (6.251) also comprise a *sufficient* condition. See Exercise 6.2.12. That is, if for a given pair $\bar{F}$,F (6.249) and (6.251) hold, then there must a Lorentz transformation Λ such that (6.237) holds.

Since it has been established that (6.249) and (6.251) comprise a necessary and sufficient condition, it follows that *any* Lorentz invariant electromagnetic field function must be expressible as some function of I_1 and I_2.

1.6.18. The purpose of this exercise is to study vector and tensor transformation properties. Recall that under a Lorentz transformation the contravariant components of a four vector transform according to the rule (2.34), which we repeat below:

$$\bar{x}^\alpha = \sum_\mu \Lambda^{\alpha\mu}x^\mu. \tag{1.6.272}$$

See Exercise 6.17. The covariant components of the same four vector are given by the relation

$$x_\mu = \sum_\nu g_{\mu\nu}x^\nu. \tag{1.6.273}$$

See (6.49). Your first task is to determine how the covariant components transform under the same Lorentz transformation.

Begin with some preparatory steps. Observe that (6.272) implies the differential relations

$$d\bar{x}^\alpha = \sum_\mu \Lambda^{\alpha\mu}dx^\mu. \tag{1.6.274}$$

If we view the $\bar{x}^\alpha$ as functions of the x^μ there are also the differential relations

$$d\bar{x}^\alpha = \sum_\mu (\partial\bar{x}^\alpha/\partial x^\mu)dx^\mu. \tag{1.6.275}$$

Upon comparison of (6.274) and (6.275) we see that there are the partial differential relations

$$\partial\bar{x}^\alpha/\partial x^\mu = \Lambda^{\alpha\mu}. \tag{1.6.276}$$

Consequently (6.272) can also be written in the form

$$\bar{x}^{\alpha} = \sum_{\mu} (\partial \bar{x}^{\alpha}/\partial x^{\mu}) x^{\mu}. \tag{1.6.277}$$

Next show that (6.274) can be inverted. Multiply both sides of (6.274) by $(\Lambda^{-1})^{\nu\alpha}$ and sum over α to find the result

$$\begin{aligned}\sum_{\alpha} (\Lambda^{-1})^{\nu\alpha} d\bar{x}^{\alpha} &= \sum_{\alpha} (\Lambda^{-1})^{\nu\alpha} \sum_{\mu} \Lambda^{\alpha\mu} dx^{\mu} = \sum_{\mu} \sum_{\alpha} (\Lambda^{-1})^{\nu\alpha} \Lambda^{\alpha\mu} dx^{\mu} \\ &= \sum_{\mu} (\Lambda^{-1}\Lambda)^{\nu\mu} dx^{\mu} = \sum_{\mu} (I)^{\nu\mu} dx^{\mu} = dx^{\nu}. \end{aligned} \tag{1.6.278}$$

(That Λ^{-1} exists follows from Exercise 7.3.27.) If we view the x^{ν} as functions of the $\bar{x}^{\alpha}$ there are also the differential relations

$$dx^{\nu} = \sum_{\alpha} (\partial x^{\nu}/\partial \bar{x}^{\alpha}) d\bar{x}^{\alpha}. \tag{1.6.279}$$

Upon comparison of (6.278) and (6.279) we see that there are the partial differential relations

$$\partial x^{\nu}/\partial \bar{x}^{\alpha} = (\Lambda^{-1})^{\nu\alpha}. \tag{1.6.280}$$

We are now ready to proceed with the first task. From (6.48) through (6.50) there are the relations

$$\bar{x}_{\alpha} = \sum_{\beta} g_{\alpha\beta} \bar{x}^{\beta} = \sum_{\beta} g^{\alpha\beta} \bar{x}^{\beta}, \tag{1.6.281}$$

$$x^{\mu} = \sum_{\nu} g^{\mu\nu} x_{\nu}. \tag{1.6.282}$$

Use these relations and a relation of the form (6.272) to show that

$$\bar{x}_{\alpha} = \sum_{\beta} g^{\alpha\beta} \sum_{\mu} \Lambda^{\beta\mu} \sum_{\nu} g^{\mu\nu} x_{\nu}, \tag{1.6.283}$$

which can be rewritten in the matrix form

$$\bar{x}_{\alpha} = \sum_{\beta} g^{\alpha\beta} \sum_{\mu} \Lambda^{\beta\mu} \sum_{\nu} g^{\mu\nu} x_{\nu} = \sum_{\nu} (g\Lambda g)^{\alpha\nu} x_{\nu}. \tag{1.6.284}$$

Next show from (6.236) that there is the relation

$$g\Lambda g = (\Lambda^{T})^{-1} \tag{1.6.285}$$

so that (6.284) can also be rewritten in the form

$$\bar{x}_{\alpha} = \sum_{\nu} K^{\alpha\nu} x_{\nu} \tag{1.6.286}$$

where

$$K = (\Lambda^T)^{-1} = (\Lambda^{-1})^T. \tag{1.6.287}$$

Show for comparison that (6.272), upon a suitable relabeling of indices, takes the form

$$\bar{x}^\alpha = \sum_\nu \Lambda^{\alpha\nu} x^\nu. \tag{1.6.288}$$

Evidently (6.286) and(6.288) take the same form with K given in terms of Λ by the relation (6.287). At this point we remark that it can be shown that K is a Lorentz transformation matrix if Λ is, and vice versa. See Exercise 6.2.6. For a further discussion of the relation (6.287), see Exercise 3.7.37 and (3.7.241).

You have completed the first task. You have shown that the covariant components of a four vector transform according to the rule given by (6.286) and (6.287). But now somewhat more can be said. Show that (6.286), (6.287), and (6.280) can be combined to give the relation

$$\bar{x}_\alpha = \sum_\nu K^{\alpha\nu} x_\nu = \sum_\nu (K^T)^{\nu\alpha} x_\nu = \sum_\nu (\Lambda^{-1})^{\nu\alpha} x_\nu = \sum_\nu (\partial x^\nu / \partial \bar{x}^\alpha) x_\nu. \tag{1.6.289}$$

You have shown that the covariant components of a four vector transform according to the rule

$$\bar{x}_\alpha = \sum_\nu (\partial x^\nu / \partial \bar{x}^\alpha) x_\nu. \tag{1.6.290}$$

Show for comparison that (6.277), upon a suitable relabeling of indices, takes the form

$$\bar{x}^\alpha = \sum_\nu (\partial \bar{x}^\alpha / \partial x^\nu) x^\nu. \tag{1.6.291}$$

Evidently, (6.290) and (6.291) are related by the symbol interchange $\partial x^\nu \leftrightarrow \partial \bar{x}^\alpha$. Note also that comparison of (6.286) and (6.290) gives the relation

$$\partial x^\nu / \partial \bar{x}^\alpha = K^{\alpha\nu}. \tag{1.6.292}$$

It should be compared with (6.276), which we rewrite in the form

$$\partial \bar{x}^\alpha / \partial x^\nu = \Lambda^{\alpha\nu}. \tag{1.6.293}$$

Note that again there is the symbol interchange $\partial x^\nu \leftrightarrow \partial \bar{x}^\alpha$.

Your second task is to apply what you have learned about the transformation properties of four vectors to the case of general tensors. To begin, suppose there are quantities B^μ which obey the four vector contravariant transformation rule

$$\bar{B}^\alpha = \sum_\mu \Lambda^{\alpha\mu} B^\mu = \sum_\mu (\partial \bar{x}^\alpha / \partial x^\mu) B^\mu, \tag{1.6.294}$$

and suppose there are quantities C_μ which obey the four vector covariant transformation rule

$$\bar{C}_\alpha = \sum_\mu K^{\alpha\mu} C_\mu = \sum_\mu (\partial x^\mu / \partial \bar{x}^\alpha) C_\mu. \tag{1.6.295}$$

Verify immediately from (6.48) through (6.51) the result

$$\sum_{\alpha} B^{\alpha} C_{\alpha} = \sum_{\alpha} B_{\alpha} C^{\alpha}, \tag{1.6.296}$$

that this result is independent of the transformation rules, and that analogous results hold for the $\bar{B}$ and $\bar{C}$ components. Next verify that there is the more subtle result

$$\sum_{\alpha} \bar{B}^{\alpha} \bar{C}_{\alpha} = \sum_{\alpha\mu\nu} \Lambda^{\alpha\mu} K^{\alpha\nu} B^{\mu} C_{\nu} = \sum_{\alpha\mu\nu} (\Lambda^{T})^{\mu\alpha} K^{\alpha\nu} B^{\mu} C_{\nu} =$$

$$= \sum_{\mu\nu} (\Lambda^{T} K)^{\mu\nu} B^{\mu} C_{\nu} = \sum_{\mu\nu} I^{\mu\nu} B^{\mu} C_{\nu} = \sum_{\nu} B^{\nu} C_{\nu}. \tag{1.6.297}$$

[Here we have used the dummy index principle to replace (6.295) by the equivalent statement $\bar{C}_{\alpha} = \sum_{\nu} K^{\alpha\nu} C_{\nu} = \sum_{\nu} (\partial x^{\nu} / \partial \bar{x}^{\alpha}) C_{\nu}$.] That is, the quantity $\sum_{\alpha} \bar{B}^{\alpha} \bar{C}_{\alpha}$ is *invariant* (has the value $\sum_{\nu} B^{\nu} C_{\nu}$) no matter what Lorentz transformation Λ may be. Indeed, Λ need not even be a Lorentz transformation. Evidently the invariance relation (6.297) holds for any (but nonsingular) matrix Λ in any number of dimensions. Finally, the invariance relation (6.297) holds for *all* (invertible) *maps* $\mathcal{M}$, including possibly *nonlinear* maps, that send quantities x^{ν} to quantities $\bar{x}^{\alpha}$ because the $\bar{B}^{\alpha}$ and $\bar{C}_{\alpha}$ can also be defined in terms of the B^{μ} and C_{μ} using only partial derivatives of $\mathcal{M}$ and its inverse. See the far right sides of (6.294) and (6.295). Therefore, although our discussion began in the context of Special Relativity, the results we have found may also be applicable in other contexts.

The invariance principle that the four-vector contravariant and covariant transformation properties compensate each other so that (6.297) holds can be extended to general tensors. For example, let $T^{\mu\nu\bullet\tau}_{\bullet\bullet\sigma\bullet}$ be a quantity that depends on the contravarient indices $\mu\nu\tau$ and the covariant index σ. Here, to keep track of index positions, we have placed $\bullet$ symbols below contravarient indices and above covariant indices to indicate where these indices would go if they were lowered or raised, respectively. The quantities $T^{\mu\nu\bullet\tau}_{\bullet\bullet\sigma\bullet}$ are said to comprise a (mixed rank 4) tensor if they transform according to the rule

$$\bar{T}^{\alpha\beta\bullet\delta}_{\bullet\bullet\gamma\bullet} = \sum_{\mu\nu\sigma\tau} \Lambda^{\alpha\mu} \Lambda^{\beta\nu} K^{\gamma\sigma} \Lambda^{\delta\tau} T^{\mu\nu\bullet\tau}_{\bullet\bullet\sigma\bullet}. \tag{1.6.298}$$

That is, Λ matrices are used for contravariant indices and K matrices are used for covariant indices.[50] Now pick a pair of indices associated with T, one being contravariant and one being covariant. For example, the pair could be the first contravariant index (which in this example would be μ or α) and the only covariant index (which in this example would be the third index and therefore would be σ or γ). Form the rank 2 objects $S^{\nu\tau}_{\bullet\bullet}$ and $\bar{S}^{\beta\delta}_{\bullet\bullet}$ by the rules

$$S^{\nu\tau}_{\bullet\bullet} = \sum_{\theta} T^{\theta\nu\bullet\tau}_{\bullet\bullet\theta\bullet}, \tag{1.6.299}$$

[50] Some authors write relations such as (6.288) in the form $\bar{x}^{\alpha} = \sum_{\nu} \Lambda^{\alpha\bullet}_{\bullet\nu} x^{\nu}$ and would also use both contravariant and covariant indices on the Λ and K appearing in expressions such as (6.298). Although doing so appears to neatly marry indices, we do not think such notation is a good idea because it makes Λ and K look like tensors, which they are not. They are transformation coefficients.

$$\bar{S}^{\beta\delta}_{\bullet\bullet} = \sum_{\theta} \bar{T}^{\theta\beta\bullet\delta}_{\bullet\bullet\theta\bullet}. \tag{1.6.300}$$

Verify that employing (6.298) in (6.300) and then using (6.299) yields the result

$$\begin{aligned}
\bar{S}^{\beta\delta}_{\bullet\bullet} &= \sum_{\theta} \bar{T}^{\theta\beta\bullet\delta}_{\bullet\bullet\theta\bullet} = \sum_{\theta\mu\nu\sigma\tau} \Lambda^{\theta\mu}\Lambda^{\beta\nu}K^{\theta\sigma}\Lambda^{\delta\tau}T^{\mu\nu\bullet\tau}_{\bullet\bullet\sigma\bullet} \\
&= \sum_{\theta\mu\nu\sigma\tau} (\Lambda^T)^{\mu\theta}\Lambda^{\beta\nu}K^{\theta\sigma}\Lambda^{\delta\tau}T^{\mu\nu\bullet\tau}_{\bullet\bullet\sigma\bullet} = \sum_{\theta\mu\nu\sigma\tau} \Lambda^{\beta\nu}(\Lambda^T)^{\mu\theta}K^{\theta\sigma}\Lambda^{\delta\tau}T^{\mu\nu\bullet\tau}_{\bullet\bullet\sigma\bullet} \\
&= \sum_{\mu\nu\sigma\tau} \Lambda^{\beta\nu}(\Lambda^T K)^{\mu\sigma}\Lambda^{\delta\tau}T^{\mu\nu\bullet\tau}_{\bullet\bullet\sigma\bullet} = \sum_{\mu\nu\sigma\tau} \Lambda^{\beta\nu}(I)^{\mu\sigma}\Lambda^{\delta\tau}T^{\mu\nu\bullet\tau}_{\bullet\bullet\sigma\bullet} \\
&= \sum_{\mu\nu\tau} \Lambda^{\beta\nu}\Lambda^{\delta\tau}T^{\mu\nu\bullet\tau}_{\bullet\bullet\mu\bullet} = \sum_{\nu\tau} \Lambda^{\beta\nu}\Lambda^{\delta\tau}\sum_{\mu} T^{\mu\nu\bullet\tau}_{\bullet\bullet\mu\bullet} \\
&= \sum_{\nu\tau} \Lambda^{\beta\nu}\Lambda^{\delta\tau}S^{\nu\tau}_{\bullet\bullet}.
\end{aligned} \tag{1.6.301}$$

You have shown that the quantities $S^{\nu\tau}_{\bullet\bullet}$ comprise a second rank contravariant tensor. The process you have executed is called *contraction.* Evidently contraction can be carried out as often as desired or possible, thereby yielding tensors of successively lower ranks by decrements of 2, until only contravariant or covariant indices remain (depending on which were more abundant initially) or no indices remain if contravariant and covariant indices were equally abundant initially. Also, if there are multiple ways of choosing contravariant and covariant pairs (as there are in this example), the net result generally depends on the choice(s) of pairs. Finally, to put our findings another way, we may say that the operations of tensor transformation and tensor contraction *commute.* That is, we may first contract one or some index pairs and then transform using the remaining indices, or we may first transform using all indices and then contract. Both operation orders yield the same result.

We have seen that contravariant and covariant components of vectors and tensors are characterized by their transformation properties. Your last task in this exercise is to apply this concept to the relations (6.55). Consider the differential operators $\partial/\partial\bar{x}^\alpha$ and $\partial/\partial x^\beta$. According to the chain rule they are related by the equation

$$\partial/\partial\bar{x}^\alpha = \sum_\beta (\partial x^\beta/\partial\bar{x}^\alpha)(\partial/\partial x^\beta). \tag{1.6.302}$$

As in (6.55), make the definitions and index assignments/placements

$$\bar{\partial}_\alpha = \partial/\partial\bar{x}^\alpha, \tag{1.6.303}$$

$$\partial_\beta = \partial/\partial x^\beta. \tag{1.6.304}$$

Also verify that, by relabeling indices, (6.292) can be rewritten in the form

$$K^{\alpha\beta} = \partial x^\beta/\partial\bar{x}^\alpha. \tag{1.6.305}$$

Finally, using (6.303), (6.304), and (6.302), verify that (6.302) can be rewritten in the form

$$\bar{\partial}_\alpha = \sum_\beta K^{\alpha\beta}\partial_\beta. \tag{1.6.306}$$

Observe, by comparing (6.295) and (6.306), that (6.306) is the expected transformation rule for *covariant* components, and therefore the index placements in (6.303) and (6.304) are correct.

1.7 Definition of Poisson Bracket

In subsequent chapters we will learn that Hamiltonian dynamics can be placed in a Lie-algebraic context. Key to this placement is the *Poisson bracket*.[51] In this section we will review its definition and some of its properties.

Let $H(q,p,t)$ be the Hamiltonian for some dynamical system and let f be any *dynamical variable*. That is, let $f(q,p,t)$ be any function of the phase-space variables q,p and the time t. Consider the problem of computing the *total* time rate of change of f along a trajectory generated by H. According to the chain rule, this derivative is given by the expression

$$df/dt = \partial f/\partial t + \sum_i \{(\partial f/\partial q_i)\dot{q}_i + (\partial f/\partial p_i)\dot{p}_i\}. \tag{1.7.1}$$

However, the $\dot{q}$'s and $\dot{p}$'s are given by Hamilton's equations of motion (5.11). Consequently, the expression for df/dt can also be written in the form

$$df/dt = \partial f/\partial t + \sum_i \{(\partial f/\partial q_i)(\partial H/\partial p_i) - (\partial f/\partial p_i)(\partial H/\partial q_i)\}. \tag{1.7.2}$$

The second quantity appearing on the right side of (7.2) occurs so often that it is given a special symbol and a special name in honor of Poisson. Let f and g be any two functions of the variables q,p,t. Then the Poisson bracket of f and g, denoted by the symbol $[f,g]$, is another function defined by the equation

$$[f,g] = \sum_i \{(\partial f/\partial q_i)(\partial g/\partial p_i) - (\partial f/\partial p_i)(\partial g/\partial q_i)\}. \tag{1.7.3}$$

With this new notation, (7.2) can be written in the compact form

$$df/dt = \partial f/\partial t + [f,H]. \tag{1.7.4}$$

The Poisson bracket operation has three important and obvious properties that are easily checked from its definition:

1. Distributive property,
$$[(af+bg),h] = a[f,h] + b[g,h] \tag{1.7.5}$$
for arbitrary constants a,b.

2. Antisymmetry condition,
$$[f,g] = -[g,f]. \tag{1.7.6}$$

[51] Poisson (1781-1840) was a student of Lagrange and Laplace and, at age 25, succeeded Fourier as a professor at the École Polytechnique.

3. Derivation with respect to multiplication,

$$[f, gh] = [f, g]h + g[f, h]. \tag{1.7.7}$$

(For those unfamiliar with the term, a *derivation* is an operation that behaves like "differentiation" in the sense that it obeys a product rule analogous to the product rule for differentiating a product in ordinary calculus.) From the definition one also easily finds the so-called *fundamental* Poisson brackets,

$$\begin{aligned} [q_i, q_j] &= 0, \\ [p_i, p_j] &= 0, \\ [q_i, p_j] &= \delta_{ij}. \end{aligned} \tag{1.7.8}$$

At this point it is convenient to introduce a more compact notation for the phase-space variables $(q_1 \cdots q_n), (p_1 \cdots p_n)$. To do this, introduce the $2n$ variables $(z_1, \ldots, z_{2n})$ by the rule

$$z = (z_1, \ldots, z_n; z_{n+1}, \ldots, z_{2n}) = (q_1, \ldots, q_n; p_1, \ldots, p_n). \tag{1.7.9}$$

That is, the first n z's are the q's and the last n z's are the p's. We will also adopt the convention of using lower case latin letters near the beginning of the alphabet to denote indices that range from 1 to $2n$.

With the definition (7.9), it is easily verified that the fundamental Poisson brackets (7.8) can also be written in the form

$$[z_a, z_b] = J_{ab}. \tag{1.7.10}$$

Here J is a $2n \times 2n$ matrix defined in block form by the equation

$$J = \begin{pmatrix} \mathbf{0} & I \\ -I & \mathbf{0} \end{pmatrix}, \tag{1.7.11}$$

where each entry in J is an $n \times n$ matrix, I denotes the $n \times n$ identity matrix, and all other entries are zero. The matrix J is sometimes called the *Poisson* matrix.

Suppose functions f and g of the variables q, p, t are written more compactly as $f(z, t)$, $g(z, t)$. Then the general Poisson bracket (7.3) can be written more compactly in the form

$$[f, g] = \sum_{a,b} (\partial f/\partial z_a) J_{ab} (\partial g/\partial z_b). \tag{1.7.12}$$

Suppose further that the $2n$ quantities $(\partial f/\partial z_a)$ are viewed as the components of a vector conveniently written as $\partial_z f$, and similarly for the quantities $(\partial g/\partial z_b)$. Then the right side of (7.12) can be viewed as a combination of two vectors and a matrix that can be written even more compactly using matrix and scalar product notation,

$$[f, g] = (\partial_z f, J\partial_z g). \tag{1.7.13}$$

Exercises

1.7.1. Verify the relations (7.5) through (7.7).

1.7.2. Derive (5.14) using (7.4) and (7.6). Show that if H does not explicitly depend on time, then it is a constant of motion and an integral of motion. See Section 5.2 for the definitions of constants and integrals of motion.

1.7.3. Verify (7.8) and (7.10).

1.7.4. Verify (7.12) and (7.13).

1.7.5. Review Exercise 6.7. Recall the Lorentz invariant Hamiltonian H_R given by (6.77) and the associated equations of motion (6.80) through (6.82). The purpose of this exercise is to study Poisson brackets in the context of a manifestly Lorentz invariant Hamiltonian formulation of the equations of motion.

a) Using the x^μ and p_μ as phase-space variables, suppose $f(x^\mu, p_\mu, \tau)$ is any dynamical variable. Repeat the steps (7.1) through (7.4) to show that in this case the Poisson bracket should be defined by the rule

$$[f, g] = \sum_\mu [(\partial f/\partial x^\mu)(\partial g/\partial p_\mu) - (\partial f/\partial p_\mu)(\partial g/\partial x^\mu)]. \tag{1.7.14}$$

As a consequence of this rule, show that

$$[x^\mu, x^\nu] = 0,\ [p_\mu, p_\nu] = 0,\ [x^\mu, p_\nu] = \delta^\mu_\nu \tag{1.7.15}$$

where δ^μ_ν is defined, as expected, by the equations

$$\begin{aligned} \delta^\mu_\nu &= 0 \text{ for } \mu \neq \nu, \\ &= 1 \text{ for } \mu = \nu. \end{aligned} \tag{1.7.16}$$

Thus, the x^μ and p_ν are canonically conjugate variables. Next, show that

$$[x^\mu, p^\nu] = g^{\mu\nu}. \tag{1.7.17}$$

b) Also show, based on (6.54), (6.59), (6.61), and (7.14), that in the presence of an electromagnetic field there are the Poisson bracket relations

$$[p^{\text{mech}}_\mu, p^{\text{mech}}_\nu] = qF_{\mu\nu} \quad \text{and} \quad [(p^{\text{mech}})^\mu, (p^{\text{mech}})^\nu] = qF^{\mu\nu}. \tag{1.7.18}$$

As a special case of (7.18), show that there are the relations

$$[p^{\text{mech}}_x, p^{\text{mech}}_y] = [(p^{\text{mech}})^1, (p^{\text{mech}})^2] = qF^{12} = -qB_z, \text{etc.} \tag{1.7.19}$$

We see that the *mechanical* momenta are *not* canonical variables because, unlike the corresponding relation in (7.15), the right sides in (7.18) are nonzero. It follows that the equations of motion (6.67) and (6.69), although first order, are *not* canonical because they involve mechanical momenta. That is, these equations of motion do not arise from any Hamiltonian. Similarly, show that converting the second-order set of equations (6.68) or (6.95) into an associated first-order set using the method of Section 1.3 yields noncanonical equations. Thus these equations of motion are not particularly useful if one wishes to exploit the symplectic (canonical) symmetry associated with Hamiltonian systems. See Exercise 6.4.11.

c) From (6.42) there is the relation

$$t = x^4/c. \tag{1.7.20}$$

Define p_t by the rule

$$p_t = -cp^4 = -cp_4. \tag{1.7.21}$$

Then, from (7.15) with $\mu = \nu = 4$, show that there is the result

$$[t, p_t] = [x^4/c, -cp_4] = [x^4, -p_4] = -1. \tag{1.7.22}$$

Also, again from (7.15), show that there are the results

$$[x^\mu, p^\nu] = -[x^\mu, p_\nu] = -\delta^{\mu\nu} \text{ for } \mu, \nu = 1 \cdots 3 \tag{1.7.23}$$

where, as again expected, $\delta^{\mu\nu}$ is defined for $\mu, \nu = 1 \cdots 3$ by the equations

$$\begin{aligned} \delta^{\mu\nu} &= 0 \text{ for } \mu \neq \nu, \\ &= 1 \text{ for } \mu = \nu. \end{aligned} \tag{1.7.24}$$

Evidently the quantities x^μ, p^ν for $\mu, \nu = 1 \cdots 3$ and t, p_t behave like canonical variables save for an annoying/alarming minus sign. We would, in fact, like to use the variables x^μ, p^ν for $\mu, \nu = 1 \cdots 3$ because then all indices are up so that we do not have to distinguish between up and down indices, and can eventually even forget about their position. Also, up index quantities are directly related to variables of interest without any additional minus signs. Contrast, for example, (6.42) and (6.52).

d) What to do? Suppose we define a new Hamiltonian $\hat{H}_R$ by the rule

$$\hat{H}_R = H_R/(-1) = -H_R. \tag{1.7.25}$$

With this definition in mind, check that the equations of motion (6.80) and (6.81) yield for the variables x^μ, p^ν for $\mu, \nu = 1 \cdots 3$ and t, p_t the results

$$(x')^\mu = \partial H_R/\partial p_\mu = -\partial H_R/\partial p^\mu = \partial \hat{H}_R/\partial p^\mu \text{ for } \mu = 1 \cdots 3, \tag{1.7.26}$$

$$t' = (1/c)(x')^4 = (1/c)\partial H_R/\partial p_4 = (1/c)\partial H_R/\partial p^4 = -\partial H_R/\partial p_t = \partial \hat{H}_R/\partial p_t; \tag{1.7.27}$$

$$(p')^\mu = -(p')_\mu = \partial H_R/\partial x^\mu = -\partial \hat{H}_R/\partial x^\mu \text{ for } \mu = 1 \cdots 3, \tag{1.7.28}$$

$$(p_t)' = -c(p')_4 = c\partial H_R/\partial x_4 = \partial H_R/\partial t = -\partial \hat{H}_R/\partial t. \tag{1.7.29}$$

Upon examining the far left and far right sides of (7.26) through (7.29) verify that, if we agree to use the Hamiltonian $\hat{H}_R$ instead of H_R, then we may *redefine* the fundamental Poisson brackets for the variables x^μ, p^ν (with $\mu, \nu = 1 \cdots 3$) and t, p_t to be the standard ones:

$$[x^\mu, t] = 0 \text{ for } \mu = 1 \cdots 3, \tag{1.7.30}$$

$$[x^\mu, x^\nu] = 0 \text{ for } \mu, \nu = 1 \cdots 3; \tag{1.7.31}$$

$$[p^\mu, p_t] = 0 \text{ for } \mu = 1 \cdots 3, \tag{1.7.32}$$

$$[p^\mu, p^\nu] = 0 \text{ for } \mu, \nu = 1 \cdots 3; \tag{1.7.33}$$

$$[t, p^\mu] = 0 \text{ for } \mu = 1 \cdots 3, \tag{1.7.34}$$

$$[x^\mu, p_t] = 0 \text{ for } \mu = 1 \cdots 3, \tag{1.7.35}$$

$$[t, p_t] = 1, \tag{1.7.36}$$

$$[x^\mu, p^\nu] = \delta^{\mu\nu} \text{ for } \mu, \nu = 1 \cdots 3. \tag{1.7.37}$$

Note that the relation between H and K given by (6.126) contains a minus sign just like the relation (7.25) between $\hat{H}_R$ and H_R. We also observe that the replacement of -1 by 1 in the Poisson bracket rules and the replacement of H_R by $\hat{H}_R = -H_R$ in the equations of motion is a special case of what we may call a *scaling* transformation. See Subsection 13.1.5.

e) Suppose there is no electromagnetic field so that all the components A^μ vanish. For the identifications (6.42), (6.101), and (6.102) show that (7.36) and (7.37) imply the relations

$$[x, p_x] = [y, p_y] = [z, p_z] = [t, -\mathcal{E}] = 1. \tag{1.7.38}$$

Observe that these relations are consistent with (7.8) and (6.105).

1.7.6. Suppose we employ the Hamiltonian H given by (5.49). Note that in this case there is no mention of up and down index quantities. There are simply the components of the vectors $\boldsymbol{r}$ and $\boldsymbol{p}^{\text{can}}$. That is, there are the dynamical variables $(x, y, z; p_x^{\text{can}}, p_y^{\text{can}}, p_z^{\text{can}})$, and the time t is treated as the independent variable. For this Hamiltonian follow the recipe of Section 1.7 to define Poisson brackets. Show that doing so yields the result that all Poisson brackets involving only components of $\boldsymbol{r}$ and $\boldsymbol{p}^{\text{can}}$ vanish save for

$$[x, p_x^{\text{can}}] = [y, p_y^{\text{can}}] = [z, p_z^{\text{can}}] = 1. \tag{1.7.39}$$

Show that for this definition of the Poisson bracket there are the results

$$[p_x^{\text{mech}}, p_y^{\text{mech}}] = qB_z, \text{ etc.} \tag{1.7.40}$$

Note that (7.19) and (7.40) differ by a sign. This difference occurs because the definition of the Poisson bracket depends on what Hamiltonian is being employed.

1.7.7. Suppose we employ the Hamiltonian K given by (6.16). In this case the dynamical variables are $(x, y, t; p_x, p_y, p_t)$ and z is the independent variable. Note that, although not indicated by our imprecise notation, the quantities (p_x, p_y, p_t) are canonical and not mechanical momenta. For this Hamiltonian follow the recipe of Section 1.7 to define Poisson brackets. Show that doing so yields the result that all poisson brackets among the dynamical variables $(x, y, t; p_x, p_y, p_t)$ vanish save for

$$[x, p_x] = [y, p_y] = [t, p_t] = 1. \tag{1.7.41}$$

Show that for this definition of the Poisson bracket there is the result

$$[p_x^{\text{mech}}, p_y^{\text{mech}}] = qB_z. \tag{1.7.42}$$

Note that (7.19) and (7.42) differ by a sign. This difference occurs because the definition of the Poisson bracket depends on what Hamiltonian is being employed

1.7.8. Suppose that

$$\psi = 0 \tag{1.7.43}$$

in (5.1), (5.49), and (6.16) so that K takes the form

$$K = -[(p_t/c)^2 - m^2c^2 - (p_x - qA_x)^2 - (p_y - qA_y)^2]^{1/2} - qA_z. \tag{1.7.44}$$

Show from (5.27) through (5.30), (6.5), and (7.43) that there are the relations

$$[(\boldsymbol{p}^{\text{mech}})^2]^{1/2} = p^{\text{mech}} = \gamma mv = \gamma\beta mc, \tag{1.7.45}$$

$$p_t = -[m^2c^4 + (\boldsymbol{p}^{\text{mech}}c)^2]^{1/2} = -\gamma mc^2, \tag{1.7.46}$$

$$p_t^2 = m^2c^4 + (\boldsymbol{p}^{\text{mech}}c)^2, \tag{1.7.47}$$

$$v = c[1 - (mc^2/p_t)^2]^{1/2}. \tag{1.7.48}$$

Here β and γ are the usual relativistic factors,

$$\beta = v/c, \tag{1.7.49}$$

$$\gamma = (1 - \beta^2)^{-1/2}. \tag{1.7.50}$$

The quantity p_t obeys the equation of motion

$$dp_t/dz = -\partial K/\partial t. \tag{1.7.51}$$

See (6.10). Therefore if $\boldsymbol{A}$ is time independent (which amounts to the case of motion in a static magnetic field), there are the relations

$$\partial K/\partial t = 0, \tag{1.7.52}$$

$$p_t = \text{constant}. \tag{1.7.53}$$

From (7.46) and (7.48) through (7.50) show that in this case β, γ, and v are also constants of motion.

In Accelerator Physics, when studying orbits in a magnetic field, it is common to introduce the quantity δ by the definition

$$p^{\text{mech}} = (1 + \delta)p_0^{\text{mech}} \tag{1.7.54}$$

where $p_0^{\rm mech} = ||\boldsymbol{p}_0^{\rm mech}||$ is the magnitude of some reference or *design* mechanical momentum and $p^{\rm mech} = ||\boldsymbol{p}^{\rm mech}||$ is the magnitude of the actual mechanical momentum. The quantity δ is called the *momentum deviation*. By combining (7.47) and (7.54) show that there are the relations

$$p_t^2 = m^2c^4 + (1+\delta)^2(p_0^{\rm mech}c)^2, \tag{1.7.55}$$

$$\delta = [(p_t^2 - m^2c^4)^{1/2}/(p_0^{\rm mech}c)] - 1. \tag{1.7.56}$$

Consider the quantity ℓ defined by

$$\ell = (p_0^{\rm mech}c)[1-(mc^2/p_t)^2]^{1/2}t. \tag{1.7.57}$$

Show from (7.48) that ℓ can also be written in the form

$$\ell = (p_0^{\rm mech})vt. \tag{1.7.58}$$

Evidently, if v is constant (which will be the case for motion in a static magnetic field), the quantity ℓ is proportional to *path length* with proportionality constant $p_0^{\rm mech}$. Note that the quantity ℓ is still defined by (7.57) in the time-dependent case, but then it has no such simple physical interpretation. Show, however, that in the extreme relativistic limit $-p_t \gg mc^2$ where $v \simeq c$ there is the relation

$$\ell \simeq (p_0^{\rm mech})ct \tag{1.7.59}$$

so that in this limit the interpretation of ℓ as being proportional to path length is regained even in the time-dependent case.

Show, starting from the known Poisson bracket relation

$$[t, p_t] = 1, \tag{1.7.60}$$

that there is the relation

$$[\delta, \ell] = 1. \tag{1.7.61}$$

Also show that there are the relations

$$[x,\delta] = [y,\delta] = [p_x,\delta] = [p_y,\delta] = 0,$$

$$[x,\ell] = [y,\ell] = [p_x,\ell] = [p_y,\ell] = 0. \tag{1.7.62}$$

Thus, δ and ℓ are *canonically conjugate* with δ being "coordinate like" and ℓ being "momentum like". See (7.8). We may therefore view the quantities $x, p_x; y, p_y; \delta, \ell$ as a set of canonical coordinates obtained from the set $x, p_x, ; y, p_y; t, p_t$ by a canonical transformation. (Recall that a canonical transformation is a transformation that preserves the fundamental Poisson brackets. See Section 6.1.2.)

Show that there are the inverse relations

$$p_t = -[m^2c^4 + (1+\delta)^2(p_0^{\rm mech}c)^2]^{1/2}, \tag{1.7.63}$$

$$t = [\ell/(p_0^{\rm mech}c)]\{1 - m^2c^4/[m^2c^4 + (1+\delta)^2(p_0^{\rm mech}c)^2]\}^{-1/2}. \tag{1.7.64}$$

If a canonical transformation does not depend explicitly on the independent variable (the quantity z in the case), then the new Hamiltonian $\bar{K}$ equals the old Hamiltonian K expressed in terms of the new variables,

$$\bar{K}\{x,p_x,y,p_y,\delta,\ell;z\} = K\{x,p_x,y,p_y,t(\delta,\ell),p_t(\delta,\ell);z\}. \tag{1.7.65}$$

(See Appendix D.) Show, using (7.55) and (7.65), that

$$\bar{K} = -[(1+\delta)^2(p_0^{\rm mech})^2 - (p_x - q\bar{A}_x)^2 - (p_y - q\bar{A}_y)^2]^{1/2} - q\bar{A}_z \tag{1.7.66}$$

where

$$\bar{\boldsymbol{A}}\{\boldsymbol{r},\delta,\ell\} = \boldsymbol{A}\{\boldsymbol{r},t(\delta,\ell)\}. \tag{1.7.67}$$

If all is well, there should be the relation

$$d\ell/dz = [\ell,\bar{K}] = -\partial\bar{K}/\partial\delta. \tag{1.7.68}$$

See (7.4). Show, from (6.10) and (7.65), that the right side of (7.68) is given by the relation

$$\partial\bar{K}/\partial\delta = (\partial K/\partial t)(\partial t/\partial\delta) + (\partial K/\partial p_t)(\partial p_t/\partial\delta) = -(dp_t/dz)(\partial t/\partial\delta) + (dt/dz)(\partial p_t/\partial\delta). \tag{1.7.69}$$

Evaluate the partial derivatives on the right side of (7.69) using (7.63) and (7.64) to find the results

$$(\partial t/\partial\delta) = -(p_0^{\rm mech})t/(mc\beta\gamma^3), \tag{1.7.70}$$

$$(\partial p_t/\partial\delta) = -(p_0^{\rm mech})v. \tag{1.7.71}$$

Evaluate the left side of (7.68) using (7.57), and verify that (7.68) is correct. Similarly, verify that

$$d\delta/dz = [\delta,\bar{K}]. \tag{1.7.72}$$

Sometimes it is convenient to introduce scaled variables P_x, P_y, and $\hat{\ell}$ by the rules

$$P_x = p_x/p_0^{\rm mech}, \tag{1.7.73}$$

$$P_y = p_y/p_0^{\rm mech}, \tag{1.7.74}$$

$$\hat{\ell} = \ell/p_0^{\rm mech} = c[1-(mc^2/p_t)^2]^{1/2}t = vt. \tag{1.7.75}$$

See Section 13.1.5. Note that P_x and P_y are dimensionless. Also now, when v is constant, $\hat{\ell}$ *is* the path length. If we now regard the pairs x, P_x; y, P_y; and $\delta, \hat{\ell}$ as canonically conjugate, their evolution will be governed by the Hamiltonian $\hat{K}$ given by

$$\hat{K} = (1/p_0^{\rm mech})\bar{K} = -[(1+\delta)^2 - (P_x - q\hat{A}_x)^2 - (P_y - q\hat{A}_y)^2]^{1/2} - q\hat{A}_z \tag{1.7.76}$$

where

$$\hat{\boldsymbol{A}}\{\boldsymbol{r},\delta,\hat{\ell}\} = (1/p_0^{\rm mech})\boldsymbol{A}\{\boldsymbol{r},t(\delta,\hat{\ell})\}. \tag{1.7.77}$$

(Again see Appendix D.)

1.7.9. Review Exercise 7.6. It treated the Cartesian-coordinate Hamiltonian (6.16). Show that the cylindrical-coordinate Hamiltonian (6.18) can be treated analogously. Conclude that in this respect there is nothing special about the use of Cartesian coordinates.

1.7.10. Review Exercise 5.1 that related mechanical and canonical momentum. Show that the *mechanical* energy $E^{\rm mech}$ is given by the relation

$$E^{\rm mech} = \gamma mc^2 = [m^2c^4 + c^2(\boldsymbol{p}^{\rm mech})^2]^{1/2} = [m^2c^4 + c^2(\boldsymbol{p} - q\boldsymbol{A})^2]^{1/2}. \tag{1.7.78}$$

Review Exercise 5.3. Using the definition (6.5), show that

$$p_t = -E^{\rm mech} - q\psi. \tag{1.7.79}$$

Make the definition

$$p_t^{\rm mech} = -E^{\rm mech}, \tag{1.7.80}$$

in which case

$$p_t = p_t^{\rm mech} - q\psi = -\gamma mc^2 - q\psi, \tag{1.7.81}$$

which is a relation analogous to those in Exercise 5.1. Also compare the above results with (6.59), those of Exercise 6.11, and (7.21). Note, using (6.45) and (6.53), that there are the relations

$$A^4 = A_4 = \psi/c. \tag{1.7.82}$$

Bibliography

Maps, Map Iteration, Chaos, and Fractals

Entering the words *dynamical systems* or *chaos* or *fractal* into the *Amazon* search window produces overwhelming lists of books on these subjects. Also, Google Dynamical Systems-Scholarpedia, and Encyclopedia Dynamical Systems-Scholarpedia, and see the Web site http://www.scholarpedia.org.

[1] S. Abdullaev, *Construction of Mappings for Hamiltonian Systems and Their Applications*, Springer (2006).

[2] P. Cvitanović, R. Artuso, R. Mainieri, G.Tanner, G. Vattay, N. Whelan, and A. Wirzba, *Chaos: Classical and Quantum*, (2016). See the Web site http://chaosbook.org/chapters/ChaosBook.pdf. See also the Web site http://chaosbook.org.

[3] R. M. May, "Simple Mathematical Models with very Complicated Dynamics", *Nature*, **261**, 459 (1976).

[4] B. B. Mandelbrot, *The Fractal Geometry of Nature*, W.H. Freeman (1983).

[5] B. B. Mandelbrot, *Fractals and Chaos: The Mandelbrot Set and Beyond*, (Springer, 2004).

[6] H. G. Schuster and W. Just, *Deterministic Chaos, An Introduction*, (Wiley-VCH, 2005).

[7] H.-O. Peitgen and D. Saupe, Eds., *The Science of Fractal Images*, Springer-Verlag (1988).

[8] H.-O. Peitgen, H. Jürgens, D. Saupe, *Chaos and Fractals: New Frontiers of Science*, Springer-Verlag (1992).

[9] G. Velo and A. S. Wightman, edit., *Regular and Chaotic Motions in Dynamic Systems*, Plenum Press (1985).

[10] D. Arrowsmith and C. Place, *An Introduction to Dynamical Systems*, Cambridge University Press (1990).

[11] D. Arrowsmith and C. Place, *Dynamical Systems: differential equations, maps, and chaotic behavior*, Chapmann & Hall (1992).

[12] M. F. Barnsley, *Fractals Everywhere*, Academic Press (1993).

[13] K. Falconer, *The Geometry of Fractal Sets*, Cambridge University Press (1985).

[14] D. R. Hofstadter, *Mathematical Themas: Questing for the Essence of Mind and Pattern*, chapter 16, (Basic Books, 1996).

[15] G. L. Baker and J. P. Gollub, *Chaotic Dynamics: an introduction*, 2nd ed., (Cambridge University Press, 1996).

[16] A. J. Lichtenberg and M.A. Lieberman, *Regular and Chaotic Dynamics*, 2nd ed., (Springer, 1992).

[17] F.C. Moon, *Chaotic and Fractal Dynamics: An Introduction for Applied Scientists and Engineers*, (Wiley, 1992).

[18] R. L. Devaney, *An Introduction to Chaotic Dynamical Systems*, (Addison-Wesley, 1989).

[19] R. L. Devaney, *An Introduction to Chaotic Dynamical Systems*, (Benjamin/Cummings, 1986).

[20] R. L. Devaney and L. Keen, edit., *Chaos and Fractals: The Mathematics Behind the Computer Graphics*, Proceedings of the Symposia in Applied Mathematics, Vol. 39, (AMS, 1989).

[21] R. L. Devaney, *A First Course in Chaotic Dynamical Systems*, (Perseus, 1992).

[22] M. Hénon, Quarterly of Applied Mathematics **27**, 291 (1969).

[23] J. Moser, *On Quadratic Symplectic Mappings*, Math. Zeitschrift **216**, 417-430 (1994).

[24] M. Lyubich, *The Quadratic Family as a Qualitatively Solvable Model of Chaos*, Notices of the American Mathematical Society **47**, p. 1042 (2000).

[25] R. A. Holmgren, *A First Course in Discrete Dynamics*, Springer (1996).

[26] G. Contopoulos, *Order and Chaos in Dynamical Astronomy*, Springer (2004).

[27] E. Ott, *Chaos in Dynamical Systems*, Cambridge (2002).

[28] D. Ruelle, *Elements of Differentiable Dynamics and Bifurcation Theory*, Academic Press (1989).

[29] D. Ruelle, Edit., *Turbulence, Strange Attractors, and Chaos*, World Scientific (1995).

[30] D. Ruelle, "What Is a Strange Attractor?", *Notices of the American Mathematical Society* **53**, p. 764, August 2006.

[31] Y. Ueda, *The Road to Chaos*, Aerial Press (1992).

[32] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press (1999).

[33] B. Hasselblatt and A. Katok, *A First Course in Dynamics: with a Panorama of Recent Developments*, Cambridge University Press (2003).

[34] B. Hasselblatt and A. Katok, Edit., *Handbook of Dynamical Systems*, Vol 1A, Elsevier (2002).

[35] B. Hasselblatt and A. Katok, Edit., *Handbook of Dynamical Systems*, Vol 1B, Elsevier (2006).

[36] B. Fiedler, Edit., *Handbook of Dynamical Systems*, Vol 2, Elsevier (2002).

[37] H. Broer, B. Hasselblatt, and F. Takens, Edit., *Handbook of Dynamical Systems*, Vol 3, Elsevier (2010).

[38] M. Brin and G. Stuck, *Introduction to Dynamical Systems*, Cambridge University Press (2002).

[39] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*, Perseus Books/WestviewPress (1994 and 2015).

[40] S. S. Abdullaev, *Construction of Mappings for Hamiltonian Systems and Their Applications*, Springer (2006).

[41] G. Zaslavsky, R. Sagdeev, D. Usikov, and A. Chernikov, *Weak chaos and quasi-regular patterns*, Cambridge University Press (1991).

[42] G. Zaslavsky, *Physics of Chaos in Hamiltonian Systems*, Imperial College Press (1998).

[43] G. Zaslavsky, *Hamiltonian Chaos and Fractional Dynamics*, Oxford University Press (2005).

[44] J. Sprott, *Elegant Chaos: Algebraically Simple Chaotic Flows*, World Scientific (2010).

[45] T. Tél and M. Gruiz, *Chaotic Dynamics: An Introduction Based on Classical Mechanics*, Cambridge University Press (2006).

Maps in the Complex Domain

[46] H.-O. Peitgen and P.H. Richter, *The Beauty of Fractals, Images of Complex Dynamical Systems*, Springer-Verlag (1986).

[47] A. Douady and J. Hubbard, *Étude dynamique des polynômes complex*, Publications mathématiques d'Orsay, Université de Paris-Sud (1984).

[48] J. Hubbard, "The Hénon mapping in the complex domain", published in *Chaotic Dynamics and Fractals*, M. Barnsley and S. Demko, Eds., p. 101, Academic Press (1986).

[49] J. H. Hubbard and R. W. Oberste-Vorth, *Hénon Mappings in the Complex Domain I: The Global Topology of Dynamical Space*, Institut Des Hautes Étude Scientifiques Publications Mathématiques 79 (1994); *II: Projective and Inductive Limits of Polynomials, in Real and Complex Dynamical Systems*, B. Branner and P. Hjorth, eds., NATO ASI Series C: Mathematical and Physical Sciences Vol. 464 (Kluwer, Amsterdam 1995).

[50] J. Hubbard, *Bulletin of the American Mathematical Society* **38**, p. 495 (2001).

[51] H. Kriete, ed., *Progress in holomorphic dynamics*, Addison Wesley Longman (1998).

[52] S. Heinemann, "Julia sets for endomorphisms of C^n", *Ergodic Theory and Dynamical Systems* **16** p. 1275 (1996).

[53] K. Schmidt, *Dynamical Systems of Algebraic Origin*, Birkhäuser (1995).

[54] J. Smillie and G. T. Buzzard, "Complex Dynamics in Several Variables", published in *Flavors of Geometry*, S. Levy, Ed., Cambridge (1997).

[55] E. Bedford and J. Smillie, "Polynomial diffeomorphisms of C^2: currents, equilibrium measure and hyperbolicity", *Invent. Math.* **103**, 69 (1991).

[56] E. Bedford, M. Lyubich, and J. Smillie, "Polynomial diffeomorphisms of C^2, IV. The measure of maximal entropy and laminar curents", *Invent. Math.* **112**, 77 (1993).

[57] S. Friedland and and J. Milnor, "Dynamical properties of plane polynomial automorphisms", *Ergodic Theory Dyn. Syst.* **9**, p. 67 (1989).

[58] J. Milnor, *Dynamics in One Complex Variable*, Third Edition, Princeton University Press (2006).

[59] C. T. McMullen, *Complex Dynamics and Renormalization*, Annals of Mathematics Studies Number 135, Princeton University Press (1994).

[60] L. Carleson and T.W. Gamelin, *Complex Dynamics*, Springer-Verlag (1993).

[61] J. E. Fornaess, *Dynamics in Several Complex Variables*, Conference Board of the Mathematical Sciences Regional Conference Series in Mathematics Number 87, American Mathematical Society (1996).

[62] J. E. Fornaess and N. Sibony, "Complex Dynamics in Higher Dimension", in *Several Complex Variables*, M. Schneider and Y.-T. Siu, Edit., Cambridge University Press (1999).

[63] B. Branner and P. Hjorth, edits., *Real and Complex Dynamical Systems*, Kluwer Academic Publishers (1995).

[64] A. F. Beardon, *Iteration of Rational Functions*, Springer-Verlag (1991).

[65] N. Steinmetz, *Rational iteration, complex analytic dynamical systems*, de Gruyter (1993).

[66] R. L. Devaney, edit., *Complex Dynamical Systems: The Mathematics Behind the Mandelbrot and Julia Sets*, Proceedings of Symposia in Applied Mathematics, Vol. 49, American Mathematical Society (1994).

[67] V. Dolotin and A. Morozov, *The Universal Mandelbrot Set: Beginning of the Story*, World Scientific (2006).

[68] Mandelbrot set Web sites. Any search engine will find several Web sites devoted to the Mandelbrot set. Search under fractal, Julia, and Mandelbrot. Two such sites are listed below:
http://aleph0.clarku.edu/~djoyce/julia/explorer.html
http://math.bu.edu/DYSYS/

Universality and Renormalization

[69] M. J. Feigenbaum, "Quantitative universality for a class of non-linear transformations", *J. Statist. Phys.* **19**, 25 (1978).

[70] O. E. Lanford III, "A Shorter Proof of the Existence of the Feigenbaum Fixed Point", *Communications in Mathematical Physics* **96**, 521 (1984).

[71] P. Cvitanović, Ed., *Universality in Chaos*, Second Edition, Adam Hilger (1989).

[72] P. Collet and J.-P. Eckmann, *Iterated Maps on the Interval as Dynamical Systems*, Birkhäuser (1980).

[73] R. S. MacKay, *Renormalization in Area Preserving Maps*, Princeton University Ph.D. Thesis (1982).

[74] J. M. Greene, R. S. MacKay, F. Vivaldi, and M.J. Feigenbaum, "Universal Behaviour of Area-Preserving Maps", *Physica* **3D**, 468 (1981).

[75] T. C. Bountis, "Period-Doubling Bifurcations and Universality in Conservative Systems", *Physica* **3D**, 577 (1981).

Differential Equations and Dynamical Systems

[76] SIAM Dynamical Systems Web Site : http://www.siam.org/activity/ds/

[77] G. D. Birkhoff, *Dynamical Systems*, American Mathematical Society (1966).

[78] E. R. Scheinerman, *Invitation to Dynamical Systems*, Dover (2012).

[79] S. Sternberg, *Dynamical Systems*, Dover (2010).

[80] Garrett Birkhoff and G-C. Rota, *Ordinary Differential Equations*, 4th Ed., Wiley (1989).

[81] A. Andronov and C. Chaikin, *Theory of Oscillations*, Princeton University Press (1949).

[82] Francis J. Murray and Kenneth S. Miller, *Existence Theorems for Ordinary Differential Equations*, (New York University Press and Interscience Publishing Co. 1954).

[83] Y. Ilyashenko and S. Yakovenko, *Lectures on Analytic Differential Equations*, American Mathematical Society (2008).

[84] W. Walter, *Ordinary Differential Equations*, Springer (1998).

[85] H. Zoladek, *The Monodromy Group*, Birkhäser Verlag (2006).

[86] Fred Brauer and John A. Nohel, *The Qualitative Theory of Ordinary Differential Equations*, Benjamin (1969).

[87] Zhang Zhi-fen, Ding Tong-ren, Huang Wen-zao, Dong Zhen-xi, *The Qualitative Theory of Differential Equations*, American Mathematical Society (1992).

[88] Earl A. Coddington and Norman Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill (1955).

[89] Philip Hartman, *Ordinary Differential Equations*, Birkhäuser (1982).

[90] I. G. Petrovski, *Ordinary Differential Equations*, Prentice-Hall (1966).

[91] V. I. Arnold, *Ordinary Differential Equations*, third edition Springer Verlag (1992).

[92] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer Verlag (1983).

[93] E. Hille, *Ordinary Differential Equations in the Complex Domain*, John Wiley (1976).

[94] E. Hille, *Lectures on Ordinary Differential Equations*, Addison-Wesley (1969).

[95] J. H. Hubbard and B. H. West, *Differential Equations: a Dynamical Systems Approach*, Springer (1971).

[96] J. Hale and H. Kocak, *Dynamics and Bifurcations*, Springer Verlag (1991).

[97] M. W. Hirsch, S. Smale, and R. L. Devaney, *Differential Equations, Dynamical Systems, and an Introduction to Chaos*, Elsevier (2013).

[98] P. Blanchard, R. L. Devaney, and G.R. Hall, *Differential Equations*, Brooks/Cole (2002).

[99] S. Lefschetz, *Differential Equations: Geometrical Theory*, 2nd Edition, Interscience (1963).

[100] J. Meiss, *Differential Dynamical Systems*, SIAM (2007).

[101] L. Perko, *Differential Equations and Dynamical Systems*, third edition, (Springer 2002).

[102] C. Chicone, *Ordinary Differential Equations with Applications*, second edition, Springer-Verlag (2006).

[103] N. Minorsky, *Nonlinear Oscillations*, Krieger Publishing Company, New York (1974).

[104] P. Hagedorn, *Non-linear Oscillations*, Clarendon Press, Oxford (1982).

[105] V. M. Starzhinskii, *Applied Methods in the Theory of Nonlinear Oscillations*, Mir Publishers, Moscow (1980).

[106] J. K. Hale, *Ordinary Differential Equations*, Wiley-Interscience (1969).

[107] R. E. Bellman, *Stability Theory of Differential Equations*, Dover (1969).

[108] R. E. Bellman, *Methods of Nonlinear Analysis*, Vols. 1 and 2, Academic Press (1970).

[109] F. Verhulst, *Nonlinear Differential Equations and Dynamical Systems*, Springer-Verlag (1990).

[110] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag (1983).

[111] J. Guckenheimer, J. Moser, and S. Newhouse, *Dynamical Systems: C.I.M.E, Lectures*, Birkhäuser (1983).

[112] R. H. Rand and D. Armbruster, *Perturbation Methods, Bifurcation Theory, and Computer Algebra*, Springer-Verlag (1987).

[113] R. H. Rand, *Topics in Nonlinear Dynamics with Computer Algebra*, Gordan and Breach (1994).

[114] J. M. T. Thompson and H. B. Stewart, *Nonlinear Dynamics and Chaos*, Second Edition, John Wiley (2002).

[115] R. C. Hilborn, *Chaos and Nonlinear Dynamics: An Introduction for Scientists and Engineers*, Second Edition, Oxford University Press (2006).

[116] R. C. Robinson, *Dynamical Systems: Stability, Symbolic Dynamics, and Chaos*, CRC Press (1995).

[117] R. C. Robinson, *An Introduction to Dynamical Systems: Continuous and Discrete*, Pearson Prentice Hall (2004).

[118] H. Stephani, *Differential Equations: Their solution using symmetries*, M. Maccallum, Edit., Cambridge University Press (1989).

[119] P. J. Olver, *Applications of Lie Groups to Differential Equations*, Springer-Verlag (1993).

[120] R. Hermann, *Lie-Theoretic ODE Numerical Analysis, Mechanics, and Differential Systems*, Math Sci Press (1994).

[121] N. Ibragimov, *Transformation Groups and Lie Algebras*, World Scientific (2013).

Existence, Uniqueness, Differentiability, and Analyticity Theorems

[122] E. Hille, *Ordinary Differential Equations in the Complex Domain*, John Wiley (1976).

[123] E. Hille, *Lectures on Ordinary Differential Equations*, Addison-Wesley (1969).

[124] Earl A. Coddington and Norman Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill (1955).

[125] Francis J. Murray and Kenneth S. Miller, *Existence Theorems for Ordinary Differential Equations*, New York University Press and Interscience Publishing Co. (1954).

[126] W. Walter, *Ordinary Differential Equations*, Springer (1998).

[127] Y. Ilyashenko and S. Yakovenko, *Lectures on Analytic Differential Equations*, American Mathematical Society (2008).

[128] Zhang Zhi-fen, Ding Tong-ren, Huang Wen-zao, Dong Zhen-xi, *The Qualitative Theory of Differential Equations*, American Mathematical Society (1992).

[129] Philip Hartman, *Ordinary Differential Equations*, Birkhäuser (1982).

[130] H. Amann, *Ordinary Differential Equations: An Introduction to Nonlinear Analysis*, Walter de Gruyter (1990).

[131] S-N. Chow and J. Hale, *Methods of Bifurcation Theory*, Springer-Verlag (1982).

[132] R. Agarwal and V. Lakshmikantham, *Uniqueness and Nonuniqueness Criteria for Ordinary Differential Equations*, World Scientific (1993).

[133] M. Irwin, *Smooth Dynamical Systems*, Academic Press (1980).

[134] V. Nemytskii and V. Stepanov, *Qualitative Theory of Differential Equations*, Princeton University Press (1960).

[135] W. Hurewicz, *Lectures on Ordinary Differential Equations*, M.I.T. Press (1963), Dover (1990).

[136] L. Piccinini, G. Stampacchia, and G. Vidossich, *Ordinary Differential Equations in R^n: Problems and Methods*, Springer-Verlag (1984).

Solution for Monomial Hamiltonian

[137] F. J. Testa, *J. Math Phys.* **14**, p. 1097 (1973).

[138] P. J. Channell, *Explicit Integration of Kick Hamiltonians in Three Degrees of Freedom*, Accelerator Theory Note AT-6:ATN-86-6, Los Alamos National Laboratory (1986).

[139] I. Gjaja, *Particle Accelerators*, vol. 43 (3), pp. 133-144 (1994).

[140] L. Michelotti, *Comment on the exact evaluation of symplectic maps*, Fermilab preprint (1992).

Original Sources and Histories

[141] I. Newton, *The Principia*, Translated by B. Cohen and A. Whitman, University of California Press (1999).

[142] S. Chandrasekhar, *Newton's Principia for the Common Reader*, Clarendon Press (1995).

[143] C. Pask, *Magnificent Principia: Exploring Isaac Newton's Masterpiece*, Prometheus Books (2013).

[144] Jed Z. Buchwald and Mordechai Feingold, *Newton and the Origin of Civilization*, Princeton University Press (2013).

[145] Rob Iliffe, *Priest of Nature: The Religious Worlds of Isaac Newton*, Oxford University Press (2017).

[146] G. Alexanderson and L. Klosinski, "The Newton-Leibniz Controversy", *Bulletin of the American Mathematical Society* **53**, p. 295 (2016).

[147] G. W. Leibniz, *Von dem Verhängnisse,* in *Hauptschriften zur Grundlegung der Philosophie*, Vol II, pp. 129-134 (Ernst Cassirer, Leibzig 1906), cited by P. Cvitanović et al. in reference 1 above.

[148] R. J. Boscovich, *A Theory of Natural Philosophy*, reprinted by MIT Press, p. 141 (1966).

[149] P. S. Laplace, *A Philosophical Essay on Probabilities*, Chapter II, On or Concerning Probability, Dover (1951) and Springer-Verlag (1995).

[150] W. R. Hamilton, *Trans. R. Irish Acad.* **15**, 69 (1828); **16**, 1 (1830); **16**, 93 (1831); **17**, 1 (1837). Reprinted in *The Mathematical Papers of Sir W. R. Hamilton, Vol. I, Geometrical Optics*, A. W. Conway and J. L. Synge, eds., Cambridge U. Press, Cambridge (1931). All the mathematical papers of Hamilton may be found at https://www.maths.tcd.ie/pub/HistMath/People/Hamilton/Papers.html

[151] H. Poincaré, *New Methods of Celestial Mechanics,* Parts 1, 2, and 3. (Originally published as Les Méthodes nouvelles de la Méchanique céleste.) American Institute of Physics *History of Modern Physics and Astronomy*, Volume 13, D. L. Goroff, Edit., American Institute of Physics (1993). See page 145 of the Editor's Introduction for a statement of Poincaré's holomorphic lemma, and Sections §20 through §27 for Poincaré's proof. The Editor's Introduction also provides a very useful summary of Poincaré's contributions to Dynamical Systems and his legacy.

[152] Lizhen Ji and Athanase Papadopoulos, Edit., *Sophus Lie and Felix Klein: The Erlangen Program and Its Impact in Mathematics and Physics*, European Mathematical Society and American Mathematical Society (2015).

[153] G. Duffing, *Erzwungene Schwingungen bei veränderlicher Eigenfrequenz und ihre technische Bedeutung*, Braunschweig, Druck und Verlag von Friedr. Vieweg und Sohn (1918).

[154] E. D. Courant and H. Snyder, "Theory of the Alternating-Gradient Synchrotron", *Annals Phys.* **3**, p. 1 (1958).

[155] S. Penner, "Calculations of Properties of Magnetic Deflection Systems", *Rev. of Sci. Instr.* **32**, p. 150 (1961).

[156] K. L. Brown, *TRANSPORT*, SLAC-75, revision 3 (1975); K.L. Brown, F. Rothacker, D. C. Carey, and Ch. Iselin, *TRANSPORT*, SLAC-91, revision 2 (1977).

[157] G. Hori, "Theory of general perturbations with unspecified canonical variables", *Publications of the Astronomical Society of Japan* **18**, p. 287 (1966).

[158] A. Deprit, *Celest. Mech.* **1**, 12 (1969).

[159] J. Barrow-Green, *Poincaré and the Three Body Problem*, American Mathematical Society (1997).

[160] F. Browder, Edit., *The Mathematical Heritage of Henri Poincaré, Proceedings of Symposia in Pure Mathematics of the American Mathematical Society* **39**, Parts 1 and 2, American Mathematical Society (1983).

[161] J. Gray, *Henri Poincaré. A Scientific Biography*, Princeton University Press (2013).

[162] É. Charpentier, É. Ghys, and A. Lesne, Edit., *The Scientific Legacy of Poincaré*, History of Mathematics Volume 36, American & London Mathematical Societies (2010).

[163] B. Duplantier and V. Rivasseau, eds., *Henri Poincaré, 1912-2012: Poincaré Seminar 2012*, Birkhäuser-Springer (2015).

[164] R. Abraham and Y. Ueda, Edit., *The Chaos Avant-Garde: Memories of the Early Days of Chaos Theory*, World Scientific (2000).

[165] D. Nolte, "The tangled tale of phase space", *Physics Today* **64**(4), 33 (April 2010).

[166] A. Motter and D. Campbell, "Chaos at fifty", *Physics Today* **66**(5), 27 (May 2013).

Stability of the Solar System

[167] E. Belbruno, *Fly Me to the Moon, an Insider's Guide to the New Science of Space Travel*, Princeton University Press (2007).

[168] E. Belbruno, *Capture Dynamics and Chaotic Motions in Celestial Mechanics: with Applications to the Construction of Low Energy Transfers*, Princeton University Press (2004).

[169] F. Diacu and P. Holmes, *Celestial Encounters: The Origins of Chaos and Stability*, Princeton University Press (1996).

[170] D. Saari and Z. Xia, "Off to Infinity in Finite Time", *Notices of the AMS* **42**, 538-546 (1995).

[171] H. Poincaré, *New Methods of Celestial Mechanics,* Parts 1, 2, and 3. (Originally published as Les Méthodes nouvelles de la Méchanique céleste.) American Institute of Physics *History of Modern Physics and Astronomy*, Volume 13, D. L. Goroff, Edit., American Institute of Physics (1993). See the Editor's Introduction for a discussion of the stability of the solar system.

[172] Y. Kozai, Edit., *The Stability of the Solar System and of Small Stellar Systems*, D. Reidel (1974).

[173] M. Suvakov and V. Dmitrasinovic, "Three Classes of Newtonian Three-Body Planar Periodic Orbits", *Phys. Rev. Lett.* **110**, 114301 (2013). See also the Web site http://suki.ipb.ac.rs/3body/.

Classical/Celestial/Galactic Mechanics and Nonlinear Dynamics

[174] G. Gallavotti, *The Elements of Mechanics*, Springer-Verlag (1983). See also the Web site http://141.108.10.74/pagine/deposito/2007/elements.pdf.

[175] A. Brizard, *An Introduction to Lagrangian Mechanics*, World Scientific (2008).

[176] H. Goldstein, *Classical Mechanics*, Addison-Wesley (1980).

[177] L. D. Landau and E.M. Lifshitz, *Mechanics*, Addison-Wesley (1969).

[178] R. Abraham and J. Marsden, *Foundations of Mechanics*, American Mathematical Society (2008).

[179] R. Abraham and C. Shaw, *Dynamics: the Geometry of Behavior*, 4 Vols., Aeriel Press (1984).

[180] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, Second Edition, Springer-Verlag (1989).

[181] V. I. Arnold, V.V. Kozlov, and A. I. Neishtadt, *Mathematical Aspects of Classical and Celestial Mechanics*, Third Edition, Springer Verlag (2006).

[182] V. I. Arnold, Ya. G. Sinai, et al., edit, *Dynamical Systems I* through *Dynamical Systems IX*, Volumes from the Encyclopedia of Mathematical Sciences, Springer Verlag (1995).

[183] V. I. Arnold and A. Avez, *Ergodic Problems of Classical Mechanics*, Benjamin (1968).

[184] H. Cabral and F. Diacu, Edit, *Classical and Celestial Mechanics*, Princeton University Press (2002).

[185] J. Danby, *Fundamentals of Celestial Mechanics*, Macmillan (1962).

[186] H. S. Dumas, K. R. Meyer, and D. S. Schmidt, *Hamiltonian Dynamical Systems: History, Theory, and Applications*, Springer Verlag (1995).

[187] H. S. Dumas, *The KAM Story: A Friendly Introduction to the Content, History, and Significance of Classical Kolmogorov-Arnold-Moser Theory*, World Scientific (2014).

[188] G. Benettin, I. Galgani, A. Giorgilli, J.-M. Strelcyn, "A Proof of Kolmogorov's Theorem on Invariant Tori Using Canonical Transformations Defined by the Lie Method", *Il Nuovo Cimento* **79 B**, 201 (1984).

[189] K. Meyer, G. Hall, and D. Offin, *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, Second Edition, Springer (2009).

[190] J. Moser, "Lectures on Hamiltonian Systems", *Mem. Am. Math. Soc.* **81**, 1-60 (1968).

[191] C. Hayashi, *Nonlinear Oscillations in Physical Systems*, McGraw-Hill (1964).

[192] E. J. Saletan and A.H. Cromer, *Theoretical Mechanics*, John Wiley (1971).

[193] J. V. Jose and E. J. Saletan, *Classical Dynamics: A Contemporary Approach*, Cambridge University Press (1998).

[194] C. L. Siegel and J. K. Moser, *Lectures on Celestial Mechanics*, Springer-Verlag (1995).

[195] J. K. Moser, *Stable and Random Motions in Dynamical Systems: with Special Emphasis on Celestial Mechanics*, Princeton University Press (1973).

[196] J. K. Moser and E. J. Zehnder, *Notes on Dynamical Systems*, American Mathematical Society (2005).

[197] G. Benettin, J. Henrard, S. Kuksin, and A. Giorgilli, *Hamiltonian Dynamics - Theory and Applications*, Springer-Verlag (2005).

[198] G. J. Sussman, J. Wisdom, and M.E. Mayer, *Structure and Interpretation of Classical Mechanics*, Second Edition, MIT Press (2014).

[199] R. Talman, *Geometric Mechanics*, John Wiley (2000).

[200] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Springer Verlag (1999).

[201] L. Michelotti, *Intermediate Classical Dynamics with Applications to Beam Physics*, John Wiley (1995).

[202] J. L. McCauley, *Classical Mechanics*, Cambridge (1997).

[203] John R. Taylor, *Classical Mechanics*, University Science Books (2005).

[204] R. L. Devaney and Z.H. Nitecki, *Classical Mechanics and Dynamical Systems*, Lecture notes in pure and applied mathematics, Vol. 70, Dekker (1981).

[205] L. A. Pars, *A Treatise on Analytical Dynamics*, Ox Bow Press (1979).

[206] E. T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies with an Introduction to the Problem of Three Bodies*, Cambridge University Press (1960).

[207] J. Lopuszanski, *The Inverse Variational Problem in Classical Mechanics*, World Scientific (1999).

[208] G. Vilasi, *Hamiltonian Dynamics*, World Scientific (2001).

[209] J. Lowenstein, *Essentials of Hamiltonian Dynamics*, Cambridge (2012).

[210] W. Thirring, *A Course in Mathematical Physics: 1. Classical Dynamical Systems*, Springer-Verlag (1978).

[211] J. Binney and S. Tremaine, *Galactic Dynamics*, Princeton University Press (1987).

[212] F. Scheck, *Mechanics: from Newton's Laws to Deterministic Chaos*, Springer-Verlag (2005).

[213] S. Sternberg, *Celestial Mechanics, Parts I and II*, W. A. Benjamin (1969).

[214] V. Szebehely, *Theory of Orbits*, Academic Press (1967).

[215] A. Nayfeh and B. Balachandran, *Applied Nonlinear Dynamics: Analytical, Computational, and Experimental Methods*, John Wiley & Sons (1995).

[216] A. Nayfeh, *Nonlinear Interactions: Analytical, Computational, and Experimental Methods*, John Wiley & Sons (2000).

[217] H. Nusse and J. Yorke, *Dynamics: Numerical Explorations*, Second revised and Enlarged Edition, Springer (1998).

[218] E. Sudarshan and M. Mukunda, *Classical Dynamics: A Modern Perspective*, Wiley (1974).

[219] F. Gantmacher, *Lectures in Analytical Mechanics*, Mir Publishers (1975).

[220] R. Matzner and L. Shepley, *Classical Mechanics*, Prentice Hall (1991).

[221] G. Benettin, J. Henrard, S. Kuksin, and A. Giorgilli (Edit.), *Hamiltonian Dynamics Theory and Applications*, Lecture Notes in Mathematics 861, Springer (2005).

[222] J.-M. Souriau, *Structure of Dynamical Systems, a Symplectic View of Physics*, Birkhäuser (1997).

[223] D. D. Holm, *Geometric Mechanics, Part I: Dynamics and Symmetry*, Imperial College Press, World Scientific (2008).

[224] D. D. Holm, *Geometric Mechanics, Part II: Rotating, Translating and Rolling*, Imperial College Press, World Scientific (2008).

[225] D. D. Holm, T. Schmah, C. Stoica, and D. C. P. Ellis, *Geometric Mechanics, from Finite to Infinite Dimensions*, Oxford (2009).

[226] W. B. Kibble and F. H. Berkshire, *Classical Mechanics*, Fifth Edition, World Scientific (2004).

[227] M. Spivak, *Physics for Mathematicians: Mechanics I*, Publish or Perish, Inc. (2010).

[228] M. Audin, *Hamiltonian Systems and Their Integrability*, American Mathematical Society (2008).

[229] J. J. Morales Ruiz, *Differential Galois Theory and Non-Integrability of Hamiltonian Systems*, Springer and Birkhäuser (1999).

[230] M. Hénon, *Generating Families in the Restricted Three-Body Problem*, Springer Verlag (1997); *Generating Families in the Restricted Three-Body Problem II. Quantitative Study of Bifurcations*, Springer Verlag (2001).

[231] D. Heggie and P. Hut, *The Gravitational Million-Body Problem: A Multidisciplinary Approach to Star Cluster Dynamics*, Cambridge University Press (2003).

[232] S. Aarseth, *Gravitational N-Body Simulations: Tools and Algorithms*, Cambridge University Press (2003).

[233] M. Levi, *Classical Mechanics with Calculus of Variations and Optimal Control*, American Mathematical Society (2014).

[234] J. Papastavridis, *Analytical Mechanics : A Comprehensive Treatise on the Dynamics of Constrained Systems*, World Scientific (2014).

[235] Richard K. Cooper and Claudio Pellegrini, *Modern Analytic Mechanics*, Kluwer Academic (2010).

[236] Kai S. Lam, *Fundamental Principles of Classical Mechanics : A Geometrical Perspective*, World Scientific (2014).

[237] H. Iro, *A Modern Approach to Classical Mechanics*, Second Edition, World Scientific (2016).

[238] L. Hand and J. Finch, *Analytical Mechanics*, Cambridge University Press (1998).

[239] D. Tong, *Classical Dynamics*, (2004-2005). See the Web site http://www.damtp.cam.ac.uk/user/tong/dynamics/one.pdf.

[240] S. Wiggins, *Chaotic Transport in Dynamical Systems*, Springer-Verlag (1992).

[241] S. Wiggins, *Normally Hyperbolic Invariant Manifolds in Dynamical Systems*, Springer-Verlag (1994).

[242] S. Wiggins, *Introduction to Applied Dynamical Systems and Chaos*, Springer-Verlag (2003).

Inverse and Implicit Function Theorems and Functional Analysis

[243] R. Courant and F. John, *Introduction to Calculus and Analysis*, Vol. I, Vol. II/1, Vol. II/2, Springer-Verlag (1998, 1999, 2000).

[244] J. E. Marsden and M. J. Hoffman, *Elementary Classical Analysis*, p. 397, W.H. Freeman (1993).

[245] R. Abraham, J. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications*, Springer-Verlag (1988).

[246] W. Rudin, *Principles of Mathematical Analysis*, Third Edition, McGraw-Hill (1976).

[247] W. Rudin, *Real and Complex Analysis*, Third Edition, McGraw-Hill (1987).

[248] A. Knapp, *Advanced Real Analysis*, Birkhäuser (2005).

[249] B. Simon, *A Comprehensive Course in Analysis* (5-volume set), American Mathematical Society (2015).

[250] S. G. Krantz and H. R. Parks, *The Implicit Function Theorem: History, Theory, and Applications*, Birkhäuser (2002).

Euler's Relation for Homogeneous Functions

[251] R. Courant and F. John, *Introduction to Calculus and Analysis*, Vol. I, Vol.. II/1, Vol. II/2, Springer-Verlag (1998, 1999, 2000). See pages 119-121 of Vol. II/1.

Electromagnetism

[252] W. R. Smythe, *Static and Dynamic Electricity*, McGraw-Hill (1939).

[253] J. A. Stratton, *Electromagnetic Theory*, McGraw-Hill (1941).

[254] J. D. Jackson, *Classical Electrodynamics*, John Wiley (1999).

[255] J.D. Jackson and L.B. Okun, "Historical roots of gauge invariance", *Reviews of Modern Physics* **73**, p. 663 (2001).

[256] W. K. H. Panofsky and M. Phillips, *Classical Electricity and Magnetism*, Addison-Wesley (1962).

[257] L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields*, Addison-Wesley (1971).

[258] J. Schwinger *et al.*, *Classical Electrodynamics*, Perseus Books (1998).

[259] E. Purcell and D. Morin, *Electricity and Magnetism*, third edition, Cambridge University Press (2013).

[260] A. Zangwill *Modern Electrodynamics*, Cambridge University Press (2013).

[261] D. J. Griffiths, *Introduction to Electrodynamics*, Prentice Hall (1999).

[262] W. Gibson, *The Method of Moments in Electromagnetics*, Chapman & Hall/CRC (2008).

Lorentz Invariant Formulation

[263] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems*, Princeton University Press (1992).

[264] J. Sipe, "New Hamiltonian for a charged particle in an applied electromagnetic field", *Phys. Rev* A **27**, p. 615 (1983).

Extended Phase Space and Variational Calculus

[265] C. Lanczos, *The Variational Principles of Mechanics*, fourth edition, Dover (1986).

[266] J. Struckmeier, "Hamiltonian dynamics on the symplectic extended phase space for autonomous and non-autonomous systems", *J. Phys. A: Math. Gen.* **38**, 1257 (2005).

[267] J. Struckmeier, "Extended Hamilton-Lagrange Formalism and Its Application to Feynman's Path Integral for Relativistic Quantum Physics", *International Journal of Modern Physics E* **18**, 79 (2009).

[268] J. Struckmeier, W. Greiner, and H. Reichau, *Extended Lagrange and Hamiltonian Formalism for Point Mechanics and Covariant Hamiltonian Field Theory*, World Scientific (2014).

[269] H. Rund, *The Hamilton-Jacobi theory in the calculus of variations*, D. Van Nostrand (1966).

[270] I. Gelfand and S. Fomin, *Calculus of Variations*, Prentice Hall (1963) and Dover (2000).

[271] M. Giaquinta and S. Hildebrandt, *Calculus of Variations I: The Lagrangian Formalism*, *Calculus of Variations II: The Hamiltonian Formalism*, Springer-Verlag (2004).

[272] M. Morse, *The Calculus of Variations in the Large*, American Mathematical Society (1934).

[273] B. van Brunt, *The Calculus of Variations*, Springer (2006).

Singular (Degenerate) Lagrangians

[274] H. Rund, *The Hamilton-Jacobi theory in the calculus of variations*, D. Van Nostrand (1966). See Chapter 3.

[275] P. A. M. Dirac, "Generalized Hamiltonian dynamics" *Can. J. Math.* **2**, 129-148 (1950).

[276] R. Abraham and J. Marsden, *Foundations of Mechanics*, American Mathematical Society (2008). See Section 3.6.

Geodesics

[277] Y. Choquet-Bruhat, C. DeWitt-Morette, and M. Dillard-Bleick, *Analysis, Manifolds, and Physics*, Elsevier North Holland (1987). See pages 302 and 320 through 324.

[278] M. Spivak, *A Comprehensive Introduction to Differential Geometry*, volume 1, second edition, Publish or Perish (1979). See Chapter 9.

[279] L. P. Eisenhart, *Riemannian Geometry*, Princeton (1960). See Section 17.

Fluid Mechanics

[280] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics*, Volume 6 of a Course of Theoretical Physics, Pergamon Press (1979).

Accelerator Physics/Charged-Particle Optics

[281] É. Forest, *Beam Dynamics: A New Attitude and Framework*, Harwood Academic Publishers (1998).

[282] A. Wolski, *Beam Dynamics in High Energy Particle Accelerators*, Imperial College Press (2014).

[283] A. Chao, K. Mess, M. Tigner, and F. Zimmermann, Edit., *Handbook of Accelerator Physics and Engineering*, Second Edition, World Scientific (2013).

[284] E. Wilson, *An Introduction to Particle Accelerators*, Oxford University Press (2001).

[285] H. Wiedemann, *Particle Accelerator Physics - Basic Principles and Linear Beam Dynamics*, Springer-Verlag (1993).

[286] H. Wiedemann, *Particle Accelerator Physics II- Nonlinear and Higher-Order Beam Dynamics*, Springer-Verlag (1995).

[287] M. Berz, *Modern Map Methods in Particle Beam Physics*, Volume 108 of *Advances in Imaging and Electron Physics*, Academic Press (1999).

[288] D. Edwards and M. Syphers, *An Introduction to the Physics of High Energy Accelerators*, Wiley (1993).

[289] D. Carey, *The Optics of Charged Particle Beams*, Harwood Academic (1987).

[290] H. Wollnik, *Optics of Charged Particles*, Academic Press (1987).

[291] M. Conte and W. MacKay, *An Introduction to the Physics of Particle Accelerators*, Second Edition, World Scientific (2008).

[292] M. Reiser, *Theory and Design of Charged Particle Beams*, John Wiley (1994).

[293] P. Bryant and K. Johnsen, *The Principles of Circular Accelerators and Storage Rings*, Cambridge University Press (1993).

[294] S. Y. Lee, *Accelerator Physics*, World Scientific (1999).

[295] N. Dikansky and D. Pestrikov, *The Physics of Intense Beams and Storage Rings*, AIP Press (1994).

[296] S. Bernal, *A Practical Introduction to Beam Physics and Particle Accelerators*, IOP Concise Physics (2016).

Chapter 2

Numerical Integration

> Nature laughs at the difficulties of integration.
>
> *Laplace*

The differential equations of motion for many systems of physical interest cannot be completely solved in terms of familiar functions. For example, there are precious few problems in Plasma Physics, Space Mechanics, or Accelerator Design that have closed-form analytical solutions. Generally, a differential equation, or a set of differential equations, should be viewed as the source of some *new* transcendental function. This fact was realized shortly after the discovery of Classical Mechanics and Differential Equations. Consequently, over the past centuries and particularly in Celestial Mechanics, considerable effort has been put into the possibility of expressing solutions not in terms of known functions, but rather in terms of *infinite* series of known functions. For example, elaborate series expansions have been worked out for the motion of the planets and their moons, and these series have been used to compute their trajectories to high precision.

The contemporary approach is somewhat different. Usually a complete knowledge of every possible "trajectory" or motion of a system is not necessary. Rather, it often suffices to have a qualitative description of the types of allowed motion supplemented by a detailed knowledge of a few representative "orbits". Detailed knowledge of specific orbits is today most easily obtained by numerical integration using digital computers.[1] The types of allowed orbits can usually be determined best by analytical and topological methods, although even here numerical studies often precede and suggest later analytical results. Contemporary mechanics is thus an interplay between both analytical and numerical methods.

Even a survey of numerical methods is outside the scope of this text. It would require a text in itself. However, we hope that the brief discussion we are about to give will impart some of the flavor of numerical techniques, and perhaps entice the reader to explore further on his or her own. We hasten to add that numerical methods are also important outside classical mechanics, and that the techniques learned here can be applied to other situations in which ordinary differential equations arise. They also serve as a background for related methods in the numerical treatment of partial differential equations.

[1] Recently, however, there has been renewed interest in series expansions with the new twist that these expansions are produced by computers programmed to perform algebraic manipulations. In some cases it is advantageous to use series expansions to transform the equations of motion, and then to integrate these transformed equations numerically.

2.1 The General Problem

2.1.1 Integrating Forward in Time

Consider a set of first-order differential equations of the form (1.3.4). For compactness of notation we shall group together the quantities $(y_1 \cdots y_m)$ and $(f_1 \cdots f_m)$, and regard them as the components of two m-dimensional vectors: $\boldsymbol{y}$ and $\boldsymbol{f}$. Thus, we rewrite (1.3.4) in the form

$$\dot{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{y}, t). \tag{2.1.1}$$

Suppose t^0 is some initial time and we wish to integrate *forward* to the time $t^0 + T$. Divide up the time axis into N equal steps, each of duration h, so that

$$Nh = T. \tag{2.1.2}$$

Define successive times t^n by writing[2]

$$t^n = t^0 + nh. \tag{2.1.3}$$

See Figure 1.1 below. The time step, h, is taken to be small compared to the characteristic time scale or period of the physical system we are studying. For example, in solving a pendulum problem, h should be much less than the period of oscillation. Our goal is to compute the vectors $\boldsymbol{y}^n$, where

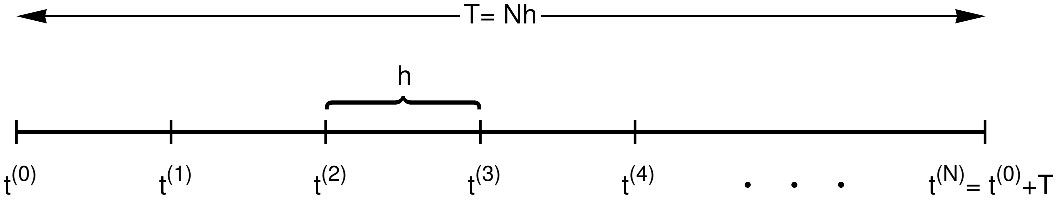


Figure 2.1.1: The Time Axis

$$\boldsymbol{y}^n = \boldsymbol{y}(t^n), \tag{2.1.4}$$

starting from the vector $\boldsymbol{y}^0$. The vector $\boldsymbol{y}^0$ is assumed given as a set of definite numbers, i.e. the initial conditions at t^0. To complete our notation, we make the definition

$$\boldsymbol{f}^n = \boldsymbol{f}(\boldsymbol{y}^n, t^n). \tag{2.1.5}$$

2.1.2 Integrating Backwards in Time

In the next several sections we will describe various methods for integrating forward in time to times later than t^0. Suppose we instead wish to integrate backwards to times earlier than t^0 so that $T < 0$. According to Theorem 1.3.1 this should be possible. After a few moments reflection we see that this problem has already been solved if we have found how to integrate forward. To integrate backward, we simply change the sign of h. That is, once an integration method has been selected, execute it with $h < 0$.

[2] Warning! Here n is a superscript, not an exponent. Sometimes, however, n will be an exponent. There should be enough clues from the context for you to decide what is meant.

2.2 A Crude Solution Due to Euler

2.2.1 Procedure

Theorem 1.3.1 guarantees that the solution vectors $\boldsymbol{y}^n$ exist and are uniquely specified by $\boldsymbol{y}^0$. The question is how to find them. Proceed one step at a time! By Taylor's theorem,

$$\boldsymbol{y}^1 = \boldsymbol{y}(t^1) = \boldsymbol{y}(t^0 + h) = \boldsymbol{y}^0 + h\dot{\boldsymbol{y}}^0 + O(h^2) \tag{2.2.1}$$

or

$$\boldsymbol{y}^1 = \boldsymbol{y}^0 + h\boldsymbol{f}^0 + O(h^2). \tag{2.2.2}$$

(Here, and in what follows, we assume analyticity in t, or at least the existence of several derivatives, as guaranteed by the theorems and discussion of Section 1.3.) Since $\boldsymbol{y}^0$ and t^0 are definite numbers, $\boldsymbol{f}^0$ is explicitly computable. Let us ignore the $O(h^2)$ error in (2.2) for the moment and accept (2.2) as an exact result for $\boldsymbol{y}^1$. Then using this $\boldsymbol{y}^1$ we can compute $\boldsymbol{f}^1$, and from that $\boldsymbol{y}^1$ and $\boldsymbol{f}^1$ proceed in similar fashion to compute $\boldsymbol{y}^2$ and $\boldsymbol{f}^2$, etc. In summary, we march forward step by step using the rule

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\boldsymbol{f}^n. \tag{2.2.3}$$

Suppose we march to the time $t^0 + T$. This requires $N = T/h$ steps. At each step we make a *local* error of order h^2. Consequently the *cumulative* error, barring cancellations that could only reduce it, is of order[3]

$$Nh^2 = Th. \tag{2.2.4}$$

We see that if the step size h is made sufficiently small and correspondingly the number of steps N sufficiently large, the error made in computing $\boldsymbol{y}(t^0 + T)$ using (2.3) can be made arbitrarily small.

2.2.2 Numerical Example

Consider the differential equation

$$\ddot{x} + x = 2t \tag{2.2.5}$$

with the initial conditions

$$x(0) = 0 \text{ and } \dot{x}(0) = 1. \tag{2.2.6}$$

We convert (2.5) into a first-order set by writing

$$y_1 = x, \quad y_2 = \dot{x}, \tag{2.2.7}$$

and find

$$f_1(\boldsymbol{y}, t) = \dot{y}_1 = y_2, \tag{2.2.8}$$

$$f_2(\boldsymbol{y}, t) = \dot{y}_2 = 2t - y_1. \tag{2.2.9}$$

[3] By "order Th" we mean proportional to Th with a bounded but *unspecified* proportionality constant.

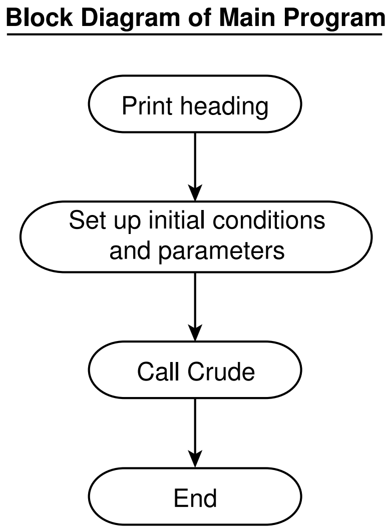
The simple computer program diagramed, listed, and annotated in Exhibit 2.1 below implements the Euler method (2.3) to integrate this set. The step size is $h = 1/10$. The differential equation we have selected is sufficiently simple that it also can be integrated analytically to give the exact result

$$y_1 = x(t) = 2t - \sin\ t, \tag{2.2.10}$$

$$y_2 = \dot{x}(t) = 2 - \cos\ t. \tag{2.2.11}$$

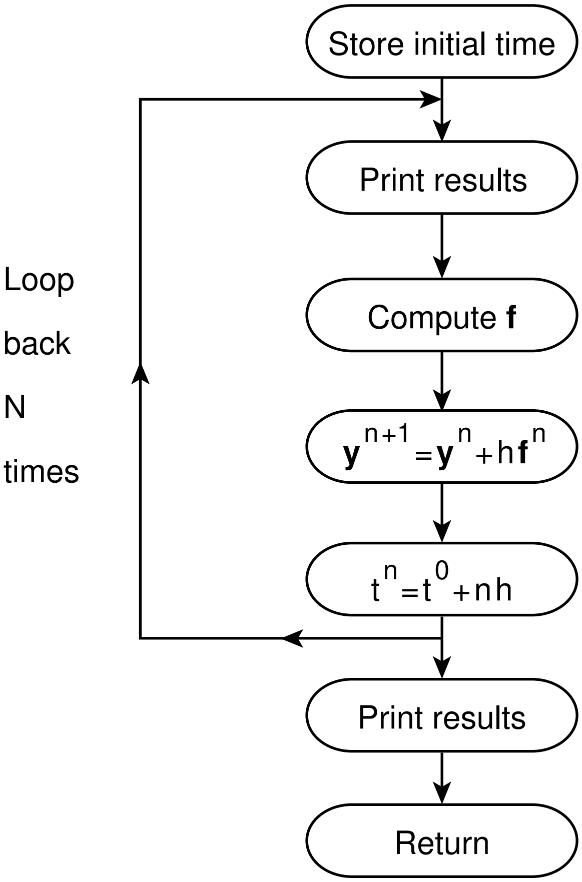
Note that the characteristic period of the solution is 2π so that the choice $h = 1/10$ is considerably smaller than the period as required. For comparison, both the Euler result and the exact result are tabulated.

Exhibit 2.2.1: Crude Euler Integration


Block Diagram of Main Program
Print heading
Set up initial conditions and parameters
Call Crude
End

```
c This is the main program for illustrating the crude Euler method
c of numerical integration.
c
c Print heading.
c
      write(6,100)
  100 format
     & (1h ,'time',4x,'y1comp',10x,'y2comp',10x,'y1true',10x,'y2true',/)
c
c Set up initial conditions and parameters. n is the number of integration
c steps we wish to make.
c
      t=0.
      h=.1
      n=15
      y1=0.
      y2=1.
c
      call crude(t,h,n,y1,y2)
c
      end
```

Block Diagram of Integration Routine



```
c This is the crude Euler integration subroutine
c
      subroutine crude(t,h,n,y1,y2)
c
c Store initial time.
c
      tint=t
c
c Printing and integration loop.
c
      do 100 i=1,n
      call prints(t,y1,y2,y1true(t),y2true(t),0)
c
c Compute f, the right side of the differential equation.
c
      call eval(y1,y2,t,f1,f2)
c
c Make integration step and update time.
c
```

```
      y1=y1+h*f1
      y2=y2+h*f2
      t=tint+float(i)*h
c
  100 continue
c
c Print final results.
c
      call prints(t,y1,y2,y1true(t),y2true(t),0)
c
      return
      end
```

Auxiliary Programs

```
c This subroutine evaluates f, the right side of the
c differential equation.
c
      subroutine eval(y1,y2,t,f1,f2)
c
      f1=y2
      f2=2.*t-y1
c
      return
      end

c
c Function for computing the exact value of y1.
c
      function y1true(t)
      y1true=2.*t-sin(t)
      return
      end

c
c Function for computing the exact value of y2.
c
      function y2true(t)
      y2true=2.-cos(t)
      return
      end

c
c Subroutine to handle printing. It need not concern the reader.
c
      Subroutine prints(t,y1,y2,y1t,y2t,iflag)
c
      if (iflag .eq. 0) then
      write(6,100) t,y1,y2,y1t,y2t
  100 format (1h ,f6.4,2x,4(e14.8,2x))
      return
      endif
c
      if (iflag .ne. 0) then
      write(6,200) y1,y2
  200 format (1h ,8x,2(e14.8,2x))
      return
      endif
c
      end
```

Numerical Results

```
time    y1comp          y2comp          y1true          y2true

0.0000  0.00000000E+00  0.10000000E+01  0.00000000E+00  0.10000000E+01
0.1000  0.10000000E+00  0.10000000E+01  0.10016658E+00  0.10049958E+01
0.2000  0.20000000E+00  0.10100000E+01  0.20133068E+00  0.10199335E+01
0.3000  0.30100000E+00  0.10300000E+01  0.30447981E+00  0.10446635E+01
0.4000  0.40400001E+00  0.10598999E+01  0.41058168E+00  0.10789391E+01
0.5000  0.50999004E+00  0.10994999E+01  0.52057445E+00  0.11224174E+01
0.6000  0.61994004E+00  0.11485009E+01  0.63535756E+00  0.11746644E+01
0.7000  0.73479015E+00  0.12065070E+01  0.75578231E+00  0.12351578E+01
0.8000  0.85544086E+00  0.12730279E+01  0.88264394E+00  0.13032933E+01
0.9000  0.98274368E+00  0.13474839E+01  0.10166732E+01  0.13783901E+01
1.0000  0.11174921E+01  0.14292095E+01  0.11585290E+01  0.14596977E+01
1.1000  0.12604131E+01  0.15174602E+01  0.13087927E+01  0.15464039E+01
1.2000  0.14121591E+01  0.16114190E+01  0.14679611E+01  0.16376423E+01
1.3000  0.15733010E+01  0.17102031E+01  0.16364419E+01  0.17325013E+01
1.4000  0.17443212E+01  0.18128730E+01  0.18145503E+01  0.18300328E+01
1.5000  0.19256085E+01  0.19184409E+01  0.20025051E+01  0.19292628E+01
```

We conclude that with $h = 1/10$, the Euler method integrates (2.5) over the range $t = 0$ to $t = 1.5$ with an accuracy of somewhat less than two signficant figures.

Exercises

2.2.1. Consider the differential equation

$$\ddot{x} + x = 0. \tag{2.2.12}$$

a) Show that in this case Euler's method amounts to solving the set of difference equations

$$y_1^{n+1} = y_1^n + h y_2^n, \tag{2.2.13}$$

$$y_2^{n+1} = y_2^n - h y_1^n. \tag{2.2.14}$$

b) Show that the difference equations have the solution

$$\boldsymbol{y}^n = M^n \boldsymbol{y}^0 \tag{2.2.15}$$

where M is the matrix

$$M = \begin{pmatrix} 1 & h \\ -h & 1 \end{pmatrix}. \tag{2.2.16}$$

c) Show by explicit computation that M has two linearly independent eigenvectors $\boldsymbol{a}$ and $\boldsymbol{b}$ with eigenvalues α and β. Expand $\boldsymbol{y}^0$ in terms of $\boldsymbol{a}$ and $\boldsymbol{b}$. That is, write

$$\boldsymbol{y}^0 = A\boldsymbol{a} + B\boldsymbol{b} \tag{2.2.17}$$

where A and B are expansion coefficients. Show that

$$\boldsymbol{y}^n = \alpha^n A\boldsymbol{a} + \beta^n B\boldsymbol{b}. \tag{2.2.18}$$

d) Study how $\boldsymbol{y}(t^0 + T)$, as computed by Euler's method, converges to the exact result as $h \to 0$.

e) Show that when $h \neq 0$, the length of $\boldsymbol{y}^n$ grows (exponentially) without bound as $n \to \infty$! What happens to the length of the true solution as $t \to \infty$?

f) Make a similar analysis for the differential equation (2.5). (Hint: find a particular solution, and then use the solution of the homogeneous equation to fit the initial conditions.)

2.2.2. Consider the differential equation

$$dx/dt = A + Bx + Cx^2, \tag{2.2.19}$$

which is a variant of the logistic/Verhulst differential equation. See (1.2.114). Solve this differential equation exactly.

Show that applying Euler's method to this differential equation produces the quadratic difference equation

$$x_{n+1} = x_n + hA + hBx_n + hC(x_n)^2, \tag{2.2.20}$$

which is a quadratic map of the form (1.2.114). Compare the behavior of the solutions of the differential equation (2.19) to that of the quadratic difference equation (2.20). Consider the cases of both small and large step size h. At what value of h does chaotic behavior set in? Chaotic behavior would be a bad thing because you should have found that the solutions to (2.19) are well behaved. How small must h be to avoid period doubling? Period doubling would also be a bad thing because you should have found that the solutions to (2.19) are not periodic.

2.3 Runge-Kutta Methods

2.3.1 Introduction

Now that we have the general idea, let us see what improvements can be made. The obvious need is to improve the accuracy of the stepping formula (2.3). One procedure would be to invoke the use of the first few additional derivatives. Higher derivatives are computable, and could in principle be used. For example, differentiating (1.1) and substituting it back into its derivative gives the result

$$\ddot{y}_i = \partial f_i/\partial t + \sum_j (\partial f_i/\partial y_j)\dot{y}_j$$

or

$$\ddot{y}_i = \partial f_i/\partial t + \sum_j (\partial f_i/\partial y_j) f_j. \tag{2.3.1}$$

This procedure can be effective for differential equations whose right sides are polynomial in the y_i. However it is evident that for most systems of differential equations the expressions for the higher derivatives become quite lengthy, and their use may be a bit cumbersome. What is needed is a stepping procedure that only involves evaluations of $\boldsymbol{f}$.

2.3.2 Procedure

Such procedures were originally studied by *Runge* and *Kutta*, and now generally bear their names.[4] Many are available, and we shall be only able to quote a few without derivation. The general idea is to evaluate $\boldsymbol{f}$ at several different points and to add the results together in such a way that $\boldsymbol{y}^{n+1}$ is correctly estimated up to some error that is proportional to a large power of h, and thus quite small. (Exactly what points to use in evaluating $\boldsymbol{f}$ and how to weight the results is a complicated matter. We refer the interested reader to Exercises 3.1 and 3.10 through 3.12, and then to the references.) A method called RK3, that makes *local* errors only of order h^4, i.e. is locally correct through order h^3, is given by

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + \frac{1}{6}(\boldsymbol{a} + 4\boldsymbol{b} + \boldsymbol{c}), \tag{2.3.2}$$

where at each step

$$\boldsymbol{a} = h\boldsymbol{f}(\boldsymbol{y}^n, t^n), \tag{2.3.3}$$

$$\boldsymbol{b} = h\boldsymbol{f}(\boldsymbol{y}^n + \frac{1}{2}\boldsymbol{a}, t^n + \frac{1}{2}h),$$

$$\boldsymbol{c} = h\boldsymbol{f}(\boldsymbol{y}^n + 2\boldsymbol{b} - \boldsymbol{a}, t^n + h).$$

Higher-order methods are also available. The higher the order, of course, the more work is involved. One of several fourth-order methods, and called RK4, is given by

$$\boldsymbol{a} = h\boldsymbol{f}(\boldsymbol{y}^n, t^n), \tag{2.3.4}$$

$$\boldsymbol{b} = h\boldsymbol{f}(\boldsymbol{y}^n + \frac{1}{2}\boldsymbol{a}, t^n + \frac{1}{2}h),$$

$$\boldsymbol{c} = h\boldsymbol{f}(\boldsymbol{y}^n + \frac{1}{2}\boldsymbol{b}, t^n + \frac{1}{2}h),$$

$$\boldsymbol{d} = h\boldsymbol{f}(\boldsymbol{y}^n + \boldsymbol{c}, t^n + h),$$

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + \frac{1}{6}(\boldsymbol{a} + 2\boldsymbol{b} + 2\boldsymbol{c} + \boldsymbol{d}). \tag{2.3.5}$$

[4] Ernest Courant, who co-invented the use of matrices to approximate transfer maps as described in Section 1.1.2, is the son of the mathematician Richard Courant of Courant and Hilbert and Courant Institute fame. Ernest Courant's mother was the daughter of Runge, and thus Ernest Courant is also a grandson of Runge. Runge was very athletic, and entertained his grandchildren at his 70th birthday by doing handstands, which Ernest Courant remembers.

This method is locally correct through order h^4, and makes *local* errors of order h^5.[5]

The reader should note that when we say that either *local* or *cumulative* error is of order h^ℓ, we mean that it is proportional to h^ℓ with an unspecified constant of proportionality. Unfortunately, an analytic estimation of the proportionality constant for Runge-Kutta is very complicated. See, for example, Exercises 3.1 and 5.1.

To see the advantage of higher-order methods over (2.3), suppose we use the third-order method (3.2) to integrate from t^0 to t^0+T. This time the *cumulative* error is proportional to $Nh^4 = Th^3$, which is an improvement over the earlier error by a factor of h^2. Of course, each integration step now requires about three times as much work since $\boldsymbol{f}$ must be evaluated three times for each step. But the integration error is reduced by considerably more than a factor of three. It is possible now to use a much larger step size thus actually *reducing* the total work required to remain within a specified error.

The error we have been discussing so far can in principle be made arbitrarily small by letting $h \to 0$ and $N \to \infty$. In actual practice using digital computers, the ideal is not quite realizable. This is because computers only work with a finite number of significant figures, and hence each step involves a certain unavoidable "round-off" error. If the sign of each round-off error is nearly random, their cumulative effect increases approximately as $\sqrt{N}$. (The IEEE hardware standard with regard to round-off procedures is designed with this goal in mind.) If the sign is systematic, their cumulative effect may grow like N. In any case, if N is made too large, not only the cost of computation increases. The total error, after reaching a certain minimum, also increases! To see how this can work out in a specific case, look over the next example and then study Figure 3.1.

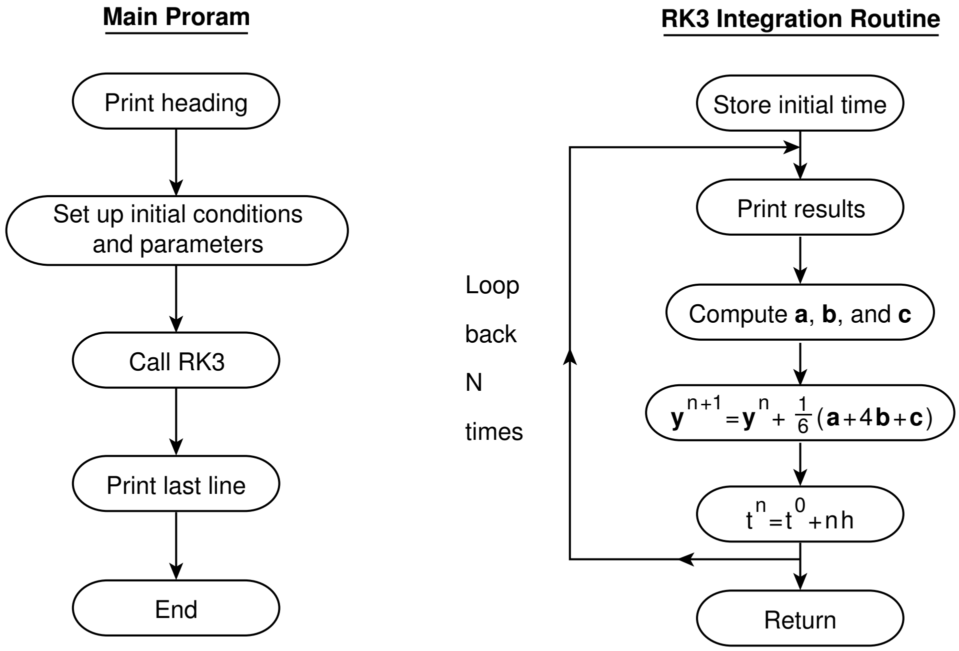
2.3.3 Numerical Example

We show below in Exhibit 3.1 a simple program that uses the third-order Runge-Kutta method (3.2) to integrate the problem of Section 2.2.2. The step size is again $h = 1/10$. We list only the main program and the subroutine RK3. The other subprograms are the same as those used in Section 2.2.2. Note that the numerical solution is now accurate to five significant figures.

In order to illustrate how the total cumulative error depends upon step size, we have also made calculations with other values of h. Figure 3.1 shows the results. Note that the cumulative error first decreases roughly as h^3 as expected, and then rises again because of round-off error.

[5] You will observe that we label a method by the order of the local accuracy. That is, an mth order method is locally correct through order h^m, and makes local errors of order h^{m+1}.

Exhibit 2.3.1: Third-Order Runge Kutta Integration



```
c This is the main program for illustrating a Runge Kutta method
c for numerical integration.
c
c Print heading.
c
      write(6,100)
  100 format
     & (1h ,'time',4x,'y1comp',10x,'y2comp',10x,'y1true',10x,'y2true',/)
c
c Set up initial conditions and parameters. n is the number of integration
c steps we wish to make.
c
      t=0.
      h=.1
      n=15
      y1=0.
      y2=1.
c
      call rk3(t,h,n,y1,y2)
      call prints(t,y1,y2,y1true(t),y2true(t),0)
c
      end

c
c This is a third-order Runge Kutta integration subroutine.
c
      subroutine rk3(t,h,n,y1,y2)
c
c Store initial time.
c
      tint=t
c
c Printing and integration loop.
c
      do 100 i=1,n
      call prints(t,y1,y2,y1true(t),y2true(t),0)
c
c Set up for integration step.
c
      call eval(y1,y2,t,f1,f2)
      a1=h*f1
      a2=h*f2
      y1t=y1+a1/2.
      y2t=y2+a2/2.
      tt=t+h/2.
      call eval(y1t,y2t,tt,f1,f2)
      b1=h*f1
      b2=h*f2
      y1t=y1+2.*b1-a1
      y2t=y2+2.*b2-a2
      tt=t+h
      call eval(y1t,y2t,tt,f1,f2)
      c1=h*f1
      c2=h*f2

c
c Make integration step and update time.
c
      y1=y1+(a1+4.*b1+c1)/6.
      y2=y2+(a2+4.*b2+c2)/6.
      t=tint+float(i)*h
c
  100 continue
c
      return
      end
```

Numerical Results

```
time    y1comp          y2comp          y1true          y2true

0.0000  0.00000000E+00  0.10000000E+01  0.00000000E+00  0.10000000E+01
0.1000  0.10016667E+00  0.10050000E+01  0.10016658E+00  0.10049958E+01
0.2000  0.20133168E+00  0.10199417E+01  0.20133068E+00  0.10199335E+01
0.3000  0.30448255E+00  0.10446757E+01  0.30447981E+00  0.10446635E+01
0.4000  0.41058692E+00  0.10789548E+01  0.41058168E+00  0.10789391E+01
0.5000  0.52058297E+00  0.11224364E+01  0.52057445E+00  0.11224174E+01
0.6000  0.63536996E+00  0.11746861E+01  0.63535756E+00  0.11746644E+01
0.7000  0.75579929E+00  0.12351816E+01  0.75578231E+00  0.12351578E+01
0.8000  0.88266593E+00  0.13033184E+01  0.88264394E+00  0.13032933E+01
0.9000  0.10167006E+01  0.13784157E+01  0.10166732E+01  0.13783901E+01
1.0000  0.11585623E+01  0.14597230E+01  0.11585290E+01  0.14596977E+01
1.1000  0.13088318E+01  0.15464280E+01  0.13087927E+01  0.15464039E+01
1.2000  0.14680060E+01  0.16376641E+01  0.14679611E+01  0.16376423E+01
1.3000  0.16364928E+01  0.17325199E+01  0.16364419E+01  0.17325013E+01
1.4000  0.18146069E+01  0.18300474E+01  0.18145503E+01  0.18300328E+01
1.5000  0.20025671E+01  0.19292722E+01  0.20025051E+01  0.19292628E+01
```

We close this subsection by remarking that the form of the Runge-Kutta program above was largely dictated by pedagogical considerations. A more compact version of this program using vector arrays and suitable for integrating any number of coupled equations is given in Appendix B. We commend this appendix to the reader who is considering more serious numerical work.

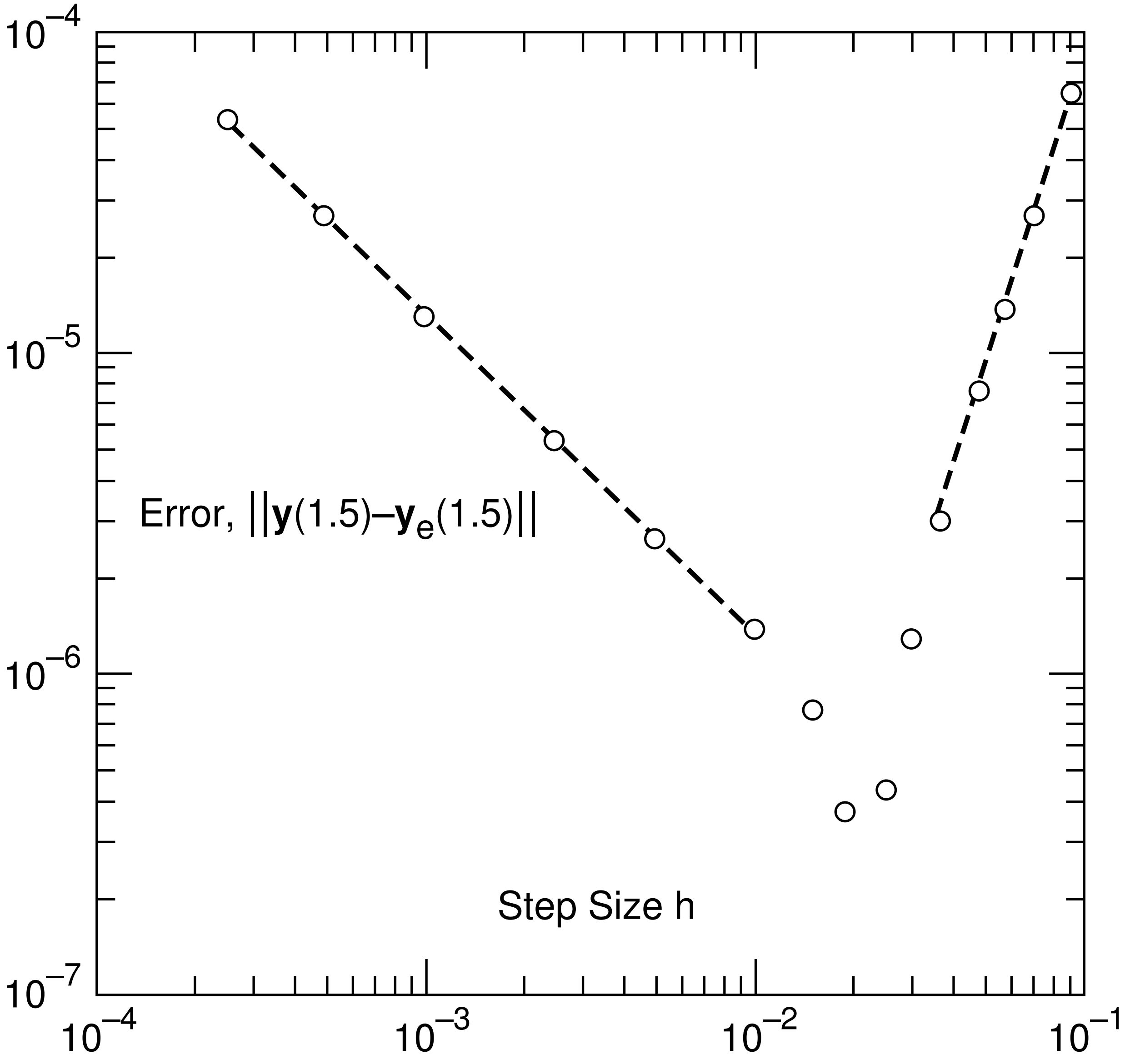


Figure 2.3.1: The result of integrating with RK3 the set (2.7) through (2.9) to $t = 1.5$ with several different step sizes to illustrate how the cumulative error depends on h. The error is measured by $\| \mathbf{y}(1.5) - \mathbf{y}_\mathrm{e}(1.5) \|$ where $\mathbf{y}_\mathrm{e}$ is the exact solution. The dashed line on the right has a slope of +3 showing that the global truncation error at first decreases as h^3. The dashed line on the left has a slope of −1 showing that in this example the global round-off error increases as the number of steps N. These calculations were made on a computer that had an accuracy of about 8 1/2 significant figures.

2.3.4 Nomenclature

Runge-Kutta methods have been studied extensively. In this subsection, as an aid to further reading, we will present briefly some of the nomenclature used to describe various Runge-Kutta concepts and methods.

Butcher Tableaux

Let b and c be s-dimensional vectors with real entries, and let a be an $s \times s$ matrix with real entries. Consider stepping formulas of the form

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h \sum_{i=1}^{s} b_i \boldsymbol{k}_i \tag{2.3.6}$$

where at each step

$$\boldsymbol{k}_i = \boldsymbol{f}(\boldsymbol{y}^n + h \sum_{j=1}^{s} a_{ij} \boldsymbol{k}_j,\ t^n + c_i h). \tag{2.3.7}$$

Observe that the integration methods RK3 and RK4 given by (3.2) through (3.5) are of this kind. The number s is called the number of *stages*. Evidently s is equal to the number of evaluations of the function $\boldsymbol{f}$ required to compute the $\boldsymbol{k}_i$ and thereby carry out one integration step using (3.6).

Before continuing on, it is sometimes useful to rewrite the relations (3.6) and (3.7) in a somewhat different form. At each step introduce intermediate times t_i and coordinates $\boldsymbol{y}_i$ by the rules

$$t_i = t^n + c_i h, \tag{2.3.8}$$

$$\boldsymbol{y}_i = \boldsymbol{y}^n + h \sum_{j=1}^{s} a_{ij} \boldsymbol{k}_j. \tag{2.3.9}$$

With this convention (3.7) can be rewritten in the form

$$\boldsymbol{k}_i = \boldsymbol{f}(\boldsymbol{y}_i, t_i). \tag{2.3.10}$$

Finally we copy (3.6) and place it last,

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h \sum_{i=1}^{s} b_i \boldsymbol{k}_i. \tag{2.3.11}$$

Evidently the relations (3.8) through (3.11) are equivalent to the relations (3.6) and (3.7), but in this expanded form it is clear that the $\boldsymbol{k}_i$ are the values of $\boldsymbol{f}$ at the intermediate points, and the stepping rule (3.11) resembles the rule (2.3) for crude Euler except that it involves a weighted sum of these $\boldsymbol{f}$ values rather than a single $\boldsymbol{f}$ value.

Continue on. The problem now is to impose various conditions on the vectors b and c and the matrix a so that the integration method will be of some particular order m, and perhaps have other desirable properties. For purposes of visualization, it is convenient to arrange the vectors b and c and the matrix a in a tableau, called a *Butcher* tableau after its author, as shown below:

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array} \tag{2.3.12}$$

The Butcher tableau for RK3 is

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1 & -1 & 2 & 0 \\ \hline & 1/6 & 4/6 & 1/6 \end{array} . \tag{2.3.13}$$

The Butcher tableau for RK4 (often called *classic* RK4) is

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 2/6 & 2/6 & 1/6 \end{array} . \tag{2.3.14}$$

There is another possible fourth-order method, also known to Kutta, sometimes called the 3/8 rule.[6] It is given by the Butcher tableau

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 0 \\ 2/3 & -1/3 & 1 & 0 & 0 \\ 1 & 1 & -1 & 1 & 0 \\ \hline & 1/8 & 3/8 & 3/8 & 1/8 \end{array} . \tag{2.3.15}$$

Two features should be noticed about the Butcher tableaux (3.13) through (3.15): The first is that the matrix a is *strictly lower triangular*. That is, all entries on or above the diagonal vanish. This feature makes these methods *explicit*. That is, each $\boldsymbol{k}_i$ is computable in terms of the $\boldsymbol{k}_j$ with $j < i$. Runge-Kutta methods without this property are called *implicit*.[7] The second feature is that the vector c is related to the matrix a by the rule

$$c_i = \sum_{j=1}^{s} a_{ij}. \tag{2.3.16}$$

[6] Although both classic RK4 and the 3/8 rule are fourth order (make local errors of order h^5), it can be shown that the 3/8 rule is somewhat more accurate because its local error terms proportional to h^5 have smaller coefficients. However, even though both classic RK4 and the 3/8 rule require the same number of function evaluations per step (namely, 4), the 3/8 rule is somewhat slower because its matrix a is somewhat more dense than that for RK4. Therefore, see (3.9), more additions and multiplications are required per step for the 3/8 rule than for RK4.

[7] *Explicit* Runge-Kutta methods are sometimes called ERK methods; and *implicit* Runge-Kutta methods are sometimes referred to as IRK methods. In the same spirit, if the matrix a has strictly lower triangular entries plus some nonzero diagonal entries but no entries above the diagonal, then the associated integration methods are called diagonally implicit Runge-Kutta (DIRK).

Pictorially, each c_i is the sum of the a's in its row. This relation is called the *consistency* condition and is, for convenience, generally required of all Runge-Kuttta methods. See Exercise 3.10. Exercise 3.9 briefly describes what further *order* conditions are required to achieve local accuracy through orders $m = 1$, $m = 2$, and $m = 3$.

Relation Between Number of Stages and Achievable Order

It is tempting to conjecture that with s stages it should be possible to find an explicit Runge-Kutta method whose order m satisfies $m = s$. This conjecture is true for $m = 1, 2, 3$, and 4, but it fails for $m \geq 5$. Table 3.1 below lists the minimum s value required to achieve order m with explicit Runge-Kutta methods. As can be seen, $s \geq 6$ is needed to achieve an explicit Runge-Kutta method with $m = 5$. Thus, there are diminishing returns in going beyond order 4, which gives fourth-order methods such as RK4 a preferred status.

Table 2.3.1: Minimum Number of Stages s Required for Explicit Runge Kutta to Achieve Order m.

m	1	2	3	4	5	6	7	8
s	1	2	3	4	6	7	9	11

With *implicit* Runge-Kutta methods it is possible for the order to even *exceed* the number of stages. Consider the one-stage method specified by the Butcher tableau

Gauss2

$$\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array} \,. \tag{2.3.17}$$

It corresponds to the *implicit midpoint rule*

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\boldsymbol{f}[(\boldsymbol{y}^n + \boldsymbol{y}^{n+1})/2, t^n + h/2], \tag{2.3.18}$$

which is known to be of order 2.[8] See Exercise 3.7. It is also related to Gaussian quadrature. See Subsection T.1.3. For this reason, and because it is second order, it is given the name Gauss2.

In fact, there are implicit Runge-Kutta methods for which $m = 2s$, and this order is the best that can be hoped for with s stages.[9] Butcher tableaux for two such methods, for the cases of two and three stages and also based on Gaussian quadrature, are given below. They have orders 4 and 6, respectively.

[8] This stepping procedure, particularly in the context of partial differential equations, is also referred to as *Crank-Nicolson*.

[9] Strictly speaking, an s-stage explicit Runge-Kutta integrator requires s function evaluations per step. Implicit Runge-Kutta methods require many more since the implicit equations involved are generally solved by multiple iteration.

Gauss4

$$\begin{array}{c|cc}
1/2-\sqrt{3}/6 & 1/4 & 1/4-\sqrt{3}/6 \\
1/2+\sqrt{3}/6 & 1/4+\sqrt{3}/6 & 1/4 \\
\hline
 & 1/2 & 1/2
\end{array}\ , \tag{2.3.19}$$

Gauss6

$$\begin{array}{c|ccc}
1/2-\sqrt{15}/10 & 5/36 & 2/9-\sqrt{15}/15 & 5/36-\sqrt{15}/30 \\
1/2 & 5/36+\sqrt{15}/24 & 2/9 & 5/36-\sqrt{15}/24 \\
1/2+\sqrt{15}/10 & 5/36+\sqrt{15}/30 & 2/9+\sqrt{15}/15 & 5/36 \\
\hline
 & 5/18 & 8/18 & 5/18
\end{array}\ . \tag{2.3.20}$$

Butcher tableaux for Gauss8 and Gauss10 are also available. See the book of *Sanz-Serna* and *Calvo* listed in the Bibliography for this chapter. For further discussion of implicit Runge-Kutta methods, see Section 12.4.

Interpolation and Dense Output

There are some situations, for example when graphical output is needed, in which one desires an accurate and efficient method for finding $\boldsymbol{y}(t^n+\theta h)$ for any $\theta \in [0,1]$. There are procedures that prepare, at each integration step, polynomials in θ for this purpose, and these procedures utilize the $\boldsymbol{k}$ vectors computed in the course of a Runge-Kutta step. See, for example, the book of *Hairer*, *Nørsett*, and *Wanner* cited at the end of this chapter.

First Same As Last

There is one final comment worth making. It is possible to construct Runge-Kutta methods for which the Butcher tableaux take the form

$$\begin{array}{c|ccccc}
0 & 0 & 0 & \cdots & 0 & 0 \\
c_2 & a_{2,1} & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
c_{s-1} & a_{s-1,1} & a_{s-1,2} & \cdots & 0 & 0 \\
1 & b_1 & b_2 & \cdots & b_{s-1} & 0 \\
\hline
 & b_1 & b_2 & \cdots & b_{s-1} & 0
\end{array} \tag{2.3.21}$$

Comparison of (3.21) with (3.12) shows that we have imposed the conditions

$$a_{ij}=0 \text{ for } j \geq i, \tag{2.3.22}$$

$$a_{sj}=b_j, \tag{2.3.23}$$

$$b_s=0, \tag{2.3.24}$$

$$c_s=1. \tag{2.3.25}$$

The condition (3.22) makes the associated integration method explicit. The condition (3.24) must hold if (3.22) and (3.23) are to be enforced. The condition (3.25) follows from the consistency condition (3.16) and the desire that the method be at least of order 1. See (3.42).

What is the virtue of the condition (3.23)? Let us compute $\boldsymbol{k}_s$ when the Butcher tableau has the form (3.21) and we are making the integration step from $t = t^n$ to $t = t^{n+1}$. From (3.7), (3.22), and (3.23) we find the result

$$\begin{aligned} \boldsymbol{k}_s|_{t=t^n} &= \boldsymbol{f}(\boldsymbol{y}^n + h\sum_{j=1}^{s} a_{sj}\boldsymbol{k}_j,\ t^n + c_s h) \\ &= \boldsymbol{f}(\boldsymbol{y}^n + h\sum_{j=1}^{s} b_j\boldsymbol{k}_j,\ t^n + h) = \boldsymbol{f}(\boldsymbol{y}^{n+1},\ t^{n+1}). \end{aligned} \tag{2.3.26}$$

Here we have used (3.6). Now let us compute $\boldsymbol{k}_1$ when the Butcher tableau has the form (3.21) and we are making the integration step from $t = t^{n+1}$ to $t = t^{n+2}$. From (3.7) and (3.22) we find the result

$$\boldsymbol{k}_1|_{t=t^{n+1}} = \boldsymbol{f}(\boldsymbol{y}^{n+1},\ t^{n+1}). \tag{2.3.27}$$

We conclude that

$$\boldsymbol{k}_1|_{t=t^{n+1}} = \boldsymbol{k}_s|_{t=t^n}, \tag{2.3.28}$$

the *first* $\boldsymbol{k}$ for a successive step is the same as the *last* $\boldsymbol{k}$ from the previous step. For this reason, a Butcher tableau of the form (3.21) is said to have a First Same As Last (FSAL) structure. We see that for a FSAL Runge-Kutta method, once an initial integration step has been completed, successive steps only require $s - 1$ function evaluations and are therefore the method effectively has $s - 1$ stages.[10] However, the price to be paid for FSAL turns out to be a reduction in order.

For example, the Butcher tableau

$$\begin{array}{c|ccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{3}{10} & \frac{3}{40} & \frac{9}{40} & 0 & 0 & 0 & 0 & 0 \\ \frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & 0 & 0 & 0 & 0 \\ \frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & 0 & 0 & 0 \\ 1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} & 0 & 0 \\ 1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\ \hline & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \end{array} \tag{2.3.29}$$

[10] Note also that, because $b_s = 0$, the final operation (3.6) is also already carried out in the evaluation of $\boldsymbol{k}_s$, which results in an additional savings.

describes a FSAL Runge-Kutta method that has $s = 7$ stages but acts like a $7-1 = 6$ stage method after the first step since 6 function evaluations are required for each subsequent step. It is a 5^{th} order method ($m = 5$). From Table 3.1 we see that this method has the optimal order that can be achieved with a 6 stage method, and has an order that is one less than the optimal order that can be achieved with a 7 stage method. Section 2.5.1 and Appendix B describe how this method can be used as part of an embedded Runge-Kuttta pair called Dormand-Prince 5(4).

Exercises

2.3.1. Consider the second-order Runge-Kutta method (sometimes called the improved Euler method or the second-order *Heun* method)

$$\boldsymbol{a} = h\boldsymbol{f}(\boldsymbol{y}^n, t^n), \tag{2.3.30}$$

$$\boldsymbol{b} = h\boldsymbol{f}(\boldsymbol{y}^n + \boldsymbol{a}, t^n + h),$$

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + \frac{1}{2}(\boldsymbol{a} + \boldsymbol{b}). \tag{2.3.31}$$

Verify that the *local* truncation error is of the form $\boldsymbol{e}h^3 + O(h^4)$ and find a formula for $\boldsymbol{e}$. Finding error estimates for Runge-Kutta methods is not easy! Hint: Use a Taylor series to write $\boldsymbol{y}^{n+1}_{\text{true}} = \boldsymbol{y}^n + h\dot{\boldsymbol{y}}^n + h^2\ddot{\boldsymbol{y}}^n/2! + h^3 \,\dddot{\boldsymbol{y}}^n\, /3! + O(h^4)$. Now expand the Runge-Kutta formula in a Taylor series and compare terms. You should find the result

$$\boldsymbol{e} = -(\dddot{\boldsymbol{y}} - 3\sum \ddot{y}_i \partial \boldsymbol{f}/\partial y_i)/(12). \tag{2.3.32}$$

2.3.2. Review Exercise 3.1. Consider the so called *explicit midpoint rule* Runge-Kutta method

$$\boldsymbol{a} = h\boldsymbol{f}(\boldsymbol{y}^n, t^n), \tag{2.3.33}$$

$$\boldsymbol{b} = h\boldsymbol{f}(\boldsymbol{y}^n + \boldsymbol{a}/2, t^n + h/2),$$

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + \boldsymbol{b}. \tag{2.3.34}$$

Show that this method is also second order. That is, verify that the local truncation error is of the form $\boldsymbol{c}h^3 + O(h^4)$. Find a formula for $\boldsymbol{e}$.

2.3.3. Show that the Euler method (2.3) is a Runge-Kutta method with Butcher tableau

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}\ . \tag{2.3.35}$$

2.3.4. Show that the Runge-Kutta method of Exercise 3.1 above has the Butcher tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array}\ . \tag{2.3.36}$$

Show that the Runge-Kutta method of Exercise 3.2 above has the Butcher tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}\;. \tag{2.3.37}$$

2.3.5. Verify that (3.13) and (3.14) are the Butcher tableaux for RK3 and RK4, respectively.

2.3.6. Show that the Runge-Kutta method with Butcher tableau

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \tag{2.3.38}$$

describes the rule

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\boldsymbol{f}(\boldsymbol{y}^{n+1}, t^n + h) = \boldsymbol{y}^n + h\boldsymbol{f}(\boldsymbol{y}^{n+1}, t^{n+1}). \tag{2.3.39}$$

This method might properly be called the implicit endpoint rule, but is more commonly called backward Euler or, simply, implicit Euler. Verify that this method has order 1 and find an estimate for the local truncation error.

2.3.7. Show that the Butcher tableau (3.17) corresponds to the implicit midpoint rule (3.18). Review Exercises 3.1 and 3.2. Verify by direct computation of Taylor series that (3.18) is of order 2, and find an estimate for the local truncation error.

2.3.8. Show that the Butcher tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}\;. \tag{2.3.40}$$

corresponds to the Runge-Kutta formula

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + (h/2)[\boldsymbol{f}(\boldsymbol{y}^n, t^n) + \boldsymbol{f}(\boldsymbol{y}^{n+1}, t^n + h)]. \tag{2.3.41}$$

This method is known as the *trapezoidal rule.* What is its order? It is interesting to note that the Butcher tableaux (3.36) and (3.40) have the same b_i and c_i, but different matrix parts a.

2.3.9. Verify, for RK3 and RK4, that there are the Butcher tableau relations

Order 1:

$$\sum_i b_i = 1, \tag{2.3.42}$$

Order 2:

$$\sum_i b_i c_i = 1/2, \tag{2.3.43}$$

Order 3:

$$\sum_i b_i c_i^2 = 1/3, \tag{2.3.44}$$

$$\sum_{ij} b_i a_{ij} c_j = 1/6. \tag{2.3.45}$$

These relations are called *order conditions*.

Verify that (3.42) is necessary for a Runge-Kutta method to at least be of order 1. It can be shown that (3.42) and (3.43) are necessary for a Runge-Kutta method to at least be of order 2. Finally, all the conditions (3.42) through (3.45) are required for a Runge-Kutta method to at least be of order 3.

Verify that (3.42) holds for the Butcher tableaux (3.35) and (3.38), but (3.43) does not. Verify that (3.42) and (3.43), but not (3.44) and (3.45), hold for the Butcher tableaux (3.17), (3.36), (3.37), and (3.40).

2.3.10. Runge-Kutta methods, particularly those of high order, are difficult to discover. To simplify the problem, it is convenient to begin with the autonomous case of differential equations of the form

$$\dot{\boldsymbol{z}} = \boldsymbol{g}(\boldsymbol{z}), \tag{2.3.46}$$

and search for stepping formulas of the form

$$\boldsymbol{z}^{n+1} = \boldsymbol{z}^n + h\sum_{i=1}^{s} b_i \boldsymbol{\ell}_i \tag{2.3.47}$$

where at each step

$$\boldsymbol{\ell}_i = \boldsymbol{g}(\boldsymbol{z}^n + h\sum_{j=1}^{s} a_{ij}\boldsymbol{\ell}_j). \tag{2.3.48}$$

In this case there is no vector c so that the Butcher tableau takes the simpler form

$$\begin{array}{|ccc}
a_{11} & \cdots & a_{1s} \\
\vdots & & \vdots \\
a_{s1} & \cdots & a_{ss} \\
\hline
b_1 & \cdots & b_s
\end{array}\ . \tag{2.3.49}$$

Suppose that such a Runge-Kutta method of some desired order has been found for the autonomous case. We will now see that it can be parlayed into a Runge-Kutta method of the same order for the non-autonomous case (1.1).

To accomplish this feat, we will convert (1.1), which is a set of m non-autonomous equations, into a set of $m+1$ autonomous differential equations of the form (3.46). We will then apply the method of (3.49) to these equations thereby producing an associated method for (1.1).

With reference to the set (1.1), let τ be a new independent variable and treat t as a dependent variable by adding the differential equation

$$dt/d\tau = 1 \tag{2.3.50}$$

to the set. That is, introduce a new set of $(m+1)$ variables $\boldsymbol{z}$ by the rule

$$\text{first } m \text{ components of } \boldsymbol{z} = \text{first } m \text{ components of } \boldsymbol{y}, \tag{2.3.51}$$

$$(m+1)^{\text{th}} \text{ component of } \boldsymbol{z} = t; \tag{2.3.52}$$

and define an $m+1$-dimensional vector of functions $\boldsymbol{g}(\boldsymbol{z})$ by the rule

$$\text{first } m \text{ components of } \boldsymbol{g}(\boldsymbol{z}) = \text{first } m \text{ components of } \boldsymbol{f}(\boldsymbol{y}, t), \tag{2.3.53}$$

$$(m+1)^{\text{th}} \text{ component of } \boldsymbol{g}(\boldsymbol{z}) = 1. \tag{2.3.54}$$

So doing produces a set of $m+1$ autonomous (τ independent) equations of the form (3.46) where a dot now indicates $d/d\tau$. A solution of this autonomous set, after making the identification $t = \tau$, evidently produces a solution of the non-autonomous set (1.1).

Let us now apply the method (3.49) to (3.46) and examine the values of t at each stage. By the construction (3.53) and (3.54), the $(m+1)^{\text{th}}$ component of $\boldsymbol{g}$ is always 1. Next, using (3.48), show that the $(m+1)^{\text{th}}$ component of every $\boldsymbol{\ell}_i$ is also 1. Conclude that the $(m+1)^{\text{th}}$ component of the argument of $\boldsymbol{g}$ in (3.48) will be

$$(m{+}1)^{\text{th}} \text{ component of } (\boldsymbol{z}^n{+}h\sum_{j=1}^{s} a_{ij}\boldsymbol{\ell}_j) = [(m{+}1)^{\text{th}} \text{ component of } \boldsymbol{z}^n]{+}h\sum_{j=1}^{s} a_{ij}. \tag{2.3.55}$$

Moreover, if the integration method is at least of order 1, it will have integrated the equation (3.50) exactly so that

$$(m+1)^{\text{th}} \text{ component of } \boldsymbol{z}^n = t^n. \tag{2.3.56}$$

Thus, verify the result

$$(m+1)^{\text{th}} \text{ component of } (\boldsymbol{z}^n + h\sum_{j=1}^{s} a_{ij}\boldsymbol{\ell}_j) \;= t^n + h\sum_{j=1}^{s} a_{ij}. \tag{2.3.57}$$

Verify that the corresponding temporal argument on the right side of (3.7) is $t^n + c_i h$. Therefore, for consistency, verify that there must be the relation

$$t^n + c_i h = t^n + h\sum_{j=1}^{s} a_{ij}, \tag{2.3.58}$$

from which the consistency condition (3.16) follows.

2.3.11. As already mentioned in Exercise 3.10, Runge-Kutta methods, particularly those of high order, are difficult to discover. The purpose of this exercise is to explore some of the relations between Runge-Kutta formulas and quadrature formulas. For a review of quadrature formulas, see Section T.1.

Suppose the general Runge-Kutta method given by (3.6) and (3.7) is applied to the differential equation (1.1) in the special case that the right side is *independent* of $\boldsymbol{y}$. That is, consider differential equations of the special form

$$\dot{\boldsymbol{y}} = \boldsymbol{g}(t). \tag{2.3.59}$$

Show that in this case the relations (3.7) become

$$\boldsymbol{k}_i = \boldsymbol{g}(t^n + c_i h), \tag{2.3.60}$$

and the relation (3.6) becomes the stepping rule

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\sum_{i=1}^{s} b_i \boldsymbol{g}(t^n + c_i h). \tag{2.3.61}$$

Suppose further that $t^0 = 0$ and $\boldsymbol{y}^0 = 0$, and set $n = 0$ so that (3.61) takes the form

$$\boldsymbol{y}^1 = h\sum_{i=1}^{s} b_i \boldsymbol{g}(c_i h). \tag{2.3.62}$$

Finally, suppose that $\boldsymbol{g}(t)$ has the special form

$$\boldsymbol{g}(t) = \boldsymbol{\alpha}_\ell t^\ell \tag{2.3.63}$$

where $\boldsymbol{\alpha}_\ell$ is some fixed vector. Then (3.62) becomes

$$\boldsymbol{y}^1 = h\sum_{i=1}^{s} b_i \boldsymbol{\alpha}_\ell (c_i h)^\ell = h^{\ell+1}\boldsymbol{\alpha}_\ell \sum_{i=1}^{s} b_i (c_i)^\ell. \tag{2.3.64}$$

Next verify that the *exact* solution to (3.59), with $t^0 = 0$ and $\boldsymbol{y}^0 = 0$ and $\boldsymbol{g}$ given by (3.63), is

$$\boldsymbol{y}_e(t) = \boldsymbol{\alpha}_\ell t^{\ell+1}/(\ell+1), \tag{2.3.65}$$

and therefore

$$\boldsymbol{y}_e^1 = \boldsymbol{y}_e(h) = \boldsymbol{\alpha}_\ell h^{\ell+1}/(\ell+1). \tag{2.3.66}$$

Upon comparing (3.64) and (3.66), to the extent that $\boldsymbol{y}^1$ and $\boldsymbol{y}_e^1$ are to agree, we see that we should explore the possibilities

$$\sum_{i=1}^{s} b_i (c_i)^\ell \overset{?}{=} 1/(\ell+1) \tag{2.3.67}$$

for various choices of the b_i and c_i and various values of ℓ. Evidently the conditions (3.67) are the conditions for a quadrature formula with the b_i playing the role of the weights w_i and the c_i playing the role of the sampling points x_i. See equation (T.1.2) in Appendix T. Note that the order conditions (3.42) through (3.44) are special cases of (3.67).

Verify that the b_i and c_i for the RK3 method (3.13) are those for the Newton-Cotes Simpson's rule $1-4-1$ formula, see (T.1.15) and (T.1.16), and therefore (3.67) holds for $\ell = 1, 2, 3$ but not $\ell > 3$. Verify that RK3, although only third-order accurate for differential equations of the general form (1.1), is fourth-order accurate for differential equations of the special form (3.59). See (T.1.66).

Verify that the b_i and c_i for the fourth-order method (3.15) are those for the Newton-Cotes Simpson's 3/8 rule, see (T.1.20) and (T.1.21), and therefore (3.67) holds for $\ell = 1, 2, 3$

but not $\ell > 3$. Verify that this method, which is fourth-order accurate for differential equations of the general form (1.1), is also fourth-order (and not still higher-order) accurate for differential equations of the special form (3.59). See (T.1.69).

What about the b_i and c_i for the classic RK4 method (3.14)? Verify that the the b_i and c_i for this case are *not* those for a Newton-Cotes quadrature. Verify that in this case (3.62) becomes

$$\begin{aligned} \boldsymbol{y}(h) &= h\sum_{i=1}^{s} b_i \boldsymbol{f}(c_i h) \\ &= (1/6)\boldsymbol{f}(0) + (2/6)\boldsymbol{f}(h/2) + (2/6)\boldsymbol{f}(h/2) + (1/6)\boldsymbol{f}(h) \\ &= (1/6)\boldsymbol{f}(0) + (4/6)\boldsymbol{f}(h/2) + (1/6)\boldsymbol{f}(h). \end{aligned} \tag{2.3.68}$$

Show that the right side of (3.68) *is* the Newton-Cotes quadrature rule corresponding to Simpson's $1-4-1$ formula, and therefore (3.68) is accurate through order 4, as we would at least expect since classic RK4 is supposed to be fourth order. See (T.1.66). Correspondingly, verify that (3.67) holds for $\ell = 1, 2, 3$ but not $\ell > 3$.

What about the b_i and c_i for the Gaussian Runge-Kutta methods (3.17), (3.19), and (3.20)? Verify that for these Butcher tableaux the b_i and c_i satisfy (3.67) through the advertised order. See Subsection T.1.3.

Finally, verify that the b_i and c_i for the Butcher tableaux (3.36) and (3.40) correspond to $k = 2$ closed Newton Cotes.

2.3.12. This exercise is a continuation of Exercise 3.11, which you should read. We have found and explored conditions to be satisfied by the b_i and the c_i. What can be said about the remaining matrix a_{ij} in the Butcher tableau (3.12)?

We will not consider the general case, but will describe a specific case. There is a class of Runge-Kutta methods that arises from a concept called *collocation*. For these methods, collocation is used to provide a stepping rule from $\boldsymbol{y}^n$ to $\boldsymbol{y}^{n+1}$. Remarkably, for these methods, there is a formula that specifies the matrix a_{ij} in terms of the coefficients c_i. This formula makes possible the construction of a class of Runge-Kutta methods of arbitrary order.

We now describe the use of collocation to provide a stepping rule. Select s *distinct* quantities c_i with $i = 1, 2, \cdots s$. Let $\boldsymbol{P}_n(t)$ be a vector-valued polynomial in t of degree s specified by the $s+1$ requirements that

$$\boldsymbol{P}_n(t^n) = \boldsymbol{y}^n, \tag{2.3.69}$$

$$\dot{\boldsymbol{P}}_n(t^n + c_i h) = \boldsymbol{f}[\boldsymbol{P}_n(t^n + c_i h), t^n + c_i h],\ i = 1, 2, \cdots, s. \tag{2.3.70}$$

The points $t^n + c_i h$ are called collocation points. According to dictionaries, *collocation* is defined as the result of "arranging" together. Here we have required that the time derivative of $\boldsymbol{P}_n(t)$ and the value of $\boldsymbol{f}$ be equal at the collocation points. Since a polynomial of degree s requires $s+1$ conditions for its specification, we have indeed specified $\boldsymbol{P}_n(t)$.

Moreover since, according to (3.69) and (3.70), $\boldsymbol{P}_n(t)$ satisfies $s+1$ relations that are also satisfied by $\boldsymbol{y}(t)$, we expect that $\boldsymbol{P}_n(t)$ will be a good approximation to $\boldsymbol{y}(t)$. We therefore make the stepping rule

$$\boldsymbol{y}^{n+1} = \boldsymbol{P}_n(t^n + h). \tag{2.3.71}$$

It can be shown that if s quantities c_i and their associated b_i can be found such that (3.67) is satisfied for all $\ell < m$ (but not $\ell = m$), then use of (3.69) through (3.71) is equivalent to an s-stage Runge-Kutta method having order m. In other words,

$$m = \ell_{\max} + 1. \tag{2.3.72}$$

The Butcher tableau for this method contains the b_i and c_i. Moreover, as will be described shortly, with a knowledge of the c_i there is a recipe for constructing the matrix entries a_{ij} in the Butcher tableau. Thus, there is a procedure for constructing a class of Runge-Kutta formulas of arbitrary order.

Before describing the recipe for the matrix entries a_{ij}, we pause to elaborate briefly on the connection between collocation and Runge-Kutta. Begin with the truism that

$$\boldsymbol{y}^{n+1} - \boldsymbol{y}^n = \int_{t^n}^{t^{n+1}} dt\, \dot{\boldsymbol{y}}(t) = \int_{t^n}^{t^{n+1}} dt\, \boldsymbol{f}[\boldsymbol{y}(t), t]. \tag{2.3.73}$$

Next estimate the right side of (3.73) using a quadrature formula that employs the points $t^n + c_i h$ as sampling points and the b_i as weights. So doing gives the approximation

$$\boldsymbol{y}^{n+1} \simeq \boldsymbol{y}^n + h \sum_{i=1}^{s} b_i \boldsymbol{f}[\boldsymbol{y}(t^n + c_i h), t^n + c_i h]. \tag{2.3.74}$$

It can be shown that there is the correspondence

$$\boldsymbol{k}_i \simeq \boldsymbol{f}[\boldsymbol{y}(t^n + c_i h), t^n + c_i h], \tag{2.3.75}$$

and therefore

$$\boldsymbol{y}^{n+1} \simeq \boldsymbol{y}^n + h \sum_{i=1}^{s} b_i \boldsymbol{k}_i, \tag{2.3.76}$$

in agreement with (3.6).

We now describe the recipe for constructing a full Butcher tableau in terms of the c_i and based on the collocation Ansatz. We already know how to construct the b_i in terms of the c_i. Given the c_i, we form the associated Lagrange polynomials $L_i(x)$ and then integrate them over the interval $[0, 1]$ to find the b_i. See (T.1.4) through (T.1.9). It can be shown that the the matrix entries a_{ij} associated with the collocation Ansatz are also given in terms of integrals of Lagrange polynomials by the rule

$$a_{ij} = \int_0^{c_i} dx\, L_j(x). \tag{2.3.77}$$

Further work, based on the result (3.77), shows that equivalently the matrix entries a_{ij} can be found from the c_i by a matrix algorithm: First, define an $s \times s$ matrix u by the rule

$$u_{jk} = c_j^{k-1} \text{ with } \; j, k = 1, \cdots, s. \tag{2.3.78}$$

Next define an $s \times s$ matrix v by the rule

$$v_{ik} = c_i^k / k \text{ with } \; i, k = 1, \cdots, s. \tag{2.3.79}$$

Then the matrix a is given by the rule

$$a = vu^{-1}. \tag{2.3.80}$$

For a proof of all these results, see the book *Geometric Numerical Integration* by *Hairer* et al. cited in the Bibliography at the end of this chapter.

Evidently there are two possible complications in executing the instructions (3.78) and (3.80). First it could happen that some $c_j = 0$, in which case use of (3.78) will involve the ambiguous quantity 0^0. Indeed, this could well occur because $c_1 = 0$ for closed Newton Cotes. However, since x^0 is taken to represent the function $g(x) = 1$, we should make the choice

$$0^0 = g(0) = 1. \tag{2.3.81}$$

Second, one must verify that the matrix u has an inverse, which is equivalent to the condition

$$\det(u) \neq 0. \tag{2.3.82}$$

The c_j violate this condition if they are not all distinct. Note that the c_j for the RK4 Butcher tableau (3.14) have $c_2 = c_3$. Here we require that the c_j be distinct, and it can be shown that this condition is sufficient to guarantee the existence of u^{-1}.[11]

In summary, it can be shown that the quantities c_i, b_i, and a_{ij}, with the b_i constructed from the c_i using (T.1.5) and (T.1.9) and the matrix a given by (3.77) or (3.80), produce a Runge-Kutta method of order m provided (3.67) is satisfied for all $\ell < m$ (but not $\ell = m$). Do m values for this procedure, which according to (3.72) and (T.1.11) may be as large as $2s$, violate the claim of Table 3.1? The answer is *no* because the Runge-Kutta methods produced in this way are *implicit*.

We emphasize, of course, that not all Runge-Kutta methods are provided by this construction. In particular, the explicit Runge-Kutta methods fall outside this class.

Your task in this exercise is to use the matrix algorithm just described to construct the Butcher tableaux (3.17) for Gauss2, (3.40) for the trapezoidal rule, and (3.19) for Gauss4.

Consider first the $s = 1$ case of Gauss2 given by (3.17). In this case both u and v are 1×1 matrices. For b_1 and c_1 we use the values $b_1 = 1$ and $c_1 = 1/2$, which corresponds to the use of $k = 1$ Legendre Gauss. In this case (3.67) is satisfied for $\ell = 0$ and $\ell = 1$, but not $\ell = 2$. Thus we expect the method to have order $m = 2$. For $c_1 = 1/2$ show that

$$u_{11} = c_1^0 = (1/2)^0 = 1, \tag{2.3.83}$$

$$v_{11} = c_1^1 = (1/2)^1 = 1/2, \tag{2.3.84}$$

from which it follows that

$$a_{11} = v_{11}/u_{11} = 1/2, \tag{2.3.85}$$

in accord with the matrix entry in (3.17).

Consider next the $s = 2$ case of the trapezoidal rule given by (3.40). Since $s = 2$, we expect that u and v will be 2×2. Suppose we use $k = 2$ Newton Cotes for which

[11] Verify that $\det(u)$ is a *Vandermonde* determinant. See (17.2.23) and (17.2.29).

$b_1 = b_2 = 1/2$, $c_1 = 0$, and $c_2 = 1$. In this case we have $\ell_{\max} = 1$, see Table T.1.1, and we expect m=2. Verify the results

$$u_{11} = c_1^0 = 0^0 = 1, \tag{2.3.86}$$

$$u_{12} = c_1^1 = 0^1 = 0, \tag{2.3.87}$$

$$u_{21} = c_2^0 = 1^0 = 1, \tag{2.3.88}$$

$$u_{22} = c_2^1 = 1^1 = 1, \tag{2.3.89}$$

and therefore

$$u = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}. \tag{2.3.90}$$

Similarly, show that

$$v = \begin{pmatrix} 0 & 0 \\ 1 & 1/2 \end{pmatrix}. \tag{2.3.91}$$

Use these results to show that

$$a = vu^{-1} = \begin{pmatrix} 0 & 0 \\ 1/2 & 1/2 \end{pmatrix}, \tag{2.3.92}$$

in agreement with the matrix part of the Butcher tableau (3.40). Verify that the integrals (3.77) for the a_{ij} are easily evaluated in this case again yielding the result (3.92).

Your last challenge is to consider the $s = 2$ case of Gauss4 given by (3.19). In this case u and v are again 2×2 matrices. For the b_i and c_i choose values associated with $k = 2$ Legendre Gauss. That is, make the choice

$$(b_1, b_2) = (1/2, 1/2), \tag{2.3.93}$$

$$(c_1, c_2) = (1/2 - \sqrt{3}/6, 1/2 + \sqrt{3}/6). \tag{2.3.94}$$

For this choice (3.67) is satisfied for $\ell = 0, 1, 2,$ and 3, but not $\ell = 4$; and we expect the order to be $m = 4$. Verify the results

$$u_{11} = c_1^0 = (1/2 - \sqrt{3}/6)^0 = 1, \tag{2.3.95}$$

$$u_{12} = c_1^1 = 1/2 - \sqrt{3}/6 \tag{2.3.96}$$

$$u_{21} = c_2^0 = (1/2 + \sqrt{3}/6)^0 = 1, \tag{2.3.97}$$

$$u_{22} = c_2^1 = 1/2 + \sqrt{3}/6, \tag{2.3.98}$$

and therefore

$$u = \begin{pmatrix} 1 & 1/2 - \sqrt{3}/6 \\ 1 & 1/2 + \sqrt{3}/6 \end{pmatrix}. \tag{2.3.99}$$

Also verify the results

$$v_{11} = c_1^1 = 1/2 - \sqrt{3}/6, \tag{2.3.100}$$

$$v_{12} = c_1^2/2 = (1/2 - \sqrt{3}/6)^2/2 = 1/6 - \sqrt{3}/12, \tag{2.3.101}$$

$$v_{21} = c_2^1 = 1/2 + \sqrt{3}/6, \tag{2.3.102}$$

$$v_{22} = c_2^2/2 = (1/2 + \sqrt{3}/6)^2/2 = 1/6 + \sqrt{3}/12, \tag{2.3.103}$$

and therefore

$$v = \begin{pmatrix} 1/2 - \sqrt{3}/6 & 1/6 - \sqrt{3}/12 \\ 1/2 + \sqrt{3}/6 & 1/6 + \sqrt{3}/12 \end{pmatrix}. \tag{2.3.104}$$

Next verify that

$$u^{-1} = \sqrt{3} \begin{pmatrix} 1/2 + \sqrt{3}/6 & -1/2 + \sqrt{3}/6 \\ -1 & 1 \end{pmatrix}. \tag{2.3.105}$$

Finally, show that

$$a = vu^{-1} = \begin{pmatrix} 1/4 & 1/4 - \sqrt{3}/6 \\ 1/4 + \sqrt{3}/6 & 1/4 \end{pmatrix}, \tag{2.3.106}$$

which agrees with the matrix part of (3.19).

2.4 Finite-Difference/Multistep/Multivalue Methods

2.4.1 Background

Motivation and Terminology

In Runge-Kutta methods, one essentially begins anew at each step, and (apart from the $\boldsymbol{y}$ value that is already at hand) disregards any previously obtained information about the trajectory under study. Methods with this property are called *single-step* methods. This is fine, of course, when one is *beginning* a solution since all one has then is the initial conditions. However, once the integration is sufficiently underway, it clearly would be advantageous to make use of some of the "information" contained in previously obtained points. We now explore how this may be done.

Suppose we are willing to store results from $k = N + 1$ previous integration steps where N is an integer.[12] That is, we are willing to store k previous successive values of $\boldsymbol{y}^\ell$ and k previous successive values of $\boldsymbol{f}^\ell$. (Generally N ranges from 3 to 10. For purposes of the present discussion, N is selected *once and for all*, and then held *fixed* throughout the integration run.) With these values at hand, we consider a relation of the form

$$\begin{aligned} \alpha_{N+1}\boldsymbol{y}^{n+1} + \alpha_N\boldsymbol{y}^n + \alpha_{N-1}\boldsymbol{y}^{n-1} + \cdots + \alpha_0\boldsymbol{y}^{n-N} &= \\ h(\beta_{N+1}\boldsymbol{f}^{n+1} + \beta_N\boldsymbol{f}^n + \beta_{N-1}\boldsymbol{f}^{n-1} + \cdots + \beta_0\boldsymbol{f}^{n-N}) \end{aligned} \tag{2.4.1}$$

which we rewrite in the (marching-order) form

$$\begin{aligned} \boldsymbol{y}^{n+1} &= -\alpha_N\boldsymbol{y}^n - \alpha_{N-1}\boldsymbol{y}^{n-1} - \cdots - \alpha_0\boldsymbol{y}^{n-N} \\ &\quad + h(\beta_{N+1}\boldsymbol{f}^{n+1} + \beta_N\boldsymbol{f}^n + \beta_{N-1}\boldsymbol{f}^{n-1} + \cdots + \beta_0\boldsymbol{f}^{n-N}). \end{aligned} \tag{2.4.2}$$

Here, without loss of generality, we have rescaled the α_ℓ and the β_ℓ so that $\alpha_{N+1} = 1$. The formula (4.2), with *fixed* h independent coefficients, is to be used to determine $\boldsymbol{y}^{n+1}$ from the stored $\boldsymbol{y}^n \cdots \boldsymbol{y}^{n-N}$ and the stored $\boldsymbol{f}^n \cdots \boldsymbol{f}^{n-N}$. It is explicit if $\beta_{N+1} = 0$, and

[12] Warning! The symbol N in this context has a different meaning than in Sections 2 and 3.

implicit otherwise. Methods of the form (4.2) are called *multistep* methods since they employ information from $k = N + 1$ previous steps. More precisely, methods of the form (4.2) are called *k-step* methods. They are also called *multivalue* methods since (4.2) involves k previous values of $\boldsymbol{y}^\ell$ and the k previous values of $\boldsymbol{f}^\ell$.[13] Sometimes they are also called *linear*-multistep or *linear*-multivalue methods since the relation (4.2) involves a *linear* combination of the $\boldsymbol{y}^\ell$ and the $\boldsymbol{f}^\ell$. Finally, they are also called *finite-difference* methods because they can often be conveniently formulated in terms of finite differences.

Maximum Order

Suppose the coefficients in (4.2) are selected to obtain the highest possible local accuracy. What local accuracy can we hope to achieve? Imagine that $\boldsymbol{y}$ is expanded in a Taylor series about $t = t^n$ and this Taylor series is used to determine $\boldsymbol{y}^{n+1} = \boldsymbol{y}(t^n + h)$. If this series is to be accurate through terms of order h^m, it must contain $m + 1$ terms since it begins with the constant term $\boldsymbol{y}^n$. On the other hand, we have $2k$ pieces of information available in the explicit case, and $2k+1$ pieces of information in the implicit case. We therefore might hope, in the explicit case, to achieve a maximal local accuracy $m_{\max}$ given by

$$m_{\max} = 2k - 1, \text{ explicit case;} \tag{2.4.3}$$

and, in the implicit case, a maximum local accuracy of

$$m_{\max} = 2k, \text{ implicit case.} \tag{2.4.4}$$

For example there is the $N = 1$, and therefore $k = 2$, two-step explicit formula

$$\boldsymbol{y}^{n+1} = -4\boldsymbol{y}^n + 5\boldsymbol{y}^{n-1} + 4h\boldsymbol{f}^n + 2h\boldsymbol{f}^{n-1}, \tag{2.4.5}$$

and the two-step implicit formula

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^{n-1} + (h/3)\boldsymbol{f}^{n+1} + (4h/3)\boldsymbol{f}^n + (h/3)\boldsymbol{f}^{n-1}. \tag{2.4.6}$$

Suppose we stipulate that the monomial

$$\boldsymbol{y}(t) = \boldsymbol{a}t^j \tag{2.4.7}$$

(where $\boldsymbol{a}$ is a constant vector) be the exact solution to (1.1), from which it follows that

$$\boldsymbol{f}(\boldsymbol{y}, t) = j\boldsymbol{a}t^{j-1}. \tag{2.4.8}$$

Upon inserting (4.7) and (4.8) into (4.5) it is easily verified that (4.5) holds exactly for $j = 0, 1, 2, 3$ and fails to be exact for $j \geq 4$. The formula (4.5) is therefore locally accurate through terms of order h^3, which according to (4.3) is the highest order that might be expected in the explicit case. Indeed, it is easy to verify that (4.5) is the unique explicit two-step formula having third-order accuracy. Similarly, it can be verified that (4.6) is exact for $j = 0, 1, 2, 3, 4$ and fails for $j \geq 5$. The formula (4.6) is therefore locally accurate through terms of order h^4, which according to (4.4) is the highest order that might be expected in the implicit case.

[13] Some authors use the term *multivalue* to refer to the jet formulation described in Subsection 5.3.

Stability

At this point we can make a simple observation. Consider the polynomial $\rho(\zeta)$, which (for reasons that will become clear later) we call the *stability* polynomial, defined by the rule

$$\rho(\zeta) = \sum_{j=0}^{N+1} \alpha_j \zeta^j = \zeta^{N+1} + \sum_{j=0}^{N} \alpha_j \zeta^j. \tag{2.4.9}$$

Suppose that the marching rule (4.2) is exact for the monomial (4.7) with $j = 0$, in which case $\boldsymbol{f} = 0$. That is, we impose the requirement that (4.2) at least integrate the constant function $\boldsymbol{y} = \boldsymbol{a}$ exactly so that the Ansatz $\boldsymbol{y}^\ell = \boldsymbol{a}$ and $\boldsymbol{f}^\ell = 0$ satisfies (4.2) exactly. (This requirement is called *consistency* of order zero.) Doing so evidently yields the result

$$\boldsymbol{a} = -\boldsymbol{a}(\alpha_N + \alpha_{N-1} + \cdots + \alpha_0) \tag{2.4.10}$$

from which it follows that

$$1 + \sum_{j=0}^{N} \alpha_j = \rho(1) = 0. \tag{2.4.11}$$

Thus, the stability polynomial must have $\zeta = 1$ as a root for the method (4.2) to even be of minimal interest. In particular, if the method (4.2) has $m_{\max} \geq 1$, which are the cases of actual interest, then (4.11) must be satisfied. At this point, for convenient subsequent use and in analogy to (4.9), we also define a polynomial $\sigma(\zeta)$ by the rule

$$\sigma(\zeta) = \sum_{j=0}^{N+1} \beta_j \zeta^j. \tag{2.4.12}$$

To gain further insight into possible properties of multistep methods, let us now examine the use of the specific procedure (4.5) in more detail. Suppose it is used to integrate the scalar differential equation

$$\dot{y} = f(y, t) = \lambda y \tag{2.4.13}$$

with the initial condition

$$y(0) = 1. \tag{2.4.14}$$

(We suppose $t^0 = 0$.) The exact solution in this case is evidently

$$y(t) = \exp(\lambda t). \tag{2.4.15}$$

Let us study how the solution to the marching orders (4.5) approximates this exact solution.

For the case (4.13) we have $f^\ell = \lambda y^\ell$, and therefore the marching orders (4.5) become

$$y^{n+1} = -4y^n + 5y^{n-1} + 4h\lambda y^n + 2h\lambda y^{n-1} = (-4 + 4h\lambda)y^n + (5 + 2h\lambda)y^{n-1}. \tag{2.4.16}$$

Observe that (4.16) is a linear recursion relation. To solve it, try the Ansatz

$$y^n \propto (\zeta)^n \tag{2.4.17}$$

where the quantity ζ is to be determined and the notation is meant to be suggestive. The Ansatz (4.17), when inserted into (4.16), yields the *characteristic* equation

$$\zeta^2 + (4 - 4h\lambda)\zeta - (5 + 2h\lambda) = 0. \tag{2.4.18}$$

It follows that (4.16) has a general solution of the form

$$y^n = A[\zeta_1(h)]^n + B[\zeta_2(h)]^n \tag{2.4.19}$$

where ζ_1 and ζ_2 are the roots of (4.18), and the solution is made specific by selecting the coefficients A,B so that the conditions

$$y^0 = 1 \tag{2.4.20}$$

and

$$y^{-1} = \exp(-h\lambda) \tag{2.4.21}$$

are satisfied.

The roots of (4.18) are

$$\zeta = -2 + 2h\lambda \pm \sqrt{[9 - 6h\lambda + 4(h\lambda)^2]}, \tag{2.4.22}$$

and they have the expansions

$$\begin{aligned} \zeta_1(h) &= -2 + 2h\lambda + \sqrt{[9 - 6h\lambda + 4(h\lambda)^2]} \\ &= 1 + (h\lambda) + (h\lambda)^2/2! + (h\lambda)^3/3! + (h\lambda)^4/72 + \cdots \\ &= \exp(h\lambda) + O(h^4), \end{aligned} \tag{2.4.23}$$

$$\begin{aligned} \zeta_2(h) &= -2 + 2h\lambda - \sqrt{[9 - 6h\lambda + 4(h\lambda)^2]} \\ &= -5 + 3(h\lambda) + O(h^2). \end{aligned} \tag{2.4.24}$$

Note from (4.23) that, as $h \to 0$, the root ζ_1 becomes $\zeta_1 = 1$. That $\zeta = 1$ is a root in this limit is to be expected: From (4.2) we see that the characteristic equation (4.18) can be written in the form

$$\rho(\zeta) - h\lambda\sigma(\zeta) = 0. \tag{2.4.25}$$

In the limit $h = 0$ the characteristic equation written as (4.25) becomes the relation

$$\rho(\zeta) = 0, \tag{2.4.26}$$

and we know from our previous discussion that $\zeta = 1$ is root of (4.26) since the method (4.5) has $m_{\max} = 3$ and therefore $m_{\max} \geq 1$.

Suppose we set $A = 1$ and $B = 0$ in (4.19). Then (4.20) is satisfied exactly, and from (4.23) we see that (4.21) is satisfied through terms of order h^3. Moreover, in this case (4.19) can be rewritten in the form

$$y^n = (\zeta_1)^n = \exp[n\log(\zeta_1)] = \exp\{n[h\lambda + O(h^4)]\} = \exp(\lambda t^n)\exp[nO(h^4)]. \tag{2.4.27}$$

And, if we follow the marching orders to the time $t^n = T$ so that $n = T/h$, we obtain the result

$$y(T) = \exp(\lambda T)\exp[(T/h)O(h^4)] = \exp(\lambda T)\exp[TO(h^3)]. \tag{2.4.28}$$

Evidently, as comparison with (4.15) reveals, (4.28) becomes exact in the limit $h \to 0$.

Suppose instead we require (4.20) as before, but now require that (4.21) hold exactly. This would seem to be desirable because (4.20) and (4.21) are properties of the exact solution (4.15). Then we find the relations

$$A + B = 1, \tag{2.4.29}$$

$$A[\exp(-h\lambda) + O(h^4)] + B[-5 + O(h)]^{-1} = \exp(-h\lambda), \tag{2.4.30}$$

from which it follows that

$$A = 1 + O(h^4), \tag{2.4.31}$$

$$B = O(h^4) \neq 0. \tag{2.4.32}$$

Correspondingly, (4.19) becomes

$$y^n = A(\zeta_1)^n + B[-5 + O(h)]^n. \tag{2.4.33}$$

And, if we now if we follow the marching orders to the time T, we find the result

$$y(T) = [1 + O(h^3)]\exp(\lambda T) + O(h^4)(-5)^{T/h}. \tag{2.4.34}$$

We see that the first term in (4.34) becomes the exact solution in the limit $h \to 0$, but the second oscillates wildly with ever growing amplitude as $h \to 0$. For this reason, the method (4.5) is called *unstable.* Although the factor B in the second term of (4.33) and (4.34) vanishes as h^4 when $h \to 0$, the second factor grows (in amplitude) very rapidly because $|\zeta_2| > 1$. And this rapid growth dominates the vanishing of B so that their product also grows rapidly. On the other hand if it had happened that $|\zeta_2| < 1$, which might be the case for some other integration procedure, then both factors would vanish as $h \to 0$ so that only the first term would remain thereby producing the exact result for $y(T)$.

What have we learned from this example? First, the characteristic equation must have a root that is near $+1$, and this root produces a "desired" solution of the marching orders that approximates the exact solution of the associated differential equation. We will call this root the *good* root. In addition there are other roots, $k - 1$ in number because the characteristic equation is a polynomial of degree k, that produce other solutions. These solutions are called *parasitic* solutions. If their associated roots, which we will call parasitic roots, lie outside the unit circle in the complex plane, these solutions grow without bound and can eventually swamp the true solution. Finally, the nature of the roots can be found for small h by examining the roots of the stability polynomial $\rho(\zeta)$.

We conclude that a multistep method is generally of little interest if any roots of the stability polynomial $\rho(\zeta)$ lie outside the unit circle. A multistep method is defined to be *strongly stable* if its $\rho(\zeta)$ has $+1$ as a root and all other roots lie *inside* the unit circle. In general, unless a multistep procedure is initiated "just right", some or all of the parasitic solutions will also be present in the result. Also, even when the procedure is initiated "just right", the parasitic solutions will continually be "excited" during the march due to round-off errors. But if a method is strongly stable and h is small enough, then the parasitic-solution roots of the characteristic equation will lie within the unit circle and the parasitic solutions will decay to zero thereby leaving behind only the desired solution as $h \to 0$.

The First Dahlquist Barrier

What is the maximum local order $m_{\max}$ that can be achieved with a strongly stable k-step method? It can be shown that if strong stability is required, then there is the result

$$m_{\max} = k, \text{ explicit case;} \tag{2.4.35}$$

$$m_{\max} = k + 2, \text{ implicit case and } k \text{ even,} \tag{2.4.36}$$

$$m_{\max} = k + 1, \text{ implicit case and } k \text{ odd.} \tag{2.4.37}$$

This limit is called the *first Dahlquist barrier.*[14] A common practice is to employ order $m = k$ methods for the explicit case and order $m = k$ or $m = k + 1$ methods for the implicit case.

Strictly speaking, the order given by (4.36) cannot be reached unless all the roots of $\rho(\zeta)$ are on the unit circle, in which case it can be arranged that they are all distinct. By our definition, methods with this property are not strongly stable, but rather are a borderline case. However, they may be useful in some circumstances. In general, the first Dahlquist barrier for the implicit case, for both k even and k odd, is $m_{\max} = k + 1$.

Convergence

Again speaking strictly, our discussion of convergence so far holds for differential equations of the form (4.13). However, it can be proved that if a multistep method has local accuracy through terms of order h^m with $m \geq 1$ and is strongly stable, then the result of following the marching orders from $t = t^0$ to $t = t^0 + T$ converges to the exact result for $\boldsymbol{y}(t^0 + T)$ as $h \to 0$ for any differential equation provided $\boldsymbol{f}(\boldsymbol{y}, t)$ has sufficiently many continuous derivatives and the stored starting values are exact.

We also note, still strictly speaking, that the concept of a characteristic equation applies only to cases of linear differential equations of the general form (4.13). However, if a method cannot integrate (4.13) well, then it is unlikely to be able to integrate more complicated nonlinear equations well.

2.4.2 Adams' Method

Suppose in (4.2) we set

$$\alpha_N = -1 \tag{2.4.38}$$

and

$$\alpha_\ell = 0 \text{ for } \ell = 0, 1, \cdots, N - 1. \tag{2.4.39}$$

In this case the stability polynomial becomes

$$\rho(\zeta) = \zeta^k - \zeta^{k-1} = (\zeta - 1)\zeta^{k-1}, \tag{2.4.40}$$

which evidently has the single root $\zeta_1 = 1$ and the multiple roots $\zeta_\ell = 0$ for $\ell = 2, 3, \cdots k$. This would seem to be a highly desirable state of affairs because with this choice for the α_ℓ

[14] There is also a *second* Dahlquist barrier that arises in the integration of so-called *stiff* equations by implicit methods. Their treatment is beyond the scope of this text.

all the parasitic roots of ρ *vanish*, and one might hope correspondingly that all the parasitic roots of the characteristic equation would be well within the unit circle providing h is not too large.

Upon taking into account the Ansatz specified by (4.38) and (4.39), the marching orders (4.2) take the form

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h(\beta_{N+1}\boldsymbol{f}^{n+1} + \beta_N\boldsymbol{f}^n + \beta_{N-1}\boldsymbol{f}^{n-1} + \cdots + \beta_0\boldsymbol{f}^{n-N}).$$

The remaining task is to chose the β_ℓ in such a way that the order is maximized to bring it as close to the first Dahlquist barrier as is conveniently possible. The stepping methods thereby obtained are variously associated with the names *modified Adams*, *Adams-Bashforth*, or *Adams-Moulton*. We shall simply call them *Adams*.

Because the derivation of Adams' method is fairly involved, we shall begin our discussion by describing the procedure for its use. Then, with the procedure well understood, we will give the derivations that justify the method. For convenience of description, we will assume the required stored starting values are obtain using Runge Kutta executed with a sufficiently small step size.

As in the cases of Crude Euler and Runge-Kutta, we begin with an initial vector $\boldsymbol{y}^0$, and our task is to compute the successive vectors $\boldsymbol{y}^1, \boldsymbol{y}^2$ etc. The procedure for Adams' method is as follows:

1. Adams' method requires the storage of information about previously obtained points on a trajectory. In particular, since(4.38) and (4.39) hold but in general $\beta_\ell \neq 0$, it requires storage of the values $\boldsymbol{f}(\boldsymbol{y}, t)$ at these points and the most recent value of $\boldsymbol{y}$. As described earlier, let $N + 1$, where N is an integer, be the number of points whose "$\boldsymbol{f}$" values we are willing to store. For purposes of our present discussion, it is selected *once and for all*, and then held *fixed* throughout the integration run. Thus, there is actually a whole family of Adams' methods with each member of the family having a different N. As we expect and will see later, the choice of N governs the accuracy of the method.

2. Using a Runge-Kutta method, compute the vectors $\boldsymbol{y}^1, \boldsymbol{y}^2, \cdots \boldsymbol{y}^N$ starting with $\boldsymbol{y}^0$. At each point $\boldsymbol{y}^n$ compute the vector $\boldsymbol{f}^n = \boldsymbol{f}(\boldsymbol{y}^n, t^n)$, and store the $N + 1$ vectors $\boldsymbol{f}^0, \boldsymbol{f}^1, \cdots \boldsymbol{f}^N$ as well as $\boldsymbol{y}^N$. Since the accuracy of these "$\boldsymbol{f}$" values greatly affects the accuracy of the solution to be obtained later on, it is worth spending considerable effort on their accurate computation. One simple method is to run Runge-Kutta with a fractional step size h/m, where m is an integer, and then use every mth Runge-Kutta step for computing the desired $\boldsymbol{y}$'s and $\boldsymbol{f}$'s.

3. We are ready to switch to Adams' method. It consists of two stepping formulas called the *predictor* and the *corrector*. The predictor formula for marching from $\boldsymbol{y}^n$ to $\boldsymbol{y}^{n+1}$ reads

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\sum_{k=0}^{N} \overset{\sim N}{b_k}\ \boldsymbol{f}^{n-k}. \qquad (\textit{predictor})$$

It is an *explicit* formula. Here the $\overset{\sim N}{b_k}$ are a set of coefficients whose values will be derived and tabulated later on. Now, using the predictor formula and the stored $\boldsymbol{f}$'s, compute $\boldsymbol{y}^{N+1}$ by putting $n = N$. This step is called *Predicting*, or P for short, and its result is called the predicted value of $\boldsymbol{y}^{N+1}$. An Adams' predictor formula is sometimes called an Adams-Bashforth formula.

4. Using $\boldsymbol{y}^{N+1}$, compute $\boldsymbol{f}^{N+1} = \boldsymbol{f}(\boldsymbol{y}^{N+1}, t^{N+1})$. This step is called *Evaluating*, or E for short, since it requires an evaluation of the function $\boldsymbol{f}$.

5. The corrector formula reads

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^{n} + h \sum_{k=0}^{N} \tilde{a}{}_k^N \; \boldsymbol{f}^{n+1-k}, \qquad (corrector)$$

where the $\tilde{a}{}_k^N$ are another set of coefficients. It is an *implicit* formula and will be solved by iteration. Using the corrector formula, the stored $\boldsymbol{f}$'s, and $\boldsymbol{f}^{N+1}$ from step 4, recompute $\boldsymbol{y}^{N+1}$ by putting $n = N$ in the corrector formula. This step is called *Correcting*, or C for short, since, as we will later see, the corrector formula is more accurate. Its result is called the corrected value of $\boldsymbol{y}^{N+1}$. An Adams' corrector formula is sometimes called an Adams-Moulton formula.

6. Return to step 4, this time using the corrected value of $\boldsymbol{y}^{N+1}$. Repeat steps 4 and 5 until successive values of $\boldsymbol{y}^{N+1}$ differ by less than some preassigned amount (usually the round-off accuracy of the computer). It can be shown that this iteration procedure converges if the step size h is small enough. Indeed, the operation EC can be shown to be a *contraction* map, and the operation P provides a first guess for the fixed point of this contraction map.[15] In actual practice the sequence $PECEC$ is usually sufficient. A need for more iterations generally indicates a too large step size.

7. The procedure is finished. We have found $\boldsymbol{y}^{N+1}$. Now update the table of $\boldsymbol{f}$'s by adding to it the value of $\boldsymbol{f}^{N+1}$ obtained from the last evaluation step, and discarding $\boldsymbol{f}^0$.

8. To compute $\boldsymbol{y}^{N+2}$, relabel the $\boldsymbol{y}$'s and $\boldsymbol{f}$'s, and return to step 3. In this manner, proceed to compute $\boldsymbol{y}^{N+2}, \boldsymbol{y}^{N+3}$, etc. until the integration run is completed. Note again that, in each case, only the previous value of $\boldsymbol{y}$ and the last $N + 1$ values of the $\boldsymbol{f}$'s are used.

2.4.3 Numerical Example

We show below in Exhibit 4.1 a program that illustrates the use of Adams' method with $N = 4$ for the differential equation set (2.7) through (2.9). Subsequently we will learn that $N = 4$ Adams is of order 5. That is, it is locally exact through terms of order h^5 and makes local errors of order h^6. See (4.37) and (4.38). Therefore, it might appropriately be called Adams5.

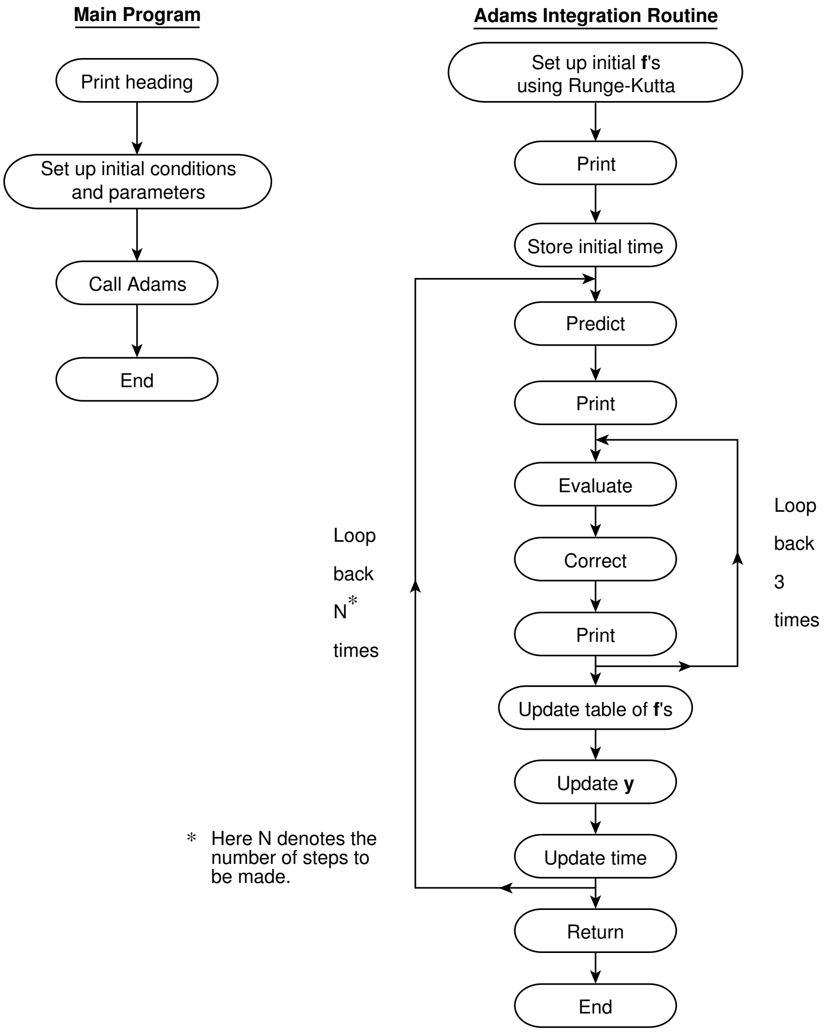
[15] For a discussion of contraction maps, see the first paragraph of Section 29.4.3 and the references at the end of Chapter 29.

The time step is again $h = 1/10$. The program is written in double precision so that round-off errors are unimportant for this step size. That is, for pedagogical simplicity, we want to avoid the need to worry about round-off errors for this example. Runge-Kutta integration, with a step size of $h/10 = 1/100$, is used as a starting procedure.

The first values in the columns labeled *y1comp* and *y2comp* printed for each time in the Adams' routine are those found by the predictor. The next three lines are the result of successive corrector iterations. That is, we have used the sequence *PECECEC*. The convergence is good, and the sequence *PECEC* would have been sufficient. *Note that the solution is now accurate to almost eight significant figures.* A more efficient version of this program using vector arrays is given in Appendix B.

In passing, let us compare the accuracy of RK3 and Adams5, and the effort involved in each, for this simple example. From Exhibit 3.1 we we saw that RK3 had an accuracy (with a step size $h = 1/10$) of five significant figures. And, according to (3.2) and (3.3), three function evaluations were required per step. By contrast, with the same step size, Adams5 has an accuracy of almost eight significant figures. And, when *PECEC* is used, only two function evaluations are required per step. Thus Adams5 is considerably more accurate and involves less effort than RK3.

Exhibit 2.4.1: Fifth-Order Adams Integration



∗ Here N denotes the number of steps to be made.

Computer Programs

```
c This is the main program for illustrating an Adams method
c for numerical integration.
c
      implicit double precision (a-h,o-z)
c
c Print heading.
c
      write(6,100)
  100 format
     & (1h ,'time',4x,'y1comp',10x,'y2comp',10x,'y1true',
     & 10x,'y2true',/)
c
c Set up initial conditions and parameters. n is the number of integration
c steps we wish to make.
c
      t=0.d0
      h=.1d0
      n=15
      y1=0.d0
      y2=1.d0
c
      call adams(t,h,n,y1,y2)
c
      end
c
c This is a fifth-order Adams integration subroutine.
c
      subroutine adams(t,h,n,y1,y2)
      implicit double precision (a-h,o-z)
      dimension f1(5),f2(5)
c
      write(6,*) 'Starting with Runge-Kutta integration'
c
c Set up initial f values.
c
      call eval(y1,y2,t,f1(1),f2(1))
      call prints(t,y1,y2,y1true(t),y2true(t),0)
      do 10 i=2,5
      call rk3(t,h/10.d0,10,y1,y2)
      call eval(y1,y2,t,f1(i),f2(i))
      call prints(t,y1,y2,y1true(t),y2true(t),0)
   10 continue
      write (6,*) 'Continuing with Adams integration'
      hdiv=h/720.d0
      n=n-4
      t=t+h
      tint=t
c
c Printing and integration loop.
c
      do 100 i=1,n
c
```

```
c Predictor step.
c
      p1=y1+hdiv*(1901.d0*f1(5)-2774.d0*f1(4)+2616.d0*f1(3)
     & -1274.d0*f1(2)+251.d0*f1(1))
      p2=y2+hdiv*(1901.d0*f2(5)-2774.d0*f2(4)+2616.d0*f2(3)
     & -1274.d0*f2(2)+251.d0*f2(1))
c
      call prints(t,p1,p2,y1true(t),y2true(t),0)
c
c Corrector steps.
c
      do 50 j=1,3
      call eval(p1,p2,t,g1,g2)
      c1=y1+hdiv*(251.d0*g1+646.d0*f1(5)-264.d0*f1(4)
     & +106.d0*f1(3)-19.d0*f1(2))
      c2=y2+hdiv*(251.d0*g2+646.d0*f2(5)-264.d0*f2(4)
     & +106.d0*f2(3)-19.d0*f2(2))
      p1=c1
      p2=c2
      call prints(t,c1,c2,0.,0.,1)
   50 continue
c
c Update
c
      do 75 j=1,4
      f1(j)=f1(j+1)
      f2(j)=f2(j+1)
   75 continue
      f1(5)=g1
      f2(5)=g2
      y1=c1
      y2=c2
      t=tint+float(i)*h
c
  100 continue
c
      return
      end
```

Numerical Results

```
time    y1comp          y2comp          y1true          y2true

Starting with Runge-Kutta integration
0.0000  0.00000000E+00  0.10000000E+01  0.00000000E+00  0.10000000E+01
0.1000  0.10016658E+00  0.10049958E+01  0.10016658E+00  0.10049958E+01
0.2000  0.20133067E+00  0.10199334E+01  0.20133067E+00  0.10199334E+01
0.3000  0.30447980E+00  0.10446635E+01  0.30447979E+00  0.10446635E+01
0.4000  0.41058166E+00  0.10789390E+01  0.41058166E+00  0.10789390E+01
Continuing with Adams integration
0.5000  0.52057439E+00  0.11224171E+01  0.52057446E+00  0.11224174E+01
        0.52057446E+00  0.11224175E+01
        0.52057448E+00  0.11224175E+01
        0.52057448E+00  0.11224175E+01
```

```
0.6000  0.63535744E+00  0.11746641E+01  0.63535753E+00  0.11746644E+01
        0.63535754E+00  0.11746644E+01
        0.63535755E+00  0.11746644E+01
        0.63535755E+00  0.11746644E+01
0.7000  0.75578220E+00  0.12351576E+01  0.75578231E+00  0.12351578E+01
        0.75578234E+00  0.12351579E+01
        0.75578235E+00  0.12351579E+01
        0.75578235E+00  0.12351579E+01
0.8000  0.88264379E+00  0.13032931E+01  0.88264391E+00  0.13032933E+01
        0.88264396E+00  0.13032934E+01
        0.88264397E+00  0.13032934E+01
        0.88264397E+00  0.13032934E+01
0.9000  0.10166730E+01  0.13783898E+01  0.10166731E+01  0.13783900E+01
        0.10166732E+01  0.13783901E+01
        0.10166732E+01  0.13783901E+01
        0.10166732E+01  0.13783901E+01
1.0000  0.11585289E+01  0.14596975E+01  0.11585290E+01  0.14596977E+01
        0.11585291E+01  0.14596978E+01
        0.11585291E+01  0.14596978E+01
        0.11585291E+01  0.14596978E+01
1.1000  0.13087925E+01  0.15464037E+01  0.13087926E+01  0.15464039E+01
        0.13087928E+01  0.15464040E+01
        0.13087928E+01  0.15464040E+01
        0.13087928E+01  0.15464040E+01
1.2000  0.14679608E+01  0.16376421E+01  0.14679609E+01  0.16376422E+01
        0.14679611E+01  0.16376423E+01
        0.14679611E+01  0.16376423E+01
        0.14679611E+01  0.16376423E+01
1.3000  0.16364417E+01  0.17325011E+01  0.16364418E+01  0.17325012E+01
        0.16364420E+01  0.17325013E+01
        0.16364420E+01  0.17325012E+01
        0.16364420E+01  0.17325012E+01
1.4000  0.18145501E+01  0.18300328E+01  0.18145503E+01  0.18300329E+01
        0.18145505E+01  0.18300329E+01
        0.18145505E+01  0.18300329E+01
        0.18145505E+01  0.18300329E+01
1.5000  0.20025049E+01  0.19292627E+01  0.20025050E+01  0.19292628E+01
        0.20025052E+01  0.19292629E+01
        0.20025052E+01  0.19292628E+01
        0.20025052E+01  0.19292628E+01
```

2.4.4 Derivation and Error Analysis

Calculus of Finite Differences

To reiterate, our remaining task is to choose the β_ℓ in such a way that the order is maximized to bring it as close to the first Dahlquist barrier as is conveniently possible. For this purpose it is useful to employ a *constructive* method based on the calculus of finite differences.

Let $\boldsymbol{y}(t)$ be any vector-valued function of t. We define a *backwards difference* operator ∇ by the rule

$$\nabla \boldsymbol{y}(t) = \boldsymbol{y}(t) - \boldsymbol{y}(t-h), \tag{2.4.41}$$

and in particular

$$\nabla \boldsymbol{y}^n = \boldsymbol{y}^n - \boldsymbol{y}^{n-1}. \tag{2.4.42}$$

Repeated applications of ∇ will be indicated by an exponent with the convention $\nabla^0 = 1$. Thus,

$$\nabla^2 \boldsymbol{y}^n = \nabla(\nabla \boldsymbol{y}^n) = \nabla \boldsymbol{y}^n - \nabla \boldsymbol{y}^{n-1} = \boldsymbol{y}^n - 2\boldsymbol{y}^{n-1} + \boldsymbol{y}^{n-2}, \tag{2.4.43}$$

and in general

$$\nabla^\ell \boldsymbol{y}^n = \sum_{k=0}^{\ell} (-1)^k \left(\begin{array}{c} \ell \\ k \end{array} \right) \boldsymbol{y}^{n-k}, \tag{2.4.44}$$

where the $\left(\begin{array}{c} \ell \\ k \end{array} \right)$ are the standard binomial coefficients.

Suppose $\boldsymbol{y}(t)$ is a polynomial in t with vector coefficients. Then it is easily checked that $\nabla \boldsymbol{y}$ is a polynomial of one order lower. We also have the relations

$$\nabla 1 = 0, \tag{2.4.45}$$

$$\nabla^k t^\ell = 0 \text{ if } k > \ell, \tag{2.4.46}$$

$$\nabla^\ell t^\ell = h^\ell \ell!, \tag{2.4.47}$$

where in this particular case t^ℓ denotes a power of t rather than the notation adopted in (1.2). Finally, we note that for $\boldsymbol{y}$ polynomial in t, not only powers of ∇ are well defined; infinite series of the form $\sum_0^\infty a_k \nabla^k$ are also defined since by (4.46) the series must always terminate when applied to a polynomial.

From Taylor's theorem we know that

$$\boldsymbol{y}^{n-1} = \boldsymbol{y}(t^n - h) = \sum_{k=0}^{\infty} [(-h)^k / k!](d^k \boldsymbol{y}^n / dt^k). \tag{2.4.48}$$

This relation can be written more compactly as

$$\boldsymbol{y}^{n-1} = e^{-hD} \boldsymbol{y}^n \tag{2.4.49}$$

where D denotes the differential operator

$$D = d/dt. \tag{2.4.50}$$

That is, if we expand e^{-hD} in a formal power series, we get (4.48). Combining (4.41) and (4.49), we find the result

$$\nabla \boldsymbol{y}^n = (1 - e^{-hD})\boldsymbol{y}^n. \tag{2.4.51}$$

Watch closely! Since (4.51) is true for any $\boldsymbol{y}$ whose functional dependence on t is polynomial, and since any continuous function can be approximated arbitrarily closely by polynomials, we can write the symbolic formula

$$\nabla = (1 - e^{-hD}) = hD - \frac{1}{2}h^2D^2 + \cdots\,. \tag{2.4.52}$$

Equation (4.52) should be viewed as a formal relation between two power series: one in ∇ and one in D. It becomes a true equation when applied to any polynomial. In this spirit, we may solve (4.52) for D to get the result

$$D = -h^{-1}\ \log(1 - \nabla). \tag{2.4.53}$$

Here $\log(1 - \nabla)$ denotes the formal series

$$\log(1 - \nabla) = -\sum_{k=1}^{\infty} \nabla^k/k. \tag{2.4.54}$$

Again, (4.53) becomes a true equation when applied to any polynomial.

Application of Finite Difference Calculus to Integration of Differential Equations

We now apply the calculus of difference operators we have just developed to the integration of differential equations. Observe that the differential equation under study, (1.1), can be written as

$$D\boldsymbol{y}^{n+1} = \boldsymbol{f}^{n+1}. \tag{2.4.55}$$

Suppose we knew how to *invert* the operator D. Then we might try writing

$$\boldsymbol{y}^{n+1} \overset{?}{=} D^{-1}\boldsymbol{f}^{n+1}. \tag{2.4.56}$$

However, we do not expect D^{-1} to be well defined since the inverse of differentiation is integration, and integration always involves the introduction of an arbitrary constant. This defect can be overcome by observing that the operator ∇D^{-1} is well defined since by (4.45) the operator ∇ annihilates any integration constant produced by D^{-1}. Thus we may convert (4.56) into the *integration formula*

$$\nabla \boldsymbol{y}^{n+1} = \nabla D^{-1}\boldsymbol{f}^{n+1}. \tag{2.4.57}$$

Now make another daring step. Since (4.53) and (4.54) express D as a formal series in ∇, we might hope to get ∇D^{-1} as another series in ∇ by the operation of division. Assuming this is possible, use of (4.53) in (4.57) gives the result

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h[-\nabla/\log(1 - \nabla)]\boldsymbol{f}^{n+1}. \tag{2.4.58}$$

We shall verify shortly that the expression $[-\nabla/\log(1-\nabla)]$ has a well-defined series expansion in ∇ so that (4.58) is formally correct, and actually true for polynomials. Before doing so, we continue on to derive another strange-looking expression. From the definition of ∇ we have the relation

$$\nabla \boldsymbol{f}^{n+1} = \boldsymbol{f}^{n+1} - \boldsymbol{f}^{n}. \tag{2.4.59}$$

Rearranging terms we find

$$\boldsymbol{f}^{n} = (1-\nabla)\boldsymbol{f}^{n+1}. \tag{2.4.60}$$

Let us solve for $\boldsymbol{f}^{n+1}$. We have symbolically

$$\boldsymbol{f}^{n+1} = (1-\nabla)^{-1}\boldsymbol{f}^{n}. \tag{2.4.61}$$

Now insert (4.61) into (4.58) to get the result

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^{n} + h\{-\nabla/[(1-\nabla)\ \log(1-\nabla)]\}\boldsymbol{f}^{n}. \tag{2.4.62}$$

How are expressions such as (4.58) and (4.62) to be understood? Consider the functions $F(z)$ and $G(z)$ defined by

$$F(z) = -z/\log(1-z), \tag{2.4.63}$$

$$G(z) = -z/[(1-z)\ \log(1-z)]. \tag{2.4.64}$$

Near $z=0$ we know that $\log(1-z) = -z + O(z^2)$ so that $F(0)$ and $G(0)$ are well defined. Furthermore, $\log(1-z)$ and $(1-z)^{-1}$ are singular only when $z=1$. We conclude that F and G are analytic in the complex z plane within the unit disk $|z|<1$. Consequently, we may write the series expansions

$$F(z) = \sum_{0}^{\infty} a_k z^k, \tag{2.4.65}$$

$$G(z) = \sum_{0}^{\infty} b_k z^k. \tag{2.4.66}$$

The first few coefficients are listed in Table 4.1 below. The ratio $|b_k/a_k|$ is also roughly tabulated for later use. The answer to our question is now clear. We use the series (4.65) and (4.66) to define the expressions in question, and in so doing obtain relations that are true for arbitrary polynomials.

Table 2.4.1: Expansion Coefficients for F and G.

k	0	1	2	3	4	5	6	7	8	9
a_k	1	$-\frac{1}{2}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$	$\frac{-3}{160}$	$\frac{-863}{60480}$	$\frac{-275}{24192}$	$\frac{-33953}{3628800}$	$\frac{-8183}{1036800}$
b_k	1	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$	$\frac{95}{288}$	$\frac{19087}{60480}$	$\frac{5257}{17280}$	$\frac{1070017}{3628800}$	$\frac{25713}{89600}$
$\lvert b_k/a_k\rvert$	1	1	5	9	~ 13	~ 17	~ 22	~ 27	~ 32	~ 36

Predictor-Corrector Formulas

With this brief explanation, we return to the problem of numerical integration. Suppose we wish to proceed from $\boldsymbol{y}^n$ to $\boldsymbol{y}^{n+1}$ on the basis of a polynomial fit in t of order $N+1$. That is, $\boldsymbol{y}(t)$ is approximated by a polynomial of order $N+1$, and we are willing to tolerate local errors of order h^{N+2}. Since $\boldsymbol{f} = \dot{\boldsymbol{y}}, \boldsymbol{f}$ will be a polynomial of order N, and according to (4.46) we need to retain only N^{th} and lower differences. Thus, we may replace (4.58) and (4.62) by the two *truncated* formulas

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\sum_0^N a_k \nabla^k \boldsymbol{f}^{n+1}, \qquad (\textit{corrector}) \qquad (2.4.67)$$

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\sum_0^N b_k \nabla^k \boldsymbol{f}^{n}. \qquad (\textit{predictor}) \qquad (2.4.68)$$

As the reader may have guessed, we have given the formulas the names *corrector* and *predictor* in anticipation of their use. We may also write (4.67) and (4.68) in the expanded form

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\sum_0^N \overset{\sim N}{a_k}\ \boldsymbol{f}^{n+1-k}, \qquad (\textit{corrector}) \qquad (2.4.69)$$

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\sum_0^N \overset{\sim N}{b_k}\ \boldsymbol{f}^{n-k} \qquad (\textit{predictor}) \qquad (2.4.70)$$

where the coefficients $\overset{\sim N}{a_k}$, $\overset{\sim N}{b_k}$ are related to the earlier set a_k, b_k using (4.44). The coefficients $\overset{\sim N}{a_k}$ and $\overset{\sim N}{b_k}$ are listed in Tables 4.2 and 4.3 at the end of this section. Note that these coefficients depend on both k and N.

Error Analysis

Both formulas (4.67) and (4.68) are correct through terms of order h^{N+1}. However, in general the *truncation errors* involved in the *corrector* (4.67) are numerically smaller than those in the *predictor* (4.68). To see this, suppose that $\boldsymbol{y}(t)$ is approximated *exactly* by a polynomial of order $N+2$,

$$\boldsymbol{y}(t) = \sum_0^{N+2} \boldsymbol{c}_j (t - t^n)^j. \qquad (2.4.71)$$

Then, using a corrector series with upper summation index $(N+1)$, we would have the exact result

$$\boldsymbol{y}^{n+1}_{\text{true}} = \boldsymbol{y}^n + h\sum_0^{N+1} a_k \nabla^k \boldsymbol{f}^{n+1}. \qquad (2.4.72)$$

[Note that in general the summation index in (4.72) should extend to infinity. However, because of the assumption (4.71), it may be cut off as indicated.] By contrast, using (4.67), the actual corrector gives the approximate result

$$\boldsymbol{y}^{n+1}_{\text{corr}} = \boldsymbol{y}^n + h\sum_{0}^{N} a_k \nabla^k \boldsymbol{f}^{n+1}. \tag{2.4.73}$$

Upon subtracting the two results, we find the relation

$$\boldsymbol{y}^{n+1}_{\text{true}} - \boldsymbol{y}^{n+1}_{\text{corr}} = h a_{N+1} \nabla^{N+1} \boldsymbol{f}^{n+1}. \tag{2.4.74}$$

The right side of (4.74) is easily evaluated using $\boldsymbol{f} = \dot{\boldsymbol{y}}$, (4.71), (4.46), and (4.47). The result is the relation

$$\boldsymbol{y}^{n+1}_{\text{true}} - \boldsymbol{y}^{n+1}_{\text{corr}} = h^{N+2} a_{N+1} (N+2)! \boldsymbol{c}_{N+2}. \tag{2.4.75}$$

Finally, we observe that

$$(N+2)! \boldsymbol{c}_{N+2} = (d^{N+2}\boldsymbol{y}/dt^{N+2}), \tag{2.4.76}$$

so that the local error involved in using the corrector formula is given by

$$\boldsymbol{y}^{n+1}_{\text{true}} - \boldsymbol{y}^{n+1}_{\text{corr}} \approx h^{N+2} a_{N+1} (d^{N+2}\boldsymbol{y}/dt^{N+2})|_{t=t^n}. \tag{2.4.77}$$

Similarly, the predictor formula local error is given by

$$\boldsymbol{y}^{n+1}_{\text{true}} - \boldsymbol{y}^{n+1}_{\text{pred}} \approx h^{N+2} b_{N+1} (d^{N+2}\boldsymbol{y}/dt^{N+2})|_{t=t^n}. \tag{2.4.78}$$

Equations (4.77) and (4.78) are exact for polynomials of order $N+2$, and approximate otherwise. Now look at Table 1. We see that, for $N > 2$, a_N is considerably smaller than b_N and therefore the corrector formula has higher accuracy.

Since the corrector formula is more accurate, why did we bother to develop a predictor formula? The answer is that (4.67), as is evident from its expanded form (4.69), is an *implicit* or *closed* formula. That is, to employ it to compute $\boldsymbol{y}^{n+1}$, we need to know $\boldsymbol{f}^{n+1}$ which itself depends on $\boldsymbol{y}^{n+1}$! By contrast the predictor formula, although less accurate, is an *explicit* or *open* formula since we already presume to know the vectors $\boldsymbol{f}^n$ back through $\boldsymbol{f}^{n-N}$ from previous integration steps.

Finally, we note that the local orders described by (4.77) and (4.78) are close to the maximum order consistent with the first Dahlquist barrier. See Exercise 4.13.

Recapitulation of Adams' Method

At this point the reader should return to the first part of this section to review once again the procedure for Adams' method. He or she will see that it exploits the explicit nature of the predictor and the higher accuracy of the corrector by the following ingenious strategy:

(a) (Step 3.) Suppose the vectors $\boldsymbol{y}^n$ and $\boldsymbol{f}^n, \boldsymbol{f}^{n-1}, \cdots \boldsymbol{f}^{n-N}$ are known. Use formula (4.70) to *predict* a preliminary value for $\boldsymbol{y}^{n+1}$.

(b) (Step 4.) Insert this $\boldsymbol{y}^{n+1}$ and t^{n+1} into $\boldsymbol{f}(\boldsymbol{y}, t)$ to *evaluate* $\boldsymbol{f}^{n+1}$.

(c) (Step 5.) With the $\boldsymbol{f}^{n+1}$ thus obtained, recompute or *correct* $\boldsymbol{y}^{n+1}$ using formula (4.69).

(d) (Step 6.) Return to (b) and repeat (b) and (c) until convergence is achieved. Generally (see the discussion at the beginning of this section), the sequence $PECEC$ should be sufficient. The net result is a value for $\boldsymbol{y}^{n+1}$ that is correct within a local error given roughly by (4.77).

(e) (Steps 7 and 8.) Update the table of $\boldsymbol{f}$'s, and go back to part (a) to compute $\boldsymbol{y}^{n+2}$, etc.

This strategy is often called the *predictor-corrector* method. Let us summarize what has been accomplished. Using (4.70) as a predictor and (4.69) as a corrector, we are able to compute $\boldsymbol{y}^{n+1}$ through order h^{N+1} by generally making two and at most three computations (evaluations) of $\boldsymbol{f}$ plus some simple additions. (That is, $PECEC$ or at worst $PECECEC$ should be sufficient. In practice it is common to use just $PECE$, and there are theoretical reasons to believe that it is best to end with an E operation.) All that is required is the storage of the previous $N+1$ values $\boldsymbol{f}^n \cdots \boldsymbol{f}^{n-N}$ and the value $\boldsymbol{y}^n$. By contrast, the Runge-Kutta method (3.2) involves three evaluations of $\boldsymbol{f}$, and is correct only through order h^3. Higher order Runge-Kutta schemes involve correspondingly more computations of $\boldsymbol{f}$. Since $\boldsymbol{f}$ is usually a complicated function of $\boldsymbol{y}$ and t, multiple computations of $\boldsymbol{f}$ are generally made at the expense of considerable machine time and round-off error. We conclude that finite-difference methods give much higher accuracy for much less work, and are generally to be preferred once a solution is underway. There is, however, a caveat that makes the matter not quite so simple. One might be tempted, with a high order finite-difference method, to increase the step size in order to gain speed. That is, one might hope to trade accuracy for speed. However, as described in Subsection 7.3, finite-difference methods can become unstable if the step size is too large.[16] See also Exercise 4.14. Consequently, Runge-Kutta may be preferable if only low accuracy is required, while finite-difference methods win if high accuracy is required.

[16] That is, if h is too large, some parasitic roots of the characteristic equation may lie outside the unit circle even when Adams is used to integrate equations of the form (4.13). And if these equations cannot be integrated well, there is doubt that more complicated nonlinear equations can be integrated well.

Table 2.4.2: The Adams' Corrector Coefficients $\tilde{a}_l^N$.

k N	2	3	4	5	6	7	8	9
0	$\frac{5}{12}$	$\frac{9}{24}$	$\frac{251}{720}$	$\frac{475}{1440}$	$\frac{19087}{60480}$	$\frac{36799}{120960}$	$\frac{1070017}{3628800}$	$\frac{2082753}{7257600}$
1	8	19	646	1427	65112	139849	4467094	9449717
2	−1	−5	−264	−798	−46461	−121797	−4604594	−11271304
3		1	106	482	37504	123133	5595358	16002320
4			−19	−173	−20211	−88547	−5033120	−17283646
5				27	6312	41499	3146338	13510082
6					−863	−11351	−1291214	−7394032
7						1375	312874	2687864
8							−33953	−583435
9								57281

The denominator of each of the coefficients of the first line is to be repeated for all the coefficients of the corresponding column.

Table 2.4.3: The Adams' Predictor Coefficients $\tilde{b}_k^{N}$.

k N	2	3	4	5	6	7	8	9
0	$\frac{23}{12}$	$\frac{55}{24}$	$\frac{1901}{720}$	$\frac{4277}{1440}$	$\frac{198721}{60480}$	$\frac{434241}{120960}$	$\frac{14097247}{3628800}$	$\frac{30277247}{7257600}$
1	−16	−59	−2774	−7923	−447288	−1152169	−43125206	−104995189
2	5	37	2616	9982	705549	2183877	95476786	265932680
3		−9	−1274	−7298	−688256	−2664477	−139855262	−454661776
4			251	2877	407139	2102243	137968480	538363838
5				−475	−134472	−1041723	−91172642	−444772162
6					19087	295767	38833486	252618224
7						−36799	−9664106	−94307320
8							1070017	20884811
9								−2082753

Again the denominator of each of the coefficients of the first line is to be repeated for all the coefficients of the corresponding column.

Exercises

2.4.1. Verify that (4.5) holds exactly for $j = 0, 1, 2, 3$ and fails to be exact for $j \geq 4$. Verify that (4.5) is the unique explicit two-step formula having third-order accuracy. Find the error when $j = 4$. Verify that (4.6) is exact for $j = 0, 1, 2, 3, 4$ and fails for $j \geq 5$.

2.4.2. Review Exercise 4.1. Verify that the accuracy of (4.5) and (4.6) exceeds the limit specified by the first Dahlquist barrier. But, consistent with this barrier, (4.34) illustrates that the method (4.5) is unstable. Verify that, in accord with the first Dahlquist barrier, the method (4.6) is also unstable.

2.4.3. Suppose that a multistep method has order $m_{\max} \geq 1$ so that, in particular, it is able to treat the case (4.7) and (4.8) exactly when $j = 1$. Show that then there is the relation

$$\rho'(1) = \sigma(1). \tag{2.4.79}$$

2.4.4. Verify (4.16), (4.18), (4.22), and the expansions (4.23) and (4.24).

2.4.5. Verify the relations (4.44) through (4.47).

2.4.6. Verify (4.67) and (4.68) by direct calculation using Table 4.1 in the case that $y(t) = t^3$. How large must N be in order to get exact results?

2.4.7. Compute the first few coefficients a_k in equation (4.65). [Hint: First try differentiating F to convince yourself that this is not a good method. Then try synthetic division using equation (4.54). Can you find any other good method?]

Show that the coefficients b_k satisfy the recursion relation

$$b_k - b_{k-1} = a_k, \tag{2.4.80}$$

and use this relation to compute the first few b's.

2.4.8. Show that the coefficients $\overset{\sim N}{a_k}$ obey the relations

$$\overset{\sim N}{a_N} = (-1)^N a_N, \tag{2.4.81}$$

$$\overset{\sim N}{a_k} = 0 \text{ if } N < k, \tag{2.4.82}$$

$$\overset{\sim N+1}{a_k} = \overset{\sim N}{a_k} + (-1)^k \binom{N+1}{k} a_{N+1}. \tag{2.4.83}$$

Compute the first few $\overset{\sim N}{a_k}$. Make a similar study of the $\tilde{b}$'s.

2.4.9. Use equation (4.77) to estimate the expected local truncation error for Example (4.1) and compare with the actual error. [Use the solution (2.10) and (2.11) to compute $(d^{N+2}\boldsymbol{y}/dt^{N+2})$.] Use both (4.37) and (4.38) to derive a formula for the corrector error that does not require a knowledge of $(d^{N+2}\boldsymbol{y}/dt^{N+2})$. Apply it to Example (4.1). [Ans: $\boldsymbol{y}_{\rm true} - \boldsymbol{y}_{\rm corr} \simeq a_{N+1}(\boldsymbol{y}_{\rm pred} - \boldsymbol{y}_{\rm corr})/(a_{N+1} - b_{N+1})$. This strategy is called the *Milne device*.]

2.4.10. Consider the differential equation set (2.7) through (2.9). You will see below a table of entries obtained from a very accurate Runge-Kutta starting routine. Using Adams, complete the table for $n = 4$. The step size is $h = 1/3$. Compare your answer with the exact result. How big do you expect your error to be?

n	t^n	y_1^n	$y_2^n = f_1^n$	f_2^n
0	0	0	1	0
1	1/3	.33947	1.05505	.32719
2	2/3	.71496	1.21412	.61837
3	1	1.15853	1.45970	.84147
4	4/3	?	?	?

2.4.11. Show that if one is integrating *linear* differential equations, then the corrector formula (4.69) can be made explicit so that it is in principle possible to integrate without a predictor. Whether or not one should actually do this is a matter of convenience and economy.

2.4.12. The use of the predictor-corrector method requires at least two evaluations of $\boldsymbol{f}$ at each step. Explore the merits of integrating with a step size of $h/2$ and using just the predictor without correcting. That is, use just PE at each step. Both methods require the same number of $\boldsymbol{f}$ evaluations to integrate over a given time interval. Which is more accurate? Answer the question first ignoring round-off error, and then taking it into account. Do not worry about stability.

2.4.13. Show that with the stored data $\boldsymbol{f}^n, \dots, \boldsymbol{f}^{n-N}$ one can set up the corrector formula

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h \sum_0^{N+1} a_k \nabla^k \boldsymbol{f}^{n+1}. \tag{2.4.84}$$

Verify that (4.84) has the expanded form

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h \sum_0^{N+1} \tilde{a}{}_k^{N+1} \, \boldsymbol{f}^{n+1-k}. \tag{2.4.85}$$

Show that the error associated with these formulas is given by the relation

$$\boldsymbol{y}_{\text{true}}^{n+1} - \boldsymbol{y}_{\text{corr}}^{n+1} \approx h^{N+3} a_{N+2} (d^{N+3}\boldsymbol{y}/dt^{N+3})|_{t=t^n}. \tag{2.4.86}$$

Comparison of (4.86) and (4.77) shows that use of (4.84), or equivalently (4.85), yields results of one order higher accuracy; and therefore we will refer to this corrector as a *higher-order corrector*.

What accuracy can be achieved if we use the corrector (4.85) in conjunction with the predictor (4.70)? Both make optimal use of the the stored data $\boldsymbol{f}^n, \dots, \boldsymbol{f}^{n-N}$. Whether or not the smaller error associated with the higher-order corrector is achieved in practice depends on the number of corrector iterations. It can be shown that PEC is not enough, but $PECEC$ may suffice. If ending on an E step is deemed desirable, then one should use at least $PECECE$.

Your next task is to compare the accuracies specified by (4.77), (4.78), and (4.86) with that specified by the first Dahlquist barrier (4.35) through (4.37). Verify that, according to (4.78), the Adams predictor (4.70) makes local errors of order h^{N+2} and therefore is exact through order h^{N+1}. We also recall that $k = N + 1$ so that the Adams predictor (4.70) is exact through order h^k. According to (4.35) the highest local error $m_{\max}$ that can be achieved by a strongly stable explicit k-step method is k. Therefore, the Adams predictor (4.70) achieves the first Dahlquist barrier limit. With regard to implicit formulas, verify that (4.77) shows that the corrector formula (4.69) is exact through order h^k. But, according to (4.36), (4.37), and the ensuing discussion, it should be possible, in the implicit case, to achieve results that are accurate through order h^{k+1}. Verify that, according to (4.86), the higher-order corrector (4.85) is exact through order h^{k+1}. Therefore in this case the first

Dahlquist barrier limit is also achieved assuming that the higher-order Adams corrector is employed.

Your last task is to consider two low-order cases. Show that use of (4.67), (4.68), and (4.84) for $N = 0$ gives the formulas

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\boldsymbol{f}^n, \quad (\textit{predictor}) \tag{2.4.87}$$

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\boldsymbol{f}^{n+1}, \quad (\textit{corrector}) \tag{2.4.88}$$

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + (h/2)(\boldsymbol{f}^{n+1} + \boldsymbol{f}^n), \quad (\textit{higher-order corrector}). \tag{2.4.89}$$

Note that (4.87) is just the Euler method (2.3). The procedure (4.88) is sometimes called *backward* Euler, and in this context (4.87) is called *forward* Euler.

Show that use of (4.67), (4.68), and (4.84) for $N = 1$ gives the formulas

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + (h/2)(3\boldsymbol{f}^n - \boldsymbol{f}^{n-1}), \quad (\textit{predictor}) \tag{2.4.90}$$

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + (h/2)(\boldsymbol{f}^{n+1} + \boldsymbol{f}^n), \quad (\textit{corrector}) \tag{2.4.91}$$

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + (h/12)(5\boldsymbol{f}^{n+1} + 8\boldsymbol{f}^n - \boldsymbol{f}^{n-1}), \quad (\textit{higher-order corrector}). \tag{2.4.92}$$

2.4.14. The purpose of this exercise is to study the stability properties of the $N = 1$ Adams routine given by (4.90) through (4.92).

Let us begin with the predictor (4.90). Show that applying it to the differential equation (4.13) produces the characteristic equation

$$\zeta^2 - [1 + (3/2)(h\lambda)]\zeta + (1/2)(h\lambda) = 0, \tag{2.4.93}$$

and verify that the characteristic equation has the roots

$$\zeta = \{[1 + (3/2)(h\lambda)] \pm \sqrt{[1 + (h\lambda) + (9/4)(h\lambda)^2]}\}/2. \tag{2.4.94}$$

Show that (4.93) has the good root

$$\begin{aligned} \zeta_1 &= \{[1 + (3/2)(h\lambda)] + \sqrt{[1 + (h\lambda) + (9/4)(h\lambda)^2]}\}/2 \\ &= 1 + h\lambda + (h\lambda)^2/2! - (1/4)(h\lambda)^3 + \cdots \\ &= \exp(h\lambda) + O(h^3) \end{aligned} \tag{2.4.95}$$

and the parasitic root

$$\begin{aligned} \zeta_2 &= \{[1 + (3/2)(h\lambda)] - \sqrt{[1 + (h\lambda) + (9/4)(h\lambda)^2]}\}/2 \\ &= (h\lambda)/2 - (h\lambda)^2/2 + O(h)^3. \end{aligned} \tag{2.4.96}$$

Note that the good root goes to 1 as h goes to 0, as required; and the parasitic root goes to 0 as h goes to 0, as expected for Adams' method. Verify that the argument of the square root appearing in (4.95) and (4.96) is always positive when the quantity $h\lambda$ is real, and therefore there is no ambiguity involved in the definition of the square root. Show that $\zeta_2 < 1/2$ when $h\lambda$ is real. Verify that ζ_2 leaves the unit disk through $\zeta_2 = -1$ when $h\lambda = -1$, and becomes ever more negative than -1 as $h\lambda$ becomes ever more negative than -1. Verify that $\zeta_1 = 1/2$ when $h\lambda = -1$.

Next consider the corrector (4.91). Show that applying it to the differential equation (4.13) produces the characteristic equation

$$\zeta^2 - \{[1 + (h\lambda)/2]/[1 - (h\lambda)/2]\}\zeta = 0. \tag{2.4.97}$$

Show that (4.97) has the good root

$$\begin{aligned}\zeta_1 &= [1 + (h\lambda)/2]/[1 - (h\lambda)/2] \\ &= 1 + h\lambda + (h\lambda)^2/2! + (1/4)(h\lambda)^3 + \cdots \\ &= \exp(h\lambda) + O(h^3)\end{aligned} \tag{2.4.98}$$

and the parasitic root

$$\zeta_2 = 0. \tag{2.4.99}$$

Evidently, in this case, the parasitic root for the corrector has the optimum value of zero for all values of h! To examine this matter further, show that applying (4.91) to the differential equation (4.13) produces the recursion relation

$$y^{n+1} = \zeta_1 y^n, \tag{2.4.100}$$

and show that this recursion relation has the unique solution

$$y^n = (\zeta_1)^n y^0. \tag{2.4.101}$$

Finally, consider the higher-order corrector (4.92). Show that applying it to the differential equation (4.13) produces the characteristic equation

$$[1 - (5/12)(h\lambda)]\zeta^2 - [1 + (2/3)(h\lambda)]\zeta + (h\lambda)/12 = 0. \tag{2.4.102}$$

Verify that (4.102) has the roots

$$\zeta = \{[1 + (2/3)(h\lambda)] \pm \sqrt{[1 + h\lambda + (7/12)(h\lambda)^2]}\}/\{2[1 - (5/12)(h\lambda)]\}. \tag{2.4.103}$$

Show that (4.102) has the good root

$$\begin{aligned}\zeta_1 &= \{[1 + (2/3)(h\lambda)] + \sqrt{[1 + h\lambda + (7/12)(h\lambda)^2]}\}/\{2[1 - (5/12)(h\lambda)]\}. \\ &= 1 + h\lambda + (h\lambda)^2/2! + (h\lambda)^3/3! + (1/12)(h\lambda)^4 + \cdots \\ &= \exp(h\lambda) + O(h^4)\end{aligned} \tag{2.4.104}$$

and the parasitic root

$$\begin{aligned}\zeta_2 &= \{[1 + (2/3)(h\lambda)] - \sqrt{[1 + h\lambda + (7/12)(h\lambda)^2]}\}/\{2[1 - (5/12)(h\lambda)]\}. \\ &= (h\lambda)/12 - (7/144)(h\lambda)^2 + O(h^3).\end{aligned} \tag{2.4.105}$$

Verify that the argument of the square root appearing in (4.104) and (4.105) is always positive when the quantity $h\lambda$ is real, and therefore there is no ambiguity involved in the definition of the square root. Verify that ζ_2 remains within the unit disk when $h\lambda > -6$, has the value $\zeta_2 = -1$ when $h\lambda = -6$, and becomes ever more negative than -1 (leaves the unit disk through -1) as $h\lambda$ becomes ever more negative than -6. Verify that $\zeta_1 = 1/7$ when $h\lambda = -6$.

2.4.15. This exercise extends that work on Finite Difference Calculus presented at the beginning of this subsection to derive what is called the *Euler-Maclaurin* formula. This formula is of use both in evaluating integrals and in summing series.

Suppose $f(t)$ is some function and we wish to calculate the integral I given by

$$I = \int_{t_0}^{t_N} f(t)dt. \tag{2.4.106}$$

If we subdivide the interval $[t_0, t_N]$ into N equal pieces as in Figure 1.1, then we may approximate the integral (4.106) by the areas of N trapezoids. each of base h, to obtain the *trapezoidal rule* result

$$\begin{aligned} I &\approx h[(1/2)(f_0 + f_1) + (1/2)(f_1 + f_2) + \cdots + (1/2)(f_{N-1} + f_N)] \\ &= h(1/2)(f_0 + f_N) + h\sum_{i=1}^{N-1} f_i = -h(1/2)(f_0 + f_N) + h\sum_{i=0}^{N} f_i. \\ . \end{aligned} \tag{2.4.107}$$

(Here we have employed subscript indices rather than the superscript indices used earlier in Subsection 4.4 because we will soon need superscript indices for another purpose.) In summary, the trapezoidal rule yields the approximation

$$\int_{t_0}^{t_N} f(t)dt \approx -h(1/2)(f_0 + f_N) + h\sum_{i=0}^{N} f_i. \tag{2.4.108}$$

We will now further develop the calculus of finite differences and employ it to improve on the accuracy of this approximation.

To begin define, in analogy to (4.4.1), a *forward difference* operator Δ by the rule

$$\Delta f(t) = f(t + h) - f(t). \tag{2.4.109}$$

Next define an operator J by the rule

$$Jf(t) = \int_{t}^{t+h} f(t')dt'. \tag{2.4.110}$$

[Here the symbol J is not to be confused with the J introduced in (1.7.11). At the expense of duplication of symbol use, we are trying to follow convention in both the subjects of Hamiltonian theory and finite difference calculus.] Show that there are the relations

$$DJf(t) = JDf(t) = f(t + h) - f(t) = \Delta f(t). \tag{2.4.111}$$

Consequently, there are the operator relations

$$DJ = JD = \Delta. \tag{2.4.112}$$

Also show, using reasoning similar to that which led to the result (4.52), that there is the operator relation

$$\Delta = \exp(hD) - 1. \tag{2.4.113}$$

Verify that (4.112) and (4.113) can be combined to produce the operator relation

$$h = \Delta^{-1} hDJ = \{hD/[\exp(hD) - 1]\}J. \tag{2.4.114}$$

How can we make sense of the right side of (4.114)? It can be shown that there is the analytic function result

$$\tau/[\exp(\tau) - 1] = \sum_{j=0}^{\infty} (B_j/j!)\tau^j \tag{2.4.115}$$

where the B_k are the *Bernoulli numbers.*[17] These numbers have the property

$$B_3 = B_5 = B_7 = \cdots = 0, \tag{2.4.116}$$

and the first few nonzero of them have the values

$$B_0 = 1,\ B_1 = -1/2,\ B_2 = 1/6,\ B_4 = -1/36,\ B_6 = 1/42,\ B_8 = -1/30. \tag{2.4.117}$$

Consequently verify that, when acting on a function, (4.114) takes the forms

$$\begin{aligned} hf(t) &= \{hD/[\exp(hD) - 1]\}Jf(t) = \sum_{j=0}^{\infty} (B_j/j!)(hD)^j Jf(t) \\ &= \int_t^{t+h} f(t')dt' + \sum_{j=1}^{\infty} (B_j/j!)(hD)^j Jf(t) \\ &= \int_t^{t+h} f(t')dt' + \sum_{j=1}^{\infty} (B_j/j!)h^j D^{j-1} DJf(t) \\ &= \int_t^{t+h} f(t')dt' + \sum_{j=1}^{\infty} (B_j/j!)h^j D^{j-1} \Delta f(t) \\ &= \int_t^{t+h} f(t')dt' + \sum_{j=1}^{\infty} (B_j/j!)h^j D^{j-1} [f(t+h) - f(t)] \\ &= \int_t^{t+h} f(t')dt' + \sum_{j=1}^{\infty} (B_j/j!)h^j [f^{(j-1)}(t+h) - f^{(j-1)}(t)]. \end{aligned} \tag{2.4.118}$$

(Here a superscript index in parenthesis denotes a derivative of that order.) Note that if $f(t)$ is a polynomial in t, then the sum on the far right side of (4.118) terminates. Therefore (4.118) is well defined and exact for any polynomial f because no convergence questions arise.

Let us further manipulate (4.118). Suppose t in (4.118) is replaced by $t + h$. Verify that so doing yields the result

$$hf(t+h) = \int_{t+h}^{t+2h} f(t')dt' + \sum_{j=1}^{\infty} (B_j/j!)h^j [f^{(j-1)}(t+2h) - f^{(j-1)}(t+h)]. \tag{2.4.119}$$

[17] Indeed, (4.115) *defines* the Bernoulli numbers.

Next add (4.119) to (4.118). Verify that so doing yields the result

$$h[f(t)+f(t+h)]=\int_t^{t+2h} f(t')dt'+\sum_{j=1}^{\infty}(B_j/j!)h^j[f^{(j-1)}(t+2h)-f^{(j-1)}(t)]. \qquad (2.4.120)$$

Verify that this process can be generalized to yield the result

$$h[f(t)+f(t+h)+\cdots+f(t+Nh-h)]=h\sum_{i=0}^{N-1} f(t+ih)$$
$$=\int_t^{t+Nh} f(t')dt'+\sum_{j=1}^{\infty}(B_j/j!)h^j[f^{(j-1)}(t+Nh)-f^{(j-1)}(t)].$$
$$(2.4.121)$$

To manipulate still further, verify that

$$(B_1/1!)h[f^{(0)}(t+Nh)-f^{(0)}(t)]=-(1/2)h[f^{(0)}(t+Nh)-f^{(0)}(t)]$$
$$=-(1/2)h[f(t+Nh)-f(t)]=-hf(t+Nh)+h(1/2)[f(t)+f(t+Nh)].$$
$$(2.4.122)$$

Verify that employing (4.122) in (4.121) yields the result

$$\int_t^{t+Nh} f(t')dt'$$
$$=-h(1/2)[f(t)+f(t+Nh)]+h\sum_{i=0}^{N} f(t+ih)$$
$$-\sum_{j=2}^{\infty}(B_j/j!)h^j[f^{(j-1)}(t+Nh)-f^{(j-1)}(t)].$$
$$(2.4.123)$$

Finally set $t=t_0$ and make use of (4.116) to convert (4.123) to the relation

$$\int_{t_0}^{t_N} f(t)dt=$$
$$-h(1/2)(f_0+f_N)+h\sum_{i=0}^{N} f_i-\sum_{k=1}^{\infty}[B_{2k}/(2k)!]h^{2k}[f_N^{(2k-1)}-f_0^{(2k-1)}].$$
$$(2.4.124)$$

It is also useful to express this result using the *initial* and *final* notation

$$t_{\text{in}}=t_0,\ t_{\text{fin}}=t_N,\ f_{\text{in}}=f_0,\ f_{\text{fin}}=f_N, \qquad (2.4.125)$$

so that (4.124) becomes

$$\int_{t_{\rm in}}^{t_{\rm fin}} f(t)dt =$$

$$-h(1/2)(f_{\rm in}+f_{\rm fin})+h\sum_{i=0}^{N} f_i-\sum_{k=1}^{\infty}[B_{2k}/(2k)!]h^{2k}[f_{\rm fin}^{(2k-1)}-f_{\rm in}^{(2k-1)}]. \tag{2.4.126}$$

We see that the accuracy of the trapezoidal rule(4.108) can be improved if one knows, in addition to the $(N+1)$ sampling-point values f_i, the end-point derivative values $f_{\rm fin}^{(2k-1)}$ and $f_{\rm in}^{(2k-1)}$.

Let us also introduce the notation

$$b_{2k}=[B_{2k}/(2k)!][f_{\rm fin}^{(2k-1)}-f_{\rm in}^{(2k-1)}]=[B_{2k}/(2k)!][f^{(2k-1)}(t)|_{t_{\rm in}}^{t_{\rm fin}}] \tag{2.4.127}$$

so that

$$\sum_{k=1}^{\infty}[B_{2k}/(2k)!]h^{2k}[f_{\rm fin}^{(2k-1)}-f_{\rm in}^{(2k-1)}]=\sum_{k=1}^{\infty} b_{2k}h^{2k} \tag{2.4.128}$$

and (4.126) can be written in the form

$$\int_{t_{\rm in}}^{t_{\rm fin}} f(t)dt=-h(1/2)(f_{\rm in}+f_{\rm fin})+h\sum_{i=0}^{N} f_i-\sum_{k=1}^{\infty} b_{2k}h^{2k}. \tag{2.4.129}$$

In this form it is manifest that the correction to the trapezoidal rule is a Taylor series in h. This series will converge within some disc about $h=0$ if the coefficients b_{2k} are sufficiently well behaved. But the convergence radius could also be zero if the coefficients b_{2k} are not sufficiently well behaved.

We can also infer from the previous discussion that (4.129) is exact when f is a polynomial in t because the sum over k then terminates. More precisely, as is evident from (4.127), in this case all the $b_{2k}=0$ once k is sufficiently large. If f is not polynomial and the sum over k is terminated when $k=m$, then may write

$$\int_{t_{\rm in}}^{t_{\rm fin}} f(t)dt=-h(1/2)(f_{\rm in}+f_{\rm fin})+h\sum_{i=0}^{N} f_i-\sum_{k=1}^{m} b_{2k}h^{2k}-E_m \tag{2.4.130}$$

where E_m is a *remainder/error* term. Note that E_m is well defined because all the other terms in (4.130) are well defined (assuming h and N are finite). Define a function $\hat{E}_m(\tau)$ by the rule

$$\begin{aligned}\hat{E}_m(\tau) &= (Nh)h^{2m+2}\{B_{2m+2}/[(2m+2)!]\}f^{(2m+2)}(\tau)\\ &= (t_{\rm fin}-t_{\rm in})h^{2m+2}\{B_{2m+2}/[(2m+2)!]\}f^{(2m+2)}(\tau).\end{aligned} \tag{2.4.131}$$

It can be shown that

$$E_m=\hat{E}_m(\tau) \tag{2.4.132}$$

for some $\tau \in (t_{\rm in}, t_{\rm fin})$. Verify that if

$$\max_{\tau \in [t_{\rm in}, t_{\rm fin}]} |\hat{E}_m(\tau)| \to 0 \text{ as } m \to \infty, \tag{2.4.133}$$

then the series over k will converge and (4.129) is well defined and exact. By contrast, verify that if

$$\min_{\tau \in [t_{\rm in}, t_{\rm fin}]} |\hat{E}_m(\tau)|$$

does not approach zero as $m \to \infty$, then the series over k is divergent.

Frequently, when f is not polynomial, the relation (4.129) has an *asymptotic* character: The E_m do not tend to zero as $m \to \infty$, but rather there is an optimum value of m for which $|E_m|$ takes on a minimum (but generally nonzero) value. For m values smaller or larger than this optimum value the remainder/error is larger.

In closing this part of the discussion we note that if f is *periodic* and analytic, and integration is to be done over a full period, then, for any k,

$$f_{\rm fin}^{(2k-1)} - f_{\rm in}^{(2k-1)} = 0 \tag{2.4.134}$$

from which it follows that $b_{2k} = 0$ for all k. In this case the only correction to the trapezoidal rule is the remainder/error term E_m. We then conclude that for this case the error associated with the trapezoidal rule vanishes as $h \to 0$ and $N \to \infty$ faster than any power of h. For further discussion of the application of the trapezoidal rule to the integration of analytic periodic functions see the paragraph on "Performing the Forward $\phi \to m$ Fourier Transform" right after (19.1.27) and Exercises 19.1.2 through 19.1.4. See also Section 19.2.4 and Exercises 19.2.2 through 19.2.4. Finally, see the references to "Angular Integrals" at the end of Chapter 19.

On some occasions it is useful to rewrite (4.124) in the form

$$\sum_{i=0}^{N} f_i =$$
$$(1/h)\int_{t_0}^{t_N} f(t)dt + (1/2)(f_0 + f_N) + \sum_{k=1}^{\infty} [B_{2k}/(2k)!] h^{2k-1} [f_N^{(2k-1)} - f_0^{(2k-1)}]. \tag{2.4.135}$$

If the integral and sum on the right side of (4.135) can be evaluated, then the sum on the left side has been computed. Even if this cannot be accomplished, verify that (4.124) can be rewritten in the form

$$\sum_{i=0}^{N} f_i =$$
$$(1/h)\int_{t_0}^{t_N} f(t)dt + (1/2)(f_0 + f_N) + \sum_{k=1}^{m} [B_{2k}/(2k)!] h^{2k-1} [f_N^{(2k-1)} - f_0^{(2k-1)}] + E_m/h. \tag{2.4.136}$$

This relation can be used to compute the sum on the left within an error that can be estimated using (4.132).

If we set $N = \infty$ and keep h finite, and also assume that

$$f_\infty = 0 \text{ and all } f_\infty^{(2k-1)} = 0, \tag{2.4.137}$$

show that then (4.136) becomes

$$\sum_{i=0}^{\infty} f_i =$$

$$(1/h)\int_{t_0}^{\infty} f(t)dt + (1/2)f_0 - \sum_{k=1}^{m}[B_{2k}/(2k)!]h^{2k-1}f_0^{(2k-1)} + E_m/h. \tag{2.4.138}$$

In this case (4.132) cannot be used to estimate E_m. However, there are other more complicated estimates that can be used, and their use provides a value for the infinite sum on the left of (4.138) within a computable error estimate.

2.5 (Automatic) Choice and Change of Step Size and Order

In our initial discussion concerning the choice of step size h, we were a bit cavalier. We merely stated that h should be small compared to the characteristic time scale of the physical system under study. This statement is somewhat vague since the time scale may be different for different parts of the trajectory. Consider, for example, the orbit of a comet about the sun. When it is far away from the sun, it nearly moves in a straight line. This part of the trajectory could be integrated with a large h. By contrast, the trajectory changes rapidly near the sun and a small time step is required for this part of the orbit.

Ideally, two things are needed: a method for automatically estimating the local truncation error at each integration step and a procedure for adjusting the step size or the order of the integration routine (or both) to keep the error within acceptable bounds. These ideals require some effort to realize in practice. Methods that accomplish at least one of these ideals are called *adaptive*.

2.5.1 Adaptive Change of Step Size in Runge-Kutta

In the case of Runge-Kutta, one can carry out a step using a step size h, and also carry out two steps using a step size $h/2$. By comparing the results of these two procedures, it is possible to estimate the local error, and then adjust the step size accordingly. See Exercise 5.1. Alternatively, as first discovered by *Fehlberg*, there are some pairs of Runge-Kutta procedures whose orders differ (usually by one) and that, in making one integration step, share many or all intermediate evaluation points. For these so-called *embedded* pairs, one can carry out both procedures simultaneously with little added expense. Then, by subtracting the higher-order result from the lower-order result, one can estimate the *local* error in the lower-order result, and adjust the step size accordingly. See Appendix B for further detail.

We have seen that it is possible to adjust the Runge-Kutta step size automatically during the course of an integration run. In principle, with more complicated procedures, it is also possible to change the order as well. This is not now done in common practice, but is a subject of current research.

In summary, there are Runge-Kutta routines for which one specifies the initial and final times (t^0 and t^0+T), the initial conditions $\boldsymbol{y}(t^0)$, and the acceptable local error. The routine then automatically selects and dynamically adjusts the step size to compute $\boldsymbol{y}(t^0+T)$ with a minimal number of integration steps and with a global error that can be estimated from the allowed local error and the number of integration steps.

2.5.2 Adaptive Finite-Difference Methods

In the case of finite-difference methods it is possible, with some effort, to adjust both the step size and the order. We will now describe how this can be done.

Change of Order

In the Adams' method we have been discussing, it is easy to raise or lower the order. Suppose we are at $t = t^n$, and wish to step to t^{n+1}. We have at our disposal $\boldsymbol{y}^n$ and the $N+1$ $\boldsymbol{f}$ values $\boldsymbol{f}^n \cdots \boldsymbol{f}^{n-N}$. To lower the order by one, throw away the stored $\boldsymbol{f}^{n-N}$, and continue the integration using the N values $\boldsymbol{f}^n \cdots \boldsymbol{f}^{n-N+1}$ with one order *lower* predictor and corrector formulas. Suppose we are at $t = t^{n+1}$ and have just completed a converged corrector step. Then the $N+2$ $\boldsymbol{f}$ values $\boldsymbol{f}^{n+1}$, $\boldsymbol{f}^n$, $\cdots$ $\boldsymbol{f}^{n-N}$ are momentarily available. To raise the order by one, keep $\boldsymbol{f}^{n-N}$ rather than discarding it, as would normally be done. Then, after relabeling the $\boldsymbol{f}$'s, we have available the $N+2$ $\boldsymbol{f}$ values $\boldsymbol{f}^n \cdots \boldsymbol{f}^{n-N-1}$, and can make all future integration steps using one order *higher* formulas.

Change of Time Step

Changing the time step is more difficult. The simplest procedure is to stop the finite-difference routine. Then a Runge-Kutta routine with a different step size is begun using the previously obtained point as an initial condition. After a few starting values have been computed, one again returns to a finite-difference method. This finite-difference method would have the modified step size, and could also have a different order. Thus, a typical integration run could consist of several finite-difference segments of various step sizes and orders joined together by short pieces of Runge-Kutta.

Is there a more sophisticated way to change the time step? There is, but it is complicated. Given the $\boldsymbol{f}^n \cdots \boldsymbol{f}^{n-N}$ at times $t^n \cdots t^{n-N}$ separated by h, it is in principle possible by interpolation to find an equivalent set of $\boldsymbol{f}'$ values $\boldsymbol{f}'^n \cdots \boldsymbol{f}'^{n-N}$ at times $t'^n \cdots t'^{n-N}$ separated by h' in such a way that the current times t^n and t'^n agree. The interpolated $\boldsymbol{f}'$ values can then be used to make Adams' steps with a step size h'.

2.5.3 Jet Formulation

Is there a reformulation of the Adams' method that would facilitate changes in the time step and, at the same time, still make it easy to change orders? There is, but its description requires some explanation and discussion. In so doing, we will also learn about *jets* and classify all finite-difference/multistep methods.

As described earlier, Adams' method is a special case of multistep/multivalue methods where some combination of both previous $\boldsymbol{f}$ values and previous $\boldsymbol{y}$ values are stored. How much information about a trajectory is contained in these stored values? Take $\boldsymbol{y}^n$ as given. Suppose there are M previously stored values (counting both $\boldsymbol{y}$ and $\boldsymbol{f}$ values). Then from this information, by suitable Taylor expansions, we might hope to compute $\dot{\boldsymbol{y}}^n$, $\ddot{\boldsymbol{y}}^n$, $\cdots$ $\boldsymbol{y}^{(M)n}$ where $\boldsymbol{y}^{(m)n}$ denotes an approximation to the m'th derivative of $\boldsymbol{y}$ evaluated at t^n. Arrange these quantities in an $M+1$ dimensional vector $\vec{\boldsymbol{\jmath}}^n$ in the form

$$\vec{\boldsymbol{\jmath}}^n = \begin{pmatrix} \boldsymbol{y}^n \\ h\dot{\boldsymbol{y}}^n \\ (h^2/2)\ddot{\boldsymbol{y}}^n \\ \vdots \\ (h^M/M!)\boldsymbol{y}^{(M)n} \end{pmatrix}. \tag{2.5.1}$$

If we wish, we can ensure that the $\dot{\boldsymbol{y}}^n$ entry in $\vec{\boldsymbol{\jmath}}^n$ is exact by using (1.1) to compute $\dot{\boldsymbol{y}}(t^n)$. The remaining derivatives will be approximate. In keeping with terminology to be employed in subsequent chapters, we will refer to $\vec{\boldsymbol{\jmath}}$ as a *jet*. More precisely we will refer to $\vec{\boldsymbol{\jmath}}^n$ as given by (5.1) as an M-jet.

Conversion of Adams' Data into Jet Data

As an example of the procedure just described, let us convert stored Adams' data into jet data.[18] Consider the case $N = 2$. Then at t^n we have the stored values $\boldsymbol{f}^{n-1}$ and $\boldsymbol{f}^{n-2}$. If we imagine that these values are exact, we may make the Taylor expansions

$$\begin{aligned} h\boldsymbol{f}^{n-1} &= h\boldsymbol{f}(t^n - h) = h\dot{\boldsymbol{y}}(t^n - h) \\ &= h\dot{\boldsymbol{y}}(t^n) - h^2\ddot{\boldsymbol{y}}(t^n) + (h^3/2)\,\dddot{\boldsymbol{y}}\,(t^n) + \cdots \\ &= h\dot{\boldsymbol{y}}^n - 2(h^2/2)\ddot{\boldsymbol{y}}^n + 3(h^3/6)\,\dddot{\boldsymbol{y}}^n + \cdots, \end{aligned} \tag{2.5.2}$$

$$\begin{aligned} h\boldsymbol{f}^{n-2} &= h\boldsymbol{f}(t^n - 2h) = h\dot{\boldsymbol{y}}(t^n - 2h) \\ &= h\dot{\boldsymbol{y}}(t^n) - 2h^2\ddot{\boldsymbol{y}}(t^n) + [(2h)^3/2]\,\dddot{\boldsymbol{y}}\,(t^n) + \cdots \\ &= h\dot{\boldsymbol{y}}^n - 4(h^2/2)\ddot{\boldsymbol{y}}^n + 12(h^3/6)\,\dddot{\boldsymbol{y}}^n + \cdots. \end{aligned} \tag{2.5.3}$$

Define a vector $\vec{\boldsymbol{s}}^n$ by writing

$$\vec{\boldsymbol{s}}^n = \begin{pmatrix} \boldsymbol{y}^n \\ h\boldsymbol{f}^n \\ h\boldsymbol{f}^{n-1} \\ h\boldsymbol{f}^{n-2} \end{pmatrix}. \tag{2.5.4}$$

[18] There is an alternate approach due to *Nordsieck* that essentially amounts to the same thing. Instead of storing Adams' data, one stores their finite differences.

We will refer to $\vec{\boldsymbol{s}}$ as *spread* data (Adams' in this case) since it refers to data at different times. Corresponding to (5.4) we expect to have a jet $\vec{\boldsymbol{\jmath}}^n$ of the form

$$\vec{\boldsymbol{\jmath}}^n = \begin{pmatrix} \boldsymbol{y}^n \\ h\dot{\boldsymbol{y}}^n \\ (h^2/2)\ddot{\boldsymbol{y}}^n \\ (h^3/6)\,\dddot{\boldsymbol{y}}^n \end{pmatrix}. \tag{2.5.5}$$

Indeed, upon neglecting higher order terms, the relations (5.2) and (5.3) along with (1.1) can be written in the form

$$\vec{\boldsymbol{s}}^n = R\vec{\boldsymbol{\jmath}}^n, \tag{2.5.6}$$

where R is the matrix

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & -2 & 3 \\ 0 & 1 & -4 & 12 \end{pmatrix}. \tag{2.5.7}$$

The matrix R has the inverse

$$R^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 3/4 & -1 & 1/4 \\ 0 & 1/6 & -1/3 & 1/6 \end{pmatrix}, \tag{2.5.8}$$

and therefore we may also write

$$\vec{\boldsymbol{\jmath}}^n = R^{-1}\vec{\boldsymbol{s}}^n. \tag{2.5.9}$$

Jet Version of Adams' Predictor Formula

The Adams' predictor formula for $N = 2$ is

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + (h/12)(23\boldsymbol{f}^n - 16\boldsymbol{f}^{n-1} + 5\boldsymbol{f}^{n-2}). \tag{2.5.10}$$

See (4.70) and Table 2.4.3. Let us also find a formula for $\boldsymbol{f}^{n+1}$ based only on Taylor expansions. We have the relation

$$\begin{aligned} h\boldsymbol{f}^{n+1} &= h\boldsymbol{f}(t^n + h) = h\dot{\boldsymbol{y}}(t^n + h) \\ &= h\dot{\boldsymbol{y}}(t^n) + h^2\ddot{\boldsymbol{y}}(t^n) + (h^3/2)\,\dddot{\boldsymbol{y}}\,(t^n) + \cdots \\ &= h\dot{\boldsymbol{y}}^n + 2(h^2/2) + \ddot{\boldsymbol{y}}^n + 3(h^3/6)\,\dddot{\boldsymbol{y}}^n + \cdots . \end{aligned} \tag{2.5.11}$$

The quantities on the right side of (5.11) are components of $\vec{\boldsymbol{\jmath}}^n$. Use (5.9) to re-express them in terms of components of $\vec{\boldsymbol{s}}^n$. Doing so gives the result

$$h\boldsymbol{f}^{n+1} = 3h\boldsymbol{f}^n - 3h\boldsymbol{f}^{n-1} + h\boldsymbol{f}^{n-2}. \tag{2.5.12}$$

According to (5.4), $\vec{\boldsymbol{s}}^{n+1}$ has the components

$$\vec{\boldsymbol{s}}^{n+1} = \begin{pmatrix} \boldsymbol{y}^{n+1} \\ h\boldsymbol{f}^{n+1} \\ h\boldsymbol{f}^n \\ h\boldsymbol{f}^{n-1} \end{pmatrix}. \tag{2.5.13}$$

We see from (5.10), (5.12), and (5.13) that the relation between $\vec{\boldsymbol{s}}^n$ and $\vec{\boldsymbol{s}}^{n+1}$ can be written in the form

$$\vec{\boldsymbol{s}}^{n+1} = A^{(2)}\vec{\boldsymbol{s}}^n \tag{2.5.14}$$

where $A^{(2)}$, the $N = 2$ *Adams' matrix*, is defined by the relation

$$A^{(2)} = \begin{pmatrix} 1 & \frac{23}{12} & -\frac{16}{12} & \frac{5}{12} \\ 0 & 3 & -3 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{2.5.15}$$

What does the Adams' predictor step (5.14) correspond to in terms of jets? Using (5.6) and (5.9) we may write (5.14) in the equivalent form

$$R^{-1}\vec{\boldsymbol{s}}^{n+1} = R^{-1}A^{(2)}\vec{\boldsymbol{s}}^n = R^{-1}A^{(2)}RR^{-1}\boldsymbol{s}^n, \tag{2.5.16}$$

or

$$\vec{\boldsymbol{\jmath}}^{n+1} = T\vec{\boldsymbol{\jmath}}^n, \tag{2.5.17}$$

where T is the matrix

$$T = R^{-1}A^{(2)}R. \tag{2.5.18}$$

From (5.7), (5.8), and (5.15) we find for T the explicit result

$$T = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.5.19}$$

Jet Version of Adams' Predictor Formula Is Simply Taylor's Theorem

Suppose we simply compute $\vec{\boldsymbol{\jmath}}^{n+1}$ from a Taylor series. For $\boldsymbol{y}^{n+1}$ we have the result

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}(t^n + h) = \boldsymbol{y}(t^n) + h\dot{\boldsymbol{y}}(t^n) + (h^2/2)\ddot{\boldsymbol{y}}(t^n) + (h^3/6)\,\dddot{\boldsymbol{y}}\,(t^n) + \cdots . \tag{2.5.20}$$

Also, from (5.11) we have the expansions

$$h\dot{\boldsymbol{y}}^{n+1} = h\dot{\boldsymbol{y}}^n + 2(h^2/2)\ddot{\boldsymbol{y}}^n + 3(h^3/6)\,\dddot{\boldsymbol{y}}^n + \cdots . \tag{2.5.21}$$

Similarly, we have the expansions

$$\begin{aligned} (h^2/2)\ddot{\boldsymbol{y}}^{n+1} &= (h^2/2)\ddot{\boldsymbol{y}}(t^n + h) = (h^2/2)\ddot{\boldsymbol{y}}(t^n) + (h^3/2)\,\dddot{\boldsymbol{y}}\,(t^n) + \cdots \\ &= (h^2/2)\ddot{\boldsymbol{y}}^n + 3(h^3/6)\,\dddot{\boldsymbol{y}}^n + \cdots , \end{aligned} \tag{2.5.22}$$

$$(h^3/6)\,\dddot{\boldsymbol{y}}^{n+1} = (h^3/6)\,\dddot{\boldsymbol{y}}^n + \cdots . \tag{2.5.23}$$

Upon comparing the coefficients in (5.17), and (5.19) through (5.23), we see that the jet relation (5.17) is simply Taylor's theorem. For this reason we will refer to T as the *Taylor* matrix. We note that the entries in T are simply related to the binomial coefficients by the formula

$$T_{k\ell} = \binom{\ell}{k}, \tag{2.5.24}$$

with the understanding that

$$\binom{\ell}{k} = 0 \text{ when } k > \ell. \tag{2.5.25}$$

[Here, for convenience, the matrix elements in T are labeled starting from 0. That is, the elements (from left to right) in the first row of T are T_{00}, T_{01}, T_{02}, T_{03}, etc.] See Exercise 5.8. Indeed, the upper triangular portion of T is just *Pascal's triangle* turned on its side.[19]

Effect of Evaluation on a Jet

So far we have seen how a jet changes under the operation of simple *prediction* P, and have found that the result (5.17) is just Taylor's theorem in disguise. Suppose we now add the *evaluation* operation E as well since it is the operation PE that is required for integration using only the predictor. See Exercise 4.12. What effect does PE have on a jet?

As before, let us first see what the E operation does to the spread vector $\vec{\boldsymbol{s}}^{n+1}$. The E operation requires that we replace the $h\boldsymbol{f}^{n+1}$ entry in the spread vector (5.13) with $h\boldsymbol{f}(\boldsymbol{y}^{n+1}, t^{n+1})$. All other entries are unchanged. For simplicity of notation let us introduce the definition

$$h\tilde{\boldsymbol{f}}^{n+1} = 3h\boldsymbol{f}^n - 3h\boldsymbol{f}^{n-1} + h\boldsymbol{f}^{n-2}. \tag{2.5.26}$$

See (5.12). Also, define a quantity $\boldsymbol{\Delta}$ by the rule

$$\boldsymbol{\Delta}(\vec{\boldsymbol{s}}^{n+1}, t^{n+1}) = h\boldsymbol{f}(\boldsymbol{y}^{n+1}, t^{n+1}) - h\tilde{\boldsymbol{f}}^{n+1}. \tag{2.5.27}$$

[Note that the vector $\boldsymbol{\Delta}$ defined by (5.27) is not to be confused with the forward-difference operator Δ employed in (4.109).] Here it is understood that the $\boldsymbol{y}^{n+1}$ in (5.27) is given by the predictor formula (5.10), and consequently also by the first component of $\vec{\boldsymbol{s}}^{n+1}$ in the relation (5.14). Note also that $h\tilde{\boldsymbol{f}}^{n+1}$ is the second component of $\vec{\boldsymbol{s}}^{n+1}$. With these definitions, we see that under the full PE operation the vector $\vec{\boldsymbol{s}}^n$ is sent to the vector $\vec{\boldsymbol{s}}^{n+1}$ according to the rule

$$\vec{\boldsymbol{s}}^{n+1} = A^{(2)}\vec{\boldsymbol{s}}^n + \vec{\boldsymbol{e}}, \tag{2.5.28}$$

where $\vec{\boldsymbol{e}}$ is the vector

$$\vec{\boldsymbol{e}} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \boldsymbol{\Delta}. \tag{2.5.29}$$

We observe that the *evaluation* vector $\vec{\boldsymbol{e}}$, as desired, changes the second entry in (5.13) and leaves all the other entries unchanged.

[19] Google *Pascal's triangle.*

Now that we know from (5.28) the effect of PE on a spread vector, we are ready to find the equivalent effect of the operation PE on a jet vector. Using (5.6) and (5.9) as before, we find that (5.28) takes the form

$$R^{-1}\vec{\boldsymbol{s}}^{n+1} = R^{-1}A^{(2)}RR^{-1}\vec{\boldsymbol{s}}^{n} + R^{-1}\vec{\boldsymbol{e}}, \tag{2.5.30}$$

and consequently we have the relation

$$\vec{\boldsymbol{\jmath}}^{n+1} = T\vec{\boldsymbol{\jmath}}^{n} + \vec{\boldsymbol{r}}, \tag{2.5.31}$$

where $\vec{\boldsymbol{r}}$ is given by

$$\vec{\boldsymbol{r}} = R^{-1}\vec{\boldsymbol{e}}. \tag{2.5.32}$$

If we use (5.7) and (5.29), we find that $\vec{\boldsymbol{r}}$ has the explicit form

$$\vec{\boldsymbol{r}} = \begin{pmatrix} 0 \\ 1 \\ 3/4 \\ 1/6 \end{pmatrix} \boldsymbol{\Delta}(\vec{\boldsymbol{\jmath}}^{n+1}, t^{n+1}). \tag{2.5.33}$$

Note that in terms of the jet $\vec{\boldsymbol{\jmath}}^{n+1}$, $\boldsymbol{\Delta}$ as given by (5.27) takes the form

$$\boldsymbol{\Delta}(\boldsymbol{\jmath}^{n+1}, t^{n+1}) = h\boldsymbol{f}(\boldsymbol{y}^{n+1}, t^{n+1}) - h\tilde{\boldsymbol{f}}^{n+1}, \tag{2.5.34}$$

where $\boldsymbol{y}^{n+1}$ and $h\boldsymbol{f}^{n+1}$ are given by the relations

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^{n} + \dot{\boldsymbol{y}}^{n} + (h^3/2)\ddot{\boldsymbol{y}}^{n} + (h^2/6)\,\dddot{\boldsymbol{y}}^{n}, \tag{2.5.35}$$

$$h\tilde{\boldsymbol{f}}^{n+1} = h\dot{\boldsymbol{y}}^{n} + 2(h^2/2)\ddot{\boldsymbol{y}}^{n} + 3(h^3/6)\,\dddot{\boldsymbol{y}}^{n}\ . \tag{2.5.36}$$

These relations follow from (5.6) and (5.10), and from (5.6) and (5.26), respectively. Note also that (5.35) and (5.36) are just the first two components of the predicted $\vec{\boldsymbol{\jmath}}^{n+1}$ given by (5.17) and (5.19).

Effect of Corrector on a Jet

We have found the effect of Adams' prediction P and evaluation E on both spread vectors $\vec{\boldsymbol{s}}$ and jets $\vec{\boldsymbol{\jmath}}$. What about the corrector operation C? The $N = 2$ corrector formula is

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^{n} + (h/12)(5\boldsymbol{f}^{n+1} + 8\boldsymbol{f}^{n} - \boldsymbol{f}^{n-1}). \tag{2.5.37}$$

See Table 4.2. As before, use (5.12) for $\boldsymbol{f}^{n+1}$. Doing so gives the result that the "$\boldsymbol{f}$" factor on the right side of (5.37) can be rewritten in the form

$$5h\boldsymbol{f}^{n+1} + 8h\boldsymbol{f}^{n} - h\boldsymbol{f}^{n-1} = 23h\boldsymbol{f}^{n} - 16h\boldsymbol{f}^{n-1} + 5h\boldsymbol{f}^{n-2}. \tag{2.5.38}$$

In view of (5.10), (5.37), and (5.38), we see that the spread vector relation (5.14) still holds with the same matrix $A^{(2)}$ given by (5.15). But now we have to take into account successive E and C operations. Their effect is to replace the $h\boldsymbol{f}^{n+1}$ in (5.37) and in the second component of the spread vector (5.13) by $\boldsymbol{f}(\boldsymbol{y}^{n+1}, t^{n+1})$, with $\boldsymbol{y}^{n+1}$ defined by (5.37). Recall that we had used (5.12) for $\boldsymbol{f}^{n+1}$. Define $\boldsymbol{\Delta}$ as before in (5.27) but with the understanding that $\boldsymbol{y}^{n+1}$ is now defined by (5.37). Then we see from (5.37) and (5.13) that the first component of $\boldsymbol{s}^{n+1}$ is altered by $(5/12)\boldsymbol{\Delta}$ and the second component, as before, is altered by $\boldsymbol{\Delta}$. Thus, when the converged correction operation is taken into account, (5.14) is modified to take the form

$$\vec{\boldsymbol{s}}^{n+1} = A^{(2)}\boldsymbol{s}^n + \vec{\boldsymbol{c}}, \tag{2.5.39}$$

where the *correction* vector $\vec{\boldsymbol{c}}$ is given by

$$\vec{\boldsymbol{c}} = \begin{pmatrix} 5/12 \\ 1 \\ 0 \\ 0 \end{pmatrix} \boldsymbol{\Delta}. \tag{2.5.40}$$

We are now ready to determine the effect of correction on jets. As before, we multiply (5.39) by R^{-1} to get the results

$$R^{-1}\vec{\boldsymbol{s}}^{n+1} = R^{-1}A^{(2)}RR^{-1}\vec{\boldsymbol{s}}^n + R^{-1}\vec{\boldsymbol{c}}, \tag{2.5.41}$$

or

$$\vec{\boldsymbol{\jmath}}^{n+1} = T\vec{\boldsymbol{\jmath}}^n + \vec{\boldsymbol{r}} \tag{2.5.42}$$

where the vector $\vec{\boldsymbol{r}}$ is now defined by the relation

$$\vec{\boldsymbol{r}} = R^{-1}\vec{\boldsymbol{c}}. \tag{2.5.43}$$

By use of (5.8) and (5.40) we find the explicit result

$$\vec{\boldsymbol{r}} = \begin{pmatrix} 5/12 \\ 1 \\ 3/4 \\ 1/6 \end{pmatrix} \boldsymbol{\Delta}. \tag{2.5.44}$$

So that there is no possible source of confusion, let us try to be perfectly clear about what is meant by $\boldsymbol{\Delta}$ in (5.44). With the use of (5.6) we have the relation

$$(h/12)(8\boldsymbol{f}^n - \boldsymbol{f}^{n-1}) = (1/12)[7h\dot{\boldsymbol{y}}^n + 2(h^2/2)\ddot{\boldsymbol{y}}^n - 3(h^3/6)\,\dddot{\boldsymbol{y}}^n]. \tag{2.5.45}$$

Also, we use (5.36). Then, from (5.34) we have the result

$$\boldsymbol{\Delta} = h\boldsymbol{f}(\boldsymbol{y}^{n+1}, t^{n+1}) - h\dot{\boldsymbol{y}}^n - 2(h^2/2)\ddot{\boldsymbol{y}}^n - 3(h^3/6)\,\dddot{\boldsymbol{y}}^n, \tag{2.5.46}$$

where, according to (5.37) and (5.45), $\boldsymbol{y}^{n+1}$ satisfies the equation

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + (h/12)\boldsymbol{f}(\boldsymbol{y}^{n+1}, t^{n+1}) + (1/12)[7h\dot{\boldsymbol{y}}^n + 2(h^2/2)\ddot{\boldsymbol{y}}^n - 3(h^3/6)\,\dddot{\boldsymbol{y}}^n]. \tag{2.5.47}$$

We note that (5.47) can be solved by iteration just as was done before in Section 2.4. We begin by putting the predicted value (5.35) into the right side of (5.47), and then iterate. If the operations $PECE$ were deemed adequate in the original spread vector variables, then the same holds true for the jet variables since (5.47) and (5.37) are actually the same equations. When the iterations have converged and the smoke has cleared, the first two components of $\vec{\boldsymbol{\jmath}}^{n+1}$ are given by $\boldsymbol{y}^{n+1}$ and $h\boldsymbol{f}(\boldsymbol{y}^{n+1}, t^{n+1})$, respectively. Now that $\boldsymbol{y}^{n+1}$ and $h\boldsymbol{f}(\boldsymbol{y}^{n+1}, t^{n+1})$ are known, $\boldsymbol{\Delta}$ and $\vec{\boldsymbol{r}}$ can be evaluated using (5.46) and (5.44). Finally, (5.42) can now also be used. By construction, it gives the same first two components of $\vec{\boldsymbol{\jmath}}^{n+1}$ as found before, namely $\boldsymbol{y}^{n+1}$ and $h\boldsymbol{f}(\boldsymbol{y}^{n+1}, t^{n+1})$. It also determines the remaining components of $\vec{\boldsymbol{\jmath}}^{n+1}$.

2.5.4 Virtues of Jet Formulation

Overview

What are the virtues of using a jet formulation? First, there is a conceptual or theoretical advantage. It can be shown that *all* multistep/multivalue methods can be brought to jet variable form, and when this is done one always obtains results of the form (5.42). As might be guessed, the matrix T is universal. The various multistep/multivalue methods differ only in the choice of $\vec{\boldsymbol{r}}$. Consequently, multistep/multivalue methods can be classified by their $\vec{\boldsymbol{r}}$ vectors. For example, the $N = 2$ Adams' predictor-evaluator method has the $\vec{\boldsymbol{r}}$ given by (5.33), and the $N = 2$ Adams' corrector method has the $\vec{\boldsymbol{r}}$ given by (5.44).

From a programming perspective, it is only necessary to build into a code the required $\vec{\boldsymbol{r}}$ vectors if it is to be able to run for various orders. (Note that if we program in the $\vec{\boldsymbol{r}}$ vectors for a variety of *methods*, the code can then also run for a variety of methods, and can even switch between methods!) Because of its simple form (5.24), the Taylor matrix can easily be computed and stored by the code itself as needed.

With regard to speed, the predictor part in the jet formulation consists in computing the first entry in $\vec{\boldsymbol{j}}^{n+1}$ as given by (5.17), which is just (5.20). This is no more difficult to compute than its spread-vector counterpart (4.70). As seen earlier, the evaluation and corrector operations required to compute the first two entries in $\vec{\boldsymbol{j}}^{n+1}$ are essentially the same in both the spread-vector and jet-vector formulations, and it is these operations that are the most time consuming. Finally, we need to compare the time required to compute the remaining entries in the spread vector $\vec{\boldsymbol{s}}^{n+1}$ with that required for the jet vector $\vec{\boldsymbol{j}}^{n+1}$. As is evident from (5.15), (5.28), and (5.29), all that is required in the spread-vector case is a simple relabelling (what we have called updating) of stored $\boldsymbol{f}$ values, which is very fast. In the jet-vector case inspection of (5.19), (5.42), and (5.44) shows that we must carry out some matrix-vector multiplies and some vector addition, which is a bit slower.

Change of Order

Changing the order for any method in the jet-vector formulation is also as easy as it was for the Adams' method in the spread-vector formulation. Suppose we wish to lower the order by one in the jet formulation. Simply delete the last component of the jet vector, and continue the integration by using the lower order version of (5.42) — which amounts to deleting the right-most column and bottom row from T and selecting the $\vec{\boldsymbol{r}}$ appropriate to the lower order. Raising the order is not much more difficult. Suppose, before the order is to be raised, that $\vec{\boldsymbol{j}}$ is an M-jet: the last entry in $\vec{\boldsymbol{j}}$ is $(h^M/M!)\boldsymbol{y}^{(M)}$. See (5.1). Store this entry for two or more successive steps and form the difference. Observe that we have the relation

$$\begin{aligned}&[1/(M+1)](h^M/M!)[\boldsymbol{y}^{(M)}(t^n) - \boldsymbol{y}^{(M)}(t^n - h)] = \\ &[1/(M+1)](h^M/M!)[h\boldsymbol{y}^{(M+1)}(t^n) - (h^2/2)\boldsymbol{y}^{(M+2)}(t^n) + \cdots] \simeq \\ &[h^{M+1}/(M+1)!]\boldsymbol{y}^{(M+1)n}.\end{aligned} \tag{2.5.48}$$

[If desired, one can use more accurate formulas involving higher order differences and based on the relations (4.53) and (4.54). For an example of the use of these relations, see Exercise

5.3.] Equation (5.48) gives a value for $[h^{M+1}/(M+1)!]\boldsymbol{y}^{(M+1)n}$ which can be appended to the end of $\vec{\boldsymbol{\jmath}}^n$ to convert it into an $(M{+}1)$-jet. We can now continue the integration by using the one order higher version of (5.42) — which amounts to enlarging the T matrix using (5.24) and selecting the $\vec{\boldsymbol{r}}$ appropriate to the higher order.

Change of Step Size

The *main* virtue of the jet formulation is that it *easy* to change the step size. Observe that the step size appears nowhere in the marching orders (5.42) except for a simple dependency in $\boldsymbol{\Delta}$ as given by (5.46) or (5.34). All the major h dependence occurs in the definition of $\vec{\boldsymbol{\jmath}}$ as given by (5.1). Suppose we wish to change the step size from h to h'. Form the diagonal *scaling* matrix S defined by

$$S = \begin{pmatrix} 1 & & & \\ & (h'/h) & & \\ & & (h'/h)^2 & \\ & & & \ddots \end{pmatrix}. \tag{2.5.49}$$

Given the jet vector $\vec{\boldsymbol{\jmath}}^n$ corresponding to step size h, form the corresponding jet vector $\vec{\boldsymbol{\jmath}}'^n$ corresponding to step size h' by the relation

$$\vec{\boldsymbol{\jmath}}'^n = S\vec{\boldsymbol{\jmath}}^n. \tag{2.5.50}$$

We are now ready to continue the integration using (5.42) with h' and the $\vec{\boldsymbol{\jmath}}'$ jet vectors.

Interpolation/Dense Output

There is yet another virtue to the jet formulation. As it runs, a numerical integration scheme only produces $\boldsymbol{y}$ values at discrete points t^n. It may happen (particularly if the step size is being controlled dynamically by the integration program) that we need to know $\boldsymbol{y}$ at some time τ that lies between two points, say t^m and t^{m+1}. This is easily done using the jet vector. Define a small quantity ϵ by the relation

$$\tau = t^m + \epsilon. \tag{2.5.51}$$

Also define an $(M+1)$ component vector $\vec{\delta}$ by the rule

$$\vec{\delta} = \begin{pmatrix} 1 \\ (\epsilon/h) \\ (\epsilon/h)^2 \\ \vdots \\ (\epsilon/h)^M \end{pmatrix}. \tag{2.5.52}$$

Then, by Taylor's theorem, we have the result

$$\begin{aligned} \boldsymbol{y}(\tau) &= \boldsymbol{y}(t^m+\epsilon) = \sum_k (\epsilon/h)^k (h^k/k!)\boldsymbol{y}^{(k)}(t^m) \\ &\simeq \vec{\delta}\cdot\vec{\boldsymbol{\jmath}}^m. \end{aligned} \tag{2.5.53}$$

Adaptive Error Control

We have seen how the use of jets makes it possible to change the order and the time step at will with a fairly modest overhead. The size of the truncation error can also be estimated. If the jet formulation is based on the Adams' method, as we have been describing in our examples, then the error estimates (4.77) and (4.78) still hold. Consequently, if the order of the predictor and the corrector are the same, the error can be estimated by comparing predictor and corrector results. See Exercise 4.9. If the corrector is one order higher than the predictor, see Exercise 4.13, then the predictor error can be estimated directly simply by subtracting the corrector result from the predictor result. Finally, the error can also be estimated directly from (4.77) or (4.78) by using finite-difference relations such as (5.48) to compute the required derivatives.

With an error estimate in hand, it is possible to construct a jet-based code that will automatically select and dynamically adjust both step size and order to achieve a solution within the allowed error and with a minimal number of integration steps. Like the Runge-Kutta codes described at the beginning of this section, all it requires in principle is a specification of initial and final times (t^0 and t^0+T), the initial condition $\boldsymbol{y}(t^0)$, and the acceptable error. A typical strategy is to have the program estimate from time to time the error currently being made at each step. If the error is too large, or if the error is too small (which means that too much effort is being spent in achieving unnecessary accuracy), the program computes what the step size should be for the error to be within the allowed bounds. This calculation is done both for the current order and for orders one higher and one lower. The program then shifts to the order that allows the largest step size, adjusts the step size to the largest value allowed, and continues to run for some time with this order and step size.

Self Starting

We observe that an integration routine having the features just described can be *self starting*. That is, unlike the finite-difference methods described in Section 2.4, such a program does not need a Runge-Kutta or other starting routine. Rather, it can begin with the $N = 0$ Adams' procedure (but in jet form) given by (4.85) and (4.86) or (4.87) since all this routine needs to start is the initial condition $\boldsymbol{y}(t^0)$. See Exercise 4.13. It can also automatically choose the step size to make sure that the accuracy of the first few steps is sufficiently high. Once the program is underway, it will then automatically adjust the order and step size to optimal values.

2.5.5 Advice to the Novice

As might be imagined, it is not a simple matter to write a variable order and variable step size program that will actually run in an optimal fashion for a wide variety of differential equations. Much time has been spent by professional mathematicians and numerical analysts in writing such programs. We have presented enough of the theory behind these programs to make them intelligible to readers and possible users; but they are advised not to try writing such programs on their own without exploring existing programs and without being prepared to expend considerable time and effort.

Exercises

2.5.1. The result of numerically integrating a differential equation from t^0 to t^0+T depends in general on the step size h. We express this fact by writing the result as $\boldsymbol{y}(t^0+T;h)$. Neglecting round-off error, we expect $\boldsymbol{y}(t^0+T;h)$ to approach the exact result as $h \to 0$. Consider an integration method that has a *cumulative* truncation error of order h^m. To be more precise, *assume* (what really requires proof and need not always be true) that we have

$$\boldsymbol{y}(t^0+T;h) = \boldsymbol{y}_{\rm e}(t^0+T) + \boldsymbol{c}h^m + O(h^{m+1}), \tag{2.5.54}$$

where the subscript "e" stands for "exact", and $\boldsymbol{c}$ is independent of h, but otherwise unknown. Show that $\boldsymbol{y}_{\rm e}$ can be approximated by the formula

$$\boldsymbol{y}_{\rm e}(t^0+T) = \boldsymbol{y}(t^0+T;h) + (1-2^{-m})^{-1}[\boldsymbol{y}(t^0+T;h/2) - \boldsymbol{y}(t^0+T;h)]. \tag{2.5.55}$$

Show that $\boldsymbol{c}$ can approximated by the formula

$$\boldsymbol{c}h^m = -(1-2^{-m})^{-1}[\boldsymbol{y}(t^0+T;h/2) - \boldsymbol{y}(t^0+T;h)]. \tag{2.5.56}$$

You see below a line of output for Example 3.1 run with a step size of $h = 1/20$.

```
  time      y1comp        y2comp
 1.5000  .20025125+01  .19292636+01
```

What should m be for RK3? Estimate $\boldsymbol{y}_{\rm e}(1.5)$ and compare with the exact result. Devise a procedure that could be used if one had results for three different step sizes. You are studying *Richardson* extrapolation.

2.5.2. Verify (5.6) through (5.9).

2.5.3. Equation (5.7) for R is a direct consequence of Taylor's theorem as used in (5.2) in (5.3). Equation (5.8) for R^{-1} was then found by inverting R. The entries in R^{-1} can also be found directly by requiring (5.9). For example, from (1.1) and (4.50) we have the result

$$\ddot{\boldsymbol{y}}^n = D\boldsymbol{f}^n. \tag{2.5.57}$$

Next use (4.53) and (4.54) to get the result

$$h\ddot{\boldsymbol{y}}^n = \sum_{k=1}^{\infty}(1/k)\nabla^k \boldsymbol{f}^n. \tag{2.5.58}$$

Discard terms in this series beyond $k=2$, and verify that doing so reproduces the third row in (5.8). Similarly, we may write $\dddot{\boldsymbol{y}}^n = D^2\boldsymbol{f}^n$. Use this result to reproduce the fourth row in (5.8).

2.5.4. Verify (5.12) using (5.9). See also Exercise 5.11.

2.5.5. Verify (5.14) and (5.15).

2.5.6. Verify (5.16) through (5.19).

2.5.7. Verify (5.20) through (5.23).

2.5.8. Verify (5.24) for the case (5.19). Let $g(t)$ be any (analytic) function. With D defined by (4.10), verify the formula

$$e^{hD}g(t) = g(t+h). \tag{2.5.59}$$

Verify the formal power series identity

$$\exp(hD)(h^i/i!)D^i = \sum_{j=0}^{\infty} \binom{j}{i} (h^j/j!)D^j = \sum_{j=0}^{\infty} T_{ij}(h^j/j!)D^j. \tag{2.5.60}$$

Apply both sides of (5.60) to $\boldsymbol{y}(t)$ to derive (5.24).

2.5.9. Verify (5.28) and (5.29).

2.5.10. Verify (5.30) through (5.36).

2.5.11. Study Exercise 4.13. Show that the higher corrector corresponding to the predictor (5.10) and the corrector (5.37) is given by the formula

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + (h/24)(9\boldsymbol{f}^{n+1} + 19\boldsymbol{f}^n - 5\boldsymbol{f}^{n-1} + \boldsymbol{f}^{n-2}). \tag{2.5.61}$$

See Table 2.2. Show that (5.12) can be derived by subtracting (5.10) from either (5.37) or (5.61). Show that (5.12) can also be dervied by subtracting (5.37) from (5.61). Let α, β, γ be any three constants satisfying $\alpha + \beta + \gamma = 0$ and $10\beta + 9\gamma \neq 0$. Form the linear combination of *equations* given by the suggestive expression

$$\alpha(5.10) + \beta(5.37) + \gamma(5.61),$$

and use the result to verify (5.12).

2.5.12. Verify (5.38).

2.5.13. Verify (5.39) through (5.47).

2.5.14. Suppose that (5.10) and (5.61) are used as a predictor-corrector pair. What will the local truncation error be in this case? Show that (5.39) and (5.40) hold in this case providing that $\vec{\boldsymbol{c}}$ is the vector

$$\vec{\boldsymbol{c}} = \begin{pmatrix} 3/8 \\ 1 \\ 0 \\ 0 \end{pmatrix} \boldsymbol{\Delta}. \tag{2.5.62}$$

Suppose this Adams' method is reformulated in terms of jets. Show that the associated vector $\vec{\boldsymbol{r}}$ for (5.42) in this case is

$$\vec{\boldsymbol{r}} = \begin{pmatrix} 3/8 \\ 1 \\ 3/4 \\ 1/6 \end{pmatrix} \boldsymbol{\Delta}. \tag{2.5.63}$$

2.5.15. Verify (5.53). Suppose we wish to estimate the entire jet $\vec{\boldsymbol{\jmath}}(\tau)$. Let $S(h', h)$ denote the matrix (5.49). Verify the result

$$\vec{\boldsymbol{\jmath}}(\tau) = S(h, \epsilon)TS(\epsilon, h)\vec{\boldsymbol{\jmath}}^m. \tag{2.5.64}$$

2.6 Extrapolation Methods

2.6.1 Overview

In the previous section we have learned how it is possible to construct multistep methods that adjust both the order and the step size dynamically. We also learned how some Runge-Kutta methods (which are single step) can be modified to include dynamic step size control. In this section we will describe a single-step method that adjusts both order and step size dynamically.

The problem we desire to solve is the same: given the differential equation (1.1) with the initial condition $\boldsymbol{y}(t^0)$ at time $t = t^0$ and some acceptable error, we wish to find the final condition $\boldsymbol{y}(t^0 + T)$ within that error. In the methods described so far we have sought to achieve this goal by a march composed of many small steps, which we will call *micro* steps, of typical size h. In the method to be described now we will try to achieve the same goal by making fewer but larger steps, which we will call *meso* steps, whose typical size will be denoted by the symbol H. The procedure for making each meso step will have the feature of being self starting in that no information will be needed about the previous step (save for its ultimate result!); the meso step procedure is therefore a single-step method. Since the meso step size H will be relatively large, we can anticipate expending considerable effort in making each such step. The first meso step will take us from $\boldsymbol{y}(t^0)$ at time t^0 to $\boldsymbol{y}(t^0 + H)$ at time $(t^0 + H)$. Subsequent meso steps, perhaps with different sizes, will take us to subsequent times.

2.6.2 Making a Meso Step

How is such a step to made? We will describe the first meso step. Subsequent meso steps are made in the same way.

Simple Micro Step Formula

As before, divide up the time axis over the interval $[t^0, t^0 + H]$ into M equal *micro* steps of duration h. Then we have the relations

$$t^m = t^0 + mh, \;\; h = H/M. \tag{2.6.1}$$

Refer back to Figure (1.1) with H now playing the role of T and intermediate times labelled as t^m. Next, in this interval, we will construct an apparently simple but actually quite subtle approximation to the values $\boldsymbol{y}^m$, the values of $\boldsymbol{y}(t)$ at the times t^m, that we will call $\boldsymbol{\eta}^m$. For $m = 0$ and $m = 1$ we will use the prescription

$$\boldsymbol{\eta}^0 = \boldsymbol{y}^0, \tag{2.6.2}$$

$$\boldsymbol{\eta}^1 = \boldsymbol{\eta}^0 + h\boldsymbol{f}(\boldsymbol{\eta}^0, t^0). \tag{2.6.3}$$

Comparison with (2.2) shows that (6.3) is simply an Euler step, and involves a local error of order h^2. With $\boldsymbol{\eta}^0$ and $\boldsymbol{\eta}^1$ in hand, we define successive $\boldsymbol{\eta}^m$ by the *midpoint rule*

$$\boldsymbol{\eta}^{m+1} = \boldsymbol{\eta}^{m-1} + 2h\boldsymbol{f}(\boldsymbol{\eta}^m, t^m). \tag{2.6.4}$$

It is easily verified that the procedure (6.4) makes local errors of order h^3. See Exercise 3.2. Continue the march (6.4) until $\boldsymbol{\eta}^{M-1}$ and $\boldsymbol{\eta}^M$ have been found. Finally, approximate $\boldsymbol{y}(t^0+H)$ using $\boldsymbol{\eta}^{M-1}$ and $\boldsymbol{\eta}^M$ by the formula

$$\boldsymbol{y}(t^0+H;M) = (1/2)[\boldsymbol{\eta}^M + \boldsymbol{\eta}^{M-1} + h\boldsymbol{f}(\boldsymbol{\eta}^M, t^M)]. \tag{2.6.5}$$

A Taylor series expansion of this last step again reveals a possible error of order h^2, just as for the first step. Here we have used the notation $\boldsymbol{y}(t^0+H;M)$ to indicate that (6.5) is an approximation to $\boldsymbol{y}(t^0+H)$ that naturally depends on the size h of the micro steps and therefore, assuming that H is held fixed, on the number of micro steps M.

What is the virtue of this process? Let us estimate the *global* error for the formula (6.5). As already described, the first and last steps involve errors of order h^2. Note that we may write

$$h^2 = (H/M)^2 = H^2(1/M)^2.$$

Here we have used (6.1). The intervening $(M-1)$ midpoint rule steps (6.4) each involve local errors of order h^3, and hence their cumulative effect should behave as

$$(M-1)h^3 \approx Mh^3 \approx Hh^2 \approx H^3(1/M)^2.$$

Thus the total global meso step error should behave as

$$\text{meso step error} \approx (1/M)^2. \tag{2.6.6}$$

Consequently, if $\boldsymbol{y}_e(t^0+H)$ denotes the "exact" solution, we might expect a relation, perhaps only asymptotic, of the form

$$\boldsymbol{y}(t^0+H;M) - \boldsymbol{y}_e(t^0+H) = \boldsymbol{c}_2(1/M)^2 + \boldsymbol{c}_3(1/M)^3 + \boldsymbol{c}_4(1/M)^4 + \cdots \tag{2.6.7}$$

where the coefficients $\boldsymbol{c}_2$, $\boldsymbol{c}_3$, $\boldsymbol{c}_4$ $\cdots$ are hoped to be independent of M. Here we reiterate that it is to be understood that H is held fixed, but h varies by changing M in (6.1).

Remarkably, it can be shown that the procedure given by (6.2) through (6.5) has the extraordinary property that the coefficients of the odd powers of $(1/M)$ in (6.7) all vanish![20] Thus, (6.7) actually has the form

$$\boldsymbol{y}(t^0+H;M) - \boldsymbol{y}_e(t^0+H) = \sum_{k=1}^{\infty} \boldsymbol{c}_{2k}(1/M)^{2k}. \tag{2.6.8}$$

To be honest, our discussion has been oversimplified. What is actually true is that there are asymptotic expansions of the form

$$\boldsymbol{y}(t^0+H;M) - \boldsymbol{y}_e(t^0+H) = \sum_{k=1}^{\infty} \boldsymbol{d}_{2k}(1/M)^{2k},\ M \text{ odd}; \tag{2.6.9}$$

[20] The choice of the integration procedure (6.2) through (6.5), called a *modified* midpoint rule because of the starting and ending steps (6.3) and (6.5), and the realization that this procedure would lead to the vanishing of all odd powers in (6.7), are due to *Gragg*.

$$\boldsymbol{y}(t^0 + H; M) - \boldsymbol{y}_e(t^0 + H) = \sum_{k=1}^{\infty} \boldsymbol{e}_{2k}(1/M)^{2k}, \; M \text{ even.} \tag{2.6.10}$$

That is, the nature of the expansion depends on whether M is odd or even. (In view of this discovery, the assumption made in Exercise 5.1 requires proof!) The proof of this result is beyond the scope of this text, as also appears to be the case for many books on numerical analysis. However, it is proved in the Extrapolation Methods references listed at the end of the chapter. See also Exercise 6.1, which treats the special case for which $\boldsymbol{f}(\boldsymbol{y}, t)$ is, in fact, not dependent on $\boldsymbol{y}$.

Extrapolation

The background has now been provided to present remarkable ideas associated variously with the names *Richardson, Gragg, Bulirsch*, and *Stoer*. According to either (6.9) or (6.10) we have the result

$$\lim_{M\to\infty} \boldsymbol{y}(t^0 + H; M) = \boldsymbol{y}_e(t^0 + H), \tag{2.6.11}$$

as is desired for any integration scheme. But now suppose we evaluate $\boldsymbol{y}(t^0 + H; M)$ for a finite number of M values (all odd or all even), and from these results try to *extrapolate* to a limiting result for $M = \infty$. This process is an example of what is called *Richardson extrapolation*. Bulirsch and Stoer originally proposed that the extrapolation be based on the sequence of (even) M values given by the list

$$M = 2, 4, 6, 8, 12, 16, 24, \cdots, \; (M_{j+2} = 2M_j \text{ when } j > 1). \tag{2.6.12}$$

Subsequent work by *Deuflhard* recommends using simply the even integers

$$M = 2, 4, 6, 8, \cdots, \; (M_j = 2j). \tag{2.6.13}$$

In some realizations of the procedure the first few integers near the beginning of either list are discarded at some stage, and the extrapolation is based on the remaining larger integers.

One possible extrapolation method is to assume a polynomial fit of the form

$$\boldsymbol{y}(t^0 + H; M) = \boldsymbol{e}_0 + \sum_{k=1}^{K} \boldsymbol{e}_{2k}(1/M)^{2k} \tag{2.6.14}$$

in the $(K + 1)$ unknowns $\boldsymbol{e}_0$, $\boldsymbol{e}_2$, $\boldsymbol{e}_4$, $\cdots$ $\boldsymbol{e}_{2K}$, which amounts to truncating the sum (6.10) at $k = K$. We then evaluate (6.14) for $(K + 1)$ different values of M (and hence h) selected from (6.12) or (6.13), and use the results to solve $\boldsymbol{e}_0$. Finally, we make the extrapolation

$$\boldsymbol{y}_e(t^0 + H) \simeq \boldsymbol{e}_0. \tag{2.6.15}$$

According to (6.10) the error involved in this extrapolation should be on the order of $\boldsymbol{e}_{(2K+2)}(1/M_{\min})^{(2K+2)}$ where $M_{\min}$ is the smallest M value used in the lists (6.12) or (6.13). Indeed, during the course of the extrapolation we have available (approximate) values for the coefficients $\boldsymbol{e}_2$, $\boldsymbol{e}_4$ $\cdots$ $\boldsymbol{e}_{2K}$, and from these we can form the quantities $\boldsymbol{e}_{2k}(1/M_{\min})^{2k}$ for $k = 1, 2, \cdots K$. These quantities should approach zero as k increases, and we can use the last few of them to estimate the error in (6.15). Alternatively (and preferably) we can solve (6.14) for $\boldsymbol{e}_0$ using successive values of K, beginning with $K = 1$, and observe how these values of $\boldsymbol{e}_0$ converge as K is allowed to increase.

The polynomial extrapolation method presupposes analyticity in h or, equivalently, analyticity in $(1/M)$. For differential equations whose right sides are analytic we expect, by Poincaré's theorem, that there will be analyticity along the real $(1/M)$ axis. However, there might be singularities somewhere off the real axis in the complex $(1/M)$ plane, and such singularities could affect the extrapolation process. Another extrapolation method, originally proposed by Bulirsch and Stoer, consists of using *rational function* or *Padé* approximation fits to $\boldsymbol{y}(t^0 + H; M)$ as a function of M rather than the fits of the form (6.14). Such a procedure should be more effective than polynomial extrapolation if there are pole singularities in the complex $(1/M)$ plane.[21] Describing the use of rational function approximation will require some additional notation. As in (1.4.4), let y_j denote the j^{th} component of $\boldsymbol{y}$. For the case of even M, as in either (6.12) or (6.13), we make fits of the form

$$\begin{aligned} y_j(t^0 + H; M) &= \frac{p_j^{(0)} + p_j^{(2)}(1/M)^2 + p_j^{(4)}(1/M)^4 + \cdots}{1 + q_j^{(2)}(1/M)^2 + q_2^{(4)}(1/M)^4 + \cdots} \\ &= \left[\sum_{k=0}^{L} p_j^{(2k)}(1/M)^{2k}\right] \Big/ \left[1 + \sum_{k=1}^{L} q_j^{(2k)}(1/M)^{2k}\right]. \qquad (2.6.16) \end{aligned}$$

These fits are called *diagonal* rational function approximations because the numerator and denominator have equal degree. For fixed L the relation (6.16) may be viewed as a fit in the $(2L+1)$ unknowns $p_j^{(0)}$, $p_j^{(2)}$, $p_j^{(4)}$, $\cdots$ $p_j^{(2K)}$, $q_j^{(2)}$, $q_j^{(4)}$, $\cdots$ $q_j^{(2K)}$. We next evaluate (6.16) for $(2L+1)$ different values of M, selected from (6.12) or (6.13), and solve for $p_j^{(0)}$. Finally, letting $\boldsymbol{p}^{(0)}$ denote a vector with components $p_j^{(0)}$, we make the extrapolation

$$\boldsymbol{y}_e(t^0 + H) \simeq \boldsymbol{p}^{(0)}. \qquad (2.6.17)$$

If the rational function (6.16) were to be expanded as a Taylor series in $(1/M)^2$, we would get an expression whose initial coefficients might be expected to agree with those of (6.14) thorugh $K = 2L$. Thus, we may expect the error in (6.17) to be of order $\boldsymbol{e}_{(4L+2)}(1/M_{\text{min}})^{(4L+2)}$. As before we can estimate the error in $\boldsymbol{p}^{(0)}$ directly by solving (6.16) for successive values of L, beginning with $L = 1$, and observing how these values of $\boldsymbol{p}^{(0)}$ converge as L is allowed to increase.

The calculation of $\boldsymbol{p}^{(0)}$ using (6.16) is obviously more work than the calculation of $\boldsymbol{e}_0$ using (6.14). However, the rational function (6.16) might be expected to be a somewhat better fit to $\boldsymbol{y}(t^0 + H; M)$ than the polynomial (6.14) for small values of M, and therefore the convergence of the $\boldsymbol{p}^{(0)}$ for successive L might be expected to be somewhat better than that of the $\boldsymbol{e}$ for corresponding K values. This has indeed been observed to be the case for a variety of differential equations. But is not yet clear whether the extra effort involved in rational function approximation is generally worth the improved convergence. Several authors find, for example applications they have examined, that it is not.[22]

[21] Padé was a student of Hermite.

[22] It is interesting to note that in the Kepler problem there are singularities in the complex t plane, but they are branch points.

Finally, we observe that in either case the convergence is remarkably *fast*. Thanks to the occurrence of only even powers of $(1/M)$ in (6.9) or (6.14), we gain an extra power of $(1/M)^2$, which is equivalent to *two* powers of h, for each unit increase in K. When (6.16) is used, we gain an extra power of $(1/M)^4$, which is equivalent to *four* powers of h, for each unit increase in L. Of course, for a given K there are $(K+1)$ values of M that must be used in (6.14) while for a given L value there are $(2L+1)$ values of M that must be used in (6.16). Thus, apart from more refined considerations concerning behavior at small M values, the convergence rates of both extrapolation methods are roughly the same.

2.6.3 Summary

Looking back over what has been described so far, we have seen that the *order* of truncation in the single meso step that takes us from t^0 to $(t^0 + H)$ can be adjusted by the choice of K or L. Also, we clearly have the choice of H at our disposal, and therefore we may also adjust the macro step size at will. Finally, we have built-in error estimates based on the observed convergence of the extrapolation procedure. Thus, we have all the ingredients for a method that can adjust both order and step size dynamically. Typically, one chooses a macro step size H, and then begins an extrapolation process. The K or L values involved are successively increased until convergence is achieved within the specified error bounds. The program has specified maximum values of K or L, call them $K_{\max}$ or $L_{\max}$, that are not allowed to be exceeded in this process in order to keep the process under control and in order to avoid excessive round-off error. If satisfactory convergence is not achieved within the allowed K or L values, the chosen meso step size is rejected, and the extrapolation process is tried again with a smaller step size. This process is repeated, if necessary, until convergence is finally achieved. When convergence is achieved, the results of this step are accepted and stored. Note is also made of the satisfactory meso step size H and the ease of convergence (the K or L values required to achieve the desired accuracy) of the extrapolation process. The size of the next meso step to be attempted is then selected based on this information, and the extrapolation process is begun anew.

2.6.4 Again, Advice to the Novice

As might be imagined (and just as for the case of the jet or multivalue methods described in the previous section), the procedure for implementing in detail the ideas of the previous paragraphs are quite involved. For this reason, potential users of extrapolation methods are advised to begin with existing programs written for this purpose; and then they should make modifications on these programs, if necessary, only when their algorithms and performance are well understood.

Exercises

2.6.1. The aim of this exercise is to examine some special cases for which the asymptotic expansion (6.10) can be verified explicitly. For this purpose we will need the following identities:

$$S(M,0)=\sum_{n=0}^{M} n^0=\sum_{n=0}^{M} 1=M+1, \tag{2.6.18}$$

$$S(M,1)=\sum_{n=0}^{M} n^1=M^2/2+M/2, \tag{2.6.19}$$

$$S(M,2)=\sum_{n=0}^{M} n^2=M^3/3+M^2/2+M/6, \tag{2.6.20}$$

$$S(M,3)=\sum_{n=0}^{M} n^3=M^4/4+M^3/2+M^2/4, \tag{2.6.21}$$

$$S(M,4)=\sum_{n=0}^{M} n^4=M^5/5+M^4/2+M^3/3-M/30. \tag{2.6.22}$$

These identities can easily be found by the method of undetermined coefficients. Show that there is the recursion relation

$$S(M+1,\ell)=S(M,\ell)+(M+1)^\ell \tag{2.6.23}$$

with the starting condition

$$S(0,\ell)=\delta_{0,\ell}. \tag{2.6.24}$$

Make, for example, the Ansatz

$$S(M,4)=AM^5+BM^4+CM^3+DM^2+EM \tag{2.6.25}$$

where the coefficients A through E are to be determined. Show that insertion of this Ansatz into the recursion relation (6.23) determines the coefficients A through E to yield the result (6.22).

In terms of the notation (1.4), the Gragg micro-step procedure (6.2) through (6.5) reads

$$h=H/M, \tag{2.6.26}$$

$$\boldsymbol{\eta}^0=\boldsymbol{y}^0, \tag{2.6.27}$$

$$\boldsymbol{\eta}^1=\boldsymbol{\eta}^0+h\boldsymbol{f}^0, \tag{2.6.28}$$

$$\boldsymbol{\eta}^{m+1}=\boldsymbol{\eta}^{m-1}+2h\boldsymbol{f}^m, \tag{2.6.29}$$

$$\boldsymbol{y}(t^0+H;M)=(1/2)[\boldsymbol{\eta}^M+\boldsymbol{\eta}^{M-1}+h\boldsymbol{f}^M]. \tag{2.6.30}$$

For M even, show that the net result of this procedure is the relation

$$\boldsymbol{y}(t^0+H;M)=\boldsymbol{y}^0-(h/2)(\boldsymbol{f}^0+\boldsymbol{f}^M)+h\sum_{n=0}^{M}\boldsymbol{f}^n. \tag{2.6.31}$$

Consider, for simplicity, the case where $\boldsymbol{f}$ is one dimensional and of the simple form

$$f(y,t)=Nt^{N-1}. \tag{2.6.32}$$

That is, f is independent of y and has only a monomial dependence on t. When f is of the form (6.32), show that the exact solution to $\dot{y} = f$ is

$$\begin{aligned} y_e(t^0 + H) &= y^0 \text{ when } N = 0, \\ &= y^0 + H^N \text{ when } N > 0. \end{aligned} \tag{2.6.33}$$

Let us examine the results of using (6.31) in the cases $N = 0$ and $N = 1$. When $N = 0$, use of (6.32) gives $f^n = 0$ so that (6.31) yields the numerical result

$$y(t^0 + H; M) = y^0, \tag{2.6.34}$$

which agrees with the exact result. When $N = 1$, verify that use of (6.32) gives $f^n = 1$ and that use of (6.31) yields the numerical result

$$y(t^0 + H; M) = y^0 + Mh = y^0 + H, \tag{2.6.35}$$

which again agrees with the exact result.

To examine the cases $N > 1$, Suppose further that

$$t^0 = 0. \tag{2.6.36}$$

Then we have the general result

$$f^n = N(nh)^{N-1} = N(H/M)^{N-1} n^{N-1} \tag{2.6.37}$$

with the particular results

$$hf^0 = 0 \tag{2.6.38}$$

and

$$hf^M = (H/M)N(H/M)^{N-1}M^{N-1} = NH^N/M. \tag{2.6.39}$$

Correspondingly, show that (6.31) then takes the form

$$\begin{aligned} y(t^0 + H; M) &= y^0 - (1/2)N(H^N/M) + (H/M)N(H/M)^{N-1} \sum_{n=0}^{M} n^{N-1} \\ &= y^0 - (1/2)N(H^N/M) + N(H/M)^N \sum_{n=0}^{M} n^{N-1} \\ &= y^0 + H^N[-(1/2)(N/M) + N/M^N \sum_{n=0}^{M} n^{N-1}]. \end{aligned} \tag{2.6.40}$$

Evaluate (6.40) for the case $N = 2$ to show that there is again the result

$$y(t^0 + H; M) = y_e(t^0 + H). \tag{2.6.41}$$

We have learned that the numerical solution is exact for all the cases $N = 0, 1, 2$.

Now consider the case $N = 3$. Verify that in this case

$$\begin{aligned} N/M^N \sum_{n=0}^{M} n^{N-1} &= (3/M^3)[M^3/3 + M^2/2 + M/6] \\ &= 1 + 3/(2M) + 1/(2M^2). \end{aligned} \tag{2.6.42}$$

Correspondingly, show that

$$-(1/2)(N/M) + N/M^N \sum_{n=0}^{M} n^{N-1} = -3/(2M) + 1 + 3/(2M) + 1/(2M^2) = 1 + 1/(2M^2). \tag{2.6.43}$$

Consequently, show that in this case (6.40) takes the form

$$y(t^0 + H; M) = y^0 + H^3 + H^3/(2M^2) = y_e(t^0 + H) + H^3/(2M^2). \tag{2.6.44}$$

Next, for the cases $N = 4$ and $N = 5$, show that (6.40) gives the results

$$y(t^0 + H; M) = y^0 + H^4 + H^4/M^2 = y_e(t^0 + H) + H^4/M^2, \tag{2.6.45}$$

and

$$\begin{aligned} y(t^0 + H; M) &= y^0 + H^5 + H^5[5/(3M^2) - 1/(6M^4)] \\ &= y_e(t^0 + H) + 5H^5/(3M^2) - H^5/(6M^4), \end{aligned} \tag{2.6.46}$$

respectively. Observe that (6.44) through (6.46) are of the claimed form (6.10).

Finally, show that the the assumption (6.36) can be dropped and that f can be any polynomial of degree 4 in t without changing the conclusions of this exercise: the result (6.10) still holds.

We close this exercise with an important remark. Observe that (6.31) can be rewritten in the form

$$\boldsymbol{y}(t^0 + H; M) = \boldsymbol{y}^0 + (h/2)(\boldsymbol{f}^0 + \boldsymbol{f}^1) + (h/2)(\boldsymbol{f}^1 + \boldsymbol{f}^2) + \cdots + (h/2)(\boldsymbol{f}^{M-1} + \boldsymbol{f}^M), \tag{2.6.47}$$

and that each term in parentheses on the right side (6.47) is the result of applying the *trapezoidal rule* over an interval of duration h. It is known, say from the *Euler-Maclaurin sum formula* (see Exercise 4.15), that the error associated with this *extended* trapezoidal rule has the properties (6.10) for any polynomial and, by extension, any analytic function. Thus, you have explicitly verified specific cases of a general result.

2.7 Things Not Covered

We have given the rudiments of numerical integration, and their mastery should provide sufficient knowledge to handle most problems. However, there are several additional topics whose study we commend to the reader who wishes to become truly expert. We list them below along with brief explanatory paragraphs. Further detail may be found in the books listed at the end of the chapter.

2.7.1 Størmer-Cowell and Nyström Methods

The differential equations of classical mechanics often contain only second derivatives with *no* first derivatives present. In this case it is possible to work directly with the second-order equations instead of converting them into a first-order set of twice the dimensionality. The result can be a saving in computer time and an increase in accuracy. Appendix A describes a predictor-corrector method due to *Størmer* and *Cowell* and modified Runge-Kutta methods due to *Nyström* that have this feature.

2.7.2 Other Starting Procedures

In this chapter we have always started Adams' solutions using Runge-Kutta. Other techniques, such as the use of Taylor series or various iterative processes, are also sometimes used. Of course, starting procedures are not required for the methods of Sections 2.5 and 2.6.

2.7.3 Stability

An introductory discussion of order, stability, and convergence was given at the beginning of Section 2.4. Much more can be found on the subject in some of the references listed at the end of this chapter. The finite-difference equations of Adams and Størmer-Cowell are special examples of the whole class of multistep/multivalue equations. To recapitulate, in general a multistep/multivalue equation (when applied to a linear differential equation) has several solutions, and only one of these solutions approximates the solution of the differential equation being integrated. It is important to be sure that the other so called *parasitic* solutions do not enter the calculation and eventually swamp the main solution. The reader should be warned that many of the numerical methods described in older books, such as *Milne's* and *Nyström's* multistep/multivalue methods, have parasitic solutions that grow exponentially. Thus, if a small amount of a parasitic solution happens to be introduced due to round-off errors or improper initial conditions, it will soon grow to the point where it completely dominates the main solution, and the accuracy of the numerical solution is completely destroyed. By contrast, the parasitic solutions in Adams and Størmer-Cowell are exponentially damped (if the step size is small enough) so that even if they happen to enter a calculation, their effect rapidly dies away. But there is a complication: the higher the order the smaller this step size must be to guarantee stability. For example, when integrating the simple harmonic oscillator with unit frequency ($x'' + x = 0$) using the adams10 method given in Appendix B, at least 50 steps per oscillation are required before stability is safely achieved and the error analysis of Section 2.4.2 becomes relevant. Finally, if a multistep/multivalue method cannot integrate a linear differential equation well, it is unlikely to be able to integrate more complicated nonlinear differential equations well.

2.7.4 Regularization, Etc.

We have already discussed in Sections 2.5 and 2.6 something about the choice of step size h and how it may be varied during the course of integration. An alternative procedure to making frequent changes in h is to analytically regularize the equations of motion before integration by the introduction of a new independent variable in place of the time. This can often be done while remaining within a Hamiltonian framework.[23] See Exercise 1.6.5 and the regularization references at the end of this chapter. It is known, for example, how to regularize the Kepler problem and the Størmer problem (the problem of finding the motion of a charged particle in the external field of a point magnetic dipole, of interest for Van Allen radiation). We also mention that in some cases it is worthwhile to change rather radically the form of the differential equation by introducing new dependent variables. For example, if a solution $y(t)$ is known to be highly oscillatory, one should try making an *eikonal* or *Madelung* transformation by writing $y(t) = a(t)\ \sin\ b(t)$ and then integrating the differential equations for a and b.

Differential equations whose solutions contain both rapidly and slowly varying terms are colorfully referred to by numerical analysts as being *stiff*. The presence of a rapidly varying part forces the integration time step to be small when the usual integration methods are used. But the features of physical interest may reside in the slowly varying part, and thus to explore these features one may be forced to integrate for many very small steps. In the case that a stiff equation cannot be regularized easily, it may be possible to use directly certain integration methods devised especially for stiff equations. These methods are beyond the scope of this text, but are described in some of the references.

2.7.5 Solutions with Few Derivatives

Our discussion has always assumed that $\boldsymbol{y}(t)$ has a large number of continuous derivatives. Although this is true for many problems, there are important examples where this is not the case. Consider a space ship outside the Earth's atmosphere. As long as it is subject only to gravitational forces, it can be shown that its trajectory vector $\boldsymbol{r}(t)$ has arbitrarily many continuous derivatives. However, suppose the space ship's rocket engine is fired at a time t_f. Then, according to Newton, $\ddot{\boldsymbol{r}}(t)$ is discontinuous at t_f. To handle this situation numerically, one possible procedure is to terminate any finite difference scheme slightly before t_f and integrate through t_f using Runge-Kutta. The Runge-Kutta routine should be used in such a way that an integration step is initiated at t_f and at any other time at which the rocket thrust either changes discontinuously or has discontinuous changes in its first few time derivatives.

2.7.6 Symplectic and Geometric/Structure-Preserving Integrators

We will see in Chapter 6 that Hamiltonian systems have special properties. Their transfer maps are symplectic. Symplectic integrators are integrators specifically constructed to preserve these properties. They produce maps that, while still approximations to the exact transfer map, are at least exactly symplectic. Symplectic integrators are an example

[23] It is also possible in some cases to arrange, by a suitable change of variables, that the final Hamiltonian will be of the form $T(p) + V(q)$. Sometimes one can also arrange that $T(p) = p \cdot p/2$. Hamiltonians of this form are desirable because there are special integration methods for them that are particularly efficient. See Chapter 12 and Appendix A.

of so-called *geometric* integrators. A second example is integration on *manifolds*. Many differential equations have the property that their solutions lie on manifolds, often manifolds associated with groups. In this case one seeks numerical integration methods that, despite truncation errors, still guarantee that the numerical solutions they generate also lie on the these manifolds. Extensive work has been done on both these aspects of geometric integration. See Chapters 11 and 12.

2.7.7 Error Analysis

In discussing Runge-Kutta errors, we have given only their expected order in h without any mention of the coefficient multiplying h^m. The analysis of the local truncation error committed in Runge-Kutta is considerably more complicated than in the case of predictor-corrector methods. Estimates, however, are available. Of course, one may also use the methods for error estimation described at the beginning of Section 2.5. A more complicated question with regard to both Runge-Kutta and finite difference methods is how an error propagates through successive time steps after its initial introduction. This question is particularly difficult with regard to round-off error, and is still a topic of study.

One way to reduce round-off error with only a small increase in machine time is to use *partial double precision*. In this method $\boldsymbol{f}$ is evaluated with the usual number of significant figures. However, in the Adams' routine for example, the addition (4.69) is carried out with additional significant figures and the $\boldsymbol{y}^n$ are stored with additional significant figures. (See Appendix B for an analogous treatment of the Runge-Kutta routine RK3.) Of course even a further reduction in round-off error is realized if all calculations are carried out in *double precision*, i.e. using twice the usual number of significant figures. But the use of full double precision may require considerably more computer time.

It is often difficult to ascertain rigorously the total error at the end of a long integration run. However, there are several informal procedures. In the case of fixed step size methods, one procedure is to make several runs with different values of h, and then study how the $\boldsymbol{y}$'s at the end of the trajectory depend upon h. The magnitude of the error should at first decrease with decreasing h, and then again increase due to round-off error. For variable step size methods, one can change the specified error to see what effect it has on the solution. Another procedure is to first integrate a trajectory forward in time, and then reintegrate it backward to see how close one comes to the original initial conditions.[24] In the case that the differential equations have known constants of motion such as energy or angular momentum, one can and always should check to see to what extent they are actually preserved by the numerical solution. Finally, the accuracy of an integration routine always should be checked on equations whose solutions are known exactly. These equations should include both those leading to oscillatory functions such as sines and cosines and those leading to growing and damped exponentials.

[24] To integrate backwards, simply replace h by $-h$. Truncation errors associated with forward integration followed by backward integration are not expected to cancel unless the integration method is *symmetric*. See Section 12.1. In any case, round-off errors are not expected to cancel.

2.7.8 Backward Error Analysis

Suppose x is some input, g is some function, and we wish to compute $g(x)$. If g is a complicated function, as is often the case, the best that we are able or willing to do is to compute some approximating function $\hat{g}$. What is then called the *forward* error associated with such a computation is the difference $[\hat{g}(x) - g(x)]$. Turn the situation around. We may ask if there is a modified input $\bar{x}$ near x such that $g(\bar{x})$ gives the result $\hat{g}(x)$. That is, there is the requirement that the exact calculation applied to the modified input should agree with the approximate calculation applied to the original input: $g(\bar{x}) = \hat{g}(x)$. We would then call the difference $[\bar{x} - x]$ the *backward* error.

Somewhat the same philosophy may be applied to the numerical integration of ordinary differential equations. Suppose, as in Section 2.1, we are given as input the vector $\boldsymbol{f}(\boldsymbol{y}, t)$ to be used as the right side of an ordinary differential equation, the vector $\boldsymbol{y}^0$ to be used as an initial condition, and the quantity h to be used as a step size. We wish to compute the vectors $\boldsymbol{y}^n$. What we actually accomplish, because of truncation error, is the computation of a set of approximate vectors $\hat{\boldsymbol{y}}^n$. (Here we assume that round-off error is negligible.) Instead of examining the forward error vectors $[\hat{\boldsymbol{y}}^n - \boldsymbol{y}^n]$, we might ask if there is a modified differential equation with right side $\bar{\boldsymbol{f}}(\boldsymbol{y}, t; h)$ (which will in general depend on h) and exact solution $\bar{\boldsymbol{y}}^n$ such that $\bar{\boldsymbol{y}}^n = \hat{\boldsymbol{y}}^n$. That is, the exact solution of the modified differential equation should agree with the approximate solution of the original differential equation at the times t^n. For some integration methods it can be shown that it is indeed possible to find such a modified differential equation. Moreover, in some cases there is the further possibility of modifying the original differential equation [its right side becomes $\tilde{\boldsymbol{f}}(\boldsymbol{y}, t; h)$] so that its approximate solution agrees with the exact solution of the original differential equation (well, not perfectly, but in principle to any desired finite order in h). That is, the original differential equation can be modified in such a way as to compensate (at least to any desired finite order in h) for the errors produced by the integration method.

These considerations are of particular interest for symplectic integrators applied to Hamiltonian differential equations. In that case it can be shown that the use of a symplectic integrator produces the exact solution of some modified Hamiltonian differential equation. Let $\mathcal{S}_{\text{exact}}$ be an integrator that solves any differential equation exactly. It could be viewed as the result of using any integrator in the limit that $h \to 0$, correspondingly the number of integration steps becomes indefinitely large, and all results are carried out with unlimited precision. Also, let $\mathcal{S}_{\text{approx}}$ be some integrator that is correct only through some order in h, but is exactly symplectic. Then, according to the discussion above, we have the result

$$\mathcal{S}_{\text{approx}}(H) \equiv \mathcal{S}_{\text{exact}}(H_{\text{mod}}). \tag{2.7.1}$$

Here the notation $\mathcal{S}_{\text{approx}}(H)$ denotes the trajectory that results from integrating the equations of motion associated with H using the integrator $\mathcal{S}_{\text{approx}}$, the notation $\mathcal{S}_{\text{exact}}(H_{\text{mod}})$ denotes the trajectory that results from integrating the equations of motion associated with H_{mod} using the integrator $\mathcal{S}_{\text{exact}}$, and the symbol $\equiv$ means *equivalent to.* If the modified Hamiltonian H_{mod} is deemed to be sufficiently close to the original Hamiltonian H in some sense (small backward error), then the results of symplectic integration might also be deemed to have some special merit.

Moreover, if $H(q, p, t)$ is the Hamiltonian of interest and some particular symplectic integrator is being used, one might try to find a modified Hamiltonian $\tilde{H}(q, p, t; h)$ such that symplectically integrating its equations of motion produces results that are closer to the exact results for the original Hamiltonian H. That is, if we can master relationships of the form (7.1), then we might be able to arrange the relation

$$\mathcal{S}_{\text{approx}}(\tilde{H}) \equiv \mathcal{S}_{\text{exact}}(H), \tag{2.7.2}$$

at least through terms of some high order in h.

2.7.9 Comparison of Methods

There is an extensive literature comparing the virtues of various integration methods and computer codes. The matter is complicated. The criteria for which method is "best" vary from problem to problem, and may also be machine dependent. Moreover, the manner in which a particular method is implemented also affects the over-all performance of a computer code. A typical discussion of such matters is given in the review article of *Shampine et al.* For relatively simple problems, those for which the characteristic time scale varies relatively little over a trajectory or those which can be regularized, the fixed step Adams' method started with Runge-Kutta is satisfactory, easy to program, and easy to use. More difficult problems may well benefit from the use of jet or extrapolation methods as described in Sections 2.5 and 2.6. In this case one is well advised to begin with professionally written programs. These programs should, however, be used with care and understanding. Some may produce unpleasant surprises. (For example, it is not uncommon that programs which automatically adjust step size have difficulties with some kinds of problems.) At present there is some indication and fairly widespread opinion that extrapolation methods (at least when high accuracy is required) may well be the method of choice for a wide variety of problems. Seek advice from a local Computer Center if it has resident experts. Much work has gone into writing good integration programs.

Bibliography

Books on General Numerical Analysis

[1] F.B. Hildebrand, *Introduction to Numerical Analysis.* (McGraw-Hill 2nd Edition, 1974) QA 297.H54. A standard reference book on numerical analysis, which has recently been reprinted by Dover.

[2] J. Todd (editor), *Survey of Numerical Analysis.* (McGraw-Hill 1962) QA 297.T6. Contains a chapter on differential equations with numerous references.

[3] S.D. Conte, *Elementary Numerical Analysis.* (McGraw-Hill 1965) QA 297.C62. Describes use of partial double precision.

[4] B. Carnahan et al., *Applied Numerical Methods.* (John Wiley 1969) QA 297.C34. Gives Fortran programs.

[5] L. Collatz, *Functional Analysis and Numerical Mathematics.* (Academic Press 1966) QA 297.C58. Discusses the theoretical aspects of numerical analysis.

[6] F.S. Acton, *Numerical Methods That (Usually) Work.* (Mathematical Association of America 1990) QA 297.A33. Discusses Madelung and other transformations.

[7] L.B. Rall, ed., *Error in Digital Computation.* Vol. 1, QA 3.U45 No. 14 (Wiley 1965). Contains a chapter by P. Henrici on error in the integration of differential equations, and gives an extensive bibliography.

[8] R.W. Hamming, *Numerical Methods for Scientists and Engineers.* (McGraw-Hill 1962) QA 297.H28.

[9] J.M. Ortega, *Numerical Analysis; a Second Course.* (Academic Press 1972) QA 297.078.

[10] G.N. Lance, *Numerical Methods for High Speed Computers.* (Iliffe and Sons 1960) QA 76.L27.

[11] A. Ralston and H.F. Wilf, ed., *Mathematical Methods for Digital Computers*, Vol. 1 and 2. (Wiley 1960) QA 76.5.R3, Vol. 1 and 2.

[12] A. Ralston, *A First Course in Numerical Analysis.* (McGraw-Hill 1965) QA 297.R3.

[13] I.S. Berezin and N.P. Zhidkov, *Computing Methods* (2 Vols), QA 297.B4213, 1965 (Pergamon Press and Addison-Wesley 1965). A thorough, readable, and scholarly presentation. Some of its predictor and corrector coefficient listings contain typographical errors.

[14] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vettering, *Numerical Recipes, the Art of Scientific Computing.* (Cambridge University Press, 1996). These authors are enthusiastic about Richardson Extrapolation and the Bulirsch-Stoer Method.

[15] D. Kahaner, C. Moler, and S. Nash, *Numerical Methods and Software*, Prentice Hall (1989).

[16] G. Forsythe, C. Moler, and M. Malcom, *Computer Methods for Mathematical Computations*, Prentice Hall (1977). See also the Web site http://www.pdas.com/fmm.html.

[17] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, third edition, Springer-Verlag (2002).

Books and Articles on the Numerical Solution of Differential Equations

[18] L. Fox, *Numerical Solution of Ordinary and Partial Differential Equations.* (Addison-Wesley 1962) QA 371.L758, 1962. Also discusses integral equations.

[19] L. Collatz, *The Numerical Treatment of Differential Equations*, Springer-Verlag (1966). A standard reference work.

[20] W.E. Milner, *Numerical Solution of Differential Equations.* (John Wiley 1953 and Dover 1970) QA 371.M57, 1970. A standard older work and a bargain in paperback.

[21] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*,Wiley (1962). P. Henrici, *Error Propagation for Difference Methods*, Wiley (1963). These two books are the standard works on round-off and truncation error and their propagation. The first book gives Nyström's versions of Runge-Kutta. They are the Runge-Kutta analog of Størmer-Cowell in that they work directly with second-order equations when first derivatives are absent.

[22] I. Babuska et al., *Numerical Processes in Differential Equations.* (John Wiley 1966) QA 371.B2313. Gives numerical examples of the effect of round-off error.

[23] F. Ceschino and J. Kuntzmann, *Numerical Solution of Initial Value Problems.* Prentice-Hall (1966). Contains useful tabulations of coefficients for various integration schemes including very high-order Runge-Kutta.

[24] G.A. Chebotarev, *Analytical and Numerical Methods of Celestial Mechanics.* (American Elsevier 1967). Describes the use of perturbation series in practical celestial mechanics.

[25] C.W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations.* (Prentice-Hall 1971). Describes Richardson extrapolation and general "extrapolation methods". Also describes methods of automatic change of order and step size.

[26] J.C. Butcher, *The numerical analysis of ordinary differential equations: Runge-Kutta and general linear methods*, John Wiley, (1987).

[27] J.C. Butcher, *Numerical Methods for Ordinary Differential Equations*, First Edition, John Wiley (2003).

[28] J.C. Butcher, *Numerical Methods for Ordinary Differential Equations*, Second Edition, John Wiley (2008). http://www.math.auckland.ac.nz/~butcher/ODE-book-2008/.

[29] J.C. Butcher, "Runge-Kutta Methods", *Scholarpedia* (2011). http://www.scholarpedia.org/article/Runge-Kutta_methods.

[30] G. Hall and J.M. Watt (eds.), *Modern numerical methods for ordinary differential equations*, Oxford University Press, Oxford (1976).

[31] L.F. Shampine and M.K. Gordon, *Computer Solution of Ordinary Differential Equations: The Initial Value Problem*, W.H. Freeman, San Francisco (1975).

[32] L.F. Shampine, *Numerical Solution of Ordinary Differential Equations*, Chapman and Hall (1994).

[33] J.D. Lambert, *Computational methods in ordinary differential equations*, Wiley, New York (1973).

[34] J.D. Lambert, *Numerical Methods for Ordinary Differential Systems: The Initial Value Problem*, Wiley (1991).

[35] H. Stetter, *Analysis of Discretization Methods for Ordinary Differential Equations*, Springer Verlag (1973).

[36] S.O. Fatunla, *Numerical methods for initial value problems in ordinary differential equations*, Academic Press, London (1989).

[37] E. Hairer, S. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I. Nonstiff Problems*, Springer (1993).

[38] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II. Stiff and Differential Algebraic Problems*, Springer (2002).

[39] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations*, Second Edition, Cambridge University Press (2009).

[40] D. F. Griffiths and D. J. Higham, *Numerical Methods for Ordinary Differential Equations: Initial Value Problems*, Springer (2010).

[41] J.B. Rosser, "A Runge-Kutta for All Seasons", *SIAM Review* **9**, 417-452 (1967). Describes an improved Runge-Kutta method.

[42] P. Deuflhard and F. Bornemann, *Scientific Computing with Ordinary Differential Equations*, Springer (2002).

[43] L.F. Shampine, H.A. Watts, and S.M. Davenport, "Solving Nonstiff Ordinary Differential Equations - The State of the Art", *SIAM Review* **18**, 376-411 (1976). Compares merits of various methods and computer codes.

[44] A.M. Stuart and A.R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press (1996).

[45] A.C. Hindmarsh, "ODEPACK: A Systematized Collection of ODE Solvers", in *Scientific Computing*, R.S. Stepleman et al. eds., North-Holland, Amsterdam (1983).

[46] L.R. Petzold, "Automatic Selection of Methods for Solving Stiff and Nonstiff Systems of Ordinary Differential Equations", *SIAM Journal of Scientific and Statistical Computing* **4**, pp. 136-148 (1983).

[47] S. Herrick, *Astrodynamics*, vols. 1 and 2, Van Nostrand Reinhold (1971).

[48] J. Dormand and P. Prince, "A family of embedded Runge-Kutta formulae", *J. Comp. Appl. Math.* **6**, 19-26 (1980).

[49] M. Sofroniou and G. Spaletta, "Construction of explicit runge-kutta pairs with stiffness detection", *Mathematical and Computer Modelling* **40**, 1157-1169 (2004).

[50] G. Wanner, "Germund Dahlquist's classical papers on Stability Theory", http://www.unige.ch/~wanner/DQsem.pdf.

[51] The program *Mathematica* provides a command NDSolve that implements many different integration methods including Runge-Kutta, Adams, and extrapolation. For a discussion of Runge-Kutta in *Mathematica*, Google "NDSolve explicit Runge-Kutta" and follow related links.

Extrapolation Methods

[52] R. Bulirsch and J. Stoer, "Numerical Treatment of Ordinary Differential Equations by Extrapolation Methods", *Numerische Mathematik* **8**, 1-13, 93-104 (1966).

[53] P. Deuflhard, "Order and Stepsize Control in Extrapolation Methods", *Numerische Mathematic* **41**, 399-422 (1983).

[54] P. Deuflhard, "Recent Progress in Extrapolation Methods for Ordinary Differential Equations", *SIAM Review* **27**, 505-535 (1985).

[55] E. Hairer, S. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I. Nonstiff Problems*, Springer (1993).

[56] E. Hairer and C. Lubich, "Asymptotic expansions of the global error of fixed stepsize methods", *Numer. Math.* **45**, 345-360 (1984).

[57] H. Stetter, *Analysis of Discretization Methods for Ordinary Differential Equations*, Springer Verlag (1973).

[58] H. Stetter, “Symmetric Two-Step Algorithms for Ordinary Differential Equations”, *Computing* **5**, 267-280 (1970).

[59] P. Deuflhard and F. Bornemann, *Scientific Computing with Ordinary Differential Equations*, Springer (2002).

References to Regularization of Kepler and Størmer Problems

[60] E.L. Stiefel and G. Scheifele, *Linear and Regular Celestial Mechanics*, Springer Verlag (1971).

[61] D. Boccaletti and G. Pucacco, *Theory of Orbits*, 2 vols., Springer-Verlag (1996). This excellent 2-volume set is full of interesting material including a discussion of Lie methods.

[62] A.J. Dragt, Trapped Orbits in a Magnetic Dipole Field, *Rev. Geophys.* **3**, p. 255-298 (1965).

[63] A.J. Dragt and J.M. Finn, Insolubility of Trapped Particle Motion in a Magnetic Dipole Field, *J. Geophys. Res.* **81**, p. 2327-2340 (1976).

[64] A.J. Dragt and J.M. Finn, Normal Form for Mirror Machine Hamiltonians, *J. Math. Physics* **20**, p. 2649-2660 (1979).

References to Symplectic Integration and Backward Error Analysis

See also the references at the end of Chapter 12.

[65] J.M. Sanz-Serna and M.P. Calvo, *Numerical Hamiltonian Problems*, Chapman and Hall (1994).

[66] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, Second Edition, Springer (2010).

[67] P. Chartier, E. Hairer, and G. Vilmart, “Numerical integrators based on modified differential equations”, Mathematics of Computation S 0025-5718(07)01967-9.

[68] Sebastian Reich, “Backward error analysis for numerical integrators”, *SIAM J. Numer. Anal.* **36**, 1549-1570 (1999).

[69] E. Hairer and C. Lubich, “The Life-Span of Backward Error Analysis for Numerical Integrators”, *Numer. Math.* **76**, 441-462 (1996).

Manuals on Computer Programming

[70] D.D. McCracken, *A Guide to FORTRAN IV Programming.* QA 76.5M1872 (Wiley 1965).

[71] E.I. Organick, *A FORTRAN IV Primer.* QA 76.5.072 (Addison-Wesley 1966).

[72] G.B. Davis and T.R. Hoffman, *FORTRAN 77, A Structured, Disciplined Style* (McGraw-Hill 1988).

[73] American National Standard Programming Language FORTRAN (77), American National Standards Institute (ANSI), New York, NY (1978).

Original Sources

[74] C. Runge, “Uber die numerische Auflösung von Differentialgleichungen”, *Math. Ann.* **46**, 167-178 (1895).

[75] G. Coriolis, “Mémoire sur le degré d'approximation qu'on obtient pour les valeurs numériques d'une variable qui satisfait à une équation différentielle, en employant pour calculer ces valeurs diverses équations aux différences plus ou moins approchées”, *Journal de Mathématiques Pures et Appliquées* **2**, 229-244 (1837). This paper reveals that some of what we now call Runge-Kutta methods were, in fact, known much earlier to Coriolis.

Chapter 3

Symplectic Matrices and Lie Algebras/Groups

> Lie theory is in the process of becoming the most important part of modern mathematics. Little by little it became obvious that the most unexpected theories, from arithmetic to quantum physics, came to encircle this Lie field like a gigantic axis.
>
> *Jean Dieudonne*

We will learn in subsequent chapters that symplectic matrices play an important role in the advanced treatment of Hamiltonian systems. Briefly put, Hamiltonian motion produces symplectic maps. Also, symplectic maps preserve the Hamiltonian form of the equations of motion. Finally, symplectic maps are characterized by symplectic matrices. The purpose of this chapter is to define symplectic matrices and to explore some of their properties in preparation for future use. This exploration also provides a context for the discussion of Lie algebras and Lie groups.

In his youth, and for publishing his first mathematical paper (1869), Sophus Lie received a travel grant from the Norwegian University of Christiania to visit the mathematical capitals of Europe. One such capital was Paris where he visited and worked with Klein, who himself was visiting there from Prussia.

While Lie and Klein thought deeply about mathematics in Paris, the political situation between France and Prussia deteriorated. The popularity of the French emperor Napoleon III was declining and he thought war with Prussia, which his advisors said the French army was sure to win, might change his political fortunes. Bismarck, the Prussian chancellor, saw a war with France as an opportunity to unite the South German states. With both sides feeling that a war was to their advantage, the Franco-Prussian war became inevitable. On July 14, 1870, Bismarck sent a telegram which infuriated the French government. On July 19, France declared war on Prussia. For Klein there was then only one possibility: he had to return quickly to Berlin.

However, Lie was a Norwegian and he was finding mathematical discussions in Paris very stimulating. He decided to remain there, but became anxious as the German offensive met with only ineffectual French response. In August, when the German army trapped part of the French army in Metz, Lie decided it was also time for him to leave Paris, and he planned to hike (on foot!) to Italy. He made as far as Fontainebleau, just south of Paris, when the French police spotted him as a suspicious-looking young man wandering in lonely places in the forest, stopping now and then to make notes and drawings in his notebook. *"He was of tall stature and had the classic Nordic appearance. A full blond beard framed his face and his grey-blue eyes sparkled behind his glasses. He gave the impression of unusual physical strength"* (Élie Cartan). The police searched him and found a map, letters in German, and papers full of mysterious formulas, complexes, diagrams, and names. He was suspected of being a German spy and imprisoned.

Lie had to stay in prison in Fontainebleau for 4 weeks before his French colleague Gaston Darboux learned about the incident and arrived on behalf of the French Academy of Sciences with a release order signed by the Minister of Home Affairs. Lie himself had taken things truly philosophically and made good use of his time in prison. For, as he recounted later, in these forced leisure days he had plenty of peace and quiet to concentrate on his problems and advance them essentially. In a letter to his Norwegian friend Ernst Motzfeldt, written directly after his release, Lie remarked: *"I think that a mathematician is well suited to be in prison."*

The French army surrendered on September 2 but, after a September 4 coup d'état against Napoleon III, France resumed the war on September 6. On September 19 the German army began to blockade Paris. This time Lie successfully fled to Italy, then from there he made his way back to Christiania via Germany so that he could again meet and discuss mathematics with Klein. Thus began the work of Sophus Lie.[1]

3.1 Definitions

To define symplectic matrices, it is first necessary to introduce a certain *fundamental* $2n \times 2n$ antisymmetric matrix J. It is defined by the equation

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}. \tag{3.1.1}$$

Here each entry in J is an $n \times n$ matrix, I denotes the $n \times n$ identity matrix, and all other entries are zero. The observant reader will recognize this J as the Poisson matrix already defined in Section 1.7 in connection with Poisson brackets.

With this background, a $2n \times 2n$ matrix M is said to be *symplectic* if

$$M^T J M = J. \tag{3.1.2}$$

Here M^T denotes the transpose of M. Observe that symplectic matrices must be of even dimension by definition. Usually we will be interested in real symplectic matrices. However, in some cases we will also be interested in symplectic matrices with complex entries.

Finally, we remark that the use of the adjective *symplectic* in this general context is due to Hermann Weyl (1885-1955). *Symplectic*, the Greek equivalent of the Latin-based word

[1] See the Web site http://www-history.mcs.st-andrews.ac.uk/Biographies/Lie.html. See also the "Overview and History of the Theory of Lie Algebras and Lie Groups" references given at the end of Chapter 27.

complex, comes from $\sigma\upsilon\mu\pi\lambda\epsilon\kappa\tau\iota\kappa\grave{o}\varsigma$, which means *intertwined* or *braided*. Weyl had in mind the *symplectic 2-form* associated with J when introducing this adjective.[2] We may view it as intertwining the components of two vectors, call them w and z, with the components of J. See (2.3). We may also view (1.2) as an intertwining of J with M^T and M.

Exercises

3.1.1. Show that the matrix J has the following properties:

$$J^T = -J, \tag{3.1.3}$$

$$J^2 = -I \quad \text{or} \quad J^{-1} = -J, \tag{3.1.4}$$

$$\det(J) = 1, \tag{3.1.5}$$

$$J^T J = J J^T = I. \tag{3.1.6}$$

3.1.2. Suppose that $n = 1$ (in which case J is 2×2) and suppose A is any 2×2 matrix. Write A in the form

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}. \tag{3.1.7}$$

Then A has determinant

$$\det(A) = ad - bc. \tag{3.1.8}$$

Verify that

$$A^{-1} = [1/\det(A)] \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}, \tag{3.1.9}$$

and

$$-JA^T J = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \tag{3.1.10}$$

Verify that

$$A^T J A = [\det(A)] J. \tag{3.1.11}$$

3.1.3. By taking the determinant of both sides of (1.2), show that any symplectic matrix M has the property

$$\det(M) = \pm 1. \tag{3.1.12}$$

It follows that symplectic matrices are always invertible.
Comment: It will be shown in Sections 3.3, 3.9, and * that $\det(M)$ actually always equals $+1$ for a symplectic matrix. Also, as is easily seen from (1.11), in the 2×2 case the necessary and sufficient condition for a matrix to be symplectic is that it have determinant $+1$.

[2] Weyl would have preferred to use the Latin-based word *complex* because the vanishing of the 2-form defines what is called a line complex, and even did so for a time. However he abandoned this usage because of the confusion it created with complex numbers. It is also of historical interest to note that the term "Lie algebra" itself was first introduced in 1934 by Weyl. Prior to that time terms like "infinitesimal group" had been employed.

3.1.4. Show that any symplectic matrix M has the following properties:

$$M^{-1} = -JM^TJ = J^{-1}M^TJ = JM^TJ^{-1}, \tag{3.1.13}$$

$$MJM^T = J, \tag{3.1.14}$$

$$(M^{-1})^T = -JMJ. \tag{3.1.15}$$

3.1.5. Show that the matrices I and J are symplectic.

3.1.6. Suppose M is a symplectic matrix. Show that M^{-1}, M^T, and $-M$ are then also symplectic matrices.

3.1.7. Suppose M and N are symplectic matrices. Show that the product MN is then also a symplectic matrix. Taken together, Exercises 1.5, 1.6, and this exercise show, among other things, that the set of all $2n \times 2n$ symplectic matrices forms a *group*. See Section 3.6.

3.1.8. Show that a symplectic matrix cannot have $\lambda = 0$ as an eigenvalue.

3.1.9. Let M be any $2n \times 2n$ matrix. Define its *symplectic transpose* M^S by the rule

$$M^S = JM^TJ^{-1}. \tag{3.1.16}$$

Show that, similar to the case for the ordinary transpose, there are the relations

$$I^S = I,\ J^S = -J,\ (M^S)^S = M,\ (MN)^S = N^SM^S. \tag{3.1.17}$$

Show that the symplectic condition (1.2) can be written in the form

$$M^SM = MM^S = I. \tag{3.1.18}$$

3.1.10. Here are some things it is assumed you know about matrices: Let A and B be any two $n \times n$ matrices. The determinant function has the properties $\det(A^T) = \det(A)$ and $\det(AB) = \det(A)\det(B)$. The trace function has the properties $\operatorname{tr}(A^T) = \operatorname{tr}(A)$ and $\operatorname{tr}(AB) = \operatorname{tr}(BA)$. The matrix A has an inverse, which is unique and is both a left and right inverse, iff $\det(A) \neq 0$. There is the relation $\det(A^{-1}) = [\det(A)]^{-1}$. The transposition, Hermitian conjugation, and inversion operations have the properties $(AB)^T = B^TA^T$, $(AB)^\dagger = B^\dagger A^\dagger$, $(AB)^{-1} = B^{-1}A^{-1}$. The operations of inversion and transposition, and inversion and Hermitian conjugation, commute: $(A^T)^{-1} = (A^{-1})^T$ and $(A^\dagger)^{-1} = (A^{-1})^\dagger$. If these results are unfamiliar to you, consult the *Matrix Theory* references provided at the end of this chapter.

3.2 Variants

There are other possible choices for the form of the matrix J. One important variant is described in this section. All possible variants are discussed in Section 3.13.

Let x and y be two n-dimensional vectors with real entries. Define a $2n$-component real vector z by the rule

$$z = (z_1 \cdots z_n, z_{n+1} \cdots z_{2n}) = (x_1 \cdots x_n, y_1 \cdots y_n). \tag{3.2.1}$$

Similarly, let u and v be another pair of real n-dimensional vectors, and define the $2n$-component real vector w by the rule

$$w = (w_1 \cdots w_n, w_{n+1} \cdots w_{2n}) = (u_1 \cdots u_n, v_1 \cdots v_n). \tag{3.2.2}$$

Then one has the relation

$$(w, Jz) = (u, y) - (v, x). \tag{3.2.3}$$

This quadratic form is called the *fundamental symplectic 2-form.*[3] Note that the inner product on the left of (2.3) is that for $2n$-dimensional vectors, and those on the right of (2.3) are for n-dimensional vectors.

Define a $2n$-component vector z' in terms of the vector z by requiring that z' have the entries

$$z' = (x_1, y_1, x_2, y_2, \cdots x_n, y_n). \tag{3.2.4}$$

Evidently, z' is related to z by a linear transformation. Indeed, the entries in z' are a *permutation* of those in z. Consequently, there is a matrix P, with entries 0 and 1, such that

$$z' = Pz. \tag{3.2.5}$$

See Exercise Let w' be defined in terms of w by an analogous relation,

$$w' = Pw. \tag{3.2.6}$$

It follows from (2.4) and its counterpart for w' that one has the relation

$$(w', z') = (u, x) + (v, y) = (w, z). \tag{3.2.7}$$

But, by (2.5) and (2.6), there is also the relation

$$(w', z') = (Pw, Pz) = (w, P^T Pz). \tag{3.2.8}$$

Comparison of (2.7) and (2.8) shows that P is *orthogonal*,

$$P^T P = I \ \text{ or } \ P^T = P^{-1}. \tag{3.2.9}$$

Let J' be the $2n \times 2n$ matrix defined by the equation

$$J' = \begin{pmatrix} J_2 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_2 \end{pmatrix}. \tag{3.2.10}$$

That is, all the entries of J' are zero save for n 2×2 blocks on the diagonal. These blocks are identical, and are specified by the equation

$$J_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{3.2.11}$$

[3] Other authors take (Jw, z) to be the fundamental symplectic 2-form. In view of (1.3), (w, Jz) and (Jw, z) differ only by a sign. A 2-form is a function that takes as inputs two vectors, is linear in both inputs, and delivers a number. It is also usually required to be odd under the interchange of the two vector inputs. See (1.3).

The matrix J' has been defined in such a way as to satisfy the relation

$$(w', J'z') = (u, y) - (v, x) = (w, Jz). \tag{3.2.12}$$

By (2.5) and (2.6) there is also the relation

$$(w', J'z') = (Pw, J'Pz) = (w, P^T J'Pz). \tag{3.2.13}$$

It follows from (2.12) and (2.13) that J and J' are related by the orthogonal similarity transformation

$$P^T J'P = J \quad \text{or} \quad J' = PJP^T. \tag{3.2.14}$$

To complete the story, suppose that M is any symplectic matrix with respect to J. See (1.2). Consider the matrix M' defined by the orthogonal similarity transformation

$$M' = PMP^T. \tag{3.2.15}$$

Then it is easily checked using (1.2) and (2.15) that M' is symplectic with respect to J',

$$(M')^T J'M' = J'. \tag{3.2.16}$$

Indeed, if M and N are any two matrices (not even necessarily symplectic), and M' and N' are their counterparts defined by relations of the form (2.15), then it follows from the orthogonality condition (2.9) that

$$(MN)' = PMNP^T = PMP^T PNP^T = M'N'. \tag{3.2.17}$$

The results of this section and the exercises below show that for the most part, *mutatis mutandis*, one may use either J or J' when defining or working with symplectic matrices. Generally we shall drop the prime notation, and use the symbol J to denote either J or J'. Sometimes, however, a particular choice of J may give simpler or more interesting results. When this is the case, we shall be more specific.

Exercises

3.2.1. Consider the properties of J given in Exercise (1.1). Show that J' has the same properties.

3.2.2. Show that I and J' are symplectic with respect to J'.

3.2.3. Exercises 1.3, 1.5, 1.6, and 1.7 describe properties of matrices symplectic with respect to J. Show that matrices symplectic with respect to J' have directly analogous properties.

3.2.4. Let M be any matrix. Define the operation of "priming" a matrix by (2.15). Show that the operations of priming and transposing commute, $(M^T)' = (M')^T$. Show that the operations of priming and inverting also commute, $(M^{-1})' = (M')^{-1}$. Finally, show that inverting and transposing also commute, $(M^{-1})^T = (M^T)^{-1}$.

3.2.5. Compute P explicitly in the 4×4 and 6×6 cases. For each case verify that P is orthogonal, and find its eigenvalues and determinant. You should find $\det(P) = -1$ in both cases. In the 4×4 case you should find the result

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{3.2.18}$$

and in the 6×6 case you should find the result

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3.2.19}$$

3.2.6. Recall the definition of the collection of variables z given by (1.7.9). By comparing (2.1) and (2.4) and identifying x_j with q_j and y_j with p_j show that z' is the collection of variables

$$z' = (q_1, p_1, q_2, p_2, \cdots q_n, p_n). \tag{3.2.20}$$

Verify that the variables z' obey the Poisson bracket relations

$$[z'_a, z'_b] = J'_{ab}. \tag{3.2.21}$$

3.2.7. Let x and y be a pair of real n-component vectors. Define a real $2n$-component vector z by the rule

$$z = (z_1 \cdots z_n, z_{n+1} \cdots z_{2n}) = (x_1 \cdots x_n, y_1 \cdots y_n). \tag{3.2.22}$$

Also define a complex n-component vector w by the rule

$$w = (w_1 \cdots w_n) = (x_1 + iy_1 \cdots x_n + iy_n) = x + iy. \tag{3.2.23}$$

Let $w' = x' + iy'$ be another such vector. Form the *complex* inner product (w, w'). Obtain the result

$$\begin{aligned} (w, w') &= (x + iy, x' + iy') \\ &= (x, x') + (y, y') + i[(x, y') - (y, x')] \\ &= (z, z') + i(z, Jz'). \end{aligned} \tag{3.2.24}$$

You have shown that the symplectic 2-form (z, Jz') may be obtained as the *imaginary* part of a complex inner product. Suppose we make the correspondence

$$w \leftrightarrow z \tag{3.2.25}$$

as described by (2.22) and (2.23). This correspondence is a bijective mapping between the complex vector space C^n and the real vector space R^{2n}. Show that there is then the correspondence

$$-iw \leftrightarrow Jz. \tag{3.2.26}$$

Thus, in some ways, J acts like $-i$. See (1.4) and Exercise 8.1. For this reason, J is said to provide phase space with an *almost complex structure*. Suppose one instead defines w by the rule

$$w = (w_1 \cdots w_n) = (y_1 + ix_1 \cdots y_n + in_n) = y + ix, \tag{3.2.27}$$

and then again makes the correspondence (2.25). Show that now there is the correspondence

$$iw \leftrightarrow Jz, \tag{3.2.28}$$

so that now J acts like $+i$.

3.2.8. Suppose two phase-space points (vectors) w and z are sent under the action of a linear symplectic map, described by the symplectic matrix M, to the points w', z':

$$w' = Mw, \tag{3.2.29}$$

$$z' = Mz. \tag{3.2.30}$$

Show that the fundamental symplectic 2-form (2.3) is preserved under a symplectic transformation. That is, the relation

$$(w', Jz') = (Mw, JMz) = (w, Jz) \tag{3.2.31}$$

holds for any real symplectic matrix M and any pair of points w,z. It follows that a real matrix is symplectic if and only if it preserves the fundamental symplectic 2-form.

3.3 Simple Symplectic Restrictions and Symplectic Factorization

3.3.1 Large-Block Formulation

Suppose M is a $2n \times 2n$ matrix. Then it can be written in the form

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \tag{3.3.1}$$

where the matrices A through D are $n \times n$ blocks. Correspondingly, M^T can be written as

$$M^T = \begin{pmatrix} A^T & C^T \\ B^T & D^T \end{pmatrix}. \tag{3.3.2}$$

Now require that M be symplectic with respect to the J of (1.1). It then follows from (1.2) that the matrices A through D must satisfy the conditions

$$A^T C = C^T A, \tag{3.3.3}$$

$$B^T D = D^T B, \tag{3.3.4}$$

$$A^T D - C^T B = I. \tag{3.3.5}$$

If M is symplectic with respect to J, so is M^T. See (1.10). From (1.10) it follows that A through D must also satisfy the conditions

$$AB^T = BA^T, \tag{3.3.6}$$

$$CD^T = DC^T, \tag{3.3.7}$$

$$AD^T - BC^T = I. \tag{3.3.8}$$

3.3.2 Symplectic Block Factorization

Consider matrices having the block forms

$$M = \begin{pmatrix} I & B \\ 0 & I \end{pmatrix}, \tag{3.3.9}$$

$$M = \begin{pmatrix} I & 0 \\ C & I \end{pmatrix}, \tag{3.3.10}$$

$$M = \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix}, \tag{3.3.11}$$

Then it is readily verified from (3.3) through (3.8) that (3.9) and (3.10) are symplectic if

$$B^T = B \text{ and } C^T = C. \tag{3.3.12}$$

Also, (3.11) is symplectic if

$$A^T D = I \text{ or } D = (A^T)^{-1}. \tag{3.3.13}$$

Observe that all matrices of the form (3.9) and (3.10) have determinant +1. Moreover the matrix M given by (3.11) also has determinant +1 if it is symplectic: Simple calculation and use of (3.13) gives the result

$$\begin{aligned} \det(M) &= \det(A)\det(D) = \det(A)\det[(A^T)^{-1}] \\ &= \det(A)\det(A^{-1}) = \det(AA^{-1}) = \det(I) = 1. \end{aligned} \tag{3.3.14}$$

See Exercise 3.2.

Let M be any matrix written in the form (3.1). Suppose A and/or D are invertible [$\det(A) \neq 0$ and/or $\det(D) \neq 0$]. Then, as can be easily checked by direct matrix multiplication, M has the block factorizations

$$M = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix}, \tag{3.3.15}$$

$$M = \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I & 0 \\ D^{-1}C & I \end{pmatrix}. \tag{3.3.16}$$

Next, suppose that M is symplectic. Then remarkably each of the factors appearing in (3.15) and (3.16) is separately symplectic. We prove this assertion for the factorization (3.15). The proof for the factorization (3.16) is similar. Before so doing, we note that the three factors appearing in (3.15) and (3.16) are of the forms (3.9) through (3.11) and therefore, if symplectic, have determinant +1. Therefore M in this case has determinant +1.

To prove that the factors in (3.15) are symplectic, begin by observing that (3.3) can be rewritten in the form

$$CA^{-1} = (CA^{-1})^T. \tag{3.3.17}$$

That is, the matrix CA^{-1} is symmetric. It follows that the first factor in (3.15) is symplectic. Similarly, observe that (3.6) can be rewritten in the form

$$A^{-1}B = (A^{-1}B)^T, \tag{3.3.18}$$

and consequently the matrix $A^{-1}B$ is also symmetric. It follows that the third factor in (3.15) is symplectic. Finally, with the aid of (3.3), the relation (3.5) can be rewritten in the form

$$A^T(D - CA^{-1}B) = I. \tag{3.3.19}$$

It follows that the second factor in (3.15) is also symplectic.

Even if a symplectic M cannot be written as a product of three symplectic factors as in (3.15) and (3.16), it can always be written as a product of a finite number of symplectic factors of the form (3.9) through (3.11). That is, symplectic matrices of the form (3.9) through (3.11) *generate* all symplectic matrices.[4]

To verify this assertion, suppose M is written in the form (3.1). We distinguish two cases: either the block A vanishes identically or it does not. Suppose A does vanish. Then we have the relations

$$\begin{pmatrix} I & I \\ 0 & I \end{pmatrix}\begin{pmatrix} 0 & B \\ C & D \end{pmatrix} = \begin{pmatrix} C & B+D \\ C & D \end{pmatrix}, \tag{3.3.20}$$

$$M = \begin{pmatrix} 0 & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & -I \\ 0 & I \end{pmatrix}\begin{pmatrix} C & B+D \\ C & D \end{pmatrix}. \tag{3.3.21}$$

Moreover, the matrix C must satisfy $\det(C) \neq 0$. For if $\det(C) = 0$, then the n columns of C must be linearly dependent, which implies that the first n columns of M must be linearly dependent, which implies $\det(M) = 0$ contrary to the result of Exercise (1.3). [The same conclusion, $\det(C) \neq 0$, also follows directly from (3.5).] Also, according to Exercise (1.6), the matrix on the right side of (3.20) is symplectic. It follows that this matrix has a factorization of the form (3.15), and correspondingly according to (3.21) M can be written as a product of four factors of the form (3.9) through (3.11).

Next suppose that A does not vanish. Let V_W denote the symplectic matrix

$$V_W = \begin{pmatrix} W & 0 \\ 0 & W^* \end{pmatrix}, \tag{3.3.22}$$

[4] The word *generate* has many meanings depending on context. Here it means that any symplectic matrix can be expressed as a product of a finite number of symplectic matrices having a specific form. In the Lie algebraic context it happens that some Lie group elements G can be written in the form $G = \exp(g)$ where g is in the associated Lie algebra. See Section 3.7. In that case, but with a different meaning, we also say that g generates G.

where

$$W^* = (W^T)^{-1}. \tag{3.3.23}$$

Pre and post multiply M by V_X and V_Y to get the result

$$M' = V_X M V_Y = \begin{pmatrix} A' & B' \\ C' & D' \end{pmatrix}, \tag{3.3.24}$$

with A' given by the relation

$$A' = XAY. \tag{3.3.25}$$

According to a standard result in matrix theory, nonsingular matrices X and Y can be selected in such a way that A' takes the block form

$$A' = \begin{pmatrix} I_\ell & 0 \\ 0 & 0_{n-\ell} \end{pmatrix}. \tag{3.3.26}$$

Here I_ℓ is the $\ell \times \ell$ identity matrix (with $\ell \geq 1$), and $0_{n-\ell}$ is a complementary zero matrix. (The integer ℓ is the rank of A.) Correspondingly, C' can be written in the block form

$$C' = \begin{pmatrix} C'_{11} & C'_{12} \\ C'_{21} & C'_{22} \end{pmatrix}. \tag{3.3.27}$$

Then use of the symplectic condition (3.3) when applied to A' and C' gives the result

$$C'_{12} = 0. \tag{3.3.28}$$

It follows that $\det(C'_{22}) \neq 0$. For if $\det(C'_{22}) = 0$, then the $(n-\ell)$ columns of C'_{22} must be linearly dependent, which implies that $(n-\ell)$ columns of M' must be linearly dependent, which implies $\det(M') = 0$ contrary to Exercises (1.3) and (1.6). Let T_λ, where λ is an arbitrary real parameter, denote the symplectic matrix

$$T_\lambda = \begin{pmatrix} I & \lambda I \\ 0 & I \end{pmatrix}. \tag{3.3.29}$$

Multiply M' by T_λ on the left to get the result

$$M'' = T_\lambda M' = \begin{pmatrix} I & \lambda I \\ 0 & I \end{pmatrix} \begin{pmatrix} A' & B' \\ C' & D' \end{pmatrix} = \begin{pmatrix} A'' & B'' \\ C'' & D'' \end{pmatrix} \tag{3.3.30}$$

with A'' given by the relation

$$A'' = A' + \lambda C' = \begin{pmatrix} I_\ell + \lambda C'_{11} & 0 \\ \lambda C'_{21} & \lambda C'_{22} \end{pmatrix}. \tag{3.3.31}$$

A little thought shows that λ can be selected in such a way that $\det(A'') \neq 0$. By inverting the relations (3.24) and (3.30), we see that M can be written in the form

$$M = V_X^{-1} T_\lambda^{-1} M'' V_Y^{-1}. \tag{3.3.32}$$

And, according to the previous discussion, M'' has a factorization of the form (3.15). Thus, M can again be written as a product of factors (this time six in number) of the form (3.9) through (3.11).

3.3.3 Symplectic Matrices Have Determinant $+1$

Moreover, as a bonus, we observe that since each factor has determinant +1, the matrix M itself must have determinant +1. We conclude that every symplectic matrix M (real or complex) must satisfy the relation

$$\det(M) = +1. \tag{3.3.33}$$

Here is a topological perspective on the relation (3.33). Suppose it can be established that symplectic matrices written in the form (3.31) and having $\det(A) \neq 0$ are *dense* in the set of all symplectic matrices. That is for any symplectic matrix M', written in the form (3.31) and having $\det(A) = 0$, there is a symplectic matrix M arbitrarily nearby with $\det(A) \neq 0$. For this matrix we know from the factorization (3.15) that $\det(M) = 1$. But from (1.8) we know that $\det(M') = \pm 1$. Since the determinant of a matrix is a continuous function of its entries, it follows from the density hypothesis that $\det(M') = +1$.

3.3.4 Small-Block Formulation

Equally interesting are the results of requiring M to be symplectic with respect to the J' of (2.10). For simplicity in this case, and for later use, we restrict our discussion to 6 × 6 matrices. Then M can be written in the form

$$M = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \tag{3.3.34}$$

where the matrices a through i are all 2 × 2. Correspondingly, M^T can be written as

$$M^T = \begin{pmatrix} a^T & d^T & g^T \\ b^T & e^T & h^T \\ c^T & f^T & i^T \end{pmatrix}. \tag{3.3.35}$$

Now require that M satisfy the condition (1.2) with J replaced by J'. We find that the matrices a through i must satisfy the conditions

$$a^T J_2 a + d^T J_2 d + g^T J_2 g = J_2, \tag{3.3.36}$$

$$\begin{gathered} b^T J_2 b + e^T J_2 e + h^T J_2 h = J_2, \\ c^T J_2 c + f^T J_2 f + i^T J_2 i = J_2, \\ b^T J_2 a + e^T J_2 d + h^T J_2 g = 0, \\ c^T J_2 a + f^T J_2 d + i^T J_2 g = 0, \\ c^T J_2 b + f^T J_2 e + i^T J_2 h = 0. \end{gathered} \tag{3.3.37}$$

Note that because of (1.7), the relations (3.36) can also be written in the form

$$\det a + \det d + \det g = 1, \tag{3.3.38}$$

$$\det b + \det e + \det h = 1,$$

$$\det c + \det f + \det i = 1.$$

As before, M must also satisfy (1.10) with J replaced by J'. As a consequence, the matrices a through i must also satisfy the conditions

$$\det a + \det b + \det c = 1, \tag{3.3.39}$$

$$\det d + \det e + \det f = 1,$$

$$\det g + \det h + \det i = 1,$$

$$dJ_2a^T + eJ_2b^T + fJ_2c^T = 0, \tag{3.3.40}$$

$$gJ_2a^T + hJ_2b^T + iJ_2c^T = 0,$$

$$gJ_2d^T + hJ_2e^T + iJ_2f^T = 0.$$

Exercises

3.3.1. Verify the relations (3.3) through (3.8).

3.3.2. Verify that M as given by (3.11) can be written in the form

$$M = \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & D \end{pmatrix}. \tag{3.3.41}$$

and therefore

$$\det(M) = \det(A)\det(D). \tag{3.3.42}$$

3.3.3. Verify the block factorizations (3.15) and (3.16). Work out in detail the proof that each factor in (3.15) and (3.16) is separately symplectic if M is symplectic.

3.3.4. Verify in detail all the steps required to show that matrices of the form (3.9) through (3.11) generate all symplectic matrices.

3.3.5. Verify the relations (3.36) through (3.40).

3.4 Eigenvalue Spectrum

Suppose a map $\mathcal{M}$ acts on some some space with coordinates z and suppose $\mathcal{M}$ has a fixed point z_f,

$$\mathcal{M}z_f = z_f.$$

What can be said about the behavior of points near this fixed point under repeated application of $\mathcal{M}$? In lowest approximation, this behavior is controlled by the matrix M that specifies the linear part of $\mathcal{M}$ when it is expanded about z_f. It can be shown, in the linear approximation, that points near z_f remain near z_f under repeated application of $\mathcal{M}$ if all the eigenvalues of M are within the unit circle in the complex plane or are on the unit circle and distinct. In this case z_f is said to be *stable*. On the other hand, if any eigenvalue lies outside the unit circle, there are points near z_f that, again in the linear approximation, are mapped away from z_f exponentially fast under repeated application of $\mathcal{M}$. In this case, z_f is said to be *unstable*. See Subsection 3.5.8.

By definition, the linear part of a symplectic map is specified by a symplectic matrix. See Section 6.1.2. We are therefore particularly interested in the eigenvalues of M when M is symplectic.

Finally, in the context of accelerator physics, suppose $\mathcal{M}$ is the one-turn map for a circular machine (ring). In this case, a fixed point of $\mathcal{M}$ corresponds to a closed orbit. In order to accelerate/store a large number of particles, it is essential that this closed orbit (fixed point) be stable. That is, if one fails to inject onto the closed orbit, as will be the case for most of any injected beam, one desires that particles near the closed orbit will remain so for very large times (in some cases equivalent in terms of the number of oscillations about the closed orbit to the number of trips of the earth around the sun since the big bang). Therefore, for successful accelerator design and operation, it is essential to know and control the eigenvalues of M.

3.4.1 Background

The *characteristic polynomial* $P(\lambda)$ of any matrix M is defined by the equation

$$P(\lambda) = \det(M - \lambda I). \tag{3.4.1}$$

Evidently $P(\lambda)$ is a polynomial with real coefficients if the matrix M is real. Also, the eigenvalues of M are the roots of the equation

$$P(\lambda) = 0. \tag{3.4.2}$$

It follows that if M is a <u>*real*</u> matrix, then its eigenvalues must also be real or must occur in complex conjugate pairs $\lambda, \overline{\lambda}$.

Suppose M is a symplectic matrix. Then it follows from (1.9) that

$$J^{-1}(M^T - \lambda I)J = M^{-1} - \lambda I = -\lambda M^{-1}(M - \lambda^{-1} I). \tag{3.4.3}$$

Since M is symplectic, we also have the relation

$$\det(M) = +1. \tag{3.4.4}$$

See Section 3.3. Now take the determinant of both sides of (4.3). The result is the relation

$$P(\lambda) = \lambda^{2n} P(1/\lambda). \tag{3.4.5}$$

It follows that if λ is an eigenvalue of a symplectic matrix, so is the reciprocal $1/\lambda$. [Note that according to Exercise 1.7, $\lambda = 0$ is not an eigenvalue, so we need not be concerned about multiplying or dividing by zero.] Consequently, the eigenvalues of a symplectic matrix must form reciprocal pairs. This property is called *reflexivity*.

The symmetry between λ and $1/\lambda$ exhibited by (4.5) can be further displayed by rewriting (4.5) in the form

$$\lambda^{-n}P(\lambda) = \lambda^n P(1/\lambda). \tag{3.4.6}$$

Now define another function $Q(\lambda)$ by writing

$$Q(\lambda) = \lambda^{-n}P(\lambda). \tag{3.4.7}$$

The functions P and Q evidently have the same zeroes. Moreover, the condition (4.6) requires that Q have the symmetry property

$$Q(\lambda) = Q(1/\lambda). \tag{3.4.8}$$

Equation (4.8) shows not only that the eigenvalues of a symplectic matrix must occur in reciprocal pairs; it shows that they must also occur with the same multiplicity. That is, if the root λ_0 has multiplicity k, so must the root $1/\lambda_0$. Indeed, the eigenvalues λ_0 and λ_0^{-1} must have the same Jordan block structure. See Exercise 4.6.

Also, if either $+1$ or -1 is a root, then this root must have *even* multiplicity. To see this, suppose for example that $\lambda = 1$ is a root. Introduce the variable μ by writing $\lambda = \exp\ \mu$. Then (4.8) shows that Q is an *even* function of the variable μ and hence near $\lambda = 1$ (near $\mu = 0$) Q must have an expansion of the form

$$Q = \sum_{m=0}^{\infty} c_m \mu^{2m}. \tag{3.4.9}$$

Moreover, when λ is near 1, λ and μ are related by the expansion

$$\mu = \log\lambda = \log[1 + (\lambda - 1)] = (\lambda - 1)[1 - (\lambda - 1)/2 + \cdots]. \tag{3.4.10}$$

Comparison of (4.9) and (4.10) shows that $\lambda = 1$ is not a root unless $c_0 = 0$. If $c_0 = 0$, then $\lambda = 1$ is a root of multiplicity 2. If $c_1 = 0$ as well, then $\lambda = 1$ is a root of multiplicity 4, etc. A similar argument holds near $\lambda = -1$ upon making the substitution $\lambda = -\exp\mu$.

In summary, it has been shown that the eigenvalues of a real symplectic matrix must satisfy the following properties:

1. They must be real or occur in complex conjugate pairs.

2. They must occur in reciprocal pairs, and each member of the pair must have the same multiplicity.

3. If either ± 1 is an eigenvalue, it must have even multiplicity.

When combined, the conditions just enumerated place strong restrictions on the possible eigenvalues of a real symplectic matrix. Among them is the fact that the the eigenvalues cannot all lie inside or all lie outside the unit circle. We will learn in later chapters that, by definition, the linear part of a symplectic map at a fixed point is a symplectic matrix. Therefore, fixed points of a symplectic map cannot be attractors or repellers. We will also learn that Hamiltonian systems produce symplectic maps. It follows that Hamiltonian systems have neither attractors nor repellers.

3.4.2 The 2×2 Case

Consider first the simplest case of a 2×2 symplectic matrix ($n = 1$). Call the eigenvalues λ_1 and λ_2. Then, by the reciprocal property, it follows that

$$\lambda_1 \lambda_2 = 1. \tag{3.4.11}$$

Suppose, now, that λ_1 is real, positive, and greater than 1. Then λ_2 is real, positive, and less than 1. Similarly, if λ_1 is real, negative, and less than -1, then λ_2 is real, negative, and greater than -1. On the other hand, if λ_1 is complex, then $\lambda_2 = \overline{\lambda}_1$. This condition, when combined with (4.11), shows that in this case λ_1 and λ_2 must lie on the unit circle in the complex plane. Finally, there are the two special cases $\lambda_1 = \lambda_2 = 1$ and $\lambda_1 = \lambda_2 = -1$.

Altogether, there are five possible cases. They are listed below along with names and designations whose significance will become clear later on. See also Figure 4.1.

1. Hyperbolic case (unstable): $\lambda_1 > 1$ and $0 < \lambda_2 < 1$.

2. Inversion hyperbolic case (unstable): $\lambda_1 < -1$ and $-1 < \lambda_2 < 0$.

3. Elliptic case (stable): $\lambda_1 = e^{i\phi}, \lambda_2 = e^{-i\phi}$. (Eigenvalues are complex conjugates and lie on the unit circle).

4. Parabolic case (generally linearly unstable): $\lambda_1 = \lambda_2 = +1$.

5. Inversion parabolic case (generally linearly unstable): $\lambda_1 = \lambda_2 = -1$.[5]

Note that in all cases both eigenvalues cannot lie inside the unit circle nor can both eigenvalues lie outside the unit circle.

3.4.3 The 4×4 and Remaining $2n \times 2n$ Cases

The next simplest case is that of a 4×4 symplectic matrix ($n = 2$). In this case, one has to deal with four possible eigenvalues and then apply reasoning analogous to the 2×2 case. Figures 4.2 illustrate the various possibilities that can occur. Analysis of the possible spectrum of the $2n$ eigenvalues for the general $2n \times 2n$ real symplectic matrix proceeds in a similar fashion. Again, in particular, it can never happen that all the eigenvalues lie inside the unit circle, nor can it happen that they all lie outside the unit circle. See Exercise 4.1.

[5] It is easily checked that inversion hyperbolic or inversion parabolic symplectic matrices can be written as the products of hyperbolic or parabolic symplectic matrices with the negative identity matrix, respectively. The negative identity matrix (which is also symplectic) acts on phase space to produce *inversion* through the origin. Some authors use the terminology *reflection* hyperbolic and *reflection* parabolic rather than *inversion* hyperbolic and *inversion* parabolic. This terminology is less precise: Reflection can mean reflection in/about the origin, in which case it is the same as inversion through the origin. In other contexts, reflections refer to transformations, like mirror reflections, that change the sign of some components of a vector, but not all components.

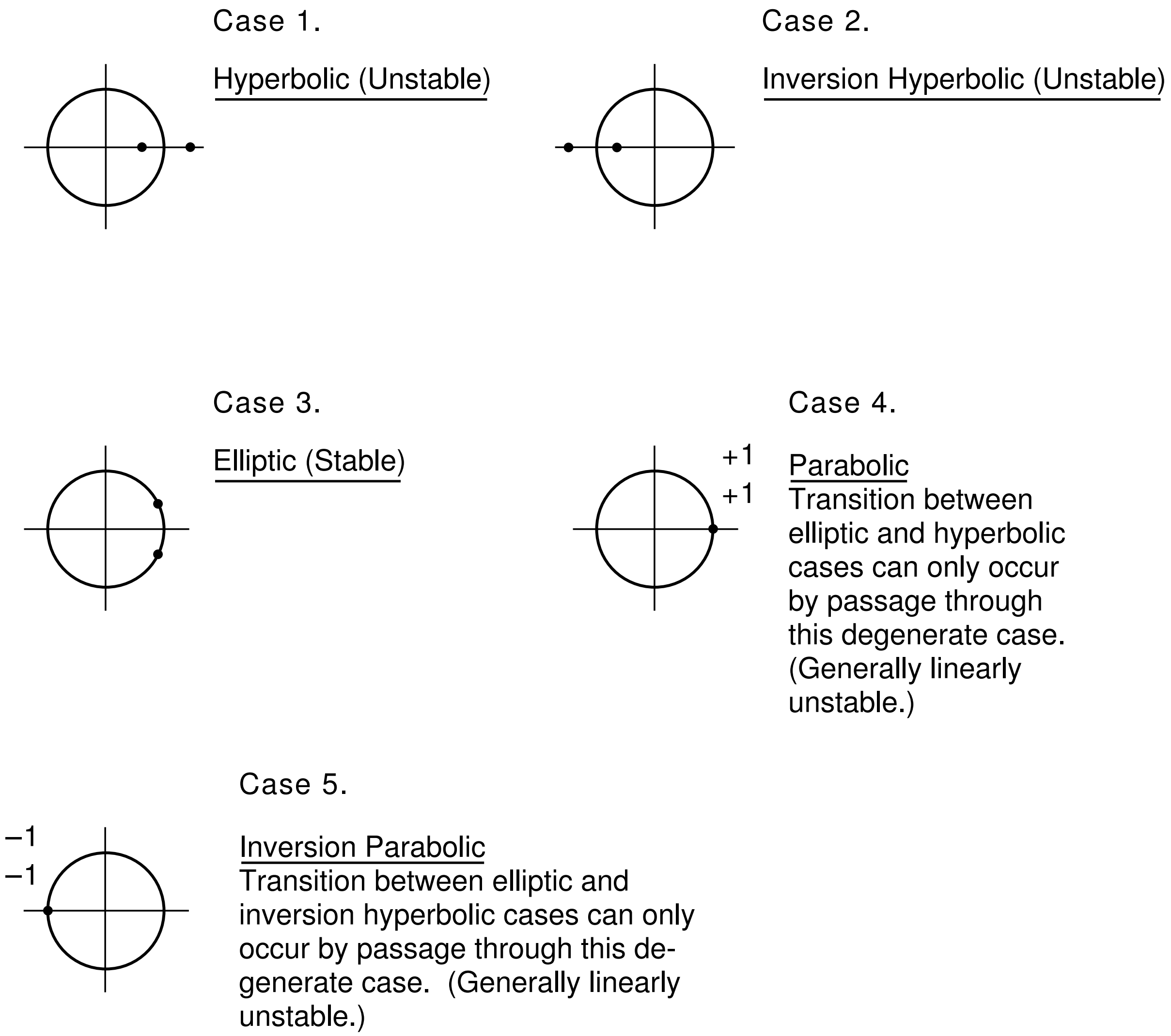
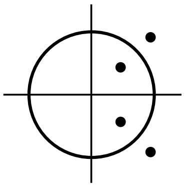

Case 1.
Hyperbolic (Unstable)
Case 2.
Inversion Hyperbolic (Unstable)
Case 3.
Elliptic (Stable)
Case 4.
+1
+1
Parabolic
Transition between elliptic and hyperbolic cases can only occur by passage through this degenerate case. (Generally linearly unstable.)
Case 5.
−1
−1
Inversion Parabolic
Transition between elliptic and inversion hyperbolic cases can only occur by passage through this degenerate case. (Generally linearly unstable.)

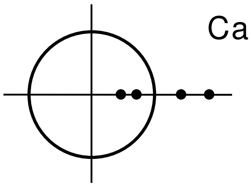
Figure 3.4.1: Possible cases for the eigenvalues of a 2×2 real symplectic matrix.

Figure 3.4.2: Possible eigenvalue configurations for a 4×4 real symplectic matrix. The mirror image of each configuration is also a possible configuration, and therefore is not shown in order to save space. Various authors have given these configurations various names. Notably, Case 1 is commonly called a *Krein quartet*.

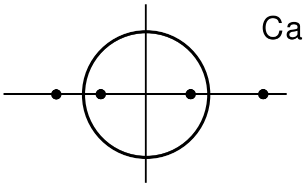
A. Generic Configurations



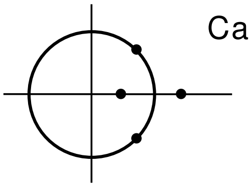
Case 1. All eigenvalues complex and off the unit circle. All eigenvalues can be obtained from a single one by the operations of complex conjugation and taking reciprocals. Unstable.



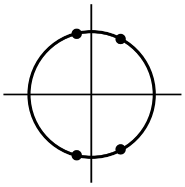
Case 2. All eigenvalues real, off the unit circle, and of same sign. Eigenvalues form reciprocal pairs. Unstable.



Case 3. All eigenvalues real, off the unit circle, and of differing sign. Eigenvalues form reciprocal pairs. Unstable.

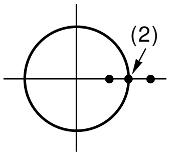


Case 4. Two eigenvalues complex and confined to unit circle. Two eigenvalues real. Eigenvalues form reciprocal pairs. Complex eigenvalues are also complex conjugate. Unstable.

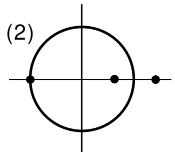


Case 5. All eigenvalues complex and confined to the unit circle. Eigenvalues form reciprocal pairs that are also complex conjugate. Stable.

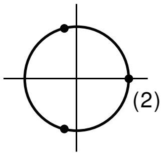
B. Degenerate Configurations. Transitions between generic configurations can only occur by passage through a degenerate configuration. Mirror image configurations are again possible, but not shown.



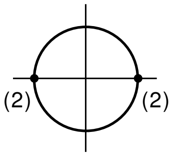
Case 1. Two eigenvalues equal, and two eigenvalues real. All of same sign. Occurs in transition between generic cases 2 and 4. Unstable.



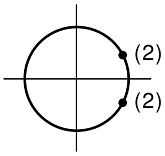
Case 2. Two eigenvalues equal, and two eigenvalues real. Signs differ. Occurs in transition between generic cases 3 and 4. Unstable.



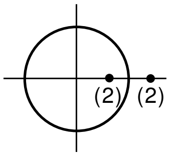
Case 3. Two eigenvalues equal, and two eigenvalues confined to unit circle. Occurs in transition between generic cases 4 and 5. Generally linearly unstable.



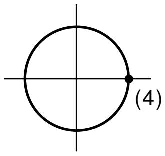
Case 4. Two eigenvalues equal +1 and two equal –1. Occurs in transition between generic cases 3 and 5, or 3 and 4, or 4 and 5. Also occurs in transition between degenerate cases 2 and 3. Generally linearly unstable.



Case 5. Two pairs of eigenvalues equal, and confined to unit circle. Occurs in transition between generic cases 1 and 5. Not, however, a sufficient condition to guarantee that such a transition is possible. Stability also undetermined in absence of further conditions.



Case 6. Two pairs of eigenvalues equal and real. Occurs in transition between generic cases 1 and 2. Unstable.



Case 7. All eigenvalues equal and have value ±1. Occurs in transitions between generic cases 1, 2, 4, and 5 and degenerate cases 1, 3, 5, and 6. Generally linearly unstable.

3.4.4 Further Symplectic Restrictions

Background

We will next see that the symplectic condition not only simplifies the computation of eigenvalues, but also influences how the eigenvalues depend on various parameters. For the general case of a $2n \times 2n$ matrix M, the characteristic polynomial (4.1) is of degree $2n$. Correspondingly, one might think that the determination of the eigenvalues from (4.2) would require finding the roots of a $2n$ degree polynomial. However, if M is symplectic, we can use the fact that P and Q have the same zeroes and the symmetry property (4.8) to reduce the problem to that of finding the roots of an n degree polynomial. Introduce a variable w by the relation

$$w = \lambda + 1/\lambda. \tag{3.4.12}$$

Equation (4.12) can be inverted to give the result

$$\lambda = [w \pm (w^2 - 4)^{1/2}]/2. \tag{3.4.13}$$

Since $P(\lambda)$ is a polynomial of degree $2n$, it follows from (4.4), (4.7), and (4.8) that Q must be of the form

$$Q(\lambda) = Q_r(w) = \sum_{m=0}^{n} b_m w^m \tag{3.4.14}$$

with

$$b_n = 1. \tag{3.4.15}$$

Note that Q_r, which we will call the *reduced* characteristic polynomial of M, has degree n. The eigenvalues of M can now be determined by finding the n roots of the equation

$$Q_r(w) = 0, \tag{3.4.16}$$

and substituting these roots into (4.13).

Let us see how the results just described work out in the cases $n = 1$ and $n = 2$.

The 2×2 Case

Suppose $n = 1$. Then, if M is symplectic, we have the result

$$P(\lambda) = \lambda^2 - A\lambda + 1, \tag{3.4.17}$$

with the coefficient (parameter) A given by the relation

$$A = \text{tr } (M). \tag{3.4.18}$$

For $Q_r(w)$ we find the result

$$Q_r(w) = w - A. \tag{3.4.19}$$

Evidently the solution to (4.16) in this case is simply $w = A$, and we find from (4.13) the eigenvalues

$$\lambda = [A \pm (A^2 - 4)^{1/2}]/2. \tag{3.4.20}$$

Note that (4.20) gives eigenvalues on the unit circle (stability) when

$$-2 < A < 2, \tag{3.4.21}$$

and real eigenvalues (instability) otherwise. Figure 4.3, which is to be compared with Figure 4.1, illustrates the nature of the eigenvalues λ as a function of A.

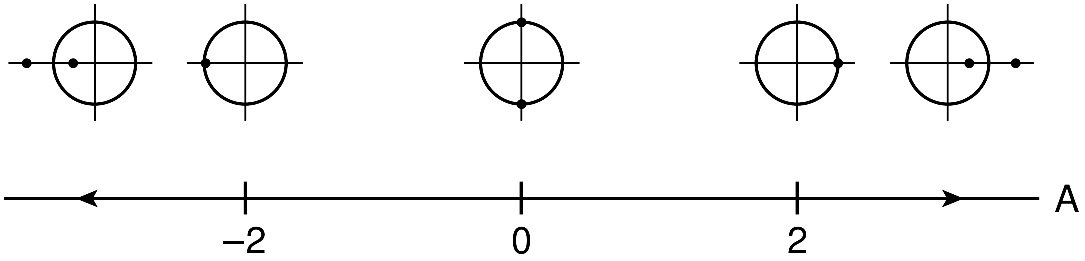


Figure 3.4.3: Eigenvalues of a 2×2 real symplectic matrix M as a function of $A = \text{tr}\ (M)$.

The 4×4 Case

Suppose $n = 2$. Then, if M is symplectic, the characteristic polynomial has the form

$$P(\lambda) = \lambda^4 - A\lambda^3 + B\lambda^2 - A\lambda + 1. \tag{3.4.22}$$

The coefficients (parameters) A and B can be found from the relations

$$A = [P(-1) - P(1)]/4, \tag{3.4.23}$$

$$B = [P(-1) + P(1)]/2 - 2. \tag{3.4.24}$$

They can also be found directly in terms of M using the relations

$$A = \text{tr}\ (M), \tag{3.4.25}$$

$$B = \{[\text{tr}\ (M)]^2 - \text{tr}\ (M^2)\}/2. \tag{3.4.26}$$

See Exercise 3.7.17. For $Q_r(w)$ we find the result

$$Q_r(w) = w^2 - Aw + B - 2. \tag{3.4.27}$$

The solutions to (4.16) in this case are

$$w = [A \pm (A^2 - 4B + 8)^{1/2}]/2. \tag{3.4.28}$$

These solutions are to be substituted into (4.13). Observe that in general there are four choices of signs to be made corresponding to the four possible eigenvalues expected for M. Figure 4.4, which is to be compared with Figure 4.2, illustrates the nature of the eigenvalues as a function of A and B. We note that the region of stability is the arrow-head shaped domain in which the following conditions are satisfied simultaneously:

$$B \geq 2A - 2\ ,\ B \geq -2A - 2,$$

$$B \le A^2/4 + 2\ ,\ B \le 6. \tag{3.4.29}$$

Transitions from stability to instability through the points $\lambda = \pm 1$ occur across the line segments

$$\lambda = +1 :\ B = 2A - 2 \text{ and } B \in [-2, 6], \tag{3.4.30}$$

$$\lambda = -1 :\ B = -2A - 2 \text{ and } B \in [-2, 6]. \tag{3.4.31}$$

Transitions to instability through Krein collisions, see Case 5 of Figure 4.2B and Section 3.5, occur across the parabolic segment

$$B = A^2/4 + 2, \tag{3.4.32}$$

with

$$A \in [-4, 4]. \tag{3.4.33}$$

The 6×6 Case

The 6×6 case can be treated in a manner analogous to the 2×2 and 4×4 cases. In the 6×6 case one needs to solve a cubic equation to find the eigenvalues. See Exercise 4.14.

Dimension Counting

We close this subsection with a remark on dimension counting. We see from (4.14) and (4.15) that the spectrum of a $2n \times 2n$ symplectic matrix is determined by the n parameters $b_0, b_1, \cdots b_{n-1}$. We will learn in Section 3.7 that symplectic matrices form a Lie group, and that the dimensionality of this group is $n(2n + 1)$. Since $n(2n + 1)$ is much larger than n, it follows that many different symplectic matrices have the same spectrum. This fact is relevant to accelerator design. As outlined at the beginning of this section, the linear stability of closed orbits in an accelerator is governed by the spectrum of the linear part of its one-turn transfer map. It is therefore important to be able to control the spectrum, and there are typically many knobs in an accelerator control room for this purpose. Despite these many knobs, accelerator operators often discover to their dismay that they are unable to adjust the spectrum at will. The dimension counting comparison tells us why. Much of the possible knob turning simply leads to different symplectic matrices having the same or nearly the same spectrum.

3.4.5 In Praise of and Gratitude for the Symplectic Condition

We have learned that the symplectic condition guarantees that there are symplectic maps $\mathcal{M}$ whose linear parts M have all their eigenvalues on the unit circle and distinct. Thus, it is in principle possible to build circular (ring) accelerators and storage rings with a stable closed orbit.

Moreover, the set of $2n \times 2n$ real symplectic matrices whose eigenvalues lie on the unit circle and are distinct is *open* in the set of all $2n \times 2n$ symplectic matrices. By this we mean that if M is a real symplectic matrix all of whose eigenvalues lie on the unit circle and are distinct, then the same is true of all real symplectic matrices sufficiently near M. To

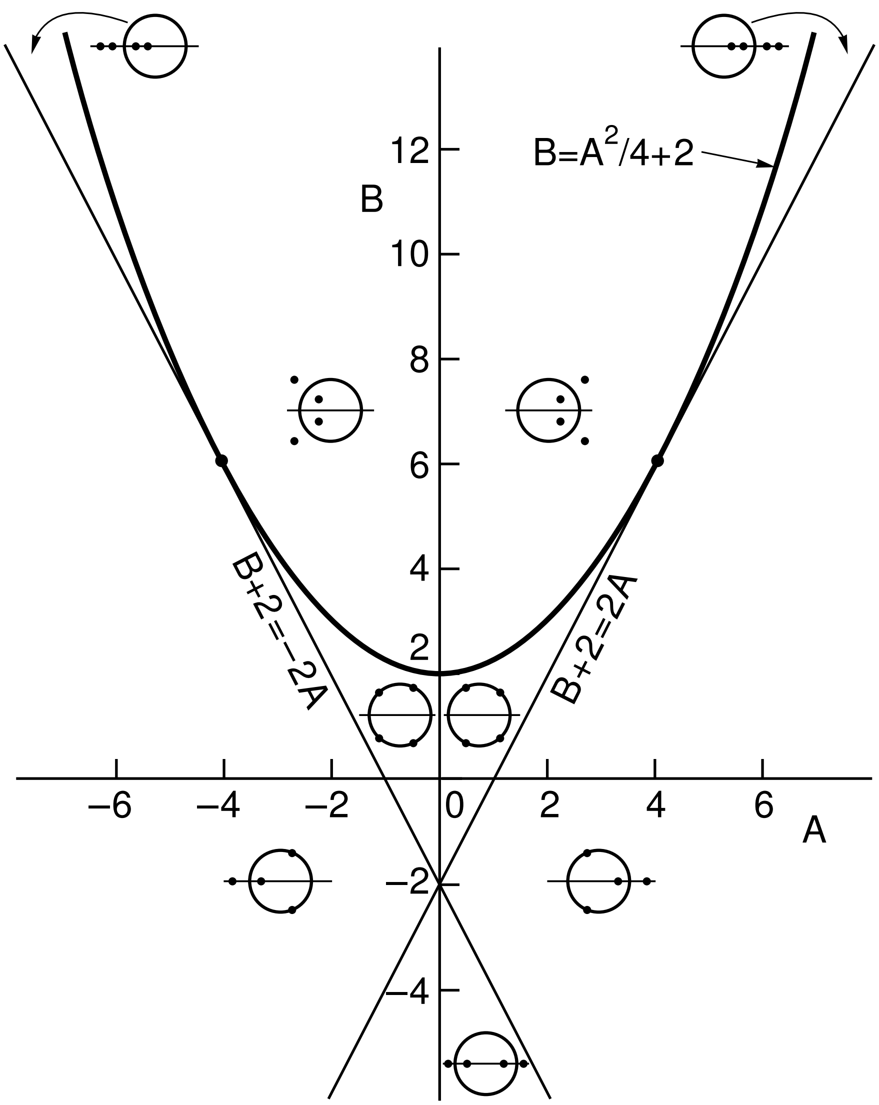


Figure 3.4.4: Eigenvalues of a 4×4 real symplectic matrix M as a function of the coefficients A and B in its characteristic polynomial.

see this, suppose that M is changed slightly, but in such a way that it remains symplectic. The eigenvalues of a matrix are the roots of a polynomial whose coefficients are continuous functions of the entries in the matrix. See (4.1) and (4.2). Also, the roots of a polynomial are continuous functions of the coefficients in the polynomial. It follows that the eigenvalues of a matrix are continuous functions of the entries in the matrix. That is, if the matrix is slightly changed, its eigenvalues are also only slightly changed. But if the eigenvalues are initially on the unit circle and distinct, there are no nearby eigenvalue configurations for a symplectic matrix where the eigenvalues are not all distinct or at least one eigenvalue is outside the unit circle. Thus, if the change in M is finite but small enough, then the eigenvalues must remain on the unit circle and must still be distinct.

The fact that this stability cannot be destroyed by small and symplectic perturbations should be of comfort to accelerator designers and builders because it means that, at least in the linear approximation, the stability of orbits will not be damaged by small errors in machine construction and operation. That is, thanks to the symplectic condition, accelerator performance is robust under small fabrication and control parameter errors.

Even more can be said. In our discussion we have implicitly assumed the existence of a closed orbit. That is, we have assumed that $\mathcal{M}$ has a fixed point z_f. It can be shown that if a symplectic map has a stable fixed point (a fixed point for which M has all its eigenvalues on the unit circle and distinct), then all nearby symplectic maps will also have a stable fixed point. See Subsection 29.4.5. Thus, if a circular accelerator or storage ring is *designed* to have a stable closed orbit, both the existence and the stability of a closed orbit for the actual machine are guaranteed even in the presence of small fabrication and control parameter errors providing these errors are not too large.

Exercises

3.4.1. Verify (4.5) starting with (4.3).

3.4.2. Show, using (3.33), that the eigenvalues of a symplectic matrix cannot all have absolute value less than 1, nor can they all have absolute value greater than 1.

3.4.3. Show that, for a real 4×4 symplectic matrix M, that all the generic eigenvalue configurations of Figure 4.2 are unchanged by small perturbations of M providing the perturbed M are also symplectic. That is why these configurations are called *generic.*

3.4.4. Show that the eigenvalues of J are all $\pm i$.

3.4.5. Suppose M' is defined in terms of M by (2.15). Show that M and M' have the same spectrum.

3.4.6. Suppose λ_0 is a complex eigenvalue of a real symplectic matrix. Show that the Jordan block structures for the eigenvalues λ_0 and $\overline{\lambda}_0$ are the same. Suppose λ_0 is an eigenvalue of a (possibly complex) symplectic matrix. Use (1.9) to show that the Jordan block structures for the eigenvalues λ_0 and λ_0^{-1} are the same.

3.4.7. Given (4.12), verify (4.13).

3.4.8. Using (4.4), (4.7), and (4.8), verify (4.14) and (4.15).

3.4.9. Verify (4.17) through (4.21). Let O and F be the matrices

$$O = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \text{ and } F = \begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix}. \tag{3.4.34}$$

The matrix O is the 2×2 version of the transfer matrix for a drift of length a, and the matrix F is the 2×2 version of the transfer matrix for a focusing element with focal length

$$f = -1/b. \tag{3.4.35}$$

Therefore we expect that $a > 0$ and (assuming F is focusing) $b < 0$. See Chapter 13. Suppose M is the 2×2 matrix

$$M = OFO. \tag{3.4.36}$$

It is the 2×2 version of the transfer matrix for an OFO cell. Verify that M and its factors in (4.37) are symplectic matrices. Verify that

$$A = \text{tr }(M) = 2(1 + ab). \tag{3.4.37}$$

By referring to Figures 4.1 and 4.3 verify that for M there are the following cases:

$$\begin{aligned} &\text{hyperbolic when } ab > 0, \\ &\text{elliptic when } -2 < ab < 0, \\ &\text{inversion hyperbolic when } ab < -2. \end{aligned} \tag{3.4.38}$$

Suppose M' is the matrix

$$M' = FOF. \tag{3.4.39}$$

It is the 2×2 version of the transfer matrix for a FOF cell. Verify that M' is symplectic. Verify that

$$A' = \text{tr }(M') = 2(1 + ab). \tag{3.4.40}$$

Verify that for M' there are also the cases (4.38).

3.4.10. Verify (4.22) through (4.24). Verify (4.27) and (4.28).

3.4.11. Study Figure 4.4. Verify the statements made in connection with (4.29) through (4.33).

3.4.12. Where do the eigenvalues λ lie when A and B are on the portion of the parabola (4.32) having $A > 4$ or $A < -4$? Where do the eigenvalues lie when A and B are on the portions of the lines $B = \pm 2A - 2$ and $B \notin [-2, 6]$? Find where the eigenvalues lie for the cases $A = 4$, $B = 6$; $A = -4$, $B = 6$; $A = 0$, $B = -2$.

3.4.13. Consider a 4-dimensional phase space. Suppose the phase-space variables are arranged according to (2.4) rather than (2.1). Verify that doing so makes no difference for the discussion of the present section. Suppose, with the arrangement (2.4), that a 4×4 symplectic matrix M is written in the 2×2 block form (3.1), and the blocks B and C are identically zero. In this case the x_1, y_1 space is mapped into itself, and the x_2, y_2 space is mapped into itself. See (2.4). With this assumption, show that the characteristic polynomial for M takes the form

$$P(\lambda) = (\lambda^2 - \alpha\lambda + 1)(\lambda^2 - \delta\lambda + 1). \tag{3.4.41}$$

Here α is the trace of the upper left 2×2 block in M, and δ is the trace of the lower right 2×2 block in M. Show that, according to (4.21), the quantities α, δ must lie in the square

$$-2 < \alpha < 2\ , -2 < \delta < 2 \tag{3.4.42}$$

in order that all eigenvalues of M lie on the unit circle (stability). Show that, when multiplied out, (4.41) takes the form

$$P(\lambda) = \lambda^4 - (\alpha + \delta)\lambda^3 + (2 + \alpha\delta)\lambda^2 - (\alpha + \delta)\lambda + 1. \tag{3.4.43}$$

Now compare (4.22) and (4.43) to get the results

$$A = \alpha + \delta, \tag{3.4.44}$$

$$B = 2 + \alpha\delta. \tag{3.4.45}$$

Show that the interior of the square (4.42) maps into the arrow-head shaped domain (4.29) of Figure 4.4 under the transformation given by (4.44) and (4.45). Also show that the exterior of the square maps to points outside the arrow-head shaped domain. Does one get all points outside the arrow-head shaped domain?

3.4.14. Consider a 6-dimensional phase space. Show that in this case $P(\lambda)$ for a symplectic matrix can be written in the form

$$P(\lambda) = \lambda^6 - A\lambda^5 + B\lambda^4 - C\lambda^3 + B\lambda^2 - A\lambda + 1, \tag{3.4.46}$$

and $Q_r(w)$ takes the form

$$Q_r(w) = w^3 - Aw^2 + (B - 3)w + (2A - C). \tag{3.4.47}$$

What region of the A, B, C parameter space gives stability (all eigenvalues on the unit circle)? That is, what is the 3-dimensional analog of the arrow-head shaped domain of Figure 4.4?

3.4.15. Look at the coefficients appearing in (4.46). Listing them from left to right, we see that they have the values $1, -A, B, -C, B, -A, 1$. This sequence is a *palindrome*. That is, when read backwards, the result is the same as reading forwards. Observe that this feature also appears in (4.17) and (4.22). Also observe that the first and last coefficients always have the value $+1$. Prove that these results hold for any phase-space dimension.

3.5 Eigenvector Structure, Normal Forms, and Stability

3.5.1 Eigenvector Basis

Let M be a $2n \times 2n$ real symplectic matrix. Suppose its eigenvalues are all distinct. Call them $\lambda_1, \lambda_2, \cdots \lambda_{2n}$, and call the associated eigenvectors $\psi_1, \psi_2, \cdots \psi_{2n}$. Then we have $2n$ relations of the form

$$M\psi_j = \lambda_j \psi_j. \tag{3.5.1}$$

Note that if any λ_j is complex, the corresponding ψ_j must also have complex entries. Finally, since the λ_j are assumed to be distinct, the $2n$ vectors ψ_j must be linearly independent, and must consequently form a basis.

3.5.2 J-Based Angular Inner Product

Let (,) denote the usual *complex* scalar product.[6] Introduce an *angular inner product* $\langle , \rangle$ by the rule

$$\langle \chi, \theta \rangle = (\chi, K\theta) \tag{3.5.2}$$

with K defined by the relation

$$K = -iJ. \tag{3.5.3}$$

Here χ and θ are any two vectors. We note that K is Hermitian,

$$K^{\dagger} = K, \tag{3.5.4}$$

with respect to the standard complex scalar product (,). Consequently, we have the relation

$$\langle \theta, \chi \rangle = \overline{\langle \chi, \theta \rangle}. \tag{3.5.5}$$

Finally we observe that for real vectors the angular inner product, apart from a factor of $-i$, is just the fundamental symplectic 2-form (2.3).[7]

3.5.3 Use of Angular Inner Product

What is the angular inner product good for? Let ψ_j and ψ_k be two eigenvectors. We have the result

$$\begin{aligned} \langle \psi_j, M\psi_k \rangle &= (\psi_j, KM\psi_k) = \lambda_k (\psi_j, K\psi_k) \\ &= \lambda_k \langle \psi_j, \psi_k \rangle. \end{aligned} \tag{3.5.6}$$

From the symplectic condition (1.2) we conclude that

$$KM = (M^T)^{-1} K. \tag{3.5.7}$$

[6] We adopt the usual physicists' convention that $(\alpha\phi, \beta\psi) = \bar{\alpha}\beta(\phi, \psi)$. Mathematicians frequently follow the convention that $(\alpha\phi, \beta\psi) = \alpha\bar{\beta}(\phi, \psi)$.

[7] Apart from a factor of $-i$, the angular inner product (5.2) is sometimes called the Lagrange bracket of $\bar{\chi}$ and θ.

Consequently, the quantity $\langle\psi_j, M\psi_k\rangle$ can also be written in the form

$$\begin{aligned}\langle\psi_j, M\psi_k\rangle &= (\psi_j, KM\psi_k) = (\psi_j, (M^T)^{-1}K\psi_k) \\ &= (M^{-1}\psi_j, K\psi_k) = \overline{\lambda}_j^{-1}(\psi_j, K\psi_k) \\ &= \overline{\lambda}_j^{-1}\langle\psi_j, \psi_k\rangle. \end{aligned} \tag{3.5.8}$$

Here we have used the relation

$$M^{-1}\psi_j = \lambda_j^{-1}\psi_j, \tag{3.5.9}$$

which follows from (5.1). Comparison of the relations (5.6) and (5.8) gives the result

$$(\overline{\lambda}_j^{-1} - \lambda_k)\langle\psi_j, \psi_k\rangle = 0. \tag{3.5.10}$$

Consequently, we have the *orthogonality* relation

$$\langle\psi_j, \psi_k\rangle = 0 \;\text{ if }\; \overline{\lambda}_j^{-1} \neq \lambda_k. \tag{3.5.11}$$

The exact consequences of the orthogonality relation depend on the nature of the spectrum. Suppose, for the purposes of this section, that all the eigenvalues of M are complex, distinct, and lie on the unit circle. Then the λ_j can be written in the form

$$\lambda_j = e^{i\phi_j} \tag{3.5.12}$$

where the phases ϕ_j are real. In this case we have the relation

$$\overline{\lambda}_j^{-1} = \lambda_j. \tag{3.5.13}$$

Correspondingly, the orthogonality relation (5.11) becomes the relation

$$\langle\psi_j, \psi_k\rangle = 0 \;\text{ if }\; \lambda_j \neq \lambda_k. \tag{3.5.14}$$

Further, we claim that

$$\langle\psi_k, \psi_k\rangle \neq 0 \;\text{ for all }\; k. \tag{3.5.15}$$

For suppose $\langle\psi_k, \psi_k\rangle$ did vanish for some k. Then, by (5.14), $\langle\psi_j, \psi_k\rangle$ would vanish for all j. Correspondingly, as a result of the definitions (5.2) and (5.3), we would conclude that

$$(\psi_j, J\psi_k) = 0 \;\text{ for all }\; j, \tag{3.5.16}$$

and consequently

$$J\psi_k = 0 \tag{3.5.17}$$

since the ψ_j form a basis. But the matrix J is invertible. Thus (5.17) implies that ψ_k itself must vanish, which is impossible because the vectors ψ_j form a basis.

3.5.4 Definition and Use of Signature

Observe that, according to (5.5), the quantities $\langle\psi_j,\psi_j\rangle$ must be real. Since they cannot vanish, they must be positive or negative. Suppose we rephase and renormalize the vectors ψ_j to produce new vectors ψ_j' defined by the relations

$$\psi_j' = r_j e^{i\chi_j}\psi_j. \tag{3.5.18}$$

Here the r_j are real, positive quantities, and the phases χ_j are arbitary. We then find the relations

$$\langle\psi_j',\psi_j'\rangle = r_j^2\langle\psi_j,\psi_j\rangle. \tag{3.5.19}$$

We see that the r_j can be selected in such a way that

$$\langle\psi_j',\psi_j'\rangle = \sigma_j \tag{3.5.20}$$

where σ_j has the (possible) values

$$\sigma_j = \pm 1. \tag{3.5.21}$$

Note that the sign of σ_j is independent of the phase χ_j. Thus, the sign is an *intrinsic* property of the vector ψ_j. It is called the *signature* of ψ_j. From now on, we drop the prime notation. With this understanding, and recalling that the λ_j are assumed to be distinct, we may require that the ψ_j be normalized in such a way that they obey the orthogonality relation

$$\langle\psi_j,\psi_k\rangle = \sigma_j\delta_{jk}. \tag{3.5.22}$$

Consider some eigenvalue λ_k. Since $\overline{\lambda}_k$ is also an eigenvalue, it must be one of the λ_j. Let $\lambda_{k'}$ denote this particular λ_j. Then, we have the relations

$$M\psi_k = \lambda_k\psi_k, \tag{3.5.23}$$

$$M\psi_{k'} = \lambda_{k'}\psi_{k'} = \overline{\lambda}_k\psi_{k'}. \tag{3.5.24}$$

Complex conjugate (5.23) to get the relation

$$M\overline{\psi}_k = \overline{\lambda}_k\overline{\psi}_k = \lambda_{k'}\overline{\psi}_k. \tag{3.5.25}$$

We observe from (5.24) and (5.25) that $\psi_{k'}$ and $\overline{\psi}_k$ are both eigenvectors of M with the same eigenvalue $\lambda_{k'}$. Consider the vector $\overline{\psi}_k$. By the same argument that led to (5.14), it must be orthogonal to all ψ_j with $j \neq k'$. Since the ψ_j form a basis, it follows that $\overline{\psi}_k$ must be proportional to $\psi_{k'}$. Thus, there is a relation of the form

$$\overline{\psi}_k = \alpha_k\psi_{k'} \tag{3.5.26}$$

where α_k is some proportionality constant yet to be determined. Consider the quantity $\sigma_k = \langle\psi_k,\psi_k\rangle$. Working it out in component form gives the result

$$\sigma_k = \langle\psi_k,\psi_k\rangle = \sum_\alpha \overline{\psi}_{k,\alpha}(K\psi_k)_\alpha = \sum_{\alpha,\beta}\overline{\psi}_{k,\alpha}K_{\alpha\beta}\psi_{k,\beta}. \tag{3.5.27}$$

Here the quantities $\psi_{k,\beta}$ are the components of ψ_k. Next consider the quantity $\langle\overline{\psi}_k,\overline{\psi}_k\rangle$. It evidently has the value

$$\begin{aligned}\langle\overline{\psi}_k,\overline{\psi}_k\rangle &= \sum_{\alpha,\beta}\psi_{k,\alpha}K_{\alpha\beta}\overline{\psi}_{k,\beta} = -\sum_{\alpha,\beta}\overline{\psi}_{k,\beta}K_{\beta\alpha}\psi_{k,\alpha} \\ &= -\langle\psi_k,\psi_k\rangle = -\sigma_k. \end{aligned} \tag{3.5.28}$$

Here use has been made of the antisymmetry of K. But from (5.26) we have the relation

$$\langle\overline{\psi}_k,\overline{\psi}_k\rangle = |\alpha_k|^2\langle\psi_{k'},\psi_{k'}\rangle = |\alpha_k|^2\sigma_{k'}. \tag{3.5.29}$$

Comparison of (5.28) and (5.29) gives the result

$$|\alpha_k|^2\sigma_{k'} = -\sigma_k. \tag{3.5.30}$$

Two relations follow from (5.30) and (5.21). First, we have the relation

$$\sigma_{k'} = -\sigma_k. \tag{3.5.31}$$

We have learned that if $\lambda_{k'} = \overline{\lambda}_k$, then ψ_k and $\psi_{k'}$ have opposite signature. It follows that half the ψ_j have signature $+1$, and half have signature -1. Second, we also have the relation

$$|\alpha_k|^2 = 1. \tag{3.5.32}$$

Thus, α_k must be just a phase factor. Since the vectors ψ_k and $\psi_{k'}$ are only defined up to overall phase factors, we may set $\alpha_k = 1$ without loss of generality to get the relation

$$\overline{\psi}_k = \psi_{k'}. \tag{3.5.33}$$

The preceding discussion makes it possible to improve our notation. Suppose the ψ_j are relabeled in such a way that the vectors ψ_ℓ with $\ell = 1, 2, \cdots n$ have positive signature. Let the corresponding ψ_j with negative signature be labeled as $\psi_{-\ell}$. That is, arrange the labeling scheme so that the following relations hold with $\ell, m = 1, 2, \cdots n$:

$$\langle\psi_\ell,\psi_m\rangle = \delta_{\ell,m}, \tag{3.5.34}$$

$$\langle\psi_{-\ell},\psi_{-m}\rangle = -\delta_{\ell,m}, \tag{3.5.35}$$

$$\langle\psi_\ell,\psi_{-m}\rangle = \langle\psi_{-\ell},\psi_m\rangle = 0, \tag{3.5.36}$$

$$\overline{\lambda}_\ell = \lambda_{-\ell}, \tag{3.5.37}$$

$$\overline{\psi}_\ell = \psi_{-\ell}. \tag{3.5.38}$$

By their association with the ψ_ℓ, the eigenvalues λ_ℓ are also said to have positive signature. Correspondingly, the eigenvalues $\lambda_{-\ell}$ are said to have negative signature.

3.5.5 Definition of Phase Advances and Tunes

Consider the eigenvalues λ_ℓ corresponding to the vectors ψ_ℓ having *positive* signature. That is, consider the eigenvalues with positive signature. Define phases ϕ_ℓ by the relation

$$\lambda_\ell = e^{i\phi_\ell}. \tag{3.5.39}$$

Evidently these phases are defined modulo 2π. For the discussion that follows, it is convenient to take them to lie in the range $(-\pi, \pi)$. The quantities ϕ_ℓ with $\ell = 1, 2, \cdots n$ are called the *phase advances* of M. Also, define corresponding quantities T_ℓ by the relations

$$T_\ell = \phi_\ell/(2\pi). \tag{3.5.40}$$

Evidently the T_ℓ are defined modulo 1, but our choice of the range $(-\pi, \pi)$ for phases places the T_ℓ in the range $(-1/2, 1/2)$. The quantities T_ℓ are called the *tunes* of M.[8]

Example 5.1: Let M be the 2×2 matrix

$$M = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix}. \tag{3.5.41}$$

Since M is 2×2 and has determinant $+1$, it must be symplectic. See Exercise 1.3. A simple calculation shows that M has the eigenvectors

$$\psi_{+1} = (1/\sqrt{2}) \begin{pmatrix} 1 \\ i \end{pmatrix}, \tag{3.5.42}$$

$$\psi_{-1} = (1/\sqrt{2}) \begin{pmatrix} 1 \\ -i \end{pmatrix}, \tag{3.5.43}$$

with eigenvalues $e^{+i\phi}$ and $e^{-i\phi}$, respectively. Also, it is easily checked that ψ_{+1} and ψ_{-1} have signatures $+1$ and -1, respectively. It follows that the phase advance of M is ϕ, and the tune is $\phi/(2\pi)$.

3.5.6 The Krein-Moser Theorem and Krein Collisions

The discussion so far has been restricted to the case for which the eigenvalues of M are distinct and lie on the unit circle. Suppose M is varied in such a way that two eigenvalues collide. In actuality (when $n \geq 2$), two pairs must collide. See Case 5 of the degenerate configurations of Figure 4.2. Then, as M is varied further, the eigenvalues can pass over each other to give Case 5 of the generic configurations, or they can leave the unit circle to give Case 1 of the generic configurations. See Figure 5.1. It can be shown that if the

[8] In Chapter 30 we will see that if M can be viewed as the product of many symplectic matrices, all of which are near the identity, then the phase advances and tunes of M can be defined in such a way that they may lie *outside* the ranges $(-\pi, \pi)$ and $(-1/2, 1/2)$, respectively. However, modulo 2π or 1, respectively, these phase advances and tunes still agree with those defined above.

colliding eigenvalues have the *same* signature, which is the case of nearly equal tunes, then they cannot leave the unit circle and must pass over each other. Also, when the eigenvalues do collide, M remains diagonalizable even though its eigenvalues are no longer distinct. This result is called the *Krein-Moser* theorem or condition.

The same signature case is the case of nearly equal tunes. By contrast, if the eigenvalues have opposite signatures, then there are small perturbations of M that will cause the eigenvalues to collide and then leave the unit circle thereby forming a Krein quartet. Such a collision is called a *Krein collision.* This is the case of nearly equal and opposite tunes. For a proof of these assertions, see Exercise 3.8.18.

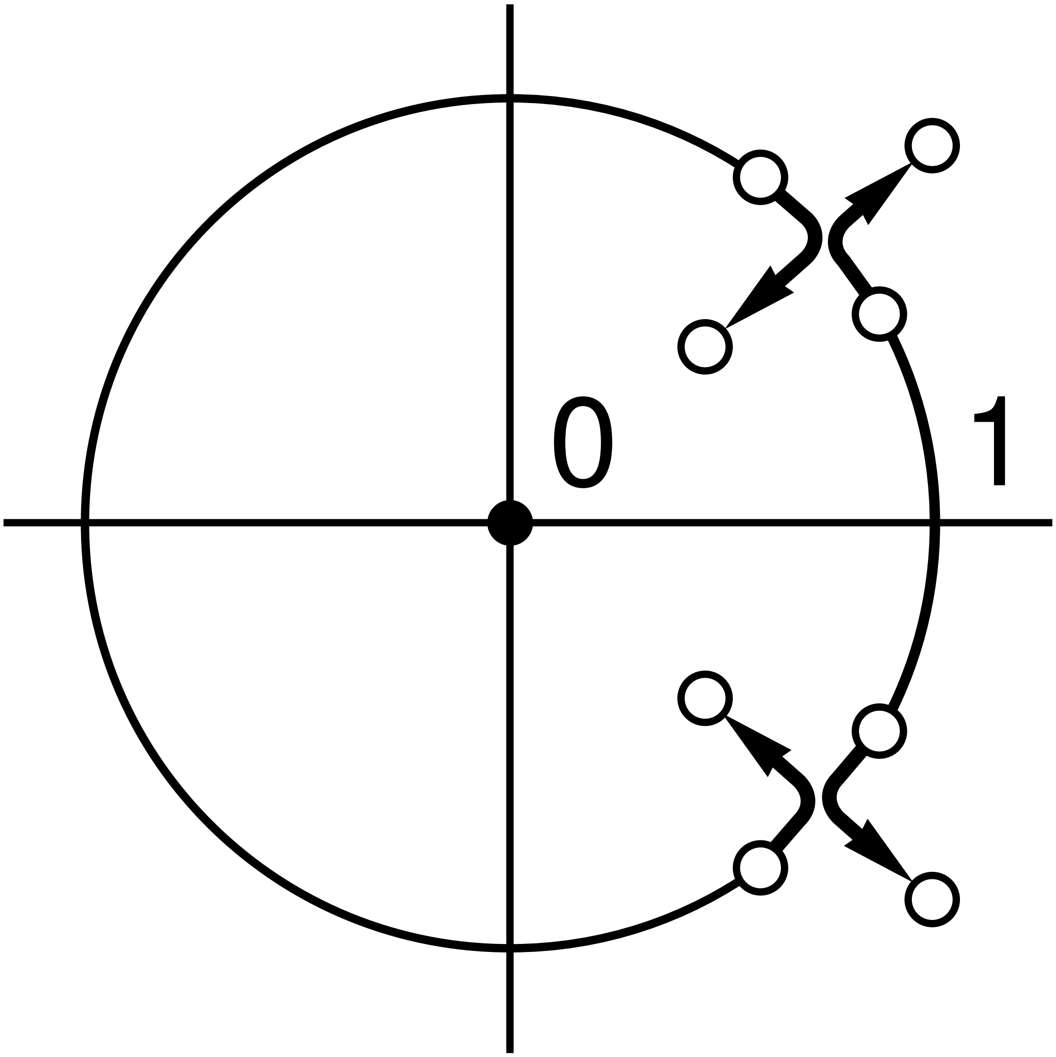


Figure 3.5.1: Illustration of eigenvalues colliding and then leaving the unit circle to form what is called a Krein quartet.

The Krein-Moser theorem is a remarkable result. Consider, for example, the case $n = 2$ corresponding to 4×4 symplectic matrices M. Suppose M is such that two eigenvalues of the *same* signature collide. In this case the values A and B given by (4.25) and (4.26) must lie on the parabolic segment specified by (4.32) and (4.33). Now consider all symplectic matrices near M. They form a 10-dimensional space. See (7.35). Compute the values of A and B for these matrices. According to the Krein-Moser theorem, *all* these values must also lie on the parabolic segment or in the arrow-head shaped domain of Figure 4.4 *below* the parabolic segment, and *none* will be *above* the parabolic segment. No matter which way we move in M space (at least locally), we *cannot* get points in the A, B image space that lie above the parabolic segment. By contrast, suppose we go to another region of M space where two eigenvalues again collide, but now have *opposite* signature. In this case the A, B values will again lie on the parabolic segment. However, if we now consider all symplectic matrices near M, we will find that their image in A, B space lies on the parabolic segment, or in the arrow-head shaped domain, *or* in the region immediately *above* the parabolic segment. Thus in this case, by making a proper move in M space, we *can* get anywhere (locally) in A, B space.

Recall that the situation in which all the eigenvalues lie on the unit circle and are distinct corresponds to stability (in the linear approximation), and the case of any eigenvalue off the unit circle corresponds to instability. See the discussion in the beginning of Section 3.4 Consequently, to achieve qualitative *insensitivity* to small perturbations, the case of nearly equal and opposite tunes should be avoided.[9]

3.5.7 Normal Forms

The last topic to be discussed in this section is that of a *normal form* for M. As will be seen, a normal form for M is a particularly simple form for M achieved by a symplectic similarity transformation. Recall that we have assumed that all eigenvalues are complex, distinct, and lie on the unit circle. Thus, we restrict our attention here to this case. (Normal forms are also known for the other cases, but their discussion is more complicated.) We also assume, without loss of generality, that M is symplectic with respect to the J of (2.10).

Suppose the eigenvectors ψ_ℓ are decomposed into real and imaginary parts by writing the relations

$$\psi_\ell = \xi_\ell + i\eta_\ell, \tag{3.5.44}$$

where the vectors ξ_ℓ and η_ℓ are real. From (5.38) we conclude that the $\psi_{-\ell}$ have the decomposition

$$\psi_{-\ell} = \xi_\ell - i\eta_\ell. \tag{3.5.45}$$

Insert the representations (5.44) and (5.45) into (5.34) and (5.36), and equate real and imaginary parts. Doing so, and use of (5.3), gives the results

$$(\xi_\ell, J\xi_m) = 0 \quad , \quad (\eta_\ell, J\eta_m) = 0, \tag{3.5.46}$$

$$2(\xi_\ell, J\eta_m) = \delta_{\ell m} \quad , \quad 2(\eta_\ell, J\xi_m) = -\delta_{\ell m}. \tag{3.5.47}$$

Also insert the representation (5.44) into the relation

$$M\psi_\ell = \lambda_\ell \psi_\ell = e^{i\phi_\ell}\psi_\ell, \tag{3.5.48}$$

[9] Instability can also occur if the eigenvalues are on the unit circle but M cannot be diagonalized. This can occur when the eigenvalues are ±1 and as well as at Krein collisions. Moreover, the eigenvalues can also leave the unit circle through these degenerate configurations. Therefore, integer and half-integer tunes should also be avoided.

and equate real and imaginary parts. Doing so gives the result

$$M\xi_\ell = (\cos\phi_\ell)\xi_\ell - (\sin\phi_\ell)\eta_\ell, \tag{3.5.49}$$

$$M\eta_\ell = (\sin\phi_\ell)\xi_\ell + (\cos\phi_\ell)\eta_\ell. \tag{3.5.50}$$

Consider the matrix A defined by the equation

$$A = \sqrt{2}(\xi_1, \eta_1, \xi_2, \eta_2, \cdots \xi_n, \eta_n). \tag{3.5.51}$$

Here each of the vectors ξ_ℓ and η_ℓ are to be viewed as column vectors so that the collection (5.51) forms a real $2n \times 2n$ matrix. Then it is easily verified that the relations (5.46) and (5.47) are equivalent to the matrix relation

$$A^T J A = J \tag{3.5.52}$$

providing the form (2.10) is employed for J. Thus, A is a symplectic matrix with respect to this J.

Finally, consider the matrix N defined by the equation

$$N = A^{-1} M A. \tag{3.5.53}$$

The matrix MA can be computed using (5.49), (5.50), and (5.51). One finds the result

$$\begin{aligned} MA &= \sqrt{2}(M\xi_1, M\eta_1, \cdots M\xi_n, M\eta_n) \\ &= \sqrt{2}(c_1\xi_1 - s_1\eta_1, s_1\xi_1 + c_1\eta_1, \cdots c_n\xi_n - s_n\eta_n, s_n\xi_n + c_n\eta_n). \end{aligned} \tag{3.5.54}$$

Here use has been made of the abbreviations

$$c_\ell = \cos\phi_\ell \quad , \quad s_\ell = \sin\phi_\ell. \tag{3.5.55}$$

Since A is symplectic, the matrix A^{-1} may be formed using (1.9),

$$A^{-1} = -JA^T J. \tag{3.5.56}$$

With this observation, we can continue the calculation. From (5.54) we find that JMA has the representation

$$JMA = \sqrt{2}(c_1J\xi_1 - s_1J\eta_1, s_1J\xi_1 + c_1J\eta_1, \cdots c_nJ\xi_n - s_nJ\eta_n, s_nJ\xi_n + c_nJ\eta_n). \tag{3.5.57}$$

The matrix $A^T JMA$ can now be computed using (5.46), (5.47), (5.51), and (5.57). The result is

$$A^T JMA = \begin{pmatrix} B_1 & & & \\ & B_2 & & \\ & & \ddots & \\ & & & B_n \end{pmatrix}. \tag{3.5.58}$$

That is, all entries are zero save for n 2×2 blocks on the diagonal. The blocks themselves are given by the equations

$$B_\ell = \begin{pmatrix} -\sin\phi_\ell & \cos\phi_\ell \\ -\cos\phi_\ell & -\sin\phi_\ell \end{pmatrix}. \tag{3.5.59}$$

Finally, $N = A^{-1}MA = -JA^TJMA$ can be computed by applying $-J$ to (5.58). The result is

$$N = \begin{pmatrix} R_1 & & & \\ & R_2 & & \\ & & \ddots & \\ & & & R_n \end{pmatrix}. \tag{3.5.60}$$

Again, all entries in N are zero save for n 2×2 blocks on the diagonal. The blocks themselves are given by the equations

$$R_\ell = \begin{pmatrix} \cos\phi_\ell & \sin\phi_\ell \\ -\sin\phi_\ell & \cos\phi_\ell \end{pmatrix}. \tag{3.5.61}$$

We conclude that, given any (real) symplectic matrix M whose eigenvalues are distinct and all lie on the unit circle, there is then a real symplectic similarity transformation (5.53) that brings M to the simple form (5.60). We call N the *normal form* of M, and say that M has been brought to normal form by the transforming matrix A. To reiterate, we have the key relations

$$N = A^{-1}MA \tag{3.5.62}$$

and

$$M = ANA^{-1}. \tag{3.5.63}$$

We also observe that the normal form is unique up to permutations of the ϕ_ℓ. There is somewhat more freedom available in the choice of the transforming matrix A. This freedom will be discussed later in Section 23.*. Finally, if we consider a two-dimensional phase space $z_\ell = (q_\ell; p_\ell)$ and define the action of R_ℓ as

$$z'_\ell = R_\ell z_\ell, \tag{3.5.64}$$

then we find the relations

$$q'_\ell = q_\ell \cos\phi_\ell + p_\ell \sin\phi_\ell, \tag{3.5.65}$$

$$p'_\ell = -q_\ell \sin\phi_\ell + p_\ell \cos\phi_\ell. \tag{3.5.66}$$

We see that the effect of R_ℓ is a clockwise rotation in the $(q_\ell; p_\ell)$ plane by the phase-advance angle ϕ_ℓ. Evidently, each R_ℓ, and therefore also N, is a real orthogonal matrix.

We close this subsection by remarking that there are also normal forms for symplectic matrices whose eigenvalues lie on the unit circle but are not distinct, or some or all of whose eigenvalues do not lie on the unit circle. The discussion of the general case is quite complicated, and falls outside the scope of this book. For a discussion of the 2×2 case, see Exercise 5.7. Further information may be found in the references listed at the end of this chapter. See also Subsection 27.2.2.

3.5.8 Stability

Suppose, as sketched at the beginning of Section 3.4, that a map $\mathcal{M}$ acts on some space with coordinates z and suppose $\mathcal{M}$ has a fixed point z_f,

$$\mathcal{M} z_f = z_f. \tag{3.5.67}$$

In this subsection we will verify some of the claims made in Section 3.4 about the repeated action of $\mathcal{M}$ on points near z_f.

A point near z_f can be written in the form $z_f + \delta$ where δ is a small vector. By the definition of *linear part* we assume the existence of an expansion of the form

$$\mathcal{M}(z_f + \delta) = z_f + M\delta + O(\delta^2) \tag{3.5.68}$$

where the matrix M describes the linear part of $\mathcal{M}$ about z_f. It follows from repeated application of (5.68) that

$$\mathcal{M}^m(z_f + \delta) = z_f + M^m\delta + O(\delta^2). \tag{3.5.69}$$

Therefore, to analyze the stability of z_f in the linear approximation, we must examine the behavior of $M^m\delta$ for large m.

It can be shown that if all the eigenvectors of M lie within the unit circle in the complex plane, then

$$\lim_{m\to\infty} M^m = 0. \tag{3.5.70}$$

See Exercise 5.10. Therefore in this case, and neglecting terms of order δ^2, we find that

$$\lim_{m\to\infty} \mathcal{M}^m(z_f + \delta) = z_f. \tag{3.5.71}$$

Thus, in this case and in linear approximation, z_f is an attractor.[10]

For the case of symplectic maps, we have seen that not all eigenvalues of M can lie within the unit circle. For symplectic maps we are interested in the next best possibility, the case where all eigenvalues lie on the unit circle. Suppose all the eigenvalues of M lie on the unit circle and are distinct. Then, employing (5.63), we may write

$$M^m = (ANA^{-1})^m = AN^mA^{-1}. \tag{3.5.72}$$

Next, with the aid of vector and matrix norms, we find that

$$||M^m\delta|| = ||AN^mA^{-1}\delta|| \le ||A||\;||N^m||\;||A^{-1}||\;||\delta||. \tag{3.5.73}$$

If we use the Euclidean norm, see Exercise 7.1, and observe that N^m is orthogonal, we find the result

$$||N^m|| = (2n)^{1/2}. \tag{3.5.74}$$

Here we have used the group property that since N is $2n \times 2n$ and orthogonal, then so is N^m. See Subsection 6.1. Combining (5.73) and (5.74) gives the estimate

$$||M^m\delta||| \le (2n)^{1/2}||A||\;\;||A^{-1}||\;||\delta||. \tag{3.5.75}$$

We conclude that $M^m\delta$ remains bounded for all m, and therefore z_f is stable in the linear approximation.[11]

[10] According to a theorem of *Hartman*, in this case z_f is also an attractor even if all nonlinear terms are taken onto account.

[11] For a discussion of what occurs when the effect of the neglected nonlinear terms is concluded, see Chapter 35.

One might wonder whether the requirement that the eigenvalues be distinct is essential. It is. There are unstable counter examples for which the eigenvalues lie on the unit circle but are not distinct. The simplest 2 × 2 case is the matrix

$$M = \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix} \tag{3.5.76}$$

where α is any nonzero real number. Since this M is 2×2 and has determinant $+1$, it is symplectic. Also, it has the non-distinct eigenvalue are $+1$, and no eigenvalues off the unit circle. However, it is easily verified that

$$M^m = \begin{pmatrix} 1 & m\alpha \\ 0 & 1 \end{pmatrix}. \tag{3.5.77}$$

Therefore, if

$$\delta = \begin{pmatrix} 0 \\ \epsilon \end{pmatrix} \tag{3.5.78}$$

where ϵ is any small number, we have the result

$$M^m\delta = \begin{pmatrix} m\alpha\epsilon \\ \epsilon \end{pmatrix}. \tag{3.5.79}$$

We see that in this case $M^m\delta$ has entries that grow *linearly* in m as m increases, and therefore we may say that the fixed point z_f is linearly unstable.

We close this subsection with a simple example for which one eigenvalue is outside the unit circle and for which z_f is manifestly unstable. Consider the symplectic 2 × 2 case

$$M = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix} \tag{3.5.80}$$

where $\lambda > 1$. In this case, if

$$\delta = \begin{pmatrix} \epsilon \\ 0 \end{pmatrix}, \tag{3.5.81}$$

we find the result

$$M^m\delta = \begin{pmatrix} \lambda^m\epsilon \\ 0 \end{pmatrix}. \tag{3.5.82}$$

Note that

$$\lambda^m = \exp(m \log \lambda) \tag{3.5.83}$$

and $\log \lambda > 0$ when $\lambda > 1$. Thus, now $M^m\delta$ has entries that grow *exponentially* in m as m increases, and therefore we may say that the fixed point z_f is exponentially unstable.

By the above examples we have demonstrated that there are cases where instability occurs when the eigenvalues are on the unit circle but not distinct, or some eigenvalue lies outside the unit circle. These result holds in general, and can be proved with the aid of normal forms for these cases.

Exercises

3.5.1. Show that if M is any matrix with distinct eigenvalues, then the corresponding eigenvectors must form a basis.

3.5.2. Carry out the calculations required for Example (5.1).

3.5.3. Show that if two tunes of a symplectic matrix M, call them T_1 and T_2, are nearly equal (modulo the integers), then there are two associated eigenvalues that are nearly equal and have the same signature, and vice versa. In this case we have a relation of the form

$$T_1 - T_2 \simeq n \tag{3.5.84}$$

where n is an integer, and say that we are dealing with a potential *difference* resonance. By the Krein-Moser theorem, we know that under perturbation M remains diagonalizable and its eigenvalues remain on the unit circle. Therefore a difference resonance is harmless as far as stability is concerned.

Show that if two tunes are nearly equal and opposite (again modulo the integers), then there are two related eigenvalues that are nearly equal and have opposite signatures, and vice versa. In this case we have a relation of the form

$$T_1 + T_2 \simeq n \tag{3.5.85}$$

where n is an integer, and say that we are dealing with a potential *sum* resonance. By the Krein-Moser theorem, we know that in this case the eigenvalues can leave the unit circle under perturbation of M, and therefore a sum resonance is likely harmful.

3.5.4. Verify (5.46), (5.47), (5.49), and (5.50).

3.5.5. Verify (5.52)

3.5.6. Verify (5.54) and (5.57) through (5.61).

3.5.7. Suppose M and N are two matrices that are related by an equation of the form (5.53) where A is yet another matrix. If such a relation exists for some (invertible) matrix A, the matrices M and N are said to be *conjugate*, and we write $M \sim N$. It can be shown that conjugacy is an *equivalence* relation. This equivalence relation can be used to partition the set of all matrices into disjoint *equivalence classes*, which in this case are called *conjugacy classes*. See Exercise 5.12.7. Suppose M and N are symplectic, and a symplectic A can be found such that (5.53) holds. Then we will say that M and N are *symplectically conjugate*. Consider the case of all 2×2 symplectic matrices. Suppose that two such matrices, call them M and N, have the same Jordan normal form. Show that they are then symplectically conjugate. Hint: See the comment following Exercises 1.2 and 1.3. Show that the matrices

$$M = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{3.5.86}$$

$$N = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \tag{3.5.87}$$

are symplectically conjugate, and find the conjugating matrix A.

3.5.8. Let χ and θ be any two (possibly complex) vectors, and let M be any real symplectic matrix. Define transformed vectors χ' and θ' by the rule

$$\chi' = M\chi, \tag{3.5.88}$$

$$\theta' = M\theta. \tag{3.5.89}$$

Show that the inner product $\langle\ ,\ \rangle$ defined by (5.2) has the invariance property

$$\langle\chi'\ ,\ \theta'\rangle = \langle\chi\ ,\ \theta\rangle. \tag{3.5.90}$$

3.5.9. Take matrix elements of (1.2) using the eigenvectors (5.1) to obtain the relation

$$(\psi_j, M^T JM\psi_k) = (\psi_j, J\psi_k). \tag{3.5.91}$$

Verify the manipulations

$$\begin{aligned}(\psi_j, M^T JM\psi_k) &= (M\psi_j, JM\psi_k) = (\lambda_j\psi_j, J\lambda_k\psi_k) = \bar{\lambda}_j\lambda_k(\psi_j, J\psi_k)\\ &= i\bar{\lambda}_j\lambda_k(\psi_j, K\psi_k) = i\bar{\lambda}_j\lambda_k\langle\psi_j, \psi_k\rangle,\end{aligned} \tag{3.5.92}$$

$$(\psi_j, J\psi_k) = i(\psi_j, K\psi_k) = i\langle\psi_j, \psi_k\rangle. \tag{3.5.93}$$

Show that (5.91) through (5.93) yield the result

$$(\bar{\lambda}_j\lambda_k - 1)\langle\psi_j, \psi_k\rangle = 0. \tag{3.5.94}$$

Verify that (5.94) is equivalent to (5.10) and (5.11).

3.5.10. Suppose that M is any matrix all of whose eigenvalues lie inside the unit circle. The aim of this exercise is to prove (5.70). Begin by assuming the eigenvalues of M are distinct. In this case there is an invertible matrix A such that

$$M = ADA^{-1} \tag{3.5.95}$$

where D is a diagonal matrix with the eigenvalues of M on its diagonal. Show from (5.95) that

$$M^m = AD^mA^{-1}. \tag{3.5.96}$$

Verify that

$$\lim_{m\to\infty} \lambda^m = 0 \text{ if } |\lambda| < 1, \tag{3.5.97}$$

and therefore

$$\lim_{m\to\infty} D^m = 0, \tag{3.5.98}$$

and thus, from (5.96), (5.70) holds.

If the eigenvalues of M are not distinct, it may not be diagonalizable. If M is not diagonalizable, it may still be brought to *Jordan* normal form,

$$M = ANA^{-1}, \tag{3.5.99}$$

so that we may write

$$M^m = AN^mA^{-1}. \tag{3.5.100}$$

Here N is a matrix having all zeroes except for possessing the eigenvalues of M on the diagonal and possibly ones just above the diagonal. For example, if M is 4×4 and not diagonalizable and all eigenvalues are the same, the most degenerate case would be that for which N has the form

$$N = \begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{pmatrix}. \tag{3.5.101}$$

In this case write

$$N = D + K \tag{3.5.102}$$

where D is diagonal and K is the matrix

$$K = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{3.5.103}$$

Verify that K is *nilpotent*, and in particular satisfies the relation

$$K^4 = 0. \tag{3.5.104}$$

By the binomial theorem for commuting entities, show that

$$\begin{aligned} N^m &= (D+K)^m \\ &= D^m + mD^{m-1}K + [m(m-1)/2!]D^{m-2}K^2 + [m(m-1)(m-2)/3!]D^{m-3}K^3. \end{aligned} \tag{3.5.105}$$

Verify that each term in (5.105) vanishes in the limit $m \to \infty$ if $|\lambda| < 1$, and thus, from (5.100), (5.70) holds. Proof of (5.70) for the general nondiagonalizable case follows similarly.

3.5.11. Scan Subsection 3.7.1 and Exercise 7.1. Verify (5.74) for the case of the Euclidean norm. Verify that for the spectral norm

$$||N^m|| = 1. \tag{3.5.106}$$

3.6 Group Properties, Dyadic and Gram Matrices, and Bases

In this section we will describe what a group is, and will show that symplectic and orthogonal matrices form groups. Closely related to the symplectic and orthogonal groups are special bases called sympletic and orthonormal bases. The treatment of bases is facilitated by the introduction of dyadic and Gram matrices. Finally, given some basis, we will explore ways of specifying associated orthonormal and symplectic bases.

3.6.1 Group Properties

Abstract Groups

Arnold once asked and answered

> What is a group? Algebraists teach that this is supposedly a set with two operations that satisfy a load of easily-forgettable axioms$\cdots$.

We will begin with the abstract definition of a group. Then we will define matrix groups.

Abstractly, a group G is a set of elements subject to some rule of combination, usually called multiplication. For the moment, let us denote multiplication by the symbol $\circ$. Then we require the following properties:

1. If M and N are in G, so is the product $M \circ N$.

2. Multiplication is associative, $L \circ (M \circ N) = (L \circ M) \circ N$.

3. G contains a unique *identity* element I such that $I \circ M = M \circ I = M$ for all M in G.

4. If M is in G, there is a unique *inverse* element M^{-1} that is also in G such that $M \circ M^{-1} = M^{-1} \circ M = I$.

We remark that requirements 3 and 4 above can be weakened. For example, requirement 3 can be weakened to just require that there are left and right identity elements. These elements can then be proven to be unique and the same. Also, requirement 4 can be weakened to just require that there are left and right inverses. These elements can then be proven to be unique and the same.

A subgroup H of G is a subset of G whose elements also satisfies the above group properties. Any group G always has the identity element I as a subgroup. Whether it has any other nontrivial subgroups depends on the nature of G.

Matrix Groups

For matrices (assumed to be $n \times n$) we may take for the combination (multiplication) rule the ordinary operation of matrix multiplication. This automatically makes multiplication associative. Also, we may take for the identity element the identity matrix I, and for the inverse element the inverse matrix. With these provisos, a set of $n \times n$ matrices G forms a group if it satisfies the following properties:

1. If M and N are in G, so is the product MN.

2. The identity matrix I is in G.

3. If M is in G, M^{-1} exists and is also in G.

Note, in the matrix case, that iff a matrix M satisfies $\det(M) \neq 0$, then there is a unique matrix denoted by M^{-1} such that $MM^{-1} = M^{-1}M = I$.

Evidently, according to Exercise 1.4, Equation (1.9), and Exercises 1.5 and 1.6, the set of all $2n \times 2n$ symplectic matrices (for any particular value of n) forms a group. This group is often denoted by the symbol $Sp(2n)$. More precisely, if we are working with *real* symplectic matrices, they form a group denoted by $Sp(2n, \mathbb{R})$; and if we are working with *complex* symplectic matrices, they form a group denoted by $Sp(2n, \mathbb{C})$. Where there is no possibility of confusion, we will use the notation $Sp(2n)$ to mean $Sp(2n, \mathbb{R})$.[12]

We remark that the symplectic condition (1.2) is a set of *algebraic* (polynomial) relations among the entries in M. For this reason, the symplectic group is an *algebraic group*.

An $n \times n$ matrix O that satisfies the condition

$$O^T O = I \tag{3.6.1}$$

is called *orthogonal*. It is easy to check that the set of all such matrices also forms a group called the orthogonal group, and denoted by the symbols $O(n, \mathbb{R})$ or $O(n, \mathbb{C})$ depending on the choice of field. Evidently, the orthogonal group is also an algebraic group. From (6.1) it follows that orthogonal matrices have the property

$$\det(O) = \pm 1. \tag{3.6.2}$$

Since the determinant of a matrix is a continuous function of the entries in the matrix, we conclude that the set of orthogonal matrices consists of two disjoint (and disconnected) subsets: those orthogonal matrices having determinant +1, and those having determinant −1. The subset of all orthogonal matrices with determinant +1 (called *proper* orthogonal matrices) forms a connected subgroup of the orthogonal group. This subgroup is called the *special* orthogonal group, and is referred to by the symbols $SO(n, \mathbb{R})$ or $SO(n, \mathbb{C})$. Note that the condition (6.1) can be written in the expanded form

$$O^T I O = I, \tag{3.6.3}$$

and this form is analogous to (1.2) with J replaced by I. Also, compare (6.1) and (3.1.14). This analogy results in some similarities in the ways that $O(n)$ and $Sp(2n)$ can be analyzed. However, in another sense, the two groups are polar opposites because I is symmetric and J is antisymmetric.

We remark for future use that the matrix J is both symplectic and special orthogonal. That is, J belongs both to $Sp(2n, \mathbb{R})$ and $SO(2n, \mathbb{R})$. See Exercises 1.1 and 1.5.

At this point one might wonder about generality. According to Exercise 2.7, the symplectic group consists of all linear transformations that preserve the fundamental symplectic 2-form (2.3). The matrix J in this 2-form has the property that it is antisymmetric and nonsingular. What happens if one replaces J by any (but real) antisymmetric nonsingular matrix? Does one still get a group, and is this group something new, or merely the symplectic group in disguise? Section 3.12 shows that one simply gets a variant of the symplectic group. It follows that the group $Sp(2n, \mathbb{R})$ is as general as might be desired.

Transformation Groups

We close this subsection with the comment that many groups arise naturally as *transformation groups*. Let $\mathcal{Z}$ be some set/space and consider mappings/transfomations $\mathcal{M}$ of $\mathcal{Z}$ into

[12] Warning! Some authors, particularly Mathematicians, use the notation $Sp(2n)$ to denote $USp(2n)$, the *unitary symplectic* group. See Section 5.10. Some other authors use $Sp(n)$ to stand for $Sp(2n, \mathbb{R})$ or $USp(2n)$.

itself. Two such mappings may be combined by letting them act on $\mathcal{Z}$ successively, and this composition operation may be taken to be a rule for multiplying mappings. By the nature of composition, this rule automatically satisfies the associative property, and thus a set of mappings of $\mathcal{Z}$ into itself has the potential of forming a group. Naturally, we will require that the product of any two mappings in the set will also be in the set. Furthermore, we may take the identity mapping $\mathcal{I}$, which leaves each element of $\mathcal{Z}$ unchanged, to be the identity element in the potential group. Finally, if require that every mapping in the set have an inverse, we may regard the set as forming a group.

By this definition we see that all matrix groups are transformation groups because each group element is also a transformation of some vector space into itself. That is, in this case, $\mathcal{Z}$ is some vector space. Moreover, in Section 5.12, we will learn that the symplectic group can also be viewed as providing a set of transformations of a generalized upper half plane, called a *Siegel space*, into itself. In this case, $\mathcal{Z}$ is a generalized upper half plane. And, as described in Chapter 6, the group of all symplectic maps is a transformation group with $\mathcal{Z}$ being phase space.

Finally, an abstract group G can always be thought of acting on itself by left or right multiplication or both:

$$h \rightarrow gh,\ h \rightarrow hg^{-1},\ h \rightarrow ghg^{-1};\ g \in G,\ h \in \mathcal{Z} = G.$$

Here g is any element in G; and h is any element in $\mathcal{Z}$, where, in fact, $\mathcal{Z}$ is also G. Thus, by any of these constructions, every group can also be viewed as being a transformation group.

3.6.2 Dyadic and Gram Matrices, Bases and Reciprocal Bases

The remaining concern of this section is a study of bases. To do so, we will first develop the tools of dyadic and Gram matrices, and then apply them in the study of various bases.

Suppose we are given a set of *real* linearly independent vectors w^1, w^2, $\cdots$ w^N. By definition, such a set constitutes a basis for an N-dimensional vector space. Let e^1,e^2, $\cdots$ e^N denote the standard column unit vectors

$$e^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix},\ \ e^2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix},\ \ e^3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix},\ \ \text{etc.} \tag{3.6.4}$$

Define a linear operator W by the rule

$$We^j = w^j. \tag{3.6.5}$$

Its matrix elements elements are given by the relation

$$W_{ij} = (e^i, We^j) = (e^i, w^j) \tag{3.6.6}$$

where $(,)$ denotes the usual scalar product but *without* complex conjugation. (Indeed, in the following, all vectors and matrices will be assumed to be *real*. And, in any symplectic context, all vector spaces will be assumed to be *even* dimensional.) In view of (6.4) and (6.5), W can be written in terms of the w^j in the form

$$W = (w^1, w^2, w^3, \cdots w^N) \tag{3.6.7}$$

where each w^j is regarded as a column vector so that the collection (6.7) forms an $N \times N$ matrix. Since the w^j are assumed to be linearly independent, we must have the relation

$$\det W \neq 0, \tag{3.6.8}$$

and we conclude that W^{-1} exists. Thus, for every *real* invertible $N \times N$ matrix W there is a basis of *real* vectors $w^1, w^2, \cdots w^N$, and vice versa.

Let us also use the notation $|w^j)$ to denote the column vector w^j, and let $(w^j|$ denote its dual row vector. With this notation we define the *dyadic* matrix $D(W)$ associated with the w^j by the rule

$$D(W) = \sum_k |w^k)(w^k|. \tag{3.6.9}$$

Simple matrix manipulation shows that D can also be written in the form

$$D(W) = WW^T. \tag{3.6.10}$$

See Exercise 6.3. Next define the *Gram* matrix $G(W)$ by the rule

$$G_{ij} = (w^i, w^j). \tag{3.6.11}$$

Matrix manipulation shows that G can also be written in the form

$$G(W) = W^T W. \tag{3.6.12}$$

See Exercise 6.4.

We have seen that for every real basis there are associated real dyadic and Gram matrices D and G. It follows from (6.10) and (6.12) that D and G are conjugate under the action of W,

$$W^{-1}DW = G \text{ and } WGW^{-1} = D. \tag{3.6.13}$$

We also note that D and G are symmetric,

$$D^T = D,\ G^T = G. \tag{3.6.14}$$

D and G are also invertible, and have positive determinant if W is real,

$$\det D = \det G = \det(W)\det(W^T) = (\det W)^2 > 0. \tag{3.6.15}$$

[We remark that $\det W$ is the (oriented) volume V of the parallelepiped with edges w^j, and consequently (6.15) is equivalent to the statement $\det D = \det G = V^2$.] Finally, we see from their forms (6.10) and (6.12) that both D and G are positive definite if W is real. That is, we have the relations

$$(v, Dv) > 0,\ (v, Gv) > 0 \tag{3.6.16}$$

for any real nonzero vector v.

We have also seen that a basis w^j specifies an associated nonsingular matrix W. Using this matrix, define the related vectors ${}^r w^j$ by the rule

$$ {}^r w^j = (W^{-1})^T e^j. \tag{3.6.17}$$

In view of (6.5) we may also write

$$ {}^r w^j = (W^{-1})^T (W^{-1}) w^j. \tag{3.6.18}$$

The ${}^r w^j$ have the pleasing property

$$ ({}^r w^i, w^j) = ([W^{-1}]^T e^i, W e^j) = (e^i, W^{-1} W e^j) = (e^i, e^j) = \delta_{ij}, \tag{3.6.19}$$

and are called the *reciprocal* basis to the original basis. Note that in analogy to (6.5) we may write

$$ {}^r W e^j = {}^r w^j \tag{3.6.20}$$

with

$$ {}^r W = (W^{-1})^T. \tag{3.6.21}$$

That is, the columns of ${}^r W$ are the ${}^r w^j$. Also, since there is the relation

$$ {}^r({}^r W) = \{[(W^{-1})^T]^{-1}\}^T = W, \tag{3.6.22}$$

it follows that the reciprocal basis of the reciprocal basis is the original basis.

It is easily verified that D and G have the property

$$ D({}^r W) = [D(W)]^{-1},\ G({}^r W) = [G(W)]^{-1}. \tag{3.6.23}$$

Also suppose U, V, and W are all invertible matrices. Define ${}^r U$ and ${}^r V$ in analogy to (6.21). Then the relation

$$ W = UV \tag{3.6.24}$$

implies the relation

$$ {}^r W = {}^r U \, {}^r V, \tag{3.6.25}$$

and vice versa. Thus, group properties are preserved under the r operation.

Simple calculation shows that the identity operator I has two dyadic representations in terms of the original and reciprocal bases,

$$ I = \sum_j |{}^r w^j)(w^j|, \tag{3.6.26}$$

$$ I = \sum_j |w^j)({}^r w^j|. \tag{3.6.27}$$

See Exercise 6.5. The representations (6.26) and (6.27) can be used to expand an arbitrary vector v in terms of either the w^j basis or the ${}^r w^j$ basis. From (6.26) we have the result

$$ v = Iv = \sum_j |{}^r w^j)(w^j, v), \tag{3.6.28}$$

which is an expansion of v in the reciprocal basis. From (6.27) we have the relation

$$v = Iv = \sum_j |w^j)({}^r w^j, v), \tag{3.6.29}$$

which is an expansion of v in the w^j basis.

Finally, it is interesting to have dyadic representations for W and W^{-1}. From (6.5) we find the representation

$$W = \sum_j |w^j)(e^j|. \tag{3.6.30}$$

Similarly, (6.17) gives the result

$$(W^{-1})^T = \sum_j |{}^r w^j)(e^j|, \tag{3.6.31}$$

from which it follows that

$$W^{-1} = \sum_j |e^j)({}^r w^j|. \tag{3.6.32}$$

3.6.3 Orthonormal and Symplectic Bases

So far we have been discussing general bases and their associated reciprocal bases. Now we want to consider two special kinds of bases: *orthonormal* bases and *symplectic* bases. A set of vectors v^j is called an orthonormal basis if it has the property

$$(v^i, v^j) = \delta_{ij}. \tag{3.6.33}$$

A set of vectors v^j (now necessarily *even* in number) is called a symplectic basis if it has the property

$$(v^i, Jv^j) = J_{ij}. \tag{3.6.34}$$

(Note that the basis set e^j, assuming the e^j are even in number, is both orthonormal and symplectic.) We shall discuss the properties of these two kinds of bases in turn.

Orthonormal Bases

We begin with orthonormal bases. Suppose a set of *real* vectors v^j satisfies (6.33). Then it follows that they are linearly independent, and therefore entitled to be called a basis. For suppose there is a relation of the form

$$\sum_j \alpha_j v^j = 0. \tag{3.6.35}$$

Then, using (6.33), we find

$$\sum_j \alpha_j (v^i, v^j) = \alpha_i = 0 \text{ for all } i. \tag{3.6.36}$$

As before, define a linear operator V by writing

$$v^j = Ve^j. \tag{3.6.37}$$

Then we find that V has the property

$$(e^i, V^T V e^j) = (Ve^i, Ve^j) = (v^i, v^j) = \delta_{ij}, \tag{3.6.38}$$

or, in matrix notation,

$$V^T V = I. \tag{3.6.39}$$

Thus, V is an orthogonal matrix. Conversely, suppose V is orthogonal, and define vectors v^j using (6.37). Then we find

$$(v^i, v^j) = (Ve^i, Ve^j) = (e^i, V^T V e^j) = (e^i, e^j) = \delta_{ij}, \tag{3.6.40}$$

and conclude that the v^j form an orthonormal basis. Put another way, the columns of an orthogonal matrix are orthonormal, and any matrix whose columns are orthonormal is orthogonal. Moreover, since the transpose of an orthogonal matrix is also orthogonal, the rows of an orthogonal matrix are orthonormal; and any matrix whose rows are orthonormal is orthogonal.

As an immediate consequence of the orthogonality condition (6.40) and the definitions of D and G we have the results

$$D(V) = G(V) = I. \tag{3.6.41}$$

Moreover, from (6.18) and (6.39) we see that an orthonormal basis is self reciprocal,

$$^r v^j = v^j. \tag{3.6.42}$$

Next suppose R is an orthogonal matrix and that the v^j form an orthonormal basis. Then the vectors u^j defined by

$$u^j = Rv^j \tag{3.6.43}$$

also form an orthonormal basis. Indeed, we find

$$(u^i, u^j) = (Rv^i, Rv^j) = (v^i, R^T R v^j) = (v^i, v^j) = \delta_{ij}. \tag{3.6.44}$$

Conversely, any two orthonormal bases u^j and v^j are related by a *unique* orthogonal transformation R. Indeed, from (6.37) and the analogous relation

$$u^j = Ue^j, \tag{3.6.45}$$

we find the result

$$u^j = Ue^j = UV^{-1} v^j = Rv^j \tag{3.6.46}$$

with

$$R = UV^{-1}. \tag{3.6.47}$$

Since orthogonal matrices form a group, we conclude that R is also orthogonal. What we have shown is that the orthogonal group acts *transitively* on the set of orthonormal bases.

Symplectic Bases

Now consider symplectic bases (in which case the dimensionality must be *even*). The discussion in this case has many parallels to the orthonormal case. Suppose a set of *real* vectors v^j satisfies (6.34). Suppose there is also an alleged linear dependency (6.35). Then use of (6.34) gives the relation

$$0 = \sum_j \alpha_j (v^i, Jv^j) = \sum_j J_{ij}\alpha_j = (J\alpha)_i \text{ for all } i. \tag{3.6.48}$$

Since J is invertible, it must again be the case that all α_j vanish, and the v^j must be linearly independent.

In analogy with the orthogonal case, there is a close connection between symplectic bases and symplectic matrices. Given a symplectic basis v^j we define V using (6.37) and find the relation

$$(e^i, V^T JV e^j) = (Ve^i, JVe^j) = (v^i, Jv^j) = J_{ij}, \tag{3.6.49}$$

and conclude that V is symplectic,

$$V^T JV = J. \tag{3.6.50}$$

Conversely, if V is symplectic and the v^j are defined by (6.37), then the v^j comprise a symplectic basis:

$$(v^i, Jv^j) = (Ve^i, JVe^j) = (e^i, V^T JV e^j) = (e^i, Je^j) = J_{ij}. \tag{3.6.51}$$

Put another way, the columns of a symplectic matrix form a symplectic basis; and any matrix whose columns form a symplectic basis is symplectic. Moreover, since the transpose of a symplectic matrix is also symplectic, the rows of a symplectic matrix form a symplectic basis; and any matrix whose rows form a symplectic basis is symplectic.

Next suppose R is a symplectic matrix and that the v^j form a symplectic basis. Then the vectors u^j defined by (6.43) also form a symplectic basis:

$$(u^i, Ju^j) = (Rv^i, JRv^j) = (v^i, R^T JRv^j) = (v^i, Jv^j) = J_{ij}. \tag{3.6.52}$$

Conversely, any two symplectic bases are related by a unique symplectic transformation. Consideration of this matter again leads to (6.45) through (6.47) with U and V now being symplectic matrices. Since symplectic matrices form a group, we conclude that the R given by (6.47) is also symplectic. What we have shown now is that the symplectic group acts *transitively* on the set of symplectic bases.

There are a few remaining observations to be made about the symplectic case. From the V analog of (6.17) we see that the ${}^r v^j$ form a symplectic basis if the v^j form a symplectic basis. Also, from their definitions and the symplectic group properties, we see that D and G are both symplectic; and (6.13) shows that they are symplectically conjugate. Finally, suppose that a basis v^j is *both* orthonormal and symplectic. In this case the V appearing in (6.37) is both orthogonal and symplectic. It will be shown in Section 3.9 that such V also form a group, which is in fact the unitary subgroup $U(n)$ of $Sp(2n, \mathbb{R})$.[13] It follows that all bases that are both orthonormal and symplectic are in one-to-one correspondence with the elements of $U(n)$, and $U(n)$ acts transitively on the set of all such bases.

[13] Unitary matrices are defined in Subsection 7.6. As illustrated in Section 3.9, there are *real* matrices whose group properties are those of $U(n)$.

3.6.4 Construction of Orthonormal Bases

The remainder of this section will be devoted to two questions: Given a set of *real* basis vectors w^j, how can one construct from them a set of basis vectors v^j that is either orthonormal or symplectic? Strictly speaking, as just posed, these questions are meaningless. Thanks to our previous discussion we already know that orthonormal and symplectic bases exist, and we pretty much know all about them. That is, using the methods to be covered later in this and subsequent chapters, we are able to manufacture all orthogonal or symplectic matrices. A better question is this: Given some *real* basis w^j, are there natural or useful ways of associating particular orthonormal or symplectic bases with the given w^j basis? We will begin with and mostly consider the orthonormal case.

Gram-Schmidt Orthogonalization

Given a set of *real* basis vectors w^j, a common procedure for constructing an orthonormal basis is to use *Gram-Schmidt* orthogonalization.[14] Starting with w^1, we construct intermediate vectors, call them u^j, and then final normalized vectors v^j by the rules

$$\begin{aligned}
&u^1 = w^1,\ v^1 = u^1/ \parallel u^1 \parallel;\\
&u^2 = w^2 - (v^1, w^2)v^1,\ v^2 = u^2/ \parallel u^2 \parallel;\\
&u^3 = w^3 - (v^1, w^3)v^1 - (v^2, w^3)v^2,\ v^3 = u^3/ \parallel u^3 \parallel;\\
&\vdots\\
&u^N = w^N - (v^1, w^N)v^1 \cdots - (v^{N-1}, w^N)v^{N-1},\ v^N = u^N/ \parallel u^N \parallel . \qquad (3.6.53)
\end{aligned}$$

Here we employ the usual notation

$$\parallel u^j \parallel = [(u^j, u^j)]^{1/2}, \qquad (3.6.54)$$

and note that all the scalar products appearing in (3.6.53) and (33,6,54) are *real* scalar products. It is easy to verify that at each step there is the relation $\parallel u^j \parallel \neq 0$ as is required for the definition of each v^j. In fact, if $\parallel u^j \parallel = 0$ for some j, then the vectors w^1 to w^j would be linearly dependent contrary to assumption. Finally, it is easy to check that the v^j are orthonormal.

Since the w^j are given and the v^j have been determined, we have their associated matrices W and V, both of which are invertible. Because the w^j form a basis, the v^j can be expanded in terms of them to given a relation of the form

$$v^i = \sum_j \alpha_{ij} w^j \qquad (3.6.55)$$

Indeed, (6.53) is just such a relation. Using the matrices W and V this relation can be written in the compact form

$$V = WA^T \qquad (3.6.56)$$

[14] The method is named after Jørgen Pedersen Gram and Erhard Schmidt, but Laplace had been familiar with it before Gram and Schmidt.

where A is the matrix given by the relation

$$A_{ij} = \alpha_{ij}. \tag{3.6.57}$$

Similarly, by considering rows rather than columns, there is a matrix B such that

$$V = BW. \tag{3.6.58}$$

Both A and B are unique and invertible, and satisfy the relations

$$I = V^T V = AW^T W A^T = AG(W)A^T, \tag{3.6.59}$$

$$I = VV^T = BWW^T B^T = BD(W)B^T. \tag{3.6.60}$$

We say that G is *congruent* to I under the action of A, and say that A is the *intertwining* transformation or matrix.[15] (Note that A is generally not orthogonal, and therefore generally $A^T \neq A^{-1}$.) Similarly, D is congruent to I under the action of B. Note that there are many pairs A,B satisfying (6.59) and (6.60) with one such pair for each orthogonalization process. Indeed, if we replace A and B by

$$A' = RA, \;\; B' = RB \tag{3.6.61}$$

where R is any orthogonal matrix, we find the relations

$$A'G(A')^T = RAGA^T R^T = RR^T = I, \tag{3.6.62}$$

$$B'D(B')^T = RBDB^T R^T = RR^T = I. \tag{3.6.63}$$

Conversely, if A and A' satisfy (6.59) and (6.62) respectively, then R defined by (6.61) is orthogonal, etc.

QR Decomposition

Closely related to Gram-Schmidt orthogonalization is what is called *QR decomposition*. It can be shown that any square nonsingular matrix W can be written in the factored form $W = QR$ where Q is orthogonal and R is upper triangular.[16] This factorization is unique if we require that the diagonal entries of R be positive. Evidently (6.56) can be rewritten in the form $W = V(A^T)^{-1}$. It can be verified that $(A^T)^{-1}$ is upper triangular with all diagonal entries positive, and we know that V is orthogonal. Thus we may make the identifications $Q = V$ and $R = (A^T)^{-1}$ to observe that the Gram-Schmidt process is one way to produce a QR decomposition. Given W there are other ways besides Gram-Schmidt to produce a QR decomposition (with the diagonal entries in R positive) including *Householder* transformations and *Givens* rotations.[17] And, by uniqueness, they all produce the same matrices Q and R that would be produced by applying Gram-Schmidt to W. Hence, by setting $V = Q$, these other ways can be also be used to produce the orthogonal matrix V that would also have resulted from applying Gram-Schmidt to W.

[15] If two matrices U and V are related by an equation of the form $V = AUA^{-1}$, they are said to be *similar* or *conjugate*. If they are related by an equation of the form $V = AUA^T$, they are said to be *congruent*. Here A is assumed to be nonsingular.

[16] Here there is an unfortunate conflict of notation. We have been using the symbol R to denote an orthogonal matrix. But in this paragraph, since QR is already standard notation in the mathematics literature, R will denote an upper triangular matrix.

[17] There are a variety of numerical QR packages.

Polar Decomposition of Real Matrices

Although Gram-Schmidt orthogonalization is straight forward and often natural, the result depends on the order in which the w^j are labeled, and does not treat all the w^j on an equal footing. For example, v^1 is always in the direction of w^1, but in general none of the other v^j are in the direction of the w^j. Sometimes it is desirable to have a procedure that treats all the w^j democratically. There are many ways to do this. The first uses *polar decomposition*.[18]

Let M be any *real* nonsingular matrix. It can be shown that any such M can be written uniquely in the form (called a polar decomposition)

$$M = PO. \tag{3.6.64}$$

Here P is a real positive-definite symmetric matrix, and O is a real orthogonal matrix. See Section 4.2. Intuitively, polar decomposition may be regarded as the matrix analog of expressing a complex number z in the polar form $z = r\exp(i\phi)$. Assuming the representation (6.64), we find that

$$D(M) = MM^T = POO^TP = P^2. \tag{3.6.65}$$

Since $D(M)$ is real, symmetric, and positive definite, it has a unique square root that is also positive definite and symmetric, and we may write

$$P = [D(M)]^{1/2}. \tag{3.6.66}$$

With this information we can solve (6.64) for O to find the result

$$O = [D(M)]^{-1/2}M. \tag{3.6.67}$$

Apply this result to the case $M = W$, where W is given by (6.5), to find the orthogonal matrix

$$O(W) = [D(W)]^{-1/2}W. \tag{3.6.68}$$

Now generate the v^j using (6.37) with

$$V = [D(W)]^{-1/2}W. \tag{3.6.69}$$

Note that (6.69) treats all the w^j on the same footing. It also has the feature (as does the Gram-Schmidt procedure) that if the w^j are already orthonormal, then

$$v^j = w^j, \tag{3.6.70}$$

for in this case $D = I$. See (6.41). In addition, the prescription (6.69) has the feature that the V it produces is the orthogonal matrix O that is *closest* to W in the sense of *minimizing* $\| W - O \|_E$ in the Euclidean matrix norm. See Exercise 7.1 and Section 4.4.2. Therefore V may be viewed as the solution to a *variational* problem.

The polar decomposition (6.64) can also be written in reverse order:

$$M = PO = OO^{-1}PO = OP', \tag{3.6.71}$$

[18] Polar decomposition was discovered by Cauchy.

where

$$P' = O^{-1}PO = O^TPO. \tag{3.6.72}$$

Evidently P' is also real positive-definite symmetric. Note that the orthogonal factor O is the same in both orders. Using the representation (6.71), we find

$$G(M) = M^TM = P'O^TOP' = (P')^2. \tag{3.6.73}$$

Thus, upon setting $M = W$, we find the equally valid relation

$$V = W[G(W)]^{-1/2}. \tag{3.6.74}$$

This relation also follows directly from (6.69) with the use of (6.13). Comparison of (6.50) and (6.58) with (6.69) and (6.74) shows that for this normalization process there are the relations

$$A = \{[G(W)]^{-1/2}\}^T = [G(W)]^{-1/2}, \tag{3.6.75}$$

$$B = [D(W)]^{-1/2}. \tag{3.6.76}$$

For the V defined by (6.69) we find the result

$$V^TD(W)V = W^T[D(W)]^{-1/2}D(W)[D(W)]^{-1/2}W = W^TW = G(W). \tag{3.6.77}$$

Since $V^T = V^{-1}$, this result can also be rewritten in the form

$$VG(W)V^T = D(W). \tag{3.6.78}$$

Thus, D and G are also conjugate under the action of the orthogonal matrix V. Compare (6.77) and (6.78) with (6.13).

Other Democratic Orthogonalizations

Having found one particular pleasing orthogonalization process, let us see if there are others. Based on the earlier discussion, without loss of generality we may consider all U of the form

$$U = VR \tag{3.6.79}$$

where R is any orthogonal matrix. If R is chosen at random, then all correlation of U with W is lost. However, if R is fixed ($R = I$ in the previous example) or is itself related to W, then U will also be related to W. Alternatively, U itself may be related to W in some direct way.

From (6.77) we find the result

$$U^TD(W)U = R^TV^TD(W)VR = R^TG(W)R. \tag{3.6.80}$$

Since G is real symmetric, we know there is an orthogonal transformation that diagonalizes it. Select R to be such a transformation,

$$R = R_G \tag{3.6.81}$$

where

$$R_G^T G(W) R_G = \Delta_G. \tag{3.6.82}$$

Here we have used the notation Δ_G to denote a diagonal form of G, and R_G to denote an orthogonal transformation that accomplishes this diagonalization. We remark that Δ_G is unique up to permutations of its diagonal entries, and R_G is unique up to (orthogonal) permutation matrices providing the entries of Δ_G (the eigenvalues of G) are distinct. Upon setting $U_D = VR_G$, and using (6.80) and (6.82), we find the result

$$U_D^T D(W) U_D = \Delta_G. \tag{3.6.83}$$

We see that U_D is an orthogonal transformation that diagonalizes D,

$$U_D^T D(W) U_D = \Delta_D, \tag{3.6.84}$$

and there is the relation

$$\Delta_D = \Delta_G. \tag{3.6.85}$$

It is interesting to recognize that an orthogonal U_D that accomplishes (6.84) may also be viewed as a solution to a variational problem. Let $\mathcal{F}_D$ be the functional

$$\mathcal{F}_D[U] = \sum_k [(U^T DU)_{kk}]^2. \tag{3.6.86}$$

(Note that $\mathcal{F}_D$ is *quartic* in the u^j. See Exercise 6.10.) There is the familiar algebraic result

$$\text{tr } \{(U^T DU)^T (U^T DU)\} = \sum_{ij} [(U^T DU)_{ij}]^2. \tag{3.6.87}$$

See Exercise 6.8. But we also find by direct evaluation the result

$$\text{tr } \{(U^T DU)^T (U^T DU)\} = \text{tr } (U^T D^T U U^T DU) = \text{tr } (U^T D^T DU) = \text{tr } (D^T D) = \text{tr } (D^2). \tag{3.6.88}$$

Here we have used the facts that U is orthogonal and D is symmetric, and standard properties of the trace operation. See Exercise 6.7. By combining (6.86) through (6.88) we find the relation

$$\text{tr } (D^2) = \sum_{ij} [(U^T DU)_{ij}]^2 = \sum_k [(U^T DU)_{kk}]^2 + \sum_{i \neq j} [(U^T DU)_{ij}]^2, \tag{3.6.89}$$

and therefore

$$\mathcal{F}_D[U] = \text{tr } (D^2) - \sum_{i \neq j} [(U^T DU)_{ij}]^2. \tag{3.6.90}$$

Evidently the maximum possible value of $\mathcal{F}_D[U]$ is tr (D^2), and this maximum can be reached if there is a $U \in SO(N)$ such that

$$(U^T DU)_{ij} = 0 \text{ for all } i, j \text{ satisfying } i \neq j. \tag{3.6.91}$$

According to (6.83) there is such a U, namely $U = U_D$. Thus we have the result

$$\max_{U \in SO(N)} \mathcal{F}_D[U] = \text{tr } (D^2), \tag{3.6.92}$$

and this maximum is achieved when

$$U = U_D = VR_G. \tag{3.6.93}$$

At this point it should be evident that there are several other possibilities for constructing orthogonal U matrices that are related to W. For example, after any construction, one could replace U by U^T. Or, one could require that U diagonalize $G(W)$ instead of $D(W)$. See Exercise 6.11.

The Complex Case and the Polar Decomposition of Complex Matrices

So far the discussion has been devoted to real vectors and real matrices. Some of it can be readily extended to the complex case. For example, Gram-Schmidt can be extended to the complex case simply by replacing the usual real scalar product with the usual complex scalar product. A second example of extension to the complex case is that there is an analogous polar decomposition (also simply called polar decomposition) for the case of complex matrices. A factorization of the form (6.64) still holds but now M is complex, P is Hermitian and positive definite, and O is unitary. See Exercise 4.2.5.

3.6.5 Construction of Symplectic Bases

We close this section with an introduction to the problem of constructing symplectic bases. Further discussion is given in Sections 4.3 through 4.8.

Darboux Symplectification

We will first describe an analog of the Gram-Schmidt procedure, which we will call *Darboux* symplectification. Suppose the w^j are a set of $2n$ linearly independent vectors and we wish to construct from them a symplectic basis v^j. For this purpose it is convenient to use the form (2.10) for J. Below is an algorithm for constructing the v^j:

1. Define v^1 by the simple rule
$$v^1 = w^1. \tag{3.6.94}$$

2. Starting with w^2, search through the w^j with $j \geq 2$ to find the first j, call it k, with the property
$$(v^1, Jw^j) \neq 0. \tag{3.6.95}$$
[Better yet, if one is working numerically and therefore only to finite precision, select j so that $|(v^1, Jw^j)|$ is maximized. The analogous choices should also be made in steps 6, 10, etc. below.] Renumber the vectors $w^2 \cdots w^{2n}$ so that w^k becomes w^2.

3. Define v^2 by the rule
$$v^2 = w^2/[(v^1, Jw^2)]. \tag{3.6.96}$$
We then have the result
$$(v^1, Jv^2) = 1 = J_{12}. \tag{3.6.97}$$
And, since J is antisymmetric, at this stage we have the result
$$(v^i, Jv^j) = J_{ij} \text{ for } i, j = 1 \text{ to } 2. \tag{3.6.98}$$

4. Using the remaining vectors $w^3 \cdots w^{2n}$, define new vectors ${}^1w^j$ with $j \geq 3$ by the rule
$${}^1w^j = w^j + (v^2, Jw^j)v^1 - (v^1, Jw^j)v^2. \tag{3.6.99}$$
As a result of this rule there are the relations
$$(v^i, J\ {}^1w^j) = 0 \text{ for } i = 1, 2 \text{ and } j = 3, 4, \cdots 2n. \tag{3.6.100}$$

5. Define v^3 by the rule
$$v^3 = {}^1w^3. \tag{3.6.101}$$

6. Starting with ${}^1w^4$, search through the ${}^1w^j$ with $j \geq 4$ to find the first j, call it k, with the property
$$(v^3, J\ {}^1w^j) \neq 0. \tag{3.6.102}$$
Renumber the vectors ${}^1w^4 \cdots {}^1w^{2n}$ so that ${}^1w^k$ becomes ${}^1w^4$.

7. Define v^4 by the rule
$$v^4 = {}^1w^4/[(v^3, J\ {}^1w^4)]. \tag{3.6.103}$$
At this stage we have the results
$$(v^i, Jv^j) = J_{ij} \text{ for } i, j = 1 \text{ to } 4. \tag{3.6.104}$$

8. Using the remaining vectors ${}^1w^5 \cdots {}^1w^{2n}$, define new vectors ${}^2w^j$ with $j \geq 5$ by the rule
$${}^2w^j = {}^1w^j + (v^4, J\ {}^1w^j)v^3 - (v^3, J\ {}^1w^j)v^4. \tag{3.6.105}$$
Now we have the relations
$$(v^i, J\ {}^2w^j) = 0 \text{ for } i = 1 \text{ to } 4 \text{ and } j = 5, 6, \cdots 2n. \tag{3.6.106}$$

9. Define v^5 by the rule
$$v^5 = {}^2w^5. \tag{3.6.107}$$

10. Starting with ${}^2w^6$, search through the ${}^2w^j$ with $j \geq 6$ to find the first j, call it k, with the property
$$(v^5, J\ {}^2w^j) \neq 0. \tag{3.6.108}$$
Renumber the vectors ${}^2w^6 \cdots {}^2\, w^{2n}$ so that ${}^2w^k$ becomes ${}^2w^6$.

11. Define v^6 by the rule
$$v^6 = {}^2w^6/[(v^5, J\ {}^2w^6)]. \tag{3.6.109}$$
At this stage we have the results
$$(v^i, Jv^j) = J_{ij} \text{ for } i,j = 1 \text{ to } 6. \tag{3.6.110}$$

12. Proceed with the obvious extension of the above process to construct $v^7, v^8, \cdots v^{2n-2}$. Then at the last stage we have
$$v^{2n-1} = {}^mw^{2n-1}, \tag{3.6.111}$$
$$v^{2n} = {}^mw^{2n}/[(v^{2n-1}, J\ {}^mw^{2n})], \tag{3.6.112}$$
with
$$m = n - 1. \tag{3.6.113}$$

At this point several comments are in order. First, if we are working only with two, four, or six-dimensional phase space, as is the case for accelerator physics, then we may terminate the algorithm at steps 3, 7, or 11. Second, how does one know that the required vectors ${}^mw^k$ described in steps 2, 6, 10, etc. exist? Finally, how does one know that the vectors $v^3, v^5, \cdots v^{2n-1}$ given in steps 5, 9, etc. are nonzero? As was the case with the Gram-Schmidt orthogonalization process, we are saved from such embarrassment because the w^j are assumed to be linearly independent and J is invertible. See Exercise 6.12.

Transitive Action of $Sp(2n)$ on Phase Space

There is also an observation that is worth making. Suppose α and β are any two nonzero vectors. Let M^α be a symplectic matrix whose first column is the vector α. We know that such a matrix exists exists because we may set $w^1 = \alpha$ in (6.94). Then we have the relations
$$\alpha = M^\alpha e^1 \text{ and } e^1 = (M^\alpha)^{-1}\alpha. \tag{3.6.114}$$

Similarly, let M^β be a symplectic matrix whose first column is the vector β. Now define a matrix M by the rule
$$M = M^\beta (M^\alpha)^{-1}. \tag{3.6.115}$$

By the group property this matrix will be symplectic, and by construction it will have the property
$$M\alpha = M^\beta (M^\alpha)^{-1}\alpha = M^\beta e^1 = \beta. \tag{3.6.116}$$

We have found the remarkable result that, with the exception of the origin, any point in phase space can be sent into any other point by a symplectic matrix. (The origin is obviously sent into itself.) Following the terminology elaborated on in Section 5.12, we say that, with the exception of the origin, $Sp(2n)$ acts *transitively* on phase space.

Other Symplectifications

Let us now explore briefly additional methods for constructing symplectic bases. One of them makes use of "symplectic" polar decomposition, and is the subject of Sections 4.3 and 4.4. To consider others introduce, in imitation of the orthogonal case, analogous dyadic and Gram matrices by the definitions

$$D_J(W) = WJW^T. \tag{3.6.117}$$

$$G_J(W) = W^TJW. \tag{3.6.118}$$

Note that both D_J and G_J are antisymmetric. It is easy to see that there are again unique nonsingular matrices A and B such that the relations (6.56) and (6.57) hold. Then, from (6.56) and (6.58) and the requirement that V be symplectic, we find the results

$$J = V^TJV = AW^TJWA^T = AG_J(W)A^T, \tag{3.6.119}$$

$$J = VJV^T = BWJW^TB^T = BD_J(W)B^T. \tag{3.6.120}$$

We see that G_J is congruent to J under the action of A, and D_J is congruent to J under the action of B.

We already know that (6.119) and (6.120) have a full infinity of solutions for A and B. One simply solves (6.56) or (6.58) for A or B using any symplectic V. This matter is considered from a broader perspective in Section 3.13.

Finally we remark that, since both D and G as given by (6.10) and (6.12) are symmetric and positive definite, it can be shown there are symplectic matrices U and V such that

$$UD(W)U^T = \text{ Williamson diagonal form,} \tag{3.6.121}$$

$$VG(W)V^T = \text{ Williamson diagonal form.} \tag{3.6.122}$$

Moreover, since J is orthogonal, the matrices $JD(W)J^T$ and $JG(W)J^T$ are also symmetric and positive definite. Therefore there are also symplectic matrices, again call them U and V, such that

$$UJD(W)J^TU^T = \text{ Williamson diagonal form,} \tag{3.6.123}$$

$$VJG(W)J^TV^T = \text{ Williamson diagonal form.} \tag{3.6.124}$$

See Section 33.6.3. The columns (or rows) of these symplectic matrices may also be regarded as symplectic bases related in a specific way to W.

Exercises

3.6.1. Show that orthogonal matrices, matrices satisfying (6.1), have the property (6.2).

3.6.2. Suppose that O is an orthogonal matrix. Show that O and O^T commute, $O^TO = OO^T$. Show that $-O$, O^T, and O^{-1} are also orthogonal matrices. Show that orthogonal matrices form a group.

3.6.3. Verify (6.10) by showing that D has the matrix elements

$$\begin{aligned} D_{ij} &= \sum_k (e^i, w^k)(w^k, e^j) = \sum_k (e^i, w^k)(e^j, w^k) \\ &= \sum_k W_{ik} W_{jk} = \sum_k W_{ik}(W^T)_{kj} = (WW^T)_{ij}. \end{aligned} \tag{3.6.125}$$

3.6.4. Verify (6.12) by showing that G can be written in the form

$$G(W) = \sum_{k\ell} |e^k)(w^k, w^\ell)(e^\ell|, \tag{3.6.126}$$

and has the matrix elements

$$G_{ij} = \sum_{k\ell} (e^i, e^k)(w^k, w^\ell)(e^\ell, e^j) = (w^i, w^j) = (We^i, We^j) = (e^i, W^T We^j). \tag{3.6.127}$$

3.6.5. Verify (6.26) and (6.27). Note that (6.5) implies the relation

$$(w^j| = (e^j|W^T. \tag{3.6.128}$$

3.6.6. Verify that the v^j given by (6.53) are orthonormal.

3.6.7. Show that the trace operation has the properties

$$\text{tr}(A) = \text{tr}(A^T),\ [\text{tr}(A)]^* = \text{tr}(A^\dagger), \tag{3.6.129}$$

$$\text{tr}(AB) = \text{tr}(BA). \tag{3.6.130}$$

Here a $*$ denotes complex conjugation.

3.6.8. Verify the relations

$$\text{tr}(A^T A) = \sum_{ij} (A_{ij})^2,\ \text{tr}(A^\dagger A) = \sum_{ij} |A_{ij}|^2. \tag{3.6.131}$$

3.6.9. Verify (6.88) through (6.90).

3.6.10. Verify that $\mathcal{F}_D$ has the explicit form

$$\mathcal{F}_D[U] = \sum_k \{\sum_\ell [(u^k, w^\ell)]^2\}^2. \tag{3.6.132}$$

3.6.11. As an alternative to (6.81), consider the option of writing

$$U = RV, \tag{3.6.133}$$

and then working with (6.80). Show that one can require that U^T diagonalize $G(W)$, and find an associated variational problem.

3.6.12. This exercise studies the Darboux (Gram-Schmidt like) method of constructing a symplectic basis. Assume that the w^j are linearly independent and recall that J is invertible. Show that the vectors ${}^m w^k$ exist and the vectors $v^3, v^5, \cdots v^{2n-1}$ are nonzero. For example, at step 2 of the algorithm, show that the possibility

$$(v^1, Jw^j) = 0 \text{ for all } j \tag{3.6.134}$$

would imply that the w^j are linearly dependent. Similarly, at steps 5 and 6, show that the vectors v^1, v^2, ${}^1w^3$, ${}^1w^4$, $\cdots {}^1 w^{2n}$, are linearly independent and the vector v^3 is nonzero. Continue on to show that steps 9, 13, $\cdots$ and steps 10, 14, $\cdots$ succeed. Alternatively, verify by induction on n that the Darboux construction is always possible:

a) Verify the case of dimension 2.

b) Assume the result holds in dimension $(2n-2)$. Consider the case of dimension $2n$ as in Section 3.6.5. Show that a w^j can be found that satisfies (6.95) because the w^i are assumed to be linearly independent.

c) Verify that the $(2n-2)$ vectors ${}^1w^j$ defined by (6.99) satisfy (6.100) and are linearly independent. Therefore, by the induction hypothesis, a symplectic basis can be found for this set of vectors. Show that these $(2n-n)$ symplectic basis vectors, along with v^1 and v^2, then form a set of $2n$ symplectic basis vectors.

3.6.13. Suppose M is a $2n \times 2n$ matrix. Regard M as a collection of column vectors m^a so that it can be written in the form

$$M = \begin{pmatrix} M_{1,1} & M_{1,2} & M_{1,3} & \cdots & M_{1,2n} \\ M_{2,1} & M_{2,2} & M_{2,3} & \cdots & M_{2,2n} \\ M_{3,1} & M_{3,2} & M_{3,3} & \cdots & M_{3,2n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ M_{2n,1} & M_{2n,2} & M_{2n,3} & \cdots & M_{2n,2n} \end{pmatrix} = (m^1, m^2, m^3, \cdots m^{2n}). \tag{3.6.135}$$

Verify that, with this convention, the column vectors m^a will have entries m^a_c given by the relations

$$m^a_c = M_{ca}. \tag{3.6.136}$$

Correspondingly, verify that there is the relation

$$m^a = Me^a. \tag{3.6.137}$$

Next, suppose M is a real symplectic matrix that satisfies (1.1) with J specified by (1.2). Verify that, in terms of matrix elements, (1.1) takes the form

$$(e^a, M^T JMe^b) = (e^a, Je^b) = J_{ab}, \tag{3.6.138}$$

from which it follows that

$$J_{ab} = (e^a, M^T JMe^b) = (Me^a, JMe^b) = (m^a, Jm^b). \tag{3.6.139}$$

Thus, as expected from the discussion of Section 3.6.3, the vectors m^a form a symplectic basis. Verify that if

$$a = i \text{ with } i \in [1, n] \tag{3.6.140}$$

and

$$b = i + n, \tag{3.6.141}$$

then

$$(m^a, Jm^b) = J_{ab} = J_{i,i+n} = 1. \tag{3.6.142}$$

Verify that

$$(m^a, m^a) = \sum_c (M_{ca})^2. \tag{3.6.143}$$

Suppose that instead M is symplectic with respect to the J' defined by (2.10). Show that the general discussion of this exercise goes through as before except that we should now set

$$a = i \text{ with } i = 1, 3, 5, \cdots \tag{3.6.144}$$

and

$$b = i + 1 \tag{3.6.145}$$

so that

$$(m^a, J'm^b) = J'_{ab} = J'_{i,i+1} = 1. \tag{3.6.146}$$

3.6.14. Show that for a symplectic basis and a J of the form (2.10) there are the relations

$${}^r v^1 = Jv^2,\ {}^r v^2 = -Jv^1;\ {}^r v^3 = Jv^4,\ {}^r v^4 = -Jv^3;\ \text{etc.} \tag{3.6.147}$$

3.6.15. Here is a curiosity: Given a set of $2n$ linearly independent vectors w^j, can one find a set of vectors ${}^{sr}w^i$ such that

$$({}^{sr}w^i, Jw^j) = J_{ij}? \tag{3.6.148}$$

The answer is yes. Such a set will be called a *symplectic reciprocal* basis. Let the vectors ${}^r w^i$ denote the ordinary reciprocal basis to the w^j. See (6.18). Define a related basis $\tilde{w}^i$ by the rule

$$\tilde{w}^i = J\ \ {}^r w^i. \tag{3.6.149}$$

This basis has the property

$$(\tilde{w}^i, Jw^j) = (J\ \ {}^r w^i, Jw^j) = ({}^r w^i, J^T J w^i) = ({}^r w^i, w^j) = \delta_{ij}. \tag{3.6.150}$$

Now it is convenient to work with the J given by (2.10). With this choice in mind, define the ${}^{sr}w^i$ by the rule

$$\begin{aligned} {}^{sr}w^1 &= \tilde{w}^2, \\ {}^{sr}w^2 &= -\tilde{w}^1, \\ {}^{sr}w^3 &= \tilde{w}^4, \\ {}^{sr}w^4 &= -\tilde{w}^3, \\ &\text{etc.} \end{aligned} \tag{3.6.151}$$

Verify that these vectors satisfy (6.148).

3.6.16. Review the discussion of transformation groups at the end of Section 3.6.1. For each of the realizations of a group acting on itself introduce the notation

$$h \to T_g h = gh,\ h \to T_g h = hg^{-1},\ h \to T_g h = ghg^{-1};\ g \in G,\ h \in \mathcal{Z} = G. \tag{3.6.152}$$

Verify that in each realization there is the relation

$$T_{g_2} T_{g_1} = T_{g_2 g_1}, \tag{3.6.153}$$

which shows that in each realization the transformations T_g form a group.

3.7 Lie Algebraic Properties

3.7.1 Matrix Exponential and Logarithm

Let B be any matrix. The *exponential* of a matrix, written variously as e^B or $\exp(B)$, is defined by the exponential series

$$e^B = \exp(B) = I + B + B^2/2! + \cdots = \sum_{n=0}^{\infty} B^n/n!. \tag{3.7.1}$$

(Here we adopt the usual convention that $B^0 = I$ for any matrix B.) Similarly, the *logarithm* of a matrix A (sufficiently near the identity) is defined by the logarithm series

$$\log(A) = \log[I - (I - A)] = -\sum_{n=1}^{\infty} (I - A)^n/n. \tag{3.7.2}$$

As might be expected, the exponential and logarithmic functions are related. Specifically, if one has

$$B = \log(A), \tag{3.7.3}$$

then it follows that

$$A = \exp(B), \tag{3.7.4}$$

and vice versa. Put another way, one has the relations

$$A = \exp[\log(A)] \text{ for } A \text{ sufficiently near the identity matrix,} \tag{3.7.5}$$

$$B = \log[\exp(B)] \text{ for } B \text{ sufficiently near the zero matrix.} \tag{3.7.6}$$

If a matrix A can be written in the form (7.4), we say that B *generates* A. It can be shown that there is the identity

$$[\exp(B/n)]^n = \exp(B) \tag{3.7.7}$$

for any integer n. See Exercise 7.5. Thus, if A is generated by B, we may also write

$$A = [\exp(B/n)]^n. \tag{3.7.8}$$

From (7.1) we see that

$$\exp(B/n) = I + B/n + O(1/n)^2.$$

Consequently, for sufficiently large n, B/n is near the zero matrix and $\exp(B/n)$ is near the identity matrix. Since B/n is near the zero matrix, we may regard it as an infinitesimal matrix. Correspondingly, in view of (7.8), we say that A is *infinitesimally generated* in that it can be written as the product of a large number of identical near identity matrices. Finally, like the ordinary exponential function, it can be verified that

$$\lim_{n\to\infty}(I + B/n)^n = \exp(B) \tag{3.7.9}$$

for any matrix B.

In some cases there may be several linearly independent matrices $B_1, B_2, \cdots, B_k$ and we consider elements of the form

$$A = \exp(s_1B_1 + s_2B_2 + \cdots + s_kB_k).$$

Here the s_j are scalars. We would then say that the B_j generate such matrices A. Even more generally, we might consider matrices that are finite products of matrices of the form A,

$$G = \exp(s_1B_1 + s_2B_2 + \cdots + s_kB_k)\exp(t_1B_1 + t_2B_2 + \cdots + t_kB_k)\cdots .$$

We would again say that the B_j generate such matrices G. However, it might not be possible to write such matrices in the form

$$G = \exp(B) \tag{3.7.10}$$

where B is some linear combination of the B_j. Consequently, if (7.10) is not possible, we would say that G is generated by the B_j, but not infinitesimally generated as defined above. Alternatively, we might broaden our definition of "infinitesimally generated" to include finite products of matrices that are themselves infinitesimally generated.

Vector and Matrix Norms

To make our discussion more precise, it is useful to introduce the concepts of vector and matrix *norms*. We will use the same notation $\|\|$ to refer to either norm with the understanding that the exact meaning of the notation depends on whether it is being applied to a vector or a matrix.

A vector norm is a rule that assigns to any vector v a real non-negative number $\|v\|$, called the norm of v, in such a way that the following properties are satisfied:

$$\|v\| \geq 0\,,\ \text{and}\ \|v\| = 0 \Leftrightarrow v = 0; \tag{3.7.11}$$

$$\|av\| = |a|\|v\|\,,\ a\ \text{any scalar}; \tag{3.7.12}$$

$$\|u + v\| \leq \|u\| + \|v\|. \tag{3.7.13}$$

Here the notation $\Leftrightarrow$ is used to denote logical implication in both directions.

Similarly, a matrix norm is a rule that assigns to any matrix A a real non-negative number $\|A\|$, called the norm of A. The matrix norm is required to satisfy properties analogous to those for a vector norm plus a property associated with matrix multiplication:

$$\|A\| \geq 0\,, \text{ and } \|A\| = 0 \Leftrightarrow A = 0; \tag{3.7.14}$$

$$\|aA\| = |a|\|A\|\,, \ a \text{ any scalar}; \tag{3.7.15}$$

$$\|A + B\| \leq \|A\| + \|B\|; \tag{3.7.16}$$

$$\|AB\| \leq \|A\|\|B\|. \tag{3.7.17}$$

Finally, a matrix norm is said to be *consistent* with a vector norm if the following condition is satisfied for any matrix A and vector v (assuming that A is $m \times m$ and v is m dimensional):

$$\|Av\| \leq \|A\|\|v\|. \tag{3.7.18}$$

Note that the norm indicated in the left side of (7.18) is a vector norm since the quantity Av is a vector. By contrast, the norms on the right side of (7.18) are matrix and vector norms, respectively.

There are several ways of defining consistent matrix and vector norms. One of the more useful is to take for the matrix norm the *maximum column sum* norm. It is defined by the rule

$$\|A\| = \max_k (\sum_j |A_{jk}|). \tag{3.7.19}$$

[Sum over j while holding k fixed to add together the values of $|A_{jk}|$ for column k. Then, for the various columns (values of k), report the largest result found.] It can be shown that this norm satisfies the requirements (7.14) through (7.17). Furthermore, it can be shown that this norm is consistent with the *component moduli sum* vector norm defined by the rule

$$\|v\| = \sum_j |v_j|. \tag{3.7.20}$$

The *strongest* matrix norm is the *spectral* norm defined by

$$\|A\|_{\text{spct}} = +(\text{maximum eigenvalue of } \mathrm{A}^\dagger \mathrm{A})^{1/2}. \tag{3.7.21}$$

Note that for any matrix A the eigenvalues of $A^\dagger A$ are guaranteed to be real and nonnegative so that (7.21) is well defined. The spectral norm is strongest in the sense that for any matrix A there is the inequality

$$||A||_{\text{spct}} \leq ||A|| \tag{3.7.22}$$

where $||A||_{\text{spct}}$ denotes the spectral norm and $||A||$ denotes any other matrix norm. However, to compute the spectral norm generally requires considerable work, and therefore it is sometimes more of theoretical value rather than suitable for frequent computation.

It can be shown that the matrix spectral norm is consistent with the *Euclidean* vector norm. Let v be a possibly complex m-dimensional vector and let $(*, *)$ denote the usual complex inner product. The Euclidean vector norm $||v||_E$ is defined by the rule

$$(||v||_E)^2 = (v, v) = \sum_j |v_j|^2. \tag{3.7.23}$$

There are also other ways of defining vector and matrix norms. See Exercise 7.1 for the definition of the Euclidean matrix norm.

Convergence of Series

With the aid of the concept of a matrix norm, it can be shown that the series (7.1) converges for *any* matrix B, and that one has the relations

$$\|\exp(B)\| \leq e^{\|B\|}, \tag{3.7.24}$$

$$\big\|[\exp(B) - I]\big\| \leq e^{\|B\|} - 1. \tag{3.7.25}$$

[There is a theorem to the effect that if a matrix power series $\sum_n c_n B^n$ converges in norm (*i.e.*, if the series converges with all coefficients c_n replaced by $|c_n|$ and with B^n replaced by $\|B\|^n$), then it also converges for each individual matrix element.] By contrast, it can be shown that in general the series (7.2) converges only when $\|(A - I)\| < 1$ for some norm, and that then one has the relation

$$\|\log(A)\| \leq -\log\big[1 - \|(A - I)\|\big]. \tag{3.7.26}$$

3.7.2 Application to Symplectic Matrices

With this background in mind, suppose that M is a real symplectic matrix near the identity. We start our analysis in a heuristic fashion by assuming that M can be written in the form

$$M = \exp(\epsilon B) \tag{3.7.27}$$

where ϵ is small so that ϵB is near the zero matrix. We then have the expansions

$$\begin{aligned} M &= I + \epsilon B + O(\epsilon^2), \\ M^T &= I + \epsilon B^T + O(\epsilon^2). \end{aligned} \tag{3.7.28}$$

Upon inserting these expansions into the symplectic condition (1.2) and equating powers of ϵ, we find the result

$$B^T J + JB = 0. \tag{3.7.29}$$

The relation (7.29) is a key result that we will now prove rigorously for $\epsilon = 1$ provided B itself is sufficiently small. Specifically, assume that M and M^{-1} are sufficiently near the identity so that $\log(M)$ and $\log(M^{-1})$ can be computed using (7.2). That is, suppose the following two series converge:

$$B = \log(M) = -\sum_{n=1}^{\infty}(I - M)^n/n, \tag{3.7.30}$$

$$-B = \log(M^{-1}) = -\sum_{n=1}^{\infty}(I - M^{-1})^n/n. \tag{3.7.31}$$

Use the series (7.30) to compute the quantity $J^{-1}B^T J$. Doing so gives the result

$$\begin{aligned} J^{-1}B^T J &= -\sum_{n=1}^{\infty}(I - J^{-1}M^T J)^n/n \\ &= -\sum_{n=1}^{\infty}(I - M^{-1})^n/n \\ &= -B. \end{aligned} \tag{3.7.32}$$

Here use has also been made of (7.31) and (1.9). Now compare the beginning and end of (7.32) to get the equivalent results

$$J^{-1}B^TJ = -B \text{ or } JB^TJ^{-1} = -B \text{ or } B^TJ + JB = 0 \text{ or } JB^T + BJ = 0. \tag{3.7.33}$$

Note that (7.29) is among these equivalent results. A matrix B that satisfies (7.33) is sometimes called *Hamiltonian* or *infinitesimally symplectic*.[19]

To understand the implications of the condition (7.33), suppose that B is written in the form

$$B = JS. \tag{3.7.34}$$

[Reader, verify that given any matrix B, because J is nonsingular, there is always a well-defined S such that (7.34) is satisfied.] Upon inserting (7.34) into (7.33), one finds the equivalent condition

$$-S^TJJ + JJS = 0 \text{ or } S^T = S. \tag{3.7.35}$$

That is, S must be a symmetric matrix. Parenthetically, we note that any Hamiltonian matrix (any matrix of the form JS with S symmetric) must be traceless. Verify this claim! It follows that any matrix M of the form $M = \exp(JS)$ must have unit determinant. See Exercise 7.10.

We have learned that any real symplectic matrix M sufficiently near the identity can be written in the form

$$M = e^B = e^{JS}, \tag{3.7.36}$$

with S small, real, and symmetric. Conversely, suppose that B is any matrix of the form (7.34) with S real and symmetric. Then, the matrix M given by (7.36) is symplectic. To verify this assertion, simply compute! One finds the results

$$M = \exp(JS), \tag{3.7.37}$$

$$M^T = \exp(-SJ),$$

$$\begin{aligned} M^TJM &= \exp(-SJ)J\exp(JS) \\ &= JJ^{-1}\exp(-SJ)J\exp(JS) \\ &= J\exp(-J^{-1}SJ^2)\exp(JS) \\ &= J\exp(-JS)\exp(JS) \\ &= J. \end{aligned} \tag{3.7.38}$$

What has been shown is that any symplectic matrix M sufficiently near the identity can be written in the form (7.36) with S small and symmetric, and vice versa.[20] Note that the

[19] This usage of the adjective *Hamiltonian* should not be confused with its usage in Quantum Mechanics where a Hamiltonian matrix would be a matrix formed by taking the matrix elements of a Hamiltonian operator with respect to an orthonormal basis. Such a matrix would generally have complex entries and would be Hermitian.

[20] Here we see the beginning of a grand theme: There is a close relation between symplectic and symmetric matrices. This theme will be developed fully in Sections 3.11, 5.13, and 6.7. We also note that in the calculation (7.38) we have used the results of Exercises 7.5 and 7.11.

symplectic condition as expressed by (1.2) is a set of *quadratic* relations among the matrix elements of M. By contrast, the conditions (7.33) or (7.35) are *linear* relations among the matrix elements of B or S, respectively. We see that the use of an exponential representation has converted a set of quadratic relations, which are generally more difficult to work with due to their nonlinearity, into a set of simple linear relations.

Finally, we remark that not every symplectic matrix can be written in single exponential form. See Exercise 7.12.

3.7.3 Matrix Lie Algebra and Lie Group: The Baker-Campbell-Hausdorff (BCH) Multiplication Theorem

The stage is now set for the introduction of a central discovery of Sophus Lie, the concept of a *Lie* algebra. We will first introduce this concept in a concrete matrix setting, and then place it in a more general abstract setting.

A set A of $m \times m$ matrices forms a Lie algebra if it satisfies the following properties:

i. If the matrix A is in the Lie algebra, then so is the matrix aA where a is any scalar.

ii. If two matrices A and B are in the Lie algebra, then so is their sum.

iii. If two matrices A and B are in the Lie algebra, then so is their *commutator* $[A, B]$. The *commutator* is defined by the relation

$$[A, B] = AB - BA. \tag{3.7.39}$$

Note that the commutator symbol $[,]$ is the same as that used earlier for a Poisson bracket. This is somewhat awkward, but unfortunately there are not always enough convenient symbols to go around. Later, when there is greater chance of confusion, we will use the symbols $\{,\}$ to denote a commutator.

At this point the reader should take pen in hand and verify that the set of matrices of the form JS with S symmetric is a Lie algebra. That is, *Hamiltonian matrices form a Lie algebra.*

That Hamiltonian matrices form a Lie algebra is no accident. It is a remarkable fact that there is a close connection between the concept of a Lie algebra and that of a group. The connection arises from a deep property of the exponential function that generally bears the names *Baker-Campbell-Hausdorff* (BCH). Their result, in a matrix setting, may be stated as follows: Let A and B be any two matrices (square and of the same dimension). Form the matrices $\exp(sA)$ and $\exp(tB)$ where s and t are parameters. Next form their product. Then, for s and t sufficiently small, it is possible to write

$$\exp(sA)\exp(tB) = \exp(C), \tag{3.7.40}$$

where C is some other matrix. The remarkable fact is that C is a member of the Lie algebra *generated* by A and B.[21] That is, C is a sum of elements formed *only* from A and B and their *multiple commutators*. Specifically, one has the relation

$$\begin{aligned} C(s,t) = sA \;&+\; tB + (st/2)[A,B] + (s^2t/12)[A,[A,B]] \\ &+\; (st^2/12)[B,[B,A]] \\ &-\; (s^2t^2/24)[A,[B,[A,B]]] + O(s^4t, s^3t^2, s^2t^3, st^4). \end{aligned} \tag{3.7.41}$$

No isolated terms of the form A^2, B^2, AB, $[A^2, B^2]$, etc. occur! Although stated in terms of matrices, this result can be extended to the case of linear operators.

In general, the series for C (called the BCH series) contains an infinite number of terms and may converge only for sufficiently small s and t. It may not converge at all if the Lie algebra generated by A and B is infinite dimensional and A and B are unbounded operators.[22]

The proof of this theorem is difficult and is given in Appendix C.[23] For present purposes, it shows that given any Lie algebra L of matrices, there exists a corresponding *Lie group* G. Furthermore, the rules for multiplying any two group elements are contained within the Lie algebra. To see the truth of this assertion, consider all matrices of the form $g(s) = \exp(s\ell)$ with ℓ contained in L. According to the previous result, one has

$$\exp(s\ell)\exp(t\ell') = \exp \ell''$$

with ℓ'' given by a relation of the form (7.41) for s, t sufficiently small. Also

$$g(0) = I \text{ and } g^{-1}(s) = g(-s).$$

Thus these matrices, at least those sufficiently near the identity, form a group. Once the group has been obtained near the identity, it can be extended to a global group by successively multiplying the different g's already obtained. We remark that if the Lie algebras of two sets of matrices are the same, it does not necessarily follow that the two corresponding groups constructed in this way are globally the same. They may only be related by a homomorphism. The groups $SU(2)$ and $SO(3,\mathbb{R})$ provide an example of this possibility.[24] Information about the matrices beyond their Lie algebra is needed to determine the global properties of the group.

It has already been shown that symplectic matrices form a group. Furthermore, it has been shown that symplectic matrices near the identity can be written as the exponentials of

[21] Here is another, and different, use of the word *generate*. Suppose one has a collection of $n \times n$ matrices B_i. Form their commutators to produce possibly new linearly independent matrices. Next, join the set of these matrices to the original set of the B_i. Now form the commutators of all these matrices, and join these matrices to the set already obtained. Repeat this process ad infinitum until no new linearly independent matrices are obtained. In the matrix case this process must terminate because there are only n^2 linearly independent $n \times n$ matrices. The net result of this procedure is a Lie algebra, which is referred to as the Lie algebra generated by the B_i. Although we have been talking about matrix Lie algebra, the same construction can be carried out for any collection of Lie elements drawn from some Lie algebra with the commutator replaced by the abstract Lie product.

[22] The BCH series can be summed in the case of $sp(2,\mathbb{C}) = s\ell(2,\mathbb{C})$ which includes $su(2)$ and $sp(2,\mathbb{R})$. See Subsection 8.7.1. For an example of divergence in the infinite-dimensional case, see Section 38.7.

[23] Also see Appendix C for a discussion of the converse *Zassenhaus* formula.

[24] The groups $SU(n)$ will be defined in Subsection 7.6.

elements of a Lie algebra. It follows that $Sp(2n)$, the group of symplectic matrices, is a Lie group. The Lie algebra associated with $Sp(2n)$, the Lie algebra of Hamiltonian matrices, is denoted by $sp(2n)$. More specifically, the Lie algebra associated with $Sp(2n, \mathbb{R})$ is denoted by $sp(2n, \mathbb{R})$, and that associated with $Sp(2n, \mathbb{C})$ is denoted by $sp(2n, \mathbb{C})$. Where there is no possibility of confusion, we will use the notation $sp(2n)$ to mean $sp(2n, \mathbb{R})$.

Properties 1 and 2 of a Lie algebra indicate that the elements of a Lie algebra form a linear vector space. It is therefore natural to speak of the *dimension* of a Lie algebra. For the case of the symplectic group, elements of the Lie algebra are of the form (7.34) where S is any symmetric matrix. The dimension of the Lie algebra in this case, therefore, is just the dimensionality of the set of all $2n \times 2n$ symmetric matrices. This number is easily computed. There are $2n$ independent entries on the diagonal of a $2n \times 2n$ symmetric matrix, and $[(2n)^2 - 2n]/2$ independent entries above the diagonal. Finally, all the entries below the diagonal are given in terms of the entries above the diagonal by the symmetry condition. Therefore, the dimension of the symplectic group Lie algebra, which will be written as $\dim sp(2n)$, is given by the relation

$$\dim sp(2n) = 2n + [(2n)^2 - 2n]/2 = n(2n + 1). \tag{3.7.42}$$

For example, the dimensions of $sp(2), sp(4)$, and $sp(6)$ are 3, 10, and 21, respectively. See Table 7.1 below.

Table 3.7.1: Dimension of $sp(2n)$.

n	$2n$	$\dim sp(2n)$	n	$2n$	$\dim sp(2n)$
1	2	3	5	10	55
2	4	10	6	12	78
3	6	21	7	14	105
4	8	36	8	16	136

Let M be some element of $Sp(2n)$ that can be written in the exponential form (7.36). To the extent that the elements of $Sp(2n)$ in some neighborhood of M can also be written in exponential form, we may say that the dimension of this neighborhood is also given by (7.42). However, in a while we will see that not all elements of $Sp(2n)$ can be written in exponential form. See Exercise 7.12 and Subsection 8.7.2. What about the general case? In Subsection 8.2 it is shown that every symplectic matrix can be written as the product of two symplectic matrices, each of which can be written in exponential form. And the *total* dimension count of the two of them together is again given by (7.42). See Exercise 9.10. Consequently we may say that $su(2n)$ as a vector space and $Sp(2n)$ as a manifold have the same dimension.

3.7.4 Abstract Definition of a Lie Algebra

For future use, it is essential to put the concept of a Lie algebra, as just defined in a matrix context, into a more general setting. We will begin by defining the concept of an algebra.

Algebra

Naively speaking, algebra has to do with the concepts of addition and multiplication. The concept of addition can be generalized to yield the concept of a linear vector space. The concept of multiplication has several possible generalizations. Formally, an *algebra* A over a field of numbers F is defined as a linear vector space supplemented by a rule for multiplying two vectors to yield a third vector. This multiplication rule must satisfy certain conditions having to do jointly with vector space properties and multiplication properties. Indicating multiplication by the symbol $\circ$, we require that to every ordered pair of elements $x, y \in A$ there corresponds a third unique element of A, denoted by $x \circ y$, and called the *product* of x and y. The product should satisfy the following requirements:

$$1.\ \ (cx) \circ y = x \circ (cy) = c(x \circ y) \tag{3.7.43}$$

$$2.\ \ (x + y) \circ z = x \circ z + y \circ z \quad \text{(right distributive)} \tag{3.7.44}$$

$$3.\ \ x \circ (y + z) = x \circ y + x \circ z \quad \text{(left distributive)} \tag{3.7.45}$$

for any $x, y, z \in A$ and $c \in F$.

Associative Algebra

An example of an algebra is the set of all $m \times m$ matrices. The set of all $m \times m$ matrices forms an m^2 dimensional vector space. It also forms an algebra if we use for the $\circ$ operation ordinary matrix multiplication. Note that in this case multiplication is *associative*, that is, the multiplication rule satisfies the property

$$(x \ \circ \ y) \ \circ \ z = x \ \circ \ (y \ \circ \ z). \tag{3.7.46}$$

Lie Algebra

A second example of an algebra is the set of all 3-vectors with the multiplication rule given by the relation

$$\boldsymbol{a} \ \circ \ \boldsymbol{b} = \boldsymbol{a} \times \boldsymbol{b}. \tag{3.7.47}$$

Here × denotes the usual cross product. This algebra is *not* associative,

$$(\boldsymbol{a} \times \boldsymbol{b}) \times \boldsymbol{c} \neq \boldsymbol{a} \times (\boldsymbol{b} \times \boldsymbol{c}).$$

A Lie algebra L is an algebra for which the multiplication rule (sometimes now called a Lie product) satisfies two *further* properties. For convenience, multiplication of x and y will now be denoted by the symbol $[x, y]$,

$$[x, y] = x \ \circ \ y.$$

In using this customary notation, however, it should be understood that the bracket [,] does not necessarily refer to a commutator (or a Poisson bracket). Rather, in this context, it refers to the Lie product abstractly, and independently of any particular realization. The two additional properties for a Lie product are the following:

$$4.\ \ [x, y] = -[y, x] \quad \text{(antisymmetry)} \tag{3.7.48}$$

$$5.\ \ [x, [y, z]] + [y, [z, x]] + [z, [x, y]] = 0 \quad \text{(Jacobi condition or identity)} \tag{3.7.49}$$

We note that a Lie algebra is not associative. Instead, the associativity condition (7.46) has been replaced by the *Jacobi condition* (7.49). We also remark, because Lie algebras are often realized in terms of matrices, two elements in a Lie algebra are said to *commute* if their Lie product vanishes.

A subalgebra K of a Lie algebra L is a subset of L whose elements also satisfy the above properties 1 through 5. Let ℓ be any element of L. Then the set of all scalar multiples of ℓ, which by definition includes the zero element, evidently forms a subalgebra of L. Whether L has any other nontrivial subalgebras depends on the nature of L.

To settle these concepts into the mind, the reader is invited to verify that the set of all 3-vectors with the multiplication rule (7.47) forms a Lie algebra. Next, she or he should verify that the set of all $m \times m$ matrices forms a Lie algebra if the Lie product is taken to be the commutator. In both cases, it is necessary to verify that properties 1 through 5 above are satisfied for the particular Lie product involved.

3.7.5 Abstract Definition of a Lie Group

At this point we should make a side comment. We have defined, and will define, various Lie groups in the context of matrix groups. However, Lie groups can also be defined abstractly. Abstractly, a Lie group is a set G with the following properties:

1. G is a *manifold.* Roughly speaking this means that G, at and near each point, looks like Euclidean space of some fixed dimension m, and there are local coordinates described by m quantities $x_1, \cdots, x_m$. For example, consider the set of all real 2×2 matrices. Since each such matrix has 4 entries, this set can be viewed as being identical to E^4, 4-dimensional Euclidean space. Within this space is the set of 2×2 matrices M that satisfy (1.2), the set $Sp(2)$ of symplectic matrices. Since (1.2) constitutes a collection of algebraic equations among the entries in M, the set of symplectic matrices forms a manifold within E^4. Elements of $Sp(2)$ sufficiently near the identity can be written in the form (7.37), and we know that the dimension m of the set of 2×2 matrices of the form JS is 3. See (7.42). Let B_1 through B_3 be a basis for this set. See (7.66) through (7.68) for one possibility. Then, near the identity, we may write $M = \exp(x_1 B_1 + x_2 B_2 + x_3 B_3)$.

2. G is also a group. See the definition of an abstract group in Section 3.6.1. Moreover, the multiplication and inversion operations are required to be *continuous.* Suppose M and N are any two group elements. Continuity means that the coordinates of the product MN are continuous functions of the coordinates of M and N, and the coordinates of M^{-1} are continuous functions of the coordinates of M.

From these assumptions it can be proved that the group operations can actually be made *analytic.*[25] That is, there is a choice of coordinates such that the coordinates of the product MN are analytic functions of the coordinates of M and N, and the coordinates of M^{-1} are analytic functions of the coordinates of M. Based on this analyticity, one can differentiate group elements with respect to their coordinates. Next, from the group elements and their

[25] See Chapter 38 for a discussion of analyticity.

derivatives, one can construct entities (vector fields) that can be shown to form a Lie algebra of dimension m. Also, the process can be turned around to reconstruct the group elements from the Lie algebra. Among other things, it can be shown that the Jacobi identity for the Lie algebra is a consequence of the associativity property assumed for the operation of group multiplication. See Appendix R.

3.7.6 Classification of Lie Algebras

Let us return to the main discussion. One of the key discoveries of modern physics is that Lie groups are important for the description of Nature. (Mathematicians already knew earlier that they were important on aesthetic grounds.) Since Lie groups are important, it would be nice to classify them. Because of the close connection between Lie groups and Lie algebras, a natural starting point is to try to classifiy Lie algebras. This classification has been substantially carried out, initially by *Wilhelm Killing* (1847-1923), and subsequently by *Élie Cartan* (1869-1951) and others. Once Lie algebras/Lie groups have been classified, a next important step is to find *representations* for them in terms of matrices or possibly nonlinear transformations acting on some space. For examples, matrix representations of $su(2)$ are familiar from the Quantum Mechanical theory of angular momentum, matrix representations of $su(3)$ are described in Section 5.8, and matrix representations of $sp(2)$ through $sp(6)$ are described in Chapter 27. Finally, Section 5.12 describes the nonlinear action of $Sp(2n)$ on Siegel space.

The first step in the classification, or even description, of Lie algebras is the introduction of the concept of *structure constants*. Suppose L is a Lie algebra. Since a Lie algebra is a vector space, it must have a basis. Suppose some basis is selected, and let the various basis elements be denoted as $B_1, B_2, \cdots, B_k$ where k is the dimension of L. Now consider the Lie product of any two basis elements. Since the Lie product is again an element in the Lie algebra, it must be expandable in the terms of the basis elements. Consequently, there must be a set of coefficients $c^{\gamma}_{\alpha\beta}$, called *structure constants*, such that one has the relations

$$[B_\alpha, B_\beta] = \sum_\gamma c^{\gamma}_{\alpha\beta} B_\gamma. \tag{3.7.50}$$

Note that once the Lie product has been specified for the basis elements as in (7.50), then the Lie product for all other elements in L follows from the right and left distributive properties 2 and 3.[26]

Simple observation shows that, as a consequence of the antisymmetry condition (7.48), the structure constants must obey the relations

$$c^{\gamma}_{\alpha\beta} = -c^{\gamma}_{\beta\alpha}. \tag{3.7.51}$$

[26] Strictly speaking, what has been defined in Subsection 7.4 is a *free* Lie algebra. That is, *no* restrictions have been placed on the Lie product save antisymmetry and the Jacobi condition. [As an example of this kind of reasoning/terminology, we may say that the BCH series (7.41) is a free Lie algebraic result because it holds for all Lie algebras no matter what the structure constants may be.] By contrast, once a basis and structure constants have been selected/determined, the Lie algebra is no longer "free" in that it is then completely specified.

Somewhat lengthier analysis shows that, as a consequence of the Jacobi condition (7.49), the structure constants must also obey the relations

$$\sum_{\sigma}(c^{\sigma}_{\alpha\beta}c^{\tau}_{\gamma\sigma} + c^{\sigma}_{\beta\gamma}c^{\tau}_{\alpha\sigma} + c^{\sigma}_{\gamma\alpha}c^{\tau}_{\beta\sigma}) = 0. \tag{3.7.52}$$

Evidently, the problem of classifying all Lie algebras is equivalent to finding all sets of structure constants satisfying (7.51) and (7.52).

Of course, the structure constants depend on the choice of basis elements. Suppose $\tilde{B}_1, \tilde{B}_2, \cdots$ is another set of basis elements. Associated with this basis set there will be a set of structure constants $\tilde{c}^{\,\gamma}_{\alpha\beta}$ with the property

$$[\tilde{B}_{\alpha}, \tilde{B}_{\beta}] = \sum_{\gamma} \tilde{c}^{\,\gamma}_{\alpha\beta}\tilde{B}_{\gamma}. \tag{3.7.53}$$

Also, since both the B_{α} and $\tilde{B}_{\beta}$ are sets of basis elements, the B_{α} can be expanded in terms of the $\tilde{B}_{\beta}$, and vice versa. That is, there must be an *invertible* matrix T with the property

$$\tilde{B}_{\alpha} = \sum_{\beta} T_{\alpha\beta}B_{\beta}, \tag{3.7.54}$$

$$B_{\alpha} = \sum_{\beta}(T^{-1})_{\alpha\beta}\tilde{B}_{\beta}. \tag{3.7.55}$$

By using (7.50), (7.53), (7.54), and (7.55), we find that the structure constants $c^{\gamma}_{\alpha\beta}$ and $\tilde{c}^{\,\gamma}_{\alpha\beta}$ are connected by the relations

$$\tilde{c}^{\,\gamma}_{\alpha\beta} = \sum_{\mu\sigma\tau} T_{\alpha\mu}T_{\beta\sigma}(T^{-1})_{\tau\gamma}\, c^{\tau}_{\mu\sigma}, \tag{3.7.56}$$

$$c^{\gamma}_{\alpha\beta} = \sum_{\mu\sigma\tau}(T^{-1})_{\alpha\mu}(T^{-1})_{\beta\sigma}(T)_{\tau\gamma}\, \tilde{c}^{\,\tau}_{\mu\sigma}. \tag{3.7.57}$$

Often two Lie algebras are deemed to be *equivalent* if their structure constants are related by a change of basis. Sometimes it is important to consider the field from which the entries of T are taken. For example, two Lie algebras may be equivalent if the entries of T are allowed to be complex, but may be inequivalent if T is required to be real. Finally we remark that, in the classification or description of a Lie algebra, it is often convenient to choose a basis in such a way that the structure constants become as neatly organized as possible. For example, one might like to arrange that all the structure constants be real (or purely imaginary). This is possible for all the so-called *simple* Lie algebras.[27] One might also like to have as many of them vanish as possible, and to have those that do not vanish

[27] Here is a wonderful definition: A Lie algebra is called *simple* if it has no *ideals*. See Section 8.9. A Lie algebra is called *semisimple* if it is the *direct sum* of simple Lie algebras. (For the purposes of this definition, these simple Lie algebras must have dimension greater than one.) By direct sum it is meant that linear combinations can be formed of the elements in the various Lie algebras, but the Lie products of elements in different Lie algebras are defined to be zero. For example, $su(2)$ is simple. And, because $so(4) = su(2) \oplus su(2)$, $so(4)$ is semisimple.

satisfy some geometric properties. As will be illustrated by examples in Section 5.8 and Chapter 27, so doing for the simple Lie algebras was one of the accomplishments of Killing and Cartan.

The classification of all Lie algebras and Lie groups is a difficult task that lies beyond the scope of our discussion. We shall be primarily interested in the symplectic group and, as will be seen in Chapters 5 and 6, the group of all symplectic maps. However, there are certain Lie groups that arise naturally as subgroups of the symplectic group, and are therefore of direct interest to us. We close this section with a brief discussion of these groups.

Consider the set of all *invertible* $n \times n$ matrices. It is easily verified that this set of matrices forms a group. This group is called the *general linear* group, and is denoted by the symbols $GL(n, \mathbb{R})$ or $GL(n, \mathbb{C})$ depending on the choice of the field to be employed (real or complex). We also use the notation $GL(n, \mathbb{R}, +)$ to indicate the subgroup of $GL(n, \mathbb{R})$ consisting of matrices with positive determinant. Next consider the set of all $n \times n$ matrices with determinant +1. This set of matrices also forms a group, called the *special* linear group. It is denoted by the symbols $SL(n, \mathbb{R})$ or $SL(n, \mathbb{C})$. Evidently, the special linear group is a subgroup of the linear group. The groups $GL(n, \mathbb{R}), GL(n, \mathbb{C}), SL(n, \mathbb{R})$, and $SL(n, \mathbb{C})$ are all Lie groups. Their associated Lie algebras are denoted by the symbols $g\ell(n, \mathbb{R}), g\ell(n, \mathbb{C}), s\ell(n, \mathbb{R})$, and $s\ell(n, \mathbb{C})$, respectively.[28]

We have already learned in Section 3.6 about the orthogonal group and its connected subgroups. The groups $SO(n, \mathbb{R})$ and $SO(n, \mathbb{C})$ are Lie groups. Their associated Lie algebras are denoted by the symbols $so(n, \mathbb{R})$ and $so(n, \mathbb{C})$.

An $n \times n$ matrix U that satisfies the condition

$$U^\dagger U = I \tag{3.7.58}$$

is called *unitary*. The set of all such matrices forms a group, called the unitary group, and is denoted by the symbol $U(n)$. (Here the field is naturally taken to be the complex field.) Next consider the subset of all $n \times n$ unitary matrices having determinant +1. This subset also forms a group [a subgroup of $U(n)$] called the *special* unitary group (or sometimes the unitary *unimodular* group), and is denoted by the symbols $SU(n)$. The groups $U(n)$ and $SU(n)$ are Lie groups. Their associated Lie algebras are denoted by the symbols $u(n)$ and $su(n)$, respectively.

The groups $SU(n)$, $Sp(2n)$, $SO(n)$ and their related Lie algebras $su(n)$, $sp(2n)$, $so(n)$ have been studied extensively. In the mathematics literature they are referred to as the *classical groups* and are given the symbols A_ℓ, B_ℓ, C_ℓ, D_ℓ.[29] To facilitate entrée to this literature, Table 7.2 below summarizes the notation and a few key properties for these groups.[30] [Contrary to what might be expected, the groups/algebras $so(2\ell + 1)$ and $so(2\ell)$ have different structures, and hence are given the different symbols B_ℓ and D_ℓ.] These groups/algebras form infinite families since they exist for each integer value of $\ell = 1, 2, \cdots$. Here, as in the table, the subscript denotes the *rank* ℓ of the Lie algebra. The concept of rank

[28] It is customary for *special* to be denoted by the symbols S or s where special means having determinant +1; and G or g means *general*, i.e. having determinant possibly $\neq 1$. The exceptions to this convention are the notations Sp and sp, where S and s stand for *symplectic*.

[29] The term *classical groups* is due to *Weyl*.

[30] We note that some authors identify A_m with $s\ell(m+1, \mathbb{R})$. It can be shown that $su(n)$ and $s\ell(n, \mathbb{R})$ are equivalent over the complex field. See Exercise 7.29.

is defined in Sections 5.8 and 17.4. It is the dimension of the so called *Cartan* subalgebra of the full Lie algebra, which is a particular subalgebra having ℓ mutually commuting elements

By their definitions, the classical Lie algebras/groups can be realized in terms of certain matrices. These realizations are called the *fundamental* or *defining representations*. What is meant by a *representation* in this context is described in Subsection 7.7. (See also Exercise 7.36.) For a given classical Lie algebra, the dimension of the vector space on which the matrices for the fundamental representation act is given within the parentheses associated with its name. For example, the fundamental representation of $sp(2\ell)$ employs $2\ell \times 2\ell$ matrices.

In addition there are a finite number, namely 5, *exceptional groups/algebras* called $G_2(14)$, $F_4(52)$, $E_6(78)$, $E_7(133)$, $E_8(248)$. [We remark that the exceptional Lie algebras are nested as subalgebras according to the relations $G_2(14) \subset F_4(52) \subset E_6(78) \subset E_7(133) \subset E_8(248)$.] Taken together, the classical and exceptional Lie algebras comprise *all* the *simple* Lie algebras.

The naming convention for the exceptional Lie algebras/groups is somewhat different. Here the number within the parentheses associated with the name of such a Lie algebra is its *dimension* and, as done for A_ℓ through D_ℓ, the subscript is its rank ℓ. For example, $E_6(78)$ has dimension 78 and rank 6.

The exceptional Lie algebras/groups can also be realized in terms of matrices, and the smallest such matrices for any given exceptional Lie algebra/group provide its fundamental representation. The construction of these matrices is quite difficult and beyond the scope of our discussion. For examples, the fundamental representation of $G_2(14)$ involves 7×7 matrices, and the fundamental representation of $E_8(248)$ involves 248×248 matrices.

Finally, for the classical Lie algebras/groups, there is some redundancy for low values of ℓ. There are the equivalencies $su(2) = so(3) = sp(2)$, $sp(4) = so(5)$, and $su(4) = so(6)$.[31] In mathematical notation, these equivalencies are $A_1 = B_1 = C_1$, $B_2 = C_2$, and $A_3 = D_3$. Moreover, $so(2)$ is one dimensional; and $so(4)$ is not simple, but rather is the direct sum of two commuting $su(2)$ algebras: $so(4) = su(2) \oplus su(2)$.[32]

It has been discovered that all Lie algebras can be constructed by putting together in various ways the simple and the so-called *solvable* and *nilpotent* Lie algebras. The solvable and nilpotent Lie algebras have more or less all been classified. And, as we have just seen, all the simple Lie algebras have been classified. Thus, after over a century of work since the time of Lie, all finite-dimensional Lie algebras and their associated Lie groups are reasonably well classified and their properties reasonably well understood. For our purposes, we are primarily interested in simple Lie algebras and the Lie algebras made out of them.

[31] The equivalence $su(2) = so(3)$ is discussed in Exercise 3.7.31. The equivalence $sp(2) = su(2)$ is treated in Exercise 7.3.24. For the equivalences $sp(4) = so(5)$ and $su(4) = so(6)$ see Exercises 27.5.4 and 8.2.12, respectively.

[32] See Exercises 4.3.19 and 4.3.20.

Table 3.7.2: Cartan Catalog of the Classical and Exceptional Lie Groups/Algebras.

Classical Lie Groups/Algebras, infinite families with an entry for each integer value of ℓ:

Symbol	Lie Algebra	Dimension	Rank
A_ℓ	$su(\ell+1)$	$\ell(\ell+2)$	ℓ
B_ℓ	$so(2\ell+1)$	$\ell(2\ell+1)$	ℓ
C_ℓ	$sp(2\ell)$	$\ell(2\ell+1)$	ℓ
D_ℓ	$so(2\ell)$	$\ell(2\ell-1)$	ℓ

Exceptional Lie Groups/Algebras:

$E_6(78)$, $E_7(133)$, $E_8(248)$, $F_4(52)$, $G_2(14)$

3.7.7 Adjoint Representation of a Lie Algebra

We close this section with a brief discussion of the subject of *representations* of Lie algebras and, in particular, the *adjoint* representation. Suppose we are given a Lie algebra L. That is, we are told that there are k basis elements (where k is the dimension of L) and we are given a set of structure constants satisfying (7.51) and (7.52). A *representation* of L is a set of $m \times m$ *matrices* $\hat{B}_\alpha$ that, in analogy to (7.50), obeys the rules

$$\{\hat{B}_\alpha, \hat{B}_\beta\} = \sum_\gamma c^\gamma_{\alpha\beta}\hat{B}_\gamma \tag{3.7.59}$$

where here, to be perfectly explicit, $\{,\}$ denotes the matrix commutator,

$$\{\hat{B}_\alpha, \hat{B}_\beta\} = \hat{B}_\alpha\hat{B}_\beta - \hat{B}_\beta\hat{B}_\alpha. \tag{3.7.60}$$

This representation is said to be of *dimension* m since the matrices $\hat{B}_\alpha$ act on an m-dimensional vector space.

At this point some clarifying comments are in order. The first comment concerns definitions. The classical Lie algebras are specified by certain matrix properties associated with their initial specifications. For example, the initially defining matrices for $sp(2n)$ obey the relation (7.29). Upon verifying that these matrices form a Lie algebra, a basis can be chosen and the structure constants associated with this basis can be found. Once the structure constants have been specified, one can search for other sets of matrices which also form a Lie algebra with the same structure constants. However, these other matrices need not satisfy the the matrix properties associated with the initial specification. For example, general representation matrices for $sp(2n)$ need not satisfy (7.29).

The second comment has to do with dimensionality. Note that the dimension m of a representation is not to be confused with the dimension k of the underlying Lie algebra.

They may be different.[33] However, since the set of $m \times m$ matrices may be viewed as a vector space of dimension m^2, there must be the relation $k \le m^2$ if we require that the $\hat{B}_\alpha$ be linearly independent.[34]

As already described, by their specification, the Classical Lie algebras $su(n)$, $so(n)$, and $sp(2n)$ have natural matrix representations which are called the fundamental or defining representations. We also remarked that the Exceptional Lie algebras have fundamental matrix representations, but that their construction is complicated. The existence of a matrix representation for an arbitrary Lie algebra is even less obvious. The purpose of the present discussion is to observe that every Lie algebra L has a matrix representation, called the *adjoint* representation, which is constructed from the structure constants. This construction turns out to be quite elementary. [According to a much more difficult theorem of *Ado*, which is far beyond the scope of our discussion, every Lie algebra over the complex field is *isomorphic* to some matrix Lie algebra. That is, every (finite-dimensional) abstract Lie algebra may be viewed as (is isomorphic to) a subalgebra of some $g\ell(n, \mathbb{C})$. However, it is not the case that every finite-dimensional Lie group is isomorphic to a subgroup of some $GL(n, \mathbb{C})$. The metaplectic group is a counter example.]

After some trial and error in the search for a representation, we hit upon the matrices $\hat{B}_\alpha$ defined in terms of the structure constants by the rules

$$(\hat{B}_\alpha)_{\mu\nu} = c^\mu_{\alpha\nu}. \tag{3.7.61}$$

Note that these matrices are $k \times k$ where k is the dimension of L. (They are also *real* if the structure constants are real.) So, in this case, we have $m = k$.[35] Let us verify that the prescription (7.61) works. Using (7.60) and the rules for matrix multiplication, we write

$$\{\hat{B}_\alpha, \hat{B}_\beta\}_{\mu\nu} = (\hat{B}_\alpha \hat{B}_\beta)_{\mu\nu} - (\hat{B}_\beta \hat{B}_\alpha)_{\mu\nu} = \sum_{\mu'} (\hat{B}_\alpha)_{\mu\mu'} (\hat{B}_\beta)_{\mu'\nu} - (\hat{B}_\beta)_{\mu\mu'} (\hat{B}_\alpha)_{\mu'\nu}. \tag{3.7.62}$$

Inserting the definition (7.61) into (7.62) gives the result

$$\begin{aligned} \{\hat{B}_\alpha, \hat{B}_\beta\}_{\mu\nu} &= \sum_{\mu'} (c^\mu_{\alpha\mu'} c^{\mu'}_{\beta\nu} - c^\mu_{\beta\mu'} c^{\mu'}_{\alpha\nu}) \\ &= \sum_{\mu'} (c^{\mu'}_{\beta\nu} c^\mu_{\alpha\mu'} - c^{\mu'}_{\alpha\nu} c^\mu_{\beta\mu'}) = \sum_{\mu'} (c^{\mu'}_{\beta\nu} c^\mu_{\alpha\mu'} + c^{\mu'}_{\nu\alpha} c^\mu_{\beta\mu'}). \end{aligned} \tag{3.7.63}$$

[33] We also note that the word *representation* can have different meanings depending on context. Here, and in Section 5.8.5 and Chapter 27 and perhaps elsewhere, it means a set of matrices having some desired commutation rules. Another possibility, as in Sections 3.8 and 3.12, is that it may mean that some matrix may be written (represented) in a useful way as some function of some other matrix or matrices.

[34] For example, the fundamental representation of $sp(2)$ involves 2×2 matrices, and the dimension of $sp(2)$ is 3.

[35] Look at Table 7.2. For the Classical Lie algebras compare the dimension m of the fundamental representation with the dimension k of the Lie algebra. For example, in the case of $su(\ell+1)$, compare $m = \ell+1$ with $k = \ell(\ell+2)$. One finds that $m < k$ save for the cases of $so(2)$ and $so(3)$. That is, with these two exceptions, for the Classical Lie algebras the dimension of the fundamental representation is less than the dimension of the adjoint representation. In the case of $so(2)$, the fundamental representation is two-dimensional, and the Lie algebra is one-dimensional. In the case of $so(3)$, $m = k = 3$, and it turns out that the fundamental representation is the adjoint representation. See Exercise 7.30. It can be shown that $m < k$ for the Exceptional Lie algebras as well save for $E_8(248)$. For $E_8(248)$ the fundamental representation is the adjoint representation.

Here we have also used (7.51). But, from (7.52) with a change of indices, we find the relation

$$\begin{aligned}\sum_{\mu'}(c^{\mu'}_{\beta\nu}c^{\mu}_{\alpha\mu'}+c^{\mu'}_{\nu\alpha}c^{\mu}_{\beta\mu'}) &= -\sum_{\mu'}c^{\mu'}_{\alpha\beta}c^{\mu}_{\nu\mu'} \\ &= \sum_{\mu'}c^{\mu'}_{\alpha\beta}c^{\mu}_{\mu'\nu}=\sum_{\gamma}c^{\gamma}_{\alpha\beta}c^{\mu}_{\gamma\nu}. \qquad (3.7.64)\end{aligned}$$

Here we have again used (7.51). Upon combining (7.63), (7.64), and (7.61) we find the final result

$$\{\hat{B}_\alpha,\hat{B}_\beta\}_{\mu\nu}=\sum_{\gamma}c^{\gamma}_{\alpha\beta}c^{\mu}_{\gamma\nu}=\sum_{\gamma}c^{\gamma}_{\alpha\beta}(\hat{B}_\gamma)_{\mu\nu}, \qquad (3.7.65)$$

and hence (7.59) is satisfied.

As a concrete example, let us construct the adjoint representation of $sp(2,\mathbb{R})$. To begin, there is the 2×2 representation of $sp(2,\mathbb{R})$ which we have agreed to call the defining or fundamental representation. According to (7.34) the Lie algebra of $sp(2,\mathbb{R})$ in the defining representation consists of 2×2 matrices of the form JS with S real and symmetric. These matrices form a 3-dimensional vector space and therefore $k=3$. See (7.42) evaluated at $n=1$. A convenient basis for this vector space is provided by the matrices B_1, B_2, and B_3 given by the relations

$$\begin{aligned}B_1 &= (1/2)F \\ &= (1/2)\begin{pmatrix}0&1\\1&0\end{pmatrix}=(1/2)\sigma^1, \qquad (3.7.66)\end{aligned}$$

$$\begin{aligned}B_2 &= (1/2)B^0 \\ &= (1/2)\begin{pmatrix}0&1\\-1&0\end{pmatrix}=(i/2)\sigma^2, \qquad (3.7.67)\end{aligned}$$

$$\begin{aligned}B_3 &= (1/2)G \\ &= (1/2)\begin{pmatrix}1&0\\0&-1\end{pmatrix}=(1/2)\sigma^3. \qquad (3.7.68)\end{aligned}$$

See Section 5.6 where the matrices B^0, F, and G are constructed and their commutation rules are derived. (Here we have also referenced the Pauli matrices σ^α. See Exercise 3.7.31. This referencing will be useful later.) Note that B_1 and B_3 are Hermitian, and B_2 is anti-Hermitian.

From the commutation rules (5.6.18) though (5.6.20) and the definitions given by the first parts of (7.66) through (7.68) it follows that the B_α obey the commutation rules

$$\{B_1,B_2\}=-B_3, \qquad (3.7.69)$$

$$\{B_2,B_3\}=-B_1, \qquad (3.7.70)$$

$$\{B_3,B_1\}=B_2, \qquad (3.7.71)$$

which are a variant of the commutation rules for $sp(2,\mathbb{R})$. From these rules we see that the only nonzero structure constants in this case are given by the relations

$$c^3_{12} = -c^3_{21} = -1, \tag{3.7.72}$$

$$c^1_{23} = -c^1_{32} = -1, \tag{3.7.73}$$

$$c^2_{31} = -c^2_{13} = 1. \tag{3.7.74}$$

Correspondingly, according to (7.61), the adjoint representation has the associated elements

$$\hat{B}_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}, \tag{3.7.75}$$

$$\hat{B}_2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \tag{3.7.76}$$

$$\hat{B}_3 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{3.7.77}$$

The reader should check that the $\hat{B}_\alpha$ do indeed satisfy (7.59), i.e., the "hatted" version of (7.69) through (7.71). Note, in accord with the comments made earlier, these matrices do *not* satisfy the original defining relation (7.29).

How could one have guessed the construction (7.61)? There is a way, which at first may also seem obscure, but which will ultimately prove to be very useful. Suppose A is some element in the Lie algebra L. We know that a Lie algebra is a vector space. We are going to associate with A a *linear* operator, denoted by the symbols (ad A) and called the *adjoint* of A, that will send L into itself.[36] The action of this operator on any element C in L is defined by the rule

$$(\text{ad } A)C = [A, C]. \tag{3.7.78}$$

Since both C and $[A, C]$ are in L, the operator (ad A) does indeed send L into itself. It is also obviously linear because of the left distributive property (7.45) of the Lie product.

What can be said about these operators? First, they too form a linear vector space. To see this, suppose (ad B) is the operator associated with the element B in L. Then, from the definition (7.78) and the right distributive property (7.44), there is the relation

$$\{\text{ad } (A + B)\}C = [A + B, C] = [A, C] + [B, C] = (\text{ad } A)C + (\text{ad } B)C. \tag{3.7.79}$$

Since C is an arbitrary element in L, we may rewrite this relation in the operator form

$$\text{ad } A + \text{ad } B = \text{ ad } (A + B), \tag{3.7.80}$$

and take this result to be the definition of operator addition.

[36] Note that in this context the term *adjoint* is not to be confused with the concept of Hermitian conjugate.

Second, these operators also form a Lie algebra with the Lie product taken to be the commutator. This fact is partly obvious since we know that linear operators may be viewed as matrices and, as seen earlier, matrices do form a linear vector space and the matrix or linear operator commutator does satisfy the requirements for a Lie product. However, we have to verify that the linear operator commutator of two adjoint operators is again the adjoint operator for some element in L. Let us check. We find the results

$$(\text{ad } A)(\text{ad } B)C = (\text{ad } A)[B,C] = [A,[B,C]], \tag{3.7.81}$$

$$(\text{ad } B)(\text{ad } A)C = (\text{ad } B)[A,C] = [B,[A,C]]. \tag{3.7.82}$$

It follows that

$$\{(\text{ad } A),(\text{ad } B)\}C = [A,[B,C]] - [B,[A,C]]. \tag{3.7.83}$$

But, from the Jacobi identity and antisymmetry, we have the result

$$\begin{aligned}[A,[B,C]] - [B,[A,C]] &= [A,[B,C]] + [B,[C,A]] = -[C,[A,B]] \\ &= [[A,B],C] = \text{ ad } ([A,B])C.\end{aligned} \tag{3.7.84}$$

[Note that (7.64) is also a result of the Jacobi identity.] Upon combining (7.83) and (7.84) and recalling that C is any element in L, we may write the operator identity

$$\{(\text{ad } A),(\text{ad } B)\} = \text{ ad } ([A,B]), \tag{3.7.85}$$

which shows that the adjoint operators do indeed form a Lie algebra.

Moreover, this Lie algebra has the *same* structure constants as L. To see this, consider the operators (ad B_α) and compute:

$$\{(\text{ad } B_\alpha),\ (\text{ad } B_\beta)\} = \text{ ad } ([B_\alpha,B_\beta]) = \text{ ad } (\sum_\gamma c^\gamma_{\alpha\beta}B_\gamma) = \sum_\gamma c^\gamma_{\alpha\beta}(\text{ad } B_\gamma). \tag{3.7.86}$$

Finally, since the adjoint operators are linear operators, let us compute the matrix elements for their equivalent matrices. Suppose D is an arbitrary element in L. Since the B_α form a basis, D has an expansion of the form

$$D = \sum_\nu d_\nu B_\nu. \tag{3.7.87}$$

Let (ad B_α) act on D to produce a "transformed" D,

$$D^{\text{tr}} = (\text{ad } B_\alpha)D. \tag{3.7.88}$$

The transformed element D^{tr} has an expansion of the form

$$D^{\text{tr}} = \sum_\mu d^{\text{tr}}_\mu B_\mu. \tag{3.7.89}$$

How are the components d^{tr}_μ related to the components d_ν? From (7.87) through (7.89) we have the relations

$$\begin{aligned}\sum_\mu d^{\text{tr}}_\mu B_\mu &= D^{\text{tr}} = (\text{ad } B_\alpha)D = [B_\alpha,D] \\ &= [B_\alpha,\sum_\nu d_\nu B_\nu] = \sum_\nu [B_\alpha,B_\nu]d_\nu = \sum_{\nu\gamma} c^\gamma_{\alpha\nu}d_\nu B_\gamma.\end{aligned} \tag{3.7.90}$$

However, since the B_γ form a basis, the relation (7.90) is equivalent to the matrix relation

$$d_\mu^{\rm tr} = \sum_\nu c_{\alpha\nu}^\mu d_\nu. \tag{3.7.91}$$

Consequently, (7.88) is logically equivalent to (7.91). Moreover, by using the definition (7.61), the relation (7.91) can be rewritten in the form

$$d_\mu^{\rm tr} = \sum_\nu (\hat{B}_\alpha)_{\mu\nu} d_\nu. \tag{3.7.92}$$

We see that $\hat{B}_\alpha$ is simply the matrix corresponding to the linear operator (ad B_α); and the fact that these operators satisfy the commutation rules (7.86) implies that their matrix representatives must do so as well.

There is one last point to be made. Suppose it happens that there is some nonzero element in L, call it A, such that the Lie product of A with any element C in L vanishes,

$$[A, C] = 0 \text{ for all } C \in L. \tag{3.7.93}$$

Then we have the results

$$\text{ad } A = 0, \tag{3.7.94}$$

$$\hat{A} = 0. \tag{3.7.95}$$

Since the B_α form a basis, and A is nonzero, A must have an expansion of the form

$$A = \sum_\alpha a_\alpha B_\alpha \tag{3.7.96}$$

where at least some of the components a_α are nonzero. As a consequence of (7.80), (7.94), and (7.95) we find the result

$$\sum_\alpha a_\alpha \hat{B}_\alpha = \hat{A} = 0, \tag{3.7.97}$$

which shows that the $\hat{B}_\alpha$ in this case are *linearly dependent*. We see that while (by the definition of a basis) the B_α are linearly independent, it can happen that the $\hat{B}_\alpha$ are not. Therefore, it may happen that the adjoint representation of L provided by the $\hat{B}_\alpha$ is *not* isomorphic to L. If the Lie algebra provided by the matrices of a representation of a Lie algebra L is isomorphic to L, then this representation is said to be *faithful*. We have learned that the adjoint representation need not be faithful.

Exercises

3.7.1. Show that the Euclidean vector norm defined by (7.23) satisfies all the requirements for a vector norm. Show that the Euclidean matrix norm defined by the rule

$$(||A||_E)^2 = \text{tr}(A^\dagger A) = \sum_{jk} |A_{jk}|^2 \tag{3.7.98}$$

satisfies all the requirements for a matrix norm. (The Euclidean matrix norm is also sometimes called the *Frobenius* norm.) Note that the Euclidean vector and matrix norms are analogous in that both involve a sum of absolute values squared. Show that the Euclidean vector and matrix norms are consistent.

The Euclidean matrix norm is easy to compute, but is weaker than the maximum column sum norm, which is also easy to compute. Show that, for example in the $m \times m$ case,

$$||I|| = \sqrt{m} \tag{3.7.99}$$

for the Euclidean norm, while

$$||I|| = 1 \tag{3.7.100}$$

for the maximum column sum and spectral norms.

Show that

$$||J|| = \sqrt{2n} \tag{3.7.101}$$

for the Euclidean norm, while

$$||J|| = 1 \tag{3.7.102}$$

for the maximum column sum and spectral norms.

Suppose u and v are any two real vectors, and let (u, v) denote the usual real Euclidean inner product,

$$(u, v) = \sum_j u_j v_j. \tag{3.7.103}$$

Verify the *Schwarz inequality*

$$|(u, v)| \leq ||u|| \, ||v|| \tag{3.7.104}$$

where here the vector norm on the right side of (7.104) is the real Euclidean norm. Verify an analogous result for complex vectors when the Euclidean complex inner product is used, in which case

$$\langle u, v \rangle = \sum_j \bar{u}_j v_j = (\bar{u}, v). \tag{3.7.105}$$

Suppose u and v are two real $2n$-dimensional vectors that are symplectically conjugate in the sense that

$$(u, Jv) = \pm 1. \tag{3.7.106}$$

Verify the chain of reasoning

$$1 = |(u, Jv)| \leq ||u|| \, ||Jv|| \leq ||u|| \, ||J|| \, ||v|| \leq ||u|| \, ||v|| \tag{3.7.107}$$

where here the matrix spectral norm (which is consistent with the Euclidean vector norm) has been used for $||J||$. Thus, (7.106) implies the inequality

$$||u|| \, ||v|| \geq 1. \tag{3.7.108}$$

3.7.2. The Schwarz *inequality* (7.104) is a relation between the absolute value of an inner product and the norms of its ingredients. The *Lagrange identity* is an *equality* that reveals what terms have been omitted to make the Schwarz inequality a true inequality. Suppose

$u = (u_1, u_2, \cdots, u_n)$ and $v = (v_1, v_2, \cdots, v_n)$ are any two n-component vectors, real or complex. Then, according to the Lagrange identity, they satisfy the relation

$$\left(\sum_j u_j^2\right)\left(\sum_k v_k^2\right) - \left(\sum_j u_j v_j\right)^2 = (1/2)\sum_{jk}(u_j v_k - u_k v_j)^2. \tag{3.7.109}$$

The relation (7.109) may also be written in the form

$$\left(\sum_j u_j v_j\right)^2 = \left(\sum_j u_j^2\right)\left(\sum_k v_k^2\right) - (1/2)\sum_{jk}(u_j v_k - u_k v_j)^2. \tag{3.7.110}$$

If the entries in u and v are real, then the last term on the right side of (7.110) can never be positive. In that case there is the inequality

$$\left(\sum_j u_j v_j\right)^2 \leq \left(\sum_j u_j^2\right)\left(\sum_k v_k^2\right), \tag{3.7.111}$$

which can be written in the more compact form

$$(u, v)^2 \leq ||u||^2\, ||v||^2. \tag{3.7.112}$$

Evidently, the relations (7.104) and (7.112) are equivalent.

Suppose $\boldsymbol{u}$ and $\boldsymbol{v}$ are two real 3-component vectors. For this case, verify the identity

$$(\boldsymbol{u}\cdot\boldsymbol{v})^2 + (\boldsymbol{u}\times\boldsymbol{v})\cdot(\boldsymbol{u}\times\boldsymbol{v}) = (\boldsymbol{u}\cdot\boldsymbol{u})(\boldsymbol{v}\cdot\boldsymbol{v}), \tag{3.7.113}$$

and show that this identity is the Lagrange identity for the instance $n = 3$. How can the Lagrange identity be verified for the case of general n? Here is one way: Define the matrix A by the rule

$$A = |u)(v| - |v)(u| \tag{3.7.114}$$

where, in the formation of dyads, complex conjugation is *not* to be employed. Show that, by this definition, A is antisymmetric and has the matrix elements

$$A_{jk} = (e^j, Ae^k) = u_j v_k - u_k v_j. \tag{3.7.115}$$

Show, for any antisymmetric matrix A, that

$$\mathrm{tr}(A^2) = -\mathrm{tr}(A^T A) = -\sum_{jk}(A_{jk})^2. \tag{3.7.116}$$

Verify the dyadic relation

$$\begin{aligned} A^2 &= [|u)(v| - |v)(u|]\,[|u)(v| - |v)(u|] \\ &= |u)(v,u)(v| - |u)(v,v)(u| - |v)(u,u)(v| + |v)(u,v)(u|. \end{aligned} \tag{3.7.117}$$

Use this dyadic result to show that

$$\begin{aligned} \mathrm{tr}(A^2) &= (v,u)^2 - (u,u)(v,v) - (u,u)(v,v) + (u,v)^2 \\ &= 2(u,v)^2 - 2(u,u)(v,v). \end{aligned} \tag{3.7.118}$$

By comparing (7.116) and (7.118), show that

$$(u,v)^2 - (u,u)(v,v) = -(1/2)\sum_{jk}(A_{jk})^2. \tag{3.7.119}$$

Verify that (7.110) and (7.119) agree.

Suppose that the usual complex inner product (7.105) is of interest rather than the usual real inner product (7.103). In this case there is the associated Lagrange identity

$$\left(\sum_j |u_j|^2\right)\left(\sum_k |v_k|^2\right) - \left|\sum_j \bar{u}_j v_j\right|^2 = (1/2)\sum_{jk}|u_j v_k - u_k v_j|^2. \tag{3.7.120}$$

The relation (7.120) can also be written in the form

$$\left|\sum_j \bar{u}_j v_j\right|^2 = \left(\sum_j |u_j|^2\right)\left(\sum_k |v_k|^2\right) - (1/2)\sum_{jk}|u_j v_k - u_k v_j|^2. \tag{3.7.121}$$

Prove this result as follows: Define A exactly as before using (7.114). Show that

$$\mathrm{tr}(A\bar{A}) = -\mathrm{tr}(AA^\dagger) = -\sum_{jk}|A_{jk}|^2. \tag{3.7.122}$$

Verify the dyadic relation

$$\begin{aligned} A\bar{A} &= [|u)(v| - |v)(u|]\,[|\bar{u})(\bar{v}| - |\bar{v})(\bar{u}|] \\ &= |u)(v,\bar{u})(\bar{v}| - |u)(v,\bar{v})(\bar{u}| - |v)(u,\bar{u})(\bar{v}| + |v)(u,\bar{v})(\bar{u}|. \end{aligned} \tag{3.7.123}$$

Use this dyadic result to show that

$$\begin{aligned} \mathrm{tr}(A\bar{A}) &= (\bar{v},u)(v,\bar{u}) - (\bar{u},u)(v,\bar{v}) - (u,\bar{u})(\bar{v},v) + (u,\bar{v})(\bar{u},v) \\ &= 2|(\bar{u},v)|^2 - 2(\bar{u},u)(\bar{v},v). \end{aligned} \tag{3.7.124}$$

Compare (7.122) and (7.124) to show that

$$|(\bar{u},v)|^2 - (\bar{u},u)(\bar{v},v) = -(1/2)\sum_{jk}|A_{jk}|^2. \tag{3.7.125}$$

Verify that (7.121) and (7.125) agree.

3.7.3. Show that the maximum column sum matrix norm defined by (7.19) satisfies the relation

$$|B_{jk}| \leq \parallel B \parallel . \tag{3.7.126}$$

Show that the series (7.1) converges for any matrix B. (Hint: Show that the set of partial sums forms a Cauchy sequence.) Consider the matrix function $F(s)$ defined by the equation

$$F(s) = \exp(sB). \tag{3.7.127}$$

Show, by term-by-term differentiation of the power series for $F(s)$, that $F(s)$ satisfies the differential equation

$$dF(s)/ds = BF(s) = F(s)B \tag{3.7.128}$$

with the initial condition

$$F(0) = I. \tag{3.7.129}$$

Justify the required interchange of the operations of (infinite) summation and differentiation.

3.7.4. Show that the series (7.2) converges for A sufficiently near the identity matrix I. Note that when A is near the identity, then $\log(A)$ is near the zero matrix. Thus, any matrix A sufficiently near the identity has a generator, namely $\log(A)$. Moreover, from the work of Section 3.7.1, we know that such an A is infinitesimally generated.

3.7.5. Suppose that B_i and B_j are any two $m \times m$ matrices that commute,

$$\{B_i, B_j\} = 0. \tag{3.7.130}$$

Equivalently, we may say that the Lie products of B_i and B_j vanish. Show from the power series definition (7.1) that in this case there is the relation

$$\exp(s_i B_i) \exp(s_j B_j) = \exp(s_i B_i + s_j B_j), \tag{3.7.131}$$

where s_i and s_j are any scalars. Verify (7.7) and (7.10). Suppose there are k linearly independent elements $B_1, B_2, \cdots, B_k$ all of which mutually commute (Lie products mutually vanish) as in (7.130). Show that these elements span a Lie algebra L. Show that all elements in L commute. A Lie algebra with this property is called *Abelian*. Consider all elements $G(s_1, \cdots, s_k)$ of the form

$$G(s_1, \cdots, s_k) = \exp(s_1 B_1 + s_2 B_2 + \cdots + s_k B_k).$$

Show that these elements form a group with the property

$$G(0, \cdots, 0) = I$$

and the group multiplication rule

$$G(s_1, \cdots, s_k) G(t_1, \cdots, t_k) = G(s_1 + t_1, \cdots, s_k + t_k).$$

Show that all elements in this group commute with respect to group multiplication,

$$G(s_1, \cdots, s_k) G(t_1, \cdots, t_k) = G(t_1, \cdots, t_k) G(s_1, \cdots, s_k).$$

A group for which all elements commute is also called Abelian. We have shown, in the context of matrices, that exponentiating an Abelian Lie algebra produces an Abelian Lie group. Conversely, again in the matrix context, the Lie algebra of an Abelian Lie group is Abelian. The same can be shown to be true for all Lie algebras and their related Lie groups.

Finally, as a special case, consider all elements of the form

$$G(s) = \exp(sB)$$

where B is some matrix. They evidently form a one-parameter Abelian Lie subgroup of $GL(m)$.

3.7.6. Verify the relations (7.24) through (7.26).

3.7.7. Verify the relations given by (7.3) through (7.6) using the definitions (7.1) and (7.2).

3.7.8. Verify (7.29) using the expansions (7.28) and the symplectic condition (1.2).

3.7.9. The calculation leading from (7.30) to (7.32) involved interchanges of the operations of matrix multiplication and transposition, and the operation of summation. Verify that these interchanges do not affect the convergence of the infinite series involved.

3.7.10. Consider two matrices A and B related by (7.4). The purpose of this exercise is to show that the determinant of A is related to the trace of B. We will do so by setting up and solving a differential equation.

Suppose that ϵ is a small parameter and B is an arbitrary matrix. Verify the expansion

$$\det(I + \epsilon B) = 1 + \epsilon\ \mathrm{tr}(B) + O(\epsilon^2). \tag{3.7.132}$$

Let $f(\lambda)$ be the function

$$f(\lambda) = \det[\exp(\lambda B)].$$

Verify the expansion

$$\begin{aligned} f(\lambda + d\lambda) &= \det\{\exp[(\lambda + d\lambda)B]\} \\ &= \det[\exp(\lambda B)\exp(d\lambda B)] \\ &= \det[\exp(\lambda B)]\det[\exp(d\lambda B)] \\ &= f(\lambda)\det\{1 + d\lambda B + O[(d\lambda)^2]\} \\ &= f(\lambda)\{1 + d\lambda \mathrm{tr}(B) + O[(d\lambda)^2]\}. \end{aligned}$$

Show that $f(\lambda)$ obeys the differential equation

$$df/d\lambda = f(\lambda)\mathrm{tr}(B)$$

with the initial condition

$$f(0) = 1.$$

Show that this differential equation has the unique solution

$$f(\lambda) = \exp[\lambda \mathrm{tr}(B)].$$

Consequently show that if A and B are any two matrices related by (7.4), then

$$\det(A) = \det[\exp(B)] = f(1) = \exp[\mathrm{tr}(B)]. \tag{3.7.133}$$

The relation (7.133), sometimes also called *Liouville's formula*, is useful and memorable. Note that, in view of the differential equation obeyed by f, this formula is a special case of the Liouville-Ostrogradski formula derived in Exercise 1.4.6.

As an application, first verify that it is always possible to find a matrix S such that (7.34) is true. Next verify that the matrix JS is traceless if S is symmetric. Finally, show that any matrix of the form $\exp(JS)$ must have determinant +1.

3.7.11. Verify the details of the calculation described in (7.37) and (7.38) using the series definition of the exponential function as given by (7.1).

3.7.12. This exercise presents two challenges:

- Show that the matrix given by the relation

$$M = \begin{pmatrix} -1 & -1 \\ 0 & -1 \end{pmatrix} \tag{3.7.134}$$

is symplectic, but cannot be written in the form (7.36).

- What happens if the −1 in the upper right corner of (7.134) is replaced by +1? See the matrix N below. Can the resulting symplectic matrix be written in the form (7.36)?

As preparatory observations, verify the symplectic conjugacy relation (see Exercises 5.7 and 8.13)

$$N = \begin{pmatrix} -1 & +1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \begin{pmatrix} -1 & -1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix}. \tag{3.7.135}$$

Note that the conjugating matrix A in this case [see (5.50)] is symplectic, but complex. Next, suppose M and N are *any* two $2n \times 2n$ symplectic matrices that are symplectically conjugate. Suppose also that M can be written in the form (7.36). Show that the same must then also be true for N.

Hint for meeting the challenges: Take them in opposite order. Now let N again be the matrix on the left side of (7.135). First prove that that N cannot be diagonalized by a similarly transformation. Next, suppose N is written in exponential form,

$$N = \exp(E). \tag{3.7.136}$$

This is possible because N is invertible. Verify this claim! The matrix E also cannot be diagonalized by a similarity transformation, because if it could be so diagonalized, then so could N. Therefore both eigenvalues of E must be identical. If we *assume* that E is of the form JS, then E must be traceless, and these eigenvalues must both be zero. Consequently, there is a similarity transformation B that brings E to the Jordan form,

$$BEB^{-1} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}. \tag{3.7.137}$$

It follows that N can be written in the form

$$\begin{aligned} N=\exp(E) &= \exp\left[B^{-1}\begin{pmatrix} 0 & 1\\ 0 & 0\end{pmatrix}B\right] = B^{-1}\left[\exp\begin{pmatrix} 0 & 1\\ 0 & 0\end{pmatrix}\right]B \\ &= B^{-1}\begin{pmatrix} 1 & 1\\ 0 & 1\end{pmatrix}B. \end{aligned} \tag{3.7.138}$$

But (7.138) is absurd because the eigenvalues of N are both -1 whereas the eigenvalues of the matrix on the right side of (7.138) are both $+1$. [Look ahead to Exercise 7.16. Compute the characteristic polynomial of the matrix product on the right side of (7.138).] Conclude that our assumption has produced a contradiction, and therefore neither N nor M can be written in the exponential form (7.36).

Show that the matrix $L=-M$, with M given by (7.134), is also symplectic, and can be written in the form (7.36). Find S for this case. See Exercise 5.6.7.

3.7.13. Verify that the set of matrices of the form JS (that is, the set of all Hamiltonian matrices) is indeed a Lie algebra by showing that properties i through iii are satisfied.

3.7.14. Let $B=JS$ be a *real* $2n\times 2n$ Hamiltonian matrix. Section 3.4 described the eigenvalue spectrum of real symplectic matrices. Derive related results for the eigenvalue spectrum of real Hamiltonian matrices. See (7.33). If $P(\lambda)=\det\,(JS-\lambda I)$ is the characteristic polynomial of a real Hamiltonian matrix, show that all its coefficients are real, and hence $\overline{P}(\lambda)=P(\overline{\lambda})$. Show that $P(\lambda)=P(-\lambda)$ and hence P contains only even powers of λ. Hint: Verify the chain of relations

$$\begin{aligned} \det\,(JS-\lambda I) &= \det\,[(JS-\lambda I)^T]=\det\,(-SJ-\lambda I) \\ &= \det\,[J(-SJ-\lambda I)J^{-1}]=\det\,(-JS-\lambda I) \\ &= \det\,[(-I)(JS+\lambda I)]=\det\,(-I)\det\,(JS+\lambda I) \\ &= \det\,(JS+\lambda I). \end{aligned}$$

Here we have used that fact that, because $-I$ is $2n\times 2n$, $\det\,(-I)=1$. Show that if λ is an eigenvalue, so are $\overline{\lambda}$ and $-\lambda$. Show that if $\lambda=0$ is an eigenvalue, it must have even multiplicity. Thus, if α is a real eigenvalue, $-\alpha$ must also be an eigenvalue, and real eigenvalues must come in $\pm\alpha$ *pairs*. Similarly, if $i\beta$ is a pure imaginary eigenvalue, $-i\beta$ must also be an eigenvalue, and pure imaginary eigenvalues must come in $\pm i\beta$ *pairs*. Finally, if $\alpha+i\beta$ is a complex eigenvalue, there must be a *quartet* of complex eigenvalues $\pm\alpha\pm i\beta$ with all signs taken independently. Section 3.4.4 showed that the problem of finding the eigenvalues of a $2n\times 2n$ symplectic matrix can be simplified to that of finding the roots of a polynomial of degree n followed by the solution of a quadratic equation. Show that the same is true for a Hamiltonian matrix. Show that the eigenvalues can be found in terms of radicals for the cases $n\leq 4$.

3.7.15. Let B be a Hamiltonian matrix. See (7.33). Show that B obeys the relation

$$KB=-B^TK \tag{3.7.139}$$

with K given by (5.3). Using the angular inner product (5.2), study the eigenvector structure of real Hamiltonian matrices in a manner similar to that done for symplectic matrices in Section 3.5. You will need the eigenvalue spectrum results of Exercise 7.14.

3.7.16. The *characteristic polynomial* $P(\lambda)$ of a matrix A is defined by the equation

$$P(\lambda) = \det(A - \lambda I). \tag{3.7.140}$$

The solutions of the equation $P(\lambda) = 0$ are the eigenvalues of A. Show that the matrices A and $A' = SAS^{-1}$, where S is any invertible matrix, have the same characteristic polynomial and hence the same eigenvalues. You have verified that the set of eigenvalues is invariant under *similarity* transformations.

3.7.17. Suppose A is $m \times m$ and let $P(\lambda)$ be its characteristic polynomial (7.140).

a) Verify that $P(\lambda)$ has the expansion

$$P(\lambda) = \sum_{\ell=0}^{m} a_\ell (-\lambda)^\ell \tag{3.7.141}$$

with

$$a_0 = \det(A) \tag{3.7.142}$$

and

$$a_m = 1. \tag{3.7.143}$$

What can be said about the other a_ℓ? Using the results (7.1) through (7.4) and the result of Exercise (7.7), we may write the relations

$$\begin{aligned} P(\lambda) &= \det(A - \lambda I) = \det[(-\lambda I)(I - A/\lambda)] \\ &= (-\lambda)^m \det(I - A/\lambda) \\ &= (-\lambda)^m \det\{\exp[\log(I - A/\lambda)]\} \\ &= (-\lambda)^m \det\{\exp[-\sum_{\ell=1}^{\infty}(A/\lambda)^\ell/\ell]\} \\ &= (-\lambda)^m \exp[-\sum_{\ell=1}^{\infty}(1/\lambda)^\ell(1/\ell)\text{tr}(A^\ell)]. \end{aligned} \tag{3.7.144}$$

b) Verify the statement made above. Note that the series employed will certainly converge for λ large enough, because $\|A/\lambda\|$ is then small.

c) Now expand out the exponential function, and collect powers of λ to get an expression of the form

$$P(\lambda) = \sum_{\ell=-\infty}^{m} a_\ell (-\lambda)^\ell. \tag{3.7.145}$$

It follows from (7.141) that $a_\ell = 0$ for $\ell < 0$. Show that this fact gives an infinite collection of identities. Find the first few coefficients a_{m-1}, a_{m-2}, $\cdots$ and verify the results

$$a_{m-1} = \text{ tr}(A), \tag{3.7.146}$$

$$a_{m-2} = \{[\text{tr}(A)]^2 - \text{ tr}(A^2)\}/2, \tag{3.7.147}$$

$$a_{m-3} = (1/3)\ \mathrm{tr}(A^3) - (1/2)[\mathrm{tr}(A)][\mathrm{tr}(A^2)] + (1/6)[\mathrm{tr}(A)]^3, \tag{3.7.148}$$

$$\begin{aligned} a_{m-4} &= (1/24)\{[\mathrm{tr}(A)]^4 - 6[\mathrm{tr}(A)]^2[\mathrm{tr}(A^2)] + 3[\mathrm{tr}(A^2)]^2 \\ &+ 8[\mathrm{tr}(A)][\mathrm{tr}(A^3)] - 6[\mathrm{tr}(A^4)]\}. \end{aligned} \tag{3.7.149}$$

Show that all the a_ℓ are functions of $[\mathrm{tr}(A^j)]^k$ for various values of j and k. Verify (4.25) and (4.26). Make a similar study of $[1/P(\lambda)]$.

d) Show that $\det(A)$ is also expressible in terms of $[\mathrm{tr}(A^j)]^k$. Verify the results

$$\det(A) = \{[\mathrm{tr}(A)]^2 - \mathrm{tr}(A^2)\}/2 \text{ when } m = 2, \tag{3.7.150}$$

$$\det(A) = [\mathrm{tr}(A^3)]/3 - [\mathrm{tr}(A)][\mathrm{tr}(A^2)]/2 + [\mathrm{tr}(A)]^3/6 \text{ when } m = 3, \tag{3.7.151}$$

$$\begin{aligned} \det(A) &= (1/24)\{[\mathrm{tr}(A)]^4 - 6[\mathrm{tr}(A)]^2[\mathrm{tr}(A^2)] + 3[\mathrm{tr}(A^2)]^2 \\ &+ 8[\mathrm{tr}(A)][\mathrm{tr}(A^3)] - 6[\mathrm{tr}(A^4)]\} \text{ when } m = 4, \text{ etc.} \end{aligned} \tag{3.7.152}$$

e) As in Exericse 7.10, let ϵ be a small parameter and C an arbitrary matrix. Verify the results

$$(I + \epsilon C) = \exp[\log(I + \epsilon C)] = \exp[-\sum_{n=1}^{\infty}(-\epsilon C)^n/n], \tag{3.7.153}$$

$$\begin{aligned} \det(I + \epsilon C) &= \exp[-\sum_{n=1}^{\infty}(1/n)(-\epsilon)^n \mathrm{tr}(C^n)] \\ &= 1 + \epsilon \mathrm{tr}(C) + (\epsilon^2/2)\{[\mathrm{tr}(C)]^2 - \mathrm{tr}(C^2)\} + \cdots . \end{aligned} \tag{3.7.154}$$

3.7.18. Let A and B be any two $m \times m$ matrices. Define matrices C and D by the equations $C = AB, D = BA$. Show that C and D have the same eigenvalue spectrum. Hint: Use Exercise (7.17) to show that they have the same characteristic polynomial.

3.7.19. Given an $n \times n$ matrix A, equation (7.140) gives a polynomial $P(\lambda)$. Show that $P(\lambda)$ has the leading term $(-\lambda)^n$. Consider the inverse problem: given a polynomial $P(\lambda)$ with leading term $(-\lambda)^n$, can one find a matrix A such that $P(\lambda)$ is the characteristic polynomial for A? Suppose the roots of $P(\lambda)$ are known. Call them λ_j. Find a *diagonal* A such that (7.140) holds. Remarkably, one does not need to know the roots to find an A that works. Show that the matrix A, called the *companion matrix* for $P(\lambda)$, given by

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & & & & 1 \\ b_1 & b_2 & \cdots & & b_n \end{pmatrix} \tag{3.7.155}$$

has the characteristic polynomial

$$P(\lambda) = (-1)^n(\lambda^n - b_n\lambda^{n-1} - b_{n-1}\lambda^{n-2} - \cdots - b_1). \tag{3.7.156}$$

Show that this A may not always be diagonalizable. Hint: Study the 2×2 case. Show that a general eigenvector is of the form $(1, \lambda)^T$, and show that there is only one (linearly independent) eigenvector if the eigenvalues are degenerate. Generalize to the $n \times n$ case and show that there are as many linearly independent eigenvectors as there are distinct eigenvalues.

3.7.20. Verify that the cross-product algebra given by (7.47) is not associative. Verify that it is a Lie algebra. In particular, check the Jacobi condition.

There is a theorem in plane Euclidean geometry to the effect that the three altitudes of a triangle intersect in a common point (called the *orthocenter*). (This point is in the interior of the triangle if the triangle is *acute*. Recall that a triangle is called acute if all its angles are less than 90 degrees.) It can be shown that the existence of a common intersection is a consequence of the Jacobi identity for the cross-product Lie algebra. Google “Jacobi identity altitudes of a triangle”.

3.7.21. Verify that the set of all $m \times m$ matrices with the multiplication rule defined by $[A, B] = AB - BA$ forms a Lie algebra. In particular, check the Jacobi condition.

3.7.22. Given any algebra, define the *associator* $\mathrm{A}(x, y, z)$ of any 3 elements x, y, z by the rule $\mathrm{A}(x, y, z) = (x \circ y) \circ z - x \circ (y \circ z)$. An algebra is called associative if the associator vanishes. Show that a Lie algebra is generally not associative. Hint: Use the Jacobi condition (and antisymmetry) to compute the associator. From this perspective, the Jacobi condition (along with antisymmetry) may be viewed as a rule that specifies the associator.

3.7.23. Suppose the vectors $\boldsymbol{e}_1 = \boldsymbol{e}_x, \boldsymbol{e}_2 = \boldsymbol{e}_y$, and $\boldsymbol{e}_3 = \boldsymbol{e}_z$ form a right-handed orthonormal triad in three-dimensional Euclidean space. Use them to form a basis for the Lie algebra (7.47). Find the structure constants $c^{\gamma}_{\alpha\beta}$ for this Lie algebra and basis. Show that these structure constants are related to the *Levi-Civita* tensor $\epsilon_{\alpha\beta\gamma}$. Consider a complex “spherical” basis $\boldsymbol{e}_{-1}, \boldsymbol{e}_0, \boldsymbol{e}_{+1}$ defined by the relations

$$\boldsymbol{e}_{+1} = -(1/\sqrt{2})(\boldsymbol{e}_x + i\boldsymbol{e}_y), \tag{3.7.157}$$

$$\boldsymbol{e}_0 = i\boldsymbol{e}_z,$$

$$\boldsymbol{e}_{-1} = (1/\sqrt{2})(\boldsymbol{e}_x - i\boldsymbol{e}_y).$$

Find the structure constants for this choice of basis.

3.7.24. Verify the relations (7.51), (7.52), (7.56), and (7.57). Verify that if (7.51) and (7.52) hold for some basis set, then the Lie algebraic properties (7.48) and (7.49) (antisymmetry and Jacobi condition) are satisfied.

3.7.25. Classify all two- and three-dimensional Lie algebras.

3.7.26. Verify that $GL(n, \mathbb{R}), GL(n, \mathbb{C}), SL(n, \mathbb{R})$, and $SL(n, \mathbb{C})$ are indeed groups. Characterize the Lie algebras $g\ell(n, \mathbb{R}), g\ell(n, \mathbb{C}), s\ell(n, \mathbb{R})$, and $s\ell(n, \mathbb{R})$. That is, what properties are satisfied by such matrices? Find the dimensions of these Lie algebras. In the complex case, find the dimension over both the real and complex fields. [Hint: Use the relation (7.133).]

3.7.27. Verify that $O(n,\mathbb{R}), O(n,\mathbb{C}), SO(n,\mathbb{R})$, and $SO(n,\mathbb{C})$ are indeed groups. Characterize the Lie algebras $so(n,\mathbb{R})$ and $so(n,\mathbb{C})$. That is, what properties are satisfied by such matrices? Find the dimensions of these Lie algebras. In the complex case, find the dimension over both the real and complex fields. [Hint: Use the relation (7.133).]

3.7.28. Show from (7.58) that $|\det(U)| = 1$. Verify that $U(n)$ and $SU(n)$ are indeed groups. Show that the set of $n \times n$ anti-Hermitian matrices forms a Lie algebra. This set is $u(n)$, the Lie algebra of $U(n)$. Find its dimension. Show that the set of all traceless matrices in $u(n)$ forms a sub Lie algebra. This sub Lie algebra is called $su(n)$. Find its dimension. Show that $su(n)$ is the Lie algebra of the group $SU(n)$. [Hint: Use the relation (7.133).]

3.7.29. By construction, because only real matrices are involved in their definitions, there are basis choices for the Lie algebras $so(n,\mathbb{R})$ and $sp(2n,\mathbb{R})$ for which the structure constants are *real*. What about the case of $su(n)$? Let $\mathcal{A}$ be the set of all real $n \times n$ antisymmetric matrices. Since all the diagonal entries in all antisymmetric matrices vanish, every antisymmetric matrix is traceless. Show that the dimension of $\mathcal{A}$ is $(n^2 - n)/2$. Let $\mathcal{S}$ be the set of all real $n \times n$ symmetric and traceless matrices. Show that the dimension of $\mathcal{S}$ is $[(n^2 - n)/2 + n - 1]$. Let the matrices A_j and S_k form bases for the sets $\mathcal{A}$ and $\mathcal{S}$, respectively. Show that the matrices A_j and iS_k form a basis for $su(n)$. See Exercise 7.28. Verify that the commutator of any two matrices is traceless. Verify that the commutator of any two antisymmetric matrices is antisymmetric. Verify that the commutator of an antisymmetric matrix and a symmetric matrix is symmetric. Verify that the commutator of any two symmetric matrices is antisymmetric. We can write these relations symbolically in the form

$$\{A, A'\} \propto A'', \tag{3.7.158}$$

$$\{A, S\} \propto S', \tag{3.7.159}$$

$$\{S, S'\} \propto A. \tag{3.7.160}$$

Correspondingly, verify that there are also the relations

$$\{A, A'\} \propto A'', \tag{3.7.161}$$

$$\{A, (iS)\} \propto (iS'), \tag{3.7.162}$$

$$\{(iS), (iS')\} \propto A. \tag{3.7.163}$$

Thus, show that in this basis, which may be viewed as a natural basis for $su(n)$, the structure constants are all real.

Consider next the Lie algebra $s\ell(n,\mathbb{R})$. Show that the matrices A_j and S_k form a basis for this Lie algebra. See Exercise 7.26. Thus show that $su(n)$ and $s\ell(n,\mathbb{R})$ have the same dimension. Show, in fact, that $su(n)$ and $s\ell(n,\mathbb{R})$ are *equivalent* over the complex field.

3.7.30. Verify that the $\hat{B}_j$ given by (7.75) through (7.77) for the $sp(2,\mathbb{R})$ case satisfy commutation rules analogous to (7.69) through (7.71). Verify that in this case the $\hat{B}_j$ are linearly independent.

3.7.31. This exercise explores the relations between the Lie algebras $su(2)$, $so(3,\mathbb{R})$, and the cross-product Lie algebra. It presumes that you have worked, or at least read, Exercises 7.23, 7.27, 7.28, and 7.29.

Define the *Pauli* matrices σ^α for $\alpha = 1,2,3$ by the rules

$$\sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{3.7.164}$$

$$\sigma^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \Leftrightarrow \sigma^2 = -iJ_2 \Leftrightarrow J_2 = i\sigma^2, \tag{3.7.165}$$

$$\sigma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{3.7.166}$$

[We remark that the Pauli matrices were discovered by Klein some 50 years prior to the time of Pauli. They came to bear Pauli's name because he was the first to use them to describe electron spin. Pauli (1900-1958) died in room *137* $\simeq 1/\alpha$ of the Red-Cross hospital at Zurich.] Verify that the Pauli matrices are traceless, Hermitian,

$$(\sigma^\alpha)^\dagger = \sigma^\alpha, \tag{3.7.167}$$

and satisfy the relations

$$(1/2)\mathrm{tr}(\sigma^\alpha\sigma^\beta) = \delta_{\alpha\beta}. \tag{3.7.168}$$

Verify also that the Pauli matrices are unitary, $(\sigma^\alpha)^\dagger\sigma^\alpha = I$. (For further properties of the Pauli matrices, see Section 5.7 and Exercises 5.7.2 and 5.7.7.)

Let K^1 through K^3 be the traceless anti-Hermitian matrices

$$K^1 = (-i/2)\sigma^1 = (-i/2)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{3.7.169}$$

$$K^2 = (-i/2)\sigma^2 = (-i/2)\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \tag{3.7.170}$$

$$K^3 = (-i/2)\sigma^3 = (-i/2)\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{3.7.171}$$

In Exercise 7.27 you should have found that the Lie algebra $su(n)$ consists of all $n\times n$ traceless anti-Hermitian matrices. Show that K^1 through K^3 form a basis for the Lie algebra $su(2)$. Show that they obey the multiplication and commutation rules

$$K^\alpha K^\beta = (1/2)K^\gamma, \tag{3.7.172}$$

$$\{K^\alpha, K^\beta\} = K^\gamma, \tag{3.7.173}$$

where α, β, γ is any cyclic permutation of $1, 2, 3$. Thus, with this choice of basis, the structure constants for $su(2)$ are the components of the Levi-Civita tensor,

$$c^\gamma_{\alpha\beta} = \epsilon_{\alpha\beta\gamma}. \tag{3.7.174}$$

We remark that we have been following what might be called a mathematician's approach to $su(2)$ [and $so(3,\mathbb{R})$] in which basis elements are chosen so that the structure constants are all *real*. For a quantum physicist's approach for which a basis is chosen to make all the structure constants *pure imaginary*, see Exercise 7.43.

Verify also that the K^α obey the anticommutation rules

$$\{K^\alpha, K^\beta\}_+ = K^\alpha K^\beta + K^\beta K^\alpha = -(1/2)\delta_{\alpha\beta} I. \tag{3.7.175}$$

Let $\boldsymbol{a}$ be a three-component vector with entries (a_1, a_2, a_3). Introduce the notation

$$\boldsymbol{a} \cdot \boldsymbol{K} = \sum_\alpha a_\alpha K^\alpha. \tag{3.7.176}$$

Show that there is the multiplication rule

$$(\boldsymbol{a} \cdot \boldsymbol{K})(\boldsymbol{b} \cdot \boldsymbol{K}) = -(1/4)(\boldsymbol{a} \cdot \boldsymbol{b})I + (1/2)(\boldsymbol{a} \times \boldsymbol{b}) \cdot \boldsymbol{K}. \tag{3.7.177}$$

Compute the matrices $(K^\alpha)^2$. Observe that they are diagonal, and hence mutually commute. Show that they sum to $-(3/4)I$.

Let L^1 through L^3 be the matrices

$$L^1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \tag{3.7.178}$$

$$L^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \tag{3.7.179}$$

$$L^3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{3.7.180}$$

In Exercise 7.27 you should have found that the Lie algebra $so(n,\mathbb{R})$ consists of all $n \times n$ real antisymmetric matrices. Show that L^1 through L^3 form a basis for the Lie algebra $so(3,\mathbb{R})$. Show that they also obey the commutation rules

$$\{L^\alpha, L^\beta\} = L^\gamma \tag{3.7.181}$$

where α, β, γ is any cyclic permutation of $1, 2, 3$. Evidently, according to (7.173) and (7.181), the Lie algebras $su(2)$ and $so(3,\mathbb{R})$ have the the same structure constants, and are therefore the same. Compute the matrices $(L^\alpha)^2$. Observe that they are diagonal, and hence mutually commute. Show that they sum to $-2I$.

Let the matrices $\hat{K}^\alpha$ be the adjoint representation matrices associated with the K^α. See (7.61). Verify the relations

$$(L^\alpha)_{\beta\gamma} = -\epsilon_{\alpha\beta\gamma} \text{ and therefore } \hat{K}^\alpha = L^\alpha. \tag{3.7.182}$$

You have shown that the L^α matrices are those for the adjoint representation of $su(2)$. Since $su(2)$ and $so(3,\mathbb{R})$ have the same structure constants, show that the adjoint representation of $so(3,\mathbb{R})$ is the fundamental representation.

Show from (7.177) that there is the relation

$$\{\boldsymbol{a}\cdot\boldsymbol{K},\boldsymbol{b}\cdot\boldsymbol{K}\}=(\boldsymbol{a}\times\boldsymbol{b})\cdot\boldsymbol{K}. \tag{3.7.183}$$

Define $\boldsymbol{a}\cdot\boldsymbol{L}$ in an analogous way to (7.176) and show that there is also the relation

$$\{\boldsymbol{a}\cdot\boldsymbol{L},\boldsymbol{b}\cdot\boldsymbol{L}\}=(\boldsymbol{a}\times\boldsymbol{b})\cdot\boldsymbol{L}. \tag{3.7.184}$$

You have shown that the cross-product Lie algebra (7.47) is intimately related to the Lie algebra for $su(2)$ and $so(3,\mathbb{R})$. Indeed, let $\boldsymbol{e}_1$ through $\boldsymbol{e}_3$ be the unit vectors of Exercise 7.22. Make the correspondences

$$\boldsymbol{e}_\alpha\leftrightarrow K^\alpha\leftrightarrow L^\alpha. \tag{3.7.185}$$

Then, you have shown that there are also the correspondences

$$(\boldsymbol{e}_\alpha\times\boldsymbol{e}_\beta)\leftrightarrow\{K^\alpha,K^\beta\}\leftrightarrow\{L^\alpha,L^\beta\}. \tag{3.7.186}$$

Finally, in Exercise 7.23 you should have found that the structure constants for the cross-product Lie algebra are the same as those for $su(2)$ and $so(3,\mathbb{R})$. Therefore the cross-product Lie algebra is the same as that of $su(2)$ and $so(3,\mathbb{R})$.

We have studied the Lie algebras $su(2)$, $so(3,\mathbb{R})$ and the cross-product Lie algebra. We now explore the relation between the groups $SU(2)$ and $SO(3,\mathbb{R})$. Begin with the case of $SU(2)$. Let $\boldsymbol{n}$ be a unit vector. Define $SU(2)$ matrices $v(\theta,\boldsymbol{n})$ by the rule

$$v(\theta,\boldsymbol{n})=\exp(\theta\boldsymbol{n}\cdot\boldsymbol{K}). \tag{3.7.187}$$

Show, in accord with Section 3.8.1, that any $SU(2)$ matrix can be written in the form (7.187). Show that

$$(\boldsymbol{n}\cdot\boldsymbol{K})^2=-(1/4)I. \tag{3.7.188}$$

Use this relation to sum the series implied by (7.187) to find the explicit result

$$v(\theta,\boldsymbol{n})=I\cos(\theta/2)+2(\boldsymbol{n}\cdot\boldsymbol{K})\sin(\theta/2). \tag{3.7.189}$$

Show that

$$v(2\pi,\boldsymbol{n})=-I, \tag{3.7.190}$$

$$v(4\pi,\boldsymbol{n})=+I. \tag{3.7.191}$$

Show that $SU(2)$ is covered once and only once when $\theta\in[0,2\pi]$ and $\boldsymbol{n}$ is allowed to be any unit vector.

As special cases of (7.187), verify the relations

$$v(\theta,\boldsymbol{e}_1)=\exp(\theta K^1)=\exp[(-i/2)\theta\sigma^1]=\begin{pmatrix}\cos(\theta/2) & -i\sin(\theta/2)\\ -i\sin(\theta/2) & \cos(\theta/2)\end{pmatrix}, \tag{3.7.192}$$

$$v(\theta,\boldsymbol{e}_2)=\exp(\theta K^2)=\exp[(-i/2)\theta\sigma^2]=\begin{pmatrix}\cos(\theta/2) & -\sin(\theta/2)\\ \sin(\theta/2) & \cos(\theta/2)\end{pmatrix}, \tag{3.7.193}$$

$$v(\theta, \boldsymbol{e}_3) = \exp(\theta K^3) = \exp[(-i/2)\theta\sigma^3] = \begin{pmatrix} \exp(-i\theta/2) & 0 \\ 0 & \exp(i\theta/2) \end{pmatrix}. \tag{3.7.194}$$

The Euler-angle parameterization of $SU(2)$ is defined by the rule

$$v(\phi, \theta, \psi) = \exp(\phi K^3)\exp(\theta K^2)\exp(\psi K^3). \tag{3.7.195}$$

Note that K^1 does not appear in the formula. Verify that every element in $SU(2)$ can be written in Euler form. Verify that $SU(2)$ is covered once, and only once, if the Euler angles lie in the ranges $\phi \in [0, 2\pi]$, $\theta \in [0, \pi]$, $\psi \in [0, 4\pi]$. By carrying out the matrix multiplications implied by (7.195), verify that $v(\phi, \theta, \psi)$ has the explicit form

$$v(\phi, \theta, \psi) = \begin{pmatrix} \cos(\theta/2)\exp[-(i/2)(\phi + \psi)] & -\sin(\theta/2)\exp[(i/2)(-\phi + \psi)] \\ \sin(\theta/2)\exp[-(i/2)(-\phi + \psi)] & \cos(\theta/2)\exp[(i/2)(\phi + \psi)] \end{pmatrix}. \tag{3.7.196}$$

The relation (7.189) is a formula for computing the 2×2 $SU(2)$ matrix v given θ and $\boldsymbol{n}$. Suppose, instead, that one is given $v \in SU(2)$ and wants to know θ and $\boldsymbol{n}$. Show that there are the formulas

$$2\cos(\theta/2) = \text{tr}(v) \tag{3.7.197}$$

and

$$4(\boldsymbol{n} \cdot \boldsymbol{K})\sin(\theta/2) = v - v^\dagger, \tag{3.7.198}$$

from which it follows that

$$n_\alpha \sin(\theta/2) = (i/4)\text{tr}[\sigma^\alpha(v - v^\dagger)]. \tag{3.7.199}$$

Consider next the case of $SO(3, \mathbb{R})$. Define $SO(3, \mathbb{R})$ matrices $R(\theta, \boldsymbol{n})$ by the rule

$$R(\theta, \boldsymbol{n}) = \exp(\theta \boldsymbol{n} \cdot \boldsymbol{L}). \tag{3.7.200}$$

Show, in accord with Section 3.8.1, that any $SO(3, \mathbb{R})$ matrix can be written in this form. For any to vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, verify the relation

$$(\boldsymbol{a} \cdot \boldsymbol{L})\boldsymbol{b} = (\boldsymbol{a} \times \boldsymbol{b}). \tag{3.7.201}$$

Use this result to show that $R(\theta, \boldsymbol{n})$ produces a rotation by angle θ about the axis $\boldsymbol{n}$.

Verify the result

$$(\boldsymbol{n} \cdot \boldsymbol{L})^3 = -\boldsymbol{n} \cdot \boldsymbol{L}. \tag{3.7.202}$$

Use this relation to sum the series implied by (7.200) to find the explicit results

$$R(\theta, \boldsymbol{n}) = I + (\boldsymbol{n} \cdot \boldsymbol{L})\sin\theta + (\boldsymbol{n} \cdot \boldsymbol{L})^2(1 - \cos\theta), \tag{3.7.203}$$

$$R(2\pi, \boldsymbol{n}) = I. \tag{3.7.204}$$

Verify that $SO(3, \mathbb{R})$ is covered once and only once when $\theta \in [0, \pi]$ and $\boldsymbol{n}$ is allowed to be any unit vector.

As special cases of (7.203), verify the relations

$$R(\theta, \boldsymbol{e}_1) = \exp(\theta L^1) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}, \tag{3.7.205}$$

$$R(\theta, \boldsymbol{e}_2) = \exp(\theta L^2) = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix}, \tag{3.7.206}$$

$$R(\theta, \boldsymbol{e}_3) = \exp(\theta L^3) = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.7.207}$$

The Euler-angle parameterization of $SO(3,\mathbb{R}$ is defined by the rule

$$R(\phi, \theta, \psi) = \exp(\phi L^3)\exp(\theta L^2)\exp(\psi L^3). \tag{3.7.208}$$

Note that L^1 does not appear in the formula. Verify that every element in $SO(3,\mathbb{R})$ can be written in Euler form. Verify that $SO(3,\mathbb{R})$ is covered once, and only once, if the Euler angles lie in the ranges $\phi \in [0, 2\pi]$, $\theta \in [0, \pi]$, $\psi \in [0, 2\pi]$. By carrying out the matrix multiplications implied by (7.208), verify that $R(\phi, \theta, \psi)$ has the explicit form

$$R(\phi, \theta, \psi) =$$
$$\begin{pmatrix} \cos\phi\cos\theta\cos\psi - \sin\phi\sin\psi & -\cos\phi\cos\theta\sin\psi - \sin\phi\cos\psi & \cos\phi\sin\theta \\ \sin\phi\cos\theta\cos\psi + \cos\phi\sin\psi & -\sin\phi\cos\theta\sin\psi + \cos\phi\cos\psi & \sin\phi\sin\theta \\ -\sin\theta\cos\psi & \sin\theta\sin\psi & \cos\theta \end{pmatrix}. \tag{3.7.209}$$

The relation (7.203) is a formula, sometimes called *Rodrigues' rotation formula*, for computing the 3×3 rotation matrix R given the rotation angle θ and the axis of rotation $\boldsymbol{n}$. Suppose, instead, that one is given R and wants to know the rotation angle θ and the axis $\boldsymbol{n}$. First verify the relation

$$\mathrm{tr}[(\boldsymbol{a} \cdot \boldsymbol{L})^2] = -2\boldsymbol{a} \cdot \boldsymbol{a}. \tag{3.7.210}$$

for any vector $\boldsymbol{a}$. Now show that there are the formulas

$$1 + 2\cos\theta = \mathrm{tr}(R), \tag{3.7.211}$$

$$2(\boldsymbol{n} \cdot \boldsymbol{L})\sin\theta = R - R^T. \tag{3.7.212}$$

Let R be any element of $SO(3,\mathbb{R})$. Show that R must have $+1$ as an eigenvalue and that $\boldsymbol{n}$ is the associated eigenvector.[37] Show that the other two eigenvalues of R are $\exp(\pm i\theta)$.

Since the Lie algebras $su(2)$ and $so(3,\mathbb{R})$ are the same, we may expect a close relation between the groups $SU(2)$ and $SO(3,\mathbb{R})$. Are they perhaps the same? The answer is *no* as can be seen by comparing (7.190) and (7.204). Although the groups have the same Lie algebra, they are not the same globally. Exercises 8.2.10 and 8.2.11 show that there is a two-to-one homomorphism between $SU(2)$ and $SO(3,\mathbb{R})$. In fact these exercises show that, given $v \in SU(2)$, there is an $R(v) \in SO(3,\mathbb{R})$ specified by the two-to-one homomorphic map

$$R_{\alpha\beta}(v) = (1/2)\mathrm{tr}(v^\dagger \sigma^\alpha v \sigma^\beta). \tag{3.7.213}$$

There is another way to highlight the distinction between $su(2)$ and $so(3,\mathbb{R})$. For a Lie algebra realized in terms of matrices, it is useful, when possible, to form the matrix for the

[37] This result is sometimes called Euler's theorem for rigid body motion.

second-order Casimir operator. The second-order Casimir matrix is defined in terms of the structure constants and the basis matrices for the Lie algebra. See Section 27.11 for details. In the case of the K^α verify that there is the matrix relation

$$(K^1)^2 + (K^2)^2 + (K^3)^2 = -(3/4)I. \tag{3.7.214}$$

In the case of the L^α verify that there is the matrix relation

$$(L^1)^2 + (L^2)^2 + (L^3)^2 = -(2)I. \tag{3.7.215}$$

In our case the structure constants are given by (7.174), and it can be shown that the quantities on the left sides of (7.214) and (7.215) are the Casimir matrices formed from the K^α and the L^α, respectively. The coefficients in parentheses on the right sides of (7.214) and (7.215) are of the form $j(j+1)$ with $j = 1/2$ in the case of the K^α and $j = 1$ in the case of the L^α. Analogous relations may be familiar to the reader from the quantum theory of spin/angular momentum. We see that $su(2)$ corresponds to the case $j = 1/2$ and $so(3, \mathbb{R})$ corresponds to the case $j = 1$.

3.7.32. Suppose that $\mathcal{R}$ is a map that sends three-dimensional Euclidean space into itself and has a cyclic action on the points $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, $\boldsymbol{e}_3$:

$$\mathcal{R}\boldsymbol{e}_1 = \boldsymbol{e}_2,\ \mathcal{R}\boldsymbol{e}_2 = \boldsymbol{e}_3,\ \mathcal{R}\boldsymbol{e}_3 = \boldsymbol{e}_1. \tag{3.7.216}$$

Extend $\mathcal{R}$ to all of three-dimensional Euclidean space by linearity so that its action can be represented by an associated matrix R. Show that $R \in SO(3, \mathbb{R})$. Use the results of Exercise 7.31 to find the axis $\boldsymbol{n}$ and angle θ for R.

3.7.33. Show that the groups $Sp(2, \mathbb{R})$ and $SL(2, \mathbb{R})$ are the same, and therefore their Lie algebras are the same: $Sp(2, \mathbb{R}) = SL(2, \mathbb{R})$ and $sp(2, \mathbb{R}) = s\ell(2, \mathbb{R})$. Show that the groups $Sp(2, \mathbb{C})$ and $SL(2, \mathbb{C})$ are the same, and therefore their Lie algebras are the same: $Sp(2, \mathbb{C}) = SL(2, \mathbb{C})$ and $sp(2, \mathbb{C}) = s\ell(2, \mathbb{C})$. See Exercises 1.2 and 1.3. [Subsequently it will be shown that the Lie algebra of the Lorentz group is the same as the Lie algebras $sp(2, \mathbb{C}) = s\ell(2, \mathbb{C})$. See Exercises 7.3.30 and 8.2.14.] Show that the Lie algebras $sp(2, \mathbb{R})$, $s\ell(2, \mathbb{R})$, $su(2)$, and $so(3, \mathbb{R})$ have the same dimension. Show that these Lie algebras are in fact the same (equivalent) over the complex field. Which of these Lie algebras are equivalent over the real field?

3.7.34. Show that the Lie algebras $sp(4, \mathbb{R})$ and $so(5, \mathbb{R})$ have the same dimension. See Exercise 27.5.4 for a demonstration that these Lie algebras are in fact the same (equivalent) over the complex field, but not the real field.

3.7.35. Show that the Lie algebras $su(4)$ and $so(6, \mathbb{R})$ have the same dimension. In fact, these Lie algebras are the same (equivalent) over the real field. Moreover, as shown in Exercise 8.2.12, there is a corresponding two-to-one homomorphism between the groups $SU(4)$ and $SO(6, \mathbb{R})$ just as there is a two-to-one homomorphism between the groups $SU(2)$ and $SO(3, \mathbb{R})$. See Exercises 7.29, 8.2.11, and 8.2.12.

3.7.36. Let $c^{\gamma}_{\alpha\beta}$ be a set of structure constants for some Lie algebra L as in (7.50). Let B_α be a set of $m \times m$ matrices that forms a basis for L thereby providing a *representation* of L. That is, the matrices satisfy the commutation rules

$$\{B_\alpha, B_\beta\} = \sum_\gamma c^{\gamma}_{\alpha\beta} B_\gamma. \tag{3.7.217}$$

Let E be any $m \times m$ invertible matrix, and define matrices B'_α by the rule (similarity transformation)

$$B'_\alpha = E B_\alpha E^{-1}. \tag{3.7.218}$$

View the B'_α as basis elements and show that the B'_α also form a representation of L. That is, the B'_α obey commutation rules identical to those of the B_α in (7.217) with the *same* structure constants. The representations provided by the B'_α and the B_α are called *equivalent*, and for many purposes may be viewed as being essentially the same.[38] Conversely, given two sets of $m \times m$ representation matrices B_α and B'_α that obey the same commutation rules, one can inquire whether there is an invertible matrix E such that (7.218) holds. If there is, the two representations are said to be equivalent.

Given the B_α, suppose we define *conjugate* matrices $\tilde{B}_\alpha$ by the "tilde" rule

$$\tilde{B}_\alpha = -B^T_\alpha. \tag{3.7.219}$$

Note that this tilde rule is an *involution*. That is, let $\tilde{\mathcal{C}}$ denote the tilde conjugacy operator defined by

$$\tilde{\mathcal{C}}(B_\alpha) = \tilde{B}_\alpha. \tag{3.7.220}$$

Verify that $\tilde{\mathcal{C}}^2$ has the property

$$\tilde{\mathcal{C}}^2(B_\alpha) = B_\alpha \tag{3.7.221}$$

so that $\tilde{\mathcal{C}}^2 = \mathcal{I}$ on every element on which it acts.

Show that the $\tilde{B}_\alpha$ also form a representation of L. This representation is called a conjugate representation. That is, the $\tilde{B}_\alpha$ obey commutation rules identical to those of the B_α in (7.217) with the *same* structure constants. Put another way, show that there is the result

$$\tilde{\mathcal{C}}(\{B_\alpha, B_\beta\}) = \{\tilde{\mathcal{C}}(B_\alpha), \tilde{\mathcal{C}}(B_\beta)\}, \tag{3.7.222}$$

which displays that $\tilde{\mathcal{C}}$ is a homomorphism for the Lie product provided by the commutator $\{*, *\}$.

Whether this conjugate representation is equivalent (in the sense defined three paragraphs above) to that provided by the B_α depends on the Lie algebra L and the representation provided by the B_α. If a representation and its conjugate are equivalent in the sense

[38] Observe that this definition of *equivalent* need not be the same as that given in Subsection 7.6. There *arbitrary* linear invertible transformations on the basis were considered with the aim of modifying the structure constants, and the structure constants were found to change according to the rules (7.56) and (7.57). But the new basis elements, being linear combinations of the B_α, are still elements of L. In the present case the structure constants are required to remain unchanged even though the basis is changed. And in this case the B_α have to be matrices or linear operators or some such things for (7.218) to make sense. Finally, unless the B_α form a basis for the set of $m \times m$ matrices, the B'_α need not be in L. For example, if we consider the fundamental representation of $sp(2n)$, the B_α are of the form JS. But the same need not be true of the B'_α.

(7.218), the representations are said to be *self conjugate*. (We remark that if the B_α are the matrices for the *adjoint* representation of L, then the matrices $\tilde{B}_\alpha$ are sometimes referred to as the *coadjoint* representation of L.)

There is a second "conjugacy" possibility. Suppose that for some choice of basis elements [see (7.56) and (7.57)] the structure constants $c^\gamma_{\alpha\beta}$ can all be made *real*. (This can be shown to be the case, for example, for all the classical and exceptional Lie algebras in Table 7.2.) Also allow the possibility that at least some of the B_α may have some complex entries. In this case, define matrices $\breve{B}_\alpha$, again called conjugate matrices, by the "accent breve" rule

$$\breve{B}_\alpha = \bar{B}_\alpha \tag{3.7.223}$$

where a bar indicates complex conjugation. Call the breve conjugacy operator $\breve{\mathcal{C}}$ so that we may write

$$\breve{\mathcal{C}}(B_\alpha) = \breve{B}_\alpha = \bar{B}_\alpha. \tag{3.7.224}$$

Verify that $\breve{\mathcal{C}}$, is also an involution. Show that the $\breve{B}_\alpha$ also form a representation of L or, equivalently, $\breve{\mathcal{C}}$ is also a commutator Lie product homomorphism. Whether this conjugate representation is equivalent to that provided by the B_α also depends on the Lie algebra L and the representation provided by the B_α. A representation involving complex matrices that is equivalent to itself under the breve operation (7.222) is called *pseudoreal*. In analogy with the tilde operation case, it may also be called self conjugate.

If the structure constants cannot be made real, show that the $\breve{B}_\alpha$ still form a Lie algebra. Whether or not this Lie algebra is different from the original one, or can be brought to the same form as the original Lie algebra by a suitable new choice of basis elements as in (7.56) and (7.57), then requires further investigation.

There is a third conjugacy operator possibility, which we will call the "accent grave" rule, that is sometimes of use. Denote it by $\grave{\mathcal{C}}$. Given basis matrices B_α for a Lie algebra L with real structure constants, define matrices $\grave{B}_\alpha$, again called conjugate matrices, by the rule

$$\grave{\mathcal{C}}(B_\alpha) = \grave{B}_\alpha = -B_\alpha^\dagger. \tag{3.7.225}$$

Verify that the grave rule is also an involution. Show that the $\grave{B}_\alpha$ also form a representation of L so that $\grave{\mathcal{C}}$ is also a commutator Lie product homomorphism. Show that $\grave{\mathcal{C}}$ consists of combining the operations of the first two conjugation rules by verifying that

$$\grave{\mathcal{C}} = \breve{\mathcal{C}}\tilde{\mathcal{C}} = \tilde{\mathcal{C}}\breve{\mathcal{C}}. \tag{3.7.226}$$

In summary, we have defined three possible conjugacy rules: tilde ˜, breve ˘, and grave ˋ. Note that, in all cases and for all three conjugacy rules, a representation and its conjugate have the same dimension.

Let us now explore conjugacy relations for some of the familiar Lie algebras: Consider the *fundamental/defining* representation of $sp(2,\mathbb{R})$ provided by the matrices (7.66) through (7.68). Find the associated tilde representation given by (7.219) above. Verify that this representation is equivalent to the fundamental representation using

$$E = J. \tag{3.7.227}$$

Consider the fundamental representation of $sp(2n,\mathbb{R})$ for *any* n. Using (7.34) show that

$$\tilde{B} = SJ. \tag{3.7.228}$$

Verify that

$$J\tilde{B}J^{-1} = JS = B, \tag{3.7.229}$$

and therefore the tilde representation is equivalent to the fundamental representation, again using (7.227), for all n. Put another way, under the tilde operation and for all n, the fundamental representation of $sp(2n,\mathbb{R})$ is *self conjugate*.

Suppose that either the breve or grave conjugacy operations are used instead. Show that, since the fundamental representation of $sp(2n,\mathbb{R})$ consists of real matrices, it is also self conjugate under both the breve and grave operation for all n.

Consider the adjoint representation of $sp(2,\mathbb{R})$. It is provided by the matrices (7.75) through (7.77). Find the associated tilde representation given by (7.219) above. Verify that this representation is equivalent to the adjoint representation using

$$E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}. \tag{3.7.230}$$

In fact, it can be shown that *all* representations of $sp(2n,\mathbb{R})$ are self conjugate for all n under all three conjugacy relations tilde, breve, and grave.

Consider the fundamental representation of $su(2)$ provided by the matrices (7.169) through (7.171). Observe that some of them are complex. Show that the associated breve representation given by (7.223) above yields the result

$$\breve{K}^1 = -K^1, \tag{3.7.231}$$

$$\breve{K}^2 = K^2, \tag{3.7.232}$$

$$\breve{K}^3 = -K^3. \tag{3.7.233}$$

Verify that this representation is equivalent to the original representation using

$$E = \exp(\pi K^2) = -J_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{3.7.234}$$

[Look ahead to see (8.2.338) or (8.2.374) through (8.2.376) if you need help.] Thus the representation of $su(2)$ provided by the matrices (7.169) through (7.171) is pseudoreal. Finally, it can be shown that *all* representations of $su(2)$ are self conjugate under the breve operation because any representation of $su(2)$ can be obtained by taking suitable real linear combinations of tensors formed from vectors (actually, *spinors*) that transform according to the fundamental representation.

Suppose that instead the tilde operation (7.219) is employed. Show that in this case there is the result

$$\tilde{K}^1 = -K^1, \tag{3.7.235}$$

$$\tilde{K}^2 = K^2, \tag{3.7.236}$$

$$\tilde{K}^3 = -K^3. \tag{3.7.237}$$

The tilde and breve operations (7.219) and (7.223), when acting on the fundamental representation of $su(2)$, give the same result. Therefore for $su(2)$ the tilde representation is also equivalent to the fundamental representation again using the E given by (7.234).

Show that the reason the tilde and breve operations give the same result in the case of $su(2)$ fundamental representation is that the K^α are anti-Hermitian, and that this "same result" conclusion will also hold for all cases in which the B_α are anti-Hermitian, including all $su(n)$ cases.

Suppose the grave operation acts on the fundamental representation of $su(n)$. Show that in this case there is the "no effect" result

$$\grave{B}_\alpha = B_\alpha \tag{3.7.238}$$

because the B_α are anti-Hermitian.

We remark that the case of $su(2)$ is special. Subsequently we will find that the fundamental representation of $su(n)$ for any $n > 2$ is not equivalent to its complex conjugate (breve) representation. Therefore these fundamental representations are not self-conjugate/pseudoreal.

Consider the orthogonal group Lie algebras $so(n, \mathbb{R})$. In this case the Lie algebra for the fundamental representation consists of all real antisymmetric matrices A. Correspondingly, the breve operation has no effect on elements in the fundamental representation. Show that, because A is real and antisymmetric, the tilde operation (7.219) and the grave operation (7.225) also have no effect for the fundamental representation. We conclude that the fundamental representation of $so(n, \mathbb{R})$ is self conjugate for all three conjugacy definitions. Since all representations can be obtained from the fundamental representation by suitable linear combinations of tensor products, all three conjugacy operations will also have no effect on any representation, and they are therefore all self conjugate for all conjugacy definitions

In summary, the representations of $sp(2n)$, $so(n)$, and $su(2)$ are all self conjugate for all three conjugacy operations. And for $su(n)$ the tilde and breve conjugacy operations yield the same result. Finally, for the case of $su(n)$ with $n > 2$, it can be shown that there are representations that are not self conjugate. See, for example, Exercise 5.8.29. In particular, the representations 3 and $\bar{3}$ for $su(3)$ are not equivalent.[39]

Subsequently we will also be interested in the Lorentz group Lie algebra and in $s\ell(2, \mathbb{C})$, which will be found to be the Lie algebra for the covering group of the Lorentz group. For these Lie algebras the three conjugacy operations will again be useful. See Exercises 7.3.27, 7.3.29 through 7.3.31, and 8.2.14.

3.7.37. Review Exercise 7.36 which described tilde, breve, and grave rules for defining conjugate matrices in Lie algebras. The purpose of this exercise is to extend these rules to the associated matrix groups. Suppose the Lie algebra L has dimension n and basis elements B_α. Let $b = (b_1, b_2, \cdots b_n)$ be a collection of n real parameters, and consider the group element

$$G(b) = \exp(\sum_{\alpha=1}^{n} b_\alpha B_\alpha). \tag{3.7.239}$$

[39] Here the "bar" notation, which is customary, is used to refer to the use of what we have called the breve conjugacy operator $\breve{\mathcal{C}}$ because this operator involves complex conjugation. Recall (7.223).

In the case of the tilde rule, define the group element $\tilde{G}(b)$ by the rule

$$\tilde{G}(b) = \exp(\sum_{\alpha=1}^{n} b_\alpha \tilde{B}_\alpha) = \exp(-\sum_{\alpha=1}^{n} b_\alpha B_\alpha^T). \tag{3.7.240}$$

Show that

$$\tilde{G}(b) = [G^T(b)]^{-1}. \tag{3.7.241}$$

[See (7.266).] Show that the tilde operation when acting as defined on group elements is again an involution. That is, for group elements define a tilde operator, which we will call $\tilde{\mathcal{D}}$, by the rule

$$\tilde{\mathcal{D}}G(b) = \tilde{G}(b) = [G^T(b)]^{-1}, \tag{3.7.242}$$

and show that

$$\tilde{\mathcal{D}}^2 G(b) = G(b). \tag{3.7.243}$$

Moreover, verify that it follows from (7.242) and (7.243) that

$$\tilde{\mathcal{D}}\tilde{G}(b) = G(b) = [\tilde{G}^T(b)]^{-1}, \tag{3.7.244}$$

which is the counterpart to (7.242). Show that $\tilde{\mathcal{D}}$ is also a group homomorphism. That is, if G and G' are any two group elements, then

$$\tilde{\mathcal{D}}(GG') = \tilde{\mathcal{D}}(G)\tilde{\mathcal{D}}(G'). \tag{3.7.245}$$

Comparison of (7.222) and (7.245) shows that we have converted a Lie product (commutator) homomorphism into a Lie group homomorphism.

In the cases of the breve and grave rules make definitions of their actions on group elements analogous that for the tilde operation. For these rules show that

$$\breve{G}(b) = \bar{G}(b) \tag{3.7.246}$$

and

$$\grave{G}(b) = [G^\dagger(b)]^{-1}. \tag{3.7.247}$$

[See (7.267).] Let $\breve{\mathcal{D}}$ and $\grave{\mathcal{D}}$ be the breve and grave operators which act on group elements so that we may write

$$\breve{\mathcal{D}}G(b) = \breve{G}(b) = \bar{G}(b) \tag{3.7.248}$$

and

$$\grave{\mathcal{D}}G(b) = \grave{G}(b) = [G^\dagger(b)]^{-1}. \tag{3.7.249}$$

Show that $\breve{\mathcal{D}}$ and $\grave{\mathcal{D}}$ are also involutions and group homomorphisms. Observe that if G is a unitary matrix ($G^\dagger = G^{-1}$) it follows from (7.249) that

$$\grave{G}(b) = [G^\dagger(b)]^{-1} = G(b). \tag{3.7.250}$$

Finally, review Exercise 1.6.18. Verify that the relation between K and Λ given by (1.6.287) is the same as the relation between $\tilde{G}(b)$ and $G(b)$ given by (7.241). According to Exercise 7.36 the B_α and the $\tilde{B}_\alpha$ obey the same Lie algebra. This fact is consistent with the result that K must be a Lorentz transformation matrix if Λ is a Lorentz transformation matrix, and vice versa.

3.7.38. Suppose g is a $(m+n)\times(m+n)$ diagonal matrix with m diagonal entries having value $+1$ followed by n diagonal entries having value -1. Show that the set of all $(m+n)\times(m+n)$ matrices O that satisfy the relation

$$O^T gO = g \tag{3.7.251}$$

forms a group. This group is called the *indefinite* orthogonal group, and is sometimes denoted by the symbol $O(m,n)$, or by the symbols $O(m,n,\mathbb{R})$ and $O(m,n,\mathbb{C})$ if the choice of field needs to be explicit. Note that (7.251) continues to hold if g is replaced by $-g$, and therefore there is no distinction between $O(m,n)$ and $O(n,m)$. Show from (7.251) that there is the result

$$\det(O) = \pm 1. \tag{3.7.252}$$

Show that indefinite orthogonal matrices with determinant $+1$ form a subgroup, called $SO(m,n)$. Find the Lie algebra $so(m,n,\mathbb{R})$, and show that it is equivalent to the Lie algebra $so(m+n,\mathbb{R})$ when working over the complex field.

3.7.39. Review Exercise 7.38 above. Let g be the matrix defined there. Show that the set of all complex $(m+n)\times(m+n)$ matrices U that satisfy the relation

$$U^\dagger gU = g \tag{3.7.253}$$

forms a group. This group is called the *indefinite* unitary group, and is denoted by the symbol $U(m,n)$. Here the field is naturally $\mathbb{C}$. Note that (7.253) continues to hold if g is replaced by $-g$, and therefore there is no distinction between $U(m,n)$ and $U(n,m)$. Show from (7.253) that there is the result

$$|\det(U)| = 1. \tag{3.7.254}$$

Verify that indefinite unitary matrices with determinant $+1$ form a subgroup. It is called $SU(m,n)$. Find the Lie algebra $su(m,n)$, and show that it is equivalent to the Lie algebra $su(m+n)$ when working over the complex field.

3.7.40. Review Exercise 7.38 above. Also look ahead and review Exercise 6.2.6. The 4×4 real matrices Λ defined there form a group, generally called the Lorentz group. From (6.2.20) we see that the Lorentz group is analogous to the rotation group $SO(4,\mathbb{R})$ except the 4×4 identity matrix I has been replaced by the 4×4 diagonal matrix g. Note that g has three equal diagonal entries with the same sign, and one with the opposite sign. For this reason, the Lorentz group is also referred to as $SO(3,1,\mathbb{R})$. Find the Lie algebra $so(3,1,\mathbb{R})$ and show that it is equivalent to the Lie algebra $so(4,\mathbb{R})$ when working over the complex field, and hence also equivalent to $su(2)\oplus su(2)$. Since the representations of $su(2)$ are well known, the finite-dimensional (and nonunitary) representations of the Lorentz Lie algebra and Lie group are also well understood. See Exercises 4.3.19, 4.3.20, and 7.3.28.

3.7.41. Suppose f and g are two elements of some group G. Using group multiplication, form the group element h defined by the rule

$$h = (gf)^{-1}fg = f^{-1}g^{-1}fg. \tag{3.7.255}$$

This element is called the *group commutator* of f and g. Note that if f and g commute, $fg = gf$, then $h = (gf)^{-1}fg = (fg)^{-1}fg = I$.

Suppose that G is a matrix Lie group and consider elements $f(s)$ and $g(s)$ of the form

$$f(s) = \exp(sa), \tag{3.7.256}$$

$$g(s) = \exp(sb), \tag{3.7.257}$$

where a and b are in the Lie algebra of G. Let $h(s)$ be the group commutator of $f(s)$ and $g(s)$,

$$h(s) = f^{-1}(s)g^{-1}(s)f(s)g(s) = \exp(-sa)\exp(-sb)\exp(sa)\exp(sb). \tag{3.7.258}$$

Show, using the BCH formula (7.41), that there is the relation

$$h(s) = \exp\left(s^2\{a,b\} + O(s^3)\right). \tag{3.7.259}$$

Thus, for a matrix Lie group, there is a relation between the group commutator and the Lie algebra commutator. It can be shown that an analogous relation holds for abstract Lie groups: There is a relation between the group commutator and the Lie product of associated elements in the Lie algebra.

For extra credit, use through third order the BCH formula (7.41) to show that

$$h(s) = \exp\left(s^2\{a,b\} - (s^3/2)\{(a+b), \{a,b\}\} + O(s^4)\right). \tag{3.7.260}$$

3.7.42. Suppose A and B are two $n \times n$ matrices that satisfy the relation

$$AB = I.$$

Show it follows that

$$BA = I,$$

and therefore A and B commute.

Suppose C and D are two commuting $n \times n$ matrices and that C is invertible. Show that then C^{-1} and D also commute. Show that C^m and D also commute for all integer values (positive, zero, and negative) of m.

Suppose that a matrix E is a function of C, and specifically is defined in terms of C as some convergent power series in C (and possibly also powers of C^{-1} if C is invertible). Show that then E and C also commute.

3.7.43. Review Exercise 7.31. There it is shown that the generators K^α for $su(2)$ and the generators L^α for $so(3,\mathbb{R})$ can be selected to be *anti-Hermitian* and, writing the generators generically as J^α, satisfy the same commutation rules

$$\{J^\alpha, J^\beta\} = \sum_\gamma \epsilon_{\alpha\beta\gamma} J^\gamma \tag{3.7.261}$$

with *real* structure constants $\epsilon_{\alpha\beta\gamma}$. Verify that the commutation rules (7.248) are consistent with the J^α being anti-Hermitian. That is, verify that the commutator of two anti-Hermitian generators is again anti-Hermitian.

Correspondingly, the associated group elements U are of the form

$$U = \exp(\sum_\gamma \lambda_\gamma J^\gamma) \tag{3.7.262}$$

with *real* parameters λ_γ. Verify that U as given by (7.262) is unitary when the λ_γ are real.

We might say this treatment of $su(2)$ and $so(3,\mathbb{R})$ is the mathematicians' approach. By contrast, in the quantum treatment of angular momentum, physicists work with *Hermitian* generators, call them $\tilde{J}^\alpha$, that are required to satisfy the commutation rules

$$\{\tilde{J}^\alpha, \tilde{J}^\beta\} = \sum_\gamma i\epsilon_{\alpha\beta\gamma}\tilde{J}^\gamma \tag{3.7.263}$$

with *purely imaginary* structure constants $i\epsilon_{\alpha\beta\gamma}$. Verify that the commutation rules (7.263) are consistent with the $\tilde{J}^\alpha$ being Hermitian. That is, verify that the commutator of two Hermitian generators is anti-Hermitian.

Correspondingly, the associated group elements U are of the form

$$U = \exp(\sum_\gamma -i\lambda_\gamma \tilde{J}^\gamma) \tag{3.7.264}$$

with *real* parameters λ_γ. Verify that U as given by (7.264) is unitary when the λ_γ are real.

Verify that the mathematicians' and quantum physicists' approaches are connected by the (*complex*) change of basis

$$J^\alpha = -i\tilde{J}^\alpha \Leftrightarrow \tilde{J}^\alpha = iJ^\alpha. \tag{3.7.265}$$

That is, the Lie algebras defined by (2.61) and (2.63) are equivalent over the complex field. Verify also that U as given by (2.62) and (2.64) are unaffected by this change of basis.

Why do quantum physicists insert what would appear to mathematicians to be superfluous factors of i? They do so because they wish to associate physical observables with Hermitian operators in order to ensure that the expectation values and eigenvalues of physical observables are *real* numbers.

3.7.44. Suppose A is *any* $n \times n$ matrix. Verify the relations

$$[\exp(A)]^T = \exp(A^T), \tag{3.7.266}$$

$$[\exp(A)]^\dagger = \exp(A^\dagger). \tag{3.7.267}$$

Suppose H is a *Hermitian* $n \times n$ matrix. Verify that then $\exp(H)$ is also Hermitian,

$$[\exp(H)]^\dagger = \exp(H). \tag{3.7.268}$$

Verify the line of reasoning below to show that $\exp(H)$ is positive definite: Let v be any nonzero n-component vector. Then there is the result

$$\begin{aligned}(v, \exp(H)v) &= (v, \exp(H/2)\exp(H/2)v) = (\exp(H/2)v, \exp(H/2)v) \\ &= (w, w) > 0 \end{aligned}\tag{3.7.269}$$

where we have employed the usual complex scalar product and

$$w = \exp(H/2)v. \tag{3.7.270}$$

3.7.45. Review Table 7.2 that provides the dimensions of the Classical Lie Algebras for each integer value of ℓ. Verify that, consistent with the Table, we may define functions $\mathcal{A}$ through $\mathcal{D}$ by the rules

$$\mathcal{A}(n) = \dim[su(n)] = n^2 - 1, \tag{3.7.271}$$

$$\mathcal{B}(n) = \dim[so(n)] = (1/2)n(n-1),\ n \text{ odd}, \tag{3.7.272}$$

$$\mathcal{C}(n) = \dim[sp(n)] = (1/2)n(n+1),\ n \text{ even}, \tag{3.7.273}$$

$$\mathcal{D}(n) = \dim[so(n)] = (1/2)n(n-1),\ n \text{ even}, \tag{3.7.274}$$

where dim stands for *dimension*. Note that $\mathcal{B}(n)$ and $\mathcal{D}(n)$ may be regarded as odd n and even n evaluations of a common formula. Recall Exercises 7.27 and 7.28. Suppose this common formula is evaluated for *negative* even values of n. Show that

$$\mathcal{D}(-n) = (1/2)(-n)(-n-1) = (1/2)n(n+1) = \mathcal{C}(n),\ n \text{ even}. \tag{3.7.275}$$

We also observe that if the right side of (7.271) is taken to define $\mathcal{A}(n)$ for all values of n, then there is the relation

$$\mathcal{A}(-n) = \mathcal{A}(n). \tag{3.7.276}$$

For a discussion of what to make of these results, see the book of P. Cvitanović cited at the end of this chapter.

3.8 Exponential Representations of Group Elements

Lie group elements that are sufficiently near the identity can be written as exponentials of elements in the corresponding Lie algebra. This rule which sends a Lie algebra element into a group element is called the *exponential map*. Can this be done globally? That is, can *every* Lie group element be written as the exponential of some element in the associated Lie algebra? In this section we will answer this question for the Lie groups $SO(n,\mathbb{R})$, $SO(n,\mathbb{C})$, $U(n)$, $SU(n)$, and $Sp(2n,\mathbb{R})$.[40] The answer for all these groups is *yes* save for $Sp(2n,\mathbb{R})$ where the matter is more complicated. Finally, we note that being global is not the same as being a bijection. As is evident from (7.187) for $SU(2)$, for example, $v(\theta,\boldsymbol{n}) = v(\theta',\boldsymbol{n})$ whenever θ and θ' differ by a multiple of 4π. In general, exponentials of various elements in the Lie algebra may result in the same group element.

[40] Exercise 8.2.16 shows that for the case of $SL(2,\mathbb{C})$, the covering group of the Lorentz group, not every element can be written in single exponential form. Nevertheless, every element of the Lorentz group can be written in single exponential form.

3.8.1 Exponential Representation of Orthogonal and Unitary Matrices

Suppose O is an orthogonal matrix with unit determinant. Then it can be shown that there is an antisymmetric matrix A such that

$$O = \exp(A). \tag{3.8.1}$$

Conversely, if A is an antisymmetric matrix, then the O given by (8.1) will be orthogonal and have unit determinant. We conclude that every element $O \in SO(n)$ can be written as the exponent of an element $A \in so(n)$. This is true when working over either the real or the complex field. In particular, if O is real, then there is a real antisymmetric A satisfying (8.1), and A will be complex if O is complex. Finally, consider all elements of the form

$$O(s) = \exp(sA).$$

We see that all elements of $SO(n)$ lie on some one-parameter subgroup of $SO(n)$.

Similarly, suppose U is a unitary matrix. Then it can be shown that there is an anti-Hermitian matrix A such that

$$U = \exp(A). \tag{3.8.2}$$

And, if U has unit determinant, then there is a traceless anti-Hermitian matrix A such that (8.2) holds. Conversely, if A is anti-Hermitian, then the U given by (8.2) will be unitary; and if A is also traceless, then U will have unit determinant as well. We conclude that every element $U \in U(n)$ can be written as the exponent of an element $A \in u(n)$, and every element $U \in SU(n)$ can be written as the exponent of an element $A \in su(n)$. Finally, consider all elements of the form

$$U(s) = \exp(sA).$$

We see that all elements of $SU(n)$ lie on some one-parameter subgroup of $SU(n)$.

In summary, the exponential maps (8.1) and (8.2) for the orthogonal and unitary groups are *global*. That is, every orthogonal and every unitary matrix can be written as the exponential of some element in the associated Lie algebra.

3.8.2 Exponential Representation of Symplectic Matrices

The case of $Sp(2n, \mathbb{R})$ is more complicated. Again the discussion so far has shown that symplectic matrices sufficiently near the identity element can be written as exponentials of elements in the symplectic group Lie algebra. But what can be said about representing symplectic matrices in general? Thanks to the work of Exercise 7.12 we know that not every symplectic matrix can be written in single exponential form. The purpose of this subsection is to study what can be accomplished.

To proceed, it is useful to employ polar decomposition. See Subsection 6.4 and Section 4.2. Any real nonsingular matrix M can be written uniquely in the form

$$M = PO, \tag{3.8.3}$$

where P is a real positive-definite symmetric matrix and O is a real orthogonal matrix. Now suppose that M is symplectic. Using (1.9), the symplectic condition can be written in the form

$$M = J^{-1}(M^T)^{-1}J. \tag{3.8.4}$$

Then, upon inserting the polar decomposition (8.3) into (8.4), one finds the relation

$$PO = J^{-1}P^{-1}JJ^{-1}OJ. \tag{3.8.5}$$

Next, observe that the matrix $J^{-1}P^{-1}J$ is real, symmetric, and positive definite; and observe that the matrix $J^{-1}OJ$ is real and orthogonal. Consequently, because polar decomposition is unique, (8.5) implies the relations

$$P = J^{-1}P^{-1}J, \tag{3.8.6}$$

$$O = J^{-1}OJ. \tag{3.8.7}$$

Using the fact that P is symmetric and O is orthogonal, (8.6) and (8.7) can also be written in form

$$P = J^{-1}(P^T)^{-1}J, \tag{3.8.8}$$

$$O = J^{-1}(O^T)^{-1}J. \tag{3.8.9}$$

It follows that each of the matrices P and O are themselves symplectic.

The next thing to do is to work with the matrices O and P. Consider first the matrix O. Since O is real orthogonal and has determinant $+1$ (O is symplectic), it can be written in the form

$$O = \exp(F), \tag{3.8.10}$$

where F is a real antisymmetric matrix,

$$F^T = -F. \tag{3.8.11}$$

Upon inserting the representation (8.10) into the condition (8.7), we find the condition

$$O = \exp(F) = \exp(J^{-1}FJ). \tag{3.8.12}$$

Note that the matrix $(J^{-1}FJ)$ is real antisymmetric if the matrix F is. Therefore, in view of (8.12), it is tempting to assume that F has the property

$$F = J^{-1}FJ \text{ or } JF = FJ. \tag{3.8.13}$$

In general this assumption need not be correct because the logarithm of an orthogonal matrix is not unique. However, it will be shown in the next section that we may indeed require (8.13) for the present problem. Using (8.11), the condition (8.13) can also be written in the form

$$F^TJ + JF = 0. \tag{3.8.14}$$

Now compare (8.14) with (7.29) or (7.33). According to the argument employed earlier, the matrix F can be written in the form

$$F = JS^c, \tag{3.8.15}$$

where S^c is a real symmetric matrix. Furthermore, since F commutes with J, see (8.13), it follows that S^c *commutes* with J,

$$S^c J = J S^c. \tag{3.8.16}$$

In summary, it has been shown that O can be written in the form

$$O = \exp(JS^c), \tag{3.8.17}$$

where S^c is a real symmetric matrix that commutes with J.

It remains to see what can be said about the matrix P. Since P is real, symmetric, and positive definite, it can be written in the form

$$P = \exp(G), \tag{3.8.18}$$

where G is real and symmetric,

$$G^T = G. \tag{3.8.19}$$

Moreover, it can be shown that the real and symmetric logarithm of a real symmetric positive definite matrix is unique. Now insert the representation (8.18) into the condition (8.8) to obtain the result

$$P = \exp(G) = \exp(-J^{-1}GJ). \tag{3.8.20}$$

Since the matrix $(-J^{-1}GJ)$ is real symmetric if the matrix G is, and since G is unique, it follows from (8.20) that G has the property

$$(-J^{-1}GJ) = G \;\text{ or }\; GJ + JG = 0. \tag{3.8.21}$$

Using (8.19), the condition (8.21) can be re-expressed in the form

$$G^T J + JG = 0. \tag{3.8.22}$$

Consequently, G can also be written in the form

$$G = JS^a, \tag{3.8.23}$$

where S^a is a real symmetric matrix. However, in this case (8.19) implies the condition

$$JS^a + S^a J = 0. \tag{3.8.24}$$

That is, S^a *anticommutes* with J. In summary, it has been shown that P can be written in the form

$$P = \exp(JS^a), \tag{3.8.25}$$

where S^a is a real symmetric matrix that anticommutes with J.

Now combine (8.3), (8.17), and (8.25). The result is that any symplectic matrix can be written in the form

$$M = \exp(JS^a)\exp(JS^c). \tag{3.8.26}$$

It has been shown that the most general symplectic matrix can be written as the product of two exponentials of elements in the symplectic group Lie algebra, and each of the elements is of a special type.

It is interesting to examine the properties of commuting and anticommuting with J in a bit more detail. Let S be any symmetric matrix. Form the matrices S^a and S^c by the rules

$$S^a = (S - J^{-1}SJ)/2,$$
$$S^c = (S + J^{-1}SJ)/2. \tag{3.8.27}$$

It is easily verified that S^a and S^c are symmetric and anticommute and commute respectively with J as the notation suggests. And, if we wish, we may express the properties of anticommuting and commuting by the relations

$$JS^aJ^{-1} = -S^a,$$
$$JS^cJ^{-1} = S^c. \tag{3.8.28}$$

Also, it is obvous by construction that

$$S = S^a + S^c. \tag{3.8.29}$$

That is, any symmetric matrix can be uniquely decomposed into a sum of two symmetric matrices that anticommute and commute with J respectively.

We have seen, according to (8.26), that any real symplectic matrix can be written as the product of two symplectic matrices, each itself written in exponential form with the exponent being a real Hamiltonian matrix. We also know, according to (7.36), that any symplectic matrix sufficiently near the identity can be written in single exponential form. Moreover, the two exponentials appearing in (8.26) can, in principle, be combined into a single exponential using the Baker-Campbell-Hausdorff formula (7.41) providing the series converges. Finally, we know from Exercise 7.12 that not every symplectic matrix can be written as the exponential of a Hamiltonian matrix. Consequently, for $Sp(2n, \mathbb{R})$, there must be cases in which the Baker-Campbell-Hausdorff series diverges.

We have learned that in the case of $Sp(2n, \mathbb{R})$ the exponential map is *not* global. It follows, unlike the case for $SO(n)$ and $SU(n)$, that not every element of $Sp(2n, \mathbb{R})$ lies on a one-parameter subgroup of $Sp(2n, \mathbb{R})$. Instead, as (8.26) shows, to reach some elements in $Sp(2n, \mathbb{R})$ from the identity requires taking a *dogleg* path.

Since the exponential map is *not* global in the case of $Sp(2n, \mathbb{R})$, it is natural to ask under what conditions a symplectic matrix can be written in single exponential form. It is known that any matrix that is invertible (has nonzero determinant, or, equivalently, all its eigenvalues are nonzero) has a logarithm. But, like the case of numbers, this logarithm may be complex even if the matrix is real. Since any symplectic matrix has determinant +1, it must have a logarithm. But this logarithm may be complex. It is known that a sufficient, but not necessary condition, for a real invertible matrix to have a real logarithm is that none of its eigenvalues be negative. It is also known that if a real invertible matrix has a real square root, then it has a real logarithm and vice versa. What we are interested in for real symplectic matrices is the possibility of the logarithm being real and Hamiltonian. The analysis required to answer this question is complicated, and beyond the scope of our present discussion. It is known that if a real symplectic matrix has a real logarithm, then this logarithm will be Hamiltonian. For an analysis of the 2×2 case, see Section 8.7.2. Basically, as one might guess from Exercise 7.12, problems can occur when -1 appears as a repeated eigenvalue in Jordan blocks. Further information may be found in the references listed at the end of this chapter.

Exercises

3.8.1. Prove the statements made in Section 3.8.1 about orthogonal matrices.

3.8.2. Prove the statements made in Section 3.8.1 about unitary matrices.

3.8.3. Show that

$$\exp(\theta J) = I\cos\theta + J\sin\theta. \tag{3.8.30}$$

3.8.4. Show that the matrices $J, -I, -J$ can be written in the form $\exp(JS^c)$. Find S^c in each case.

3.8.5. Verify (8.5).

3.8.6. Verify that $J^{-1}P^{-1}J$ is real, symmetric, and positive definite. Verify that $J^{-1}OJ$ is real and orthogonal.

3.8.7. Verify (8.8) and (8.9), and the claim that O and P are symplectic.

3.8.8. Verify (8.12) using (8.9) and the definition (7.1).

3.8.9. Verify that $(J^{-1}FJ)$ is real antisymmetric if F is.

3.8.10. Verify that $(-J^{-1}GJ)$ is real symmetric if G is.

3.8.11. Verify (8.27) through (8.29).

3.8.12. Show that every matrix of the form $M = \exp(JS^a)$ is symplectic and has all its eigenvalues on the positive real axis. Show that every matrix of the form $M = \exp(JS^c)$ is symplectic, diagonalizable, and has all its eigenvalues on the unit circle. To prove the "diagonalizable" claim you may have to read the next section, Section 9.

3.8.13. Let M and A be the symplectic matrices

$$M = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},\ A = \begin{pmatrix} \tau & 0 \\ 0 & 1/\tau \end{pmatrix}. \tag{3.8.31}$$

Show that there is the symplectic conjugacy relation

$$AMA^{-1} = \begin{pmatrix} 1 & \tau^2 \\ 0 & 1 \end{pmatrix}. \tag{3.8.32}$$

3.8.14. Show that

$$\exp(JS^c) = \cosh(JS^c) + \sinh(JS^c), \tag{3.8.33}$$

$$\exp(JS^a) = \cosh(JS^a) + \sinh(JS^a). \tag{3.8.34}$$

Using the property that J commutes with S^c, show that

$$\begin{aligned} \cosh(JS^c) &= I + (JS^c)^2/2! + (JS^c)^4/4! + \cdots \\ &= I + J^2(S^c)^2/2! + J^4(S^c)^4/4! + \cdots \\ &= I - (S^c)^2/2! + (S^c)^4/4! + \cdots = \cos(S^c), \end{aligned} \tag{3.8.35}$$

$$\begin{aligned}\sinh(JS^c) &= JS^c + (JS^c)^3/3! + \cdots \\ &= JS^c + J^3(S^c)^3/3! + \cdots \\ &= J[S^c - (S^c)^3/3! + \cdots] = J\sin(S^c).\end{aligned} \tag{3.8.36}$$

Thus, show that

$$\exp(JS^c) = \cos(S^c) + J\sin(S^c). \tag{3.8.37}$$

This relation may be viewed as a symplectic Euler formula in which J plays the role of i.

Using the property that J anticommutes with S^a, show that

$$\begin{aligned}\cosh(JS^a) &= I + (JS^a)^2/2! + (JS^a)^4/4! + \cdots \\ &= I - J^2(S^a)^2/2! + J^4(S^a)^4/4! + \cdots \\ &= I + (S^a)^2/2! + (S^a)^4/4! + \cdots = \cosh(S^a),\end{aligned} \tag{3.8.38}$$

$$\begin{aligned}\sinh(JS^a) &= JS^a + (JS^a)^3/3! + \cdots \\ &= JS^a - J^3(S^a)^3/3! + \cdots \\ &= J[S^a + (S^a)^3/3! + \cdots] = J\sinh(S^a).\end{aligned} \tag{3.8.39}$$

Thus, show that

$$\exp(JS^a) = \cosh(S^a) + J\sinh(S^a). \tag{3.8.40}$$

3.8.15. Show that N as given by (5.60) and (5.61) is symplectic. Let E^ℓ be the matrix defined by the relation

$$E^\ell = \begin{pmatrix} 0_1 & & & & & & \\ & 0_2 & & & & & \\ & & \ddots & & & & \\ & & & E^{[2]}_\ell & & & \\ & & & & \ddots & & \\ & & & & & 0_{n-1} & \\ & & & & & & 0_n \end{pmatrix}. \tag{3.8.41}$$

Here each 0_ℓ is a 2×2 null matrix, $E^{[2]}_\ell$ is the 2×2 identity matrix,

$$E^{[2]}_\ell = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{3.8.42}$$

and all other entries are zero so that the $2n \times 2n$ identity matrix has the decomposition

$$I = \sum_{\ell=1}^{n} E^\ell. \tag{3.8.43}$$

Show that N can be written in the exponential form

$$N = \exp(JS^c) \tag{3.8.44}$$

with J given by (2.10) and S^c given by

$$S^c = \sum_{\ell=1}^{n} \phi_\ell E^\ell. \tag{3.8.45}$$

Note, as the notation is meant to indicate, that S^c *commutes* with J. Suppose that (5.53) is solved for M to give the result

$$M = ANA^{-1}. \tag{3.8.46}$$

Use this result and (8.44) to show that

$$M = ANA^{-1} = A[\exp(JS^c)]A^{-1} = \exp(AJS^cA^{-1}). \tag{3.8.47}$$

Next show that

$$AJ = J(A^{-1})^T. \tag{3.8.48}$$

Finally, show that M can be written in the form $M = \exp(JS)$ with

$$S = (A^{-1})^T S^c A^{-1}. \tag{3.8.49}$$

3.8.16. This exercise presumes that you have read Exercise 8.15 above. Its purpose is to describe various invariant quadratic forms.

Suppose a real $2n \times 2n$ symplectic matrix M can be written in the exponential form (7.30) with S real and symmetric. Let $z = (z_1, z_2, \cdots z_{2n})$ be the vector formed from the $2n$ variables z_1 through z_{2n}. Define the quadratic form $Q(z)$ by the rule

$$Q(z) = (z, Sz) \tag{3.8.50}$$

where $(*, *)$ denotes the usual real vector inner product. Suppose that $\bar{z}$ is defined in terms of M and z by the rule

$$\bar{z} = Mz. \tag{3.8.51}$$

Show that Q is *invariant* under this action. That is, show that

$$Q(\bar{z}) = Q(Mz) = Q(z). \tag{3.8.52}$$

Begin by verifying that

$$Q(\bar{z}) = (\bar{z}, S\bar{z}) = (Mz, SMz) = (z, M^TSMz). \tag{3.8.53}$$

Next verify that

$$M^TSM = -M^TJJSM = -M^TJMJS = -JJS = S, \tag{3.8.54}$$

from which it follows that

$$Q(\bar{z}) = (\bar{z}, S\bar{z}) = (z, Sz) = Q(z). \tag{3.8.55}$$

Consider the quadratic forms

$$Q_k(z) = (z, J[JS]^k z) \text{ for } k = 1, 2, \cdots \tag{3.8.56}$$

and

$$Q'_k(z) = (z, [SJ]^k Jz) \text{ for } k = 1, 2, \cdots . \tag{3.8.57}$$

Let a and b be any two possibly non-commuting entities. For all integers $k > 0$ verify the identity

$$(ab)^k = a(ba)^{k-1}b. \tag{3.8.58}$$

Use this identity to verify that

$$Q_k(z) = Q'_k(z). \tag{3.8.59}$$

Show that the $Q_k(z)$ are invariant and vanish for *even* k.

Consider the quadratic forms

$$\tilde{Q}_k(z) = (z, JM^k z) \text{ for } k = 1, 2, \cdots \tag{3.8.60}$$

and

$$\tilde{Q}'_k(z) = (z, [M^T]^k Jz) \text{ for } k = 1, 2, \cdots . \tag{3.8.61}$$

Show from the symplectic condition that $\tilde{Q}'_k(z)$ can also be written in the form

$$\tilde{Q}'_k(z) = (z, JM^{-k} z) \text{ for } k = 1, 2, \cdots \tag{3.8.62}$$

so that $\tilde{Q}_k$ and $\tilde{Q}'_k$ are analogous. Show that $\tilde{Q}_k$ and $\tilde{Q}'_k$ are also invariant. Note that the construction of these invariants does not require that M can be written in exponential form.

Suppose f is some function that sends the $2n \times 2n$ matrix JS to some other $2n \times 2n$ matrix $f(JS)$, and suppose that JS and $f(JS)$ commute. Suppose also that (7.30) holds. Show that Q_f defined by

$$Q_f(z) = (z, Jf(JS)z) \tag{3.8.63}$$

is then an invariant function. Similarly, suppose g is some function that sends the $2n \times 2n$ matrix M to some other $2n \times 2n$ matrix $g(M)$, and suppose that M and $g(M)$ commute. Show that Q_g defined by

$$Q_g(z) = (z, Jg(M)z) \tag{3.8.64}$$

is then an invariant function. See Exercise 11.4 for an example of such an invariant.

3.8.17. Work Exercises 8.15 and 8.16 above if you have no already done so. One might wonder whether/how all the invariants found in Exercise 8.16 are related. With what we know so far, we can study this question for the case in which all the eigenvalues of M lie on the unit circle and are distinct. In that case we can use the normal-form results of Section 3.5.

Let us begin with the case of $Q(z)$. From (8.49) there is the result

$$Q(z) = (z, Sz) = (z, [A^{-1}]^T S^c A^{-1} z) = (A^{-1}z, S^c A^{-1} z). \tag{3.8.65}$$

Consider the *normalized* quadratic form $Q^{\text{norm}}(z)$ defined by writing

$$Q^{\text{norm}}(z) = (z, S^c z). \tag{3.8.66}$$

Show from (8.41) and (8.45) that there is the result

$$Q^{\rm norm}(z) = \sum_{\ell=1}^{n} \phi_\ell (p_\ell^2 + q_\ell^2). \tag{3.8.67}$$

Define transformed variables $\hat{z}$ by writing

$$\hat{z} = A^{-1} z. \tag{3.8.68}$$

With this definition, show that (8.65) can be rewritten in the form

$$Q(z) = (A^{-1}z, S^c A^{-1} z) = (\hat{z}, S^c \hat{z}) = Q^{\rm norm}(\hat{z}) = \sum_{\ell=1}^{n} \phi_\ell (\hat{p}_\ell^2 + \hat{q}_\ell^2). \tag{3.8.69}$$

Let us next consider the Q_k. Show from Exercise 8.15 that

$$JS = AJS^c A^{-1}, \tag{3.8.70}$$

from which it follows that

$$(JS)^k = A(JS^c)^k A^{-1} \tag{3.8.71}$$

and

$$J(JS)^k = JA(JS^c)^k A^{-1} = (A^{-1})^T J(JS^c)^k A^{-1}. \tag{3.8.72}$$

Consequently, show that

$$Q_k(z) = (z, J(JS)^k z) = (z, (A^{-1})^T J(JS^c)^k A^{-1} z) = (A^{-1}z, J(JS^c)^k A^{-1} z). \tag{3.8.73}$$

Define $Q_k^{\rm norm}(z)$ by writing

$$Q_k^{\rm norm}(z) = (z, J(JS^c)^k z). \tag{3.8.74}$$

Then, again using (8.68), show that

$$Q_k(z) = Q_k^{\rm norm}(\hat{z}). \tag{3.8.75}$$

We still have to evaluate $Q_k^{\rm norm}$. Since J and S^c commute, we may write

$$J(JS^c)^k = J^{k+1}(S^c)^k \tag{3.8.76}$$

and therefore

$$Q_k^{\rm norm}(z) = (z, J^{k+1}(S^c)^k z). \tag{3.8.77}$$

We already know that we only have to deal with the odd k case, in which case

$$J^{k+1} = (-1)^{(k+1)/2} I. \tag{3.8.78}$$

Also, show from (8.45) that

$$(S^c)^k = \sum_{\ell=1}^{n} (\phi_\ell)^k E^\ell. \tag{3.8.79}$$

Show, therefore, that for odd k there is the result

$$Q_k^{\rm norm}(z) = (-1)^{(k+1)/2} \sum_{\ell=1}^{n} (\phi_\ell)^k (p_\ell^2 + q_\ell^2). \tag{3.8.80}$$

It follows that

$$Q_k(z) = (-1)^{(k+1)/2} \sum_{\ell=1}^{n} (\phi_\ell)^k (\hat{p}_\ell^2 + \hat{q}_\ell^2). \tag{3.8.81}$$

The last task is to consider $\tilde{Q}_k$ and $\tilde{Q}'_k$. From the representation (8.46) show that

$$M^k = (ANA^{-1})^k = AN^kA^{-1} \tag{3.8.82}$$

and consequently

$$\begin{aligned} \tilde{Q}_k(z) &= (z, JM^kz) = (z, JAN^kA^{-1}z) = (z, [A^{-1}]^T JN^kA^{-1}z) \\ &= (A^{-1}z, JN^kA^{-1}z) = (\hat{z}, JN^k\hat{z}) = \tilde{Q}_k^{\rm norm}(\hat{z}) \end{aligned} \tag{3.8.83}$$

where

$$\tilde{Q}_k^{\rm norm}(z) = (z, JN^kz). \tag{3.8.84}$$

Let us write N as given by (5.60) in the more explicit form

$$N(\phi_1, \phi_2, \cdots \phi_n) = \begin{pmatrix} R_1(\phi_1) & & & \\ & R_2(\phi_2) & & \\ & & \ddots & \\ & & & R_n(\phi_n) \end{pmatrix} \tag{3.8.85}$$

to emphasize that it depends on n angles ϕ_1 through ϕ_n. Verify that

$$[N(\phi_1, \phi_2, \cdots \phi_n)]^k = N(k\phi_1, k\phi_2, \cdots k\phi_n) \tag{3.8.86}$$

so that

$$\tilde{Q}_k^{\rm norm}(z) = (z, JN^kz) = (z, JN(k\phi_1, k\phi_2, \cdots k\phi_n)z). \tag{3.8.87}$$

Show from (5.60) and (5.61) that there is the result

$$(z, JNz) = -\sum_{\ell=1}^{n} (\sin\phi_\ell)(p_\ell^2 + q_\ell^2), \tag{3.8.88}$$

and therefore

$$(z, JN(k\phi_1, k\phi_2, \cdots k\phi_n)z) = -\sum_{\ell=1}^{n} (\sin k\phi_\ell)(p_\ell^2 + q_\ell^2). \tag{3.8.89}$$

You have shown that

$$\tilde{Q}_k(z) = -\sum_{\ell=1}^{n} (\sin k\phi_\ell)(\hat{p}_\ell^2 + \hat{q}_\ell^2). \tag{3.8.90}$$

Verify also, in view of (8.62), that

$$\tilde{Q}'_k(z) = \sum_{\ell=1}^{n} (\sin k\phi_\ell)(\hat{p}_\ell^2 + \hat{q}_\ell^2). \tag{3.8.91}$$

At this point it is evident that all the invariant quadratic forms found above involve the n quantities $(\hat{p}_\ell^2 + \hat{q}_\ell^2)$. You are now to show that these n quantities themselves are also invariant under the action of M. In particular, define quadratic forms $I_\ell(z)$ by the rule

$$I_\ell(z) = (z, [A^{-1}]^T E_\ell A^{-1} z). \tag{3.8.92}$$

Your task is to show that these n quadratic forms are equal to the $(\hat{p}_\ell^2 + \hat{q}_\ell^2)$, and are also invariant under the action of M. In so doing, you will have shown that all the (infinite in number) invariant quadratic forms found above are functions of the n functionally independent invariants I_ℓ.

Begin by verifying that

$$I_\ell(z) = (z, [A^{-1}]^T E_\ell A^{-1} z) = (A^{-1}z, E_\ell A^{-1} z) = (\hat{z}, E_\ell \hat{z}) = \hat{p}_\ell^2 + \hat{q}_\ell^2. \tag{3.8.93}$$

Next, as preparatory steps, show that N commutes with each E_ℓ and that N is orthogonal. Finally, verify that the I_ℓ are invariant under the action of M by checking that

$$\begin{aligned} I_\ell(Mz) &= (Mz, [A^{-1}]^T E_\ell A^{-1} Mz) = (ANA^{-1}z, [A^{-1}]^T E_\ell A^{-1} ANA^{-1} z) \\ &= (A^{-1}z, N^T A^T [A^{-1}]^T E_\ell N A^{-1} z) = (A^{-1}z, N^T E_\ell N A^{-1} z) \\ &= (A^{-1}z, N^T N E_\ell A^{-1} z) = (A^{-1}z, E_\ell A^{-1} z) = (z, [A^{-1}]^T E_\ell A^{-1} z) \\ &= I_\ell(z). \end{aligned} \tag{3.8.94}$$

Here again you will need to use the representation (8.46).

3.8.18. This exercise is devoted to the Krein-Moser theorem. It presumes that you have worked Exercises 8.16 and 8.17 above.

Let M be a real symplectic matrix all of whose eigenvalues are distinct, lie on the unit circle, and are different from ±1. Suppose we compute the quadratic form $\tilde{Q}_1(z)$ as given by (8.60). For notational convenience we will simply call it Q. Then we know from (8.90) that it has the representation

$$Q(z) = -\sum_{\ell=1}^{n} (\sin \phi_\ell)(\hat{p}_\ell^2 + \hat{q}_\ell^2). \tag{3.8.95}$$

We have agreed to employ the range $\phi_\ell \in (-\pi, \pi)$ and to exclude the possibilities $\phi_\ell = 0$ and $\phi_\ell = \pm\pi$. Show that as a consequence there is the relation

$$\text{sign}(\sin \phi_\ell) = \text{sign}(\phi_\ell). \tag{3.8.96}$$

It follows that $Q(z)$ is a negative-definite quadratic form if all phase advances are positive, and a positive-definite quadratic form if all phase advances are negative. If some phase advances are positive and some are negative, then $Q(z)$ is an indefinite quadratic form.

Now suppose, for example, that all phase advances are positive, and that two of them, say ϕ_1 and ϕ_2, are nearly equal. Then the eigenvalues λ_1 and λ_2 associated with each of them, as given by (5.39), are very nearly equal so that they are likely to collide if M is perturbed. According to the discussion in Section 3.5, these two eigenvalues will have the same signature, namely +1. There will be another pair λ_{-1} and λ_{-2} given by (5.37). They will also have the same signature, namely -1, and they will also collide if the first pair collides. Suppose that each pair does collide under perturbation of M, and afterward, contrary to the Krein-Moser theorem, each pair leaves the unit circle to form a Krein quartet. See Figure 5.1. Then there will be two eigenvalues, call them λ_+ and $\bar{\lambda}_+$, such that

$$|\lambda_+| = |\bar{\lambda}_+| > 1. \tag{3.8.97}$$

Show that, correspondingly, there will then be an initial condition z^0 such that the distance from the origin of the points z^k given by

$$z^k = M^k z^0 \tag{3.8.98}$$

grows without bound as $k \to \infty$.

Another more delicate situation that we need to consider is that M becomes undiagonalizable when some eigenvalues coincide so that they are no longer distinct. Then the best that can be achieved for M is that it can be brought to Jordan normal form with some +1's above the diagonal. Show that there will again be an initial condition z^0 such that the distance from the origin of the points z^k given by (8.98) grows without bound as $k \to \infty$.

Under the assumptions made, Q is negative definite before M is perturbed. By continuity, it will remain negative definite under perturbation of M. See Appendix O. But now we have reached a contradiction. Show that if z^k grows without bound as $k \to \infty$ and Q is negative definite, then $Q(z^k)$ must become ever more negative as $k \to \infty$. Indeed, suppose that after M is perturbed we use the representation (O.7) for Q.[41] We know that all the σ_j will be negative because Q is negative definite. Let $s_{\min}$ be the minimum of the quantities $-\sigma_j$. Show that

$$-Q(z) \geq s_{\min}||z||^2. \tag{3.8.99}$$

But, because Q is invariant, we must also have the relation

$$Q(z^k) = Q(z^0) \tag{3.8.100}$$

so that $Q(z^k)$ must, in fact, remain constant as $k \to \infty$. It follows that the eigenvalues associated with ϕ_1 and ϕ_2 cannot leave the unit circle as M is perturbed, nor can M become undiagonalizable.

Carry out similar reasoning for the case where all phase advances are negative. Finally, show that Q is indefinite if some phase advances are positive and some are negative so that the above reasoning cannot be applied in that case. In fact, Exercise 25.2.9 provides an example for which two pairs of eigenvalues of opposite signature do indeed collide and then leave the unit circle to become a Krein quartet.

[41] The representation (8.95) cannot be employed in this case because its construction required that all the eigenvalues be on the unit circle and distinct.

Suppose that Q is indefinite before M is perturbed. Suppose also that two pairs of eigenvalues do come off the unit circle when M is perturbed. Show that Q must then remain indefinite after M is perturbed. Indeed, show that there is a contradiction if Q becomes definite.

We have used the invariant quadratic form $\tilde{Q}_1(z)$ in all our analysis above. Show that $\tilde{Q}_1'(z)$, and the $Q_c(z)$ described in Exercise 11.4 of Section 3.11, could also have been used. Note that we need an invariant form that is defined in terms of M itself since, at least without further work, we cannot presume that a suitable S, as employed to construct the $Q_k(z)$ in Exercise 8.16, can be found after M is perturbed. Recall that in that exercise our construction of S itself assumed that the eigenvalues of M were on the unit circle and distinct. What is needed, if we are not sure this is the case, is some other way of constructing S from M, say by proving the existence of $\log M$ and verifying that it has various desired properties.

3.9 Unitary Subgroup Structure

It is easily verified that the commutator of any two matrices of the form JS^c is again a matrix of the form JS^c. Consequently, matrices of the form JS^c constitute a Lie algebra all by themselves. By contrast, the commutator of a matrix of the form JS^c with that of the form JS^a is again a matrix of the form JS^a. Finally, the commutator of two matrices of the form JS^a is a matrix of the form JS^c. We summarize these results by writing the relations

$$\{JS^c, JS^{c\prime}\} \propto JS^{c\prime\prime}, \tag{3.9.1}$$

$$\{JS^c, JS^a\} \propto JS^{a\prime}, \tag{3.9.2}$$

$$\{JS^a, JS^{a\prime}\} \propto JS^c. \tag{3.9.3}$$

Since matrices of the form JS^c form a Lie algebra, their exponentials must form a group, and this group will be a subgroup of the full symplectic group. Let us call this subgroup H. We know that it is symplectic and, since it arose from polar decomposition [see (8.15)], it is also orthogonal.[42] Therefore H is in the intersection of the orthogonal and symplectic groups,

$$H \subseteq [O(2n, \mathbb{R}) \cap Sp(2n, \mathbb{R})]. \tag{3.9.4}$$

The purpose of this section is to study H. For this study it is useful to employ the form (1.1) for J. We will find that H is isomorphic to the unitary group $U(n)$.

The most general $2n \times 2n$ real symmetric matrix S can be written in the block form

$$S = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}, \tag{3.9.5}$$

where the matrices A, B, and C are $n \times n$ and real, and the matrices A and C are themselves symmetric,

$$A^T = A, \tag{3.9.6}$$

[42] Note also that matrices of the form JS^c are antisymmetric and therefore, when exponentiated, must produce orthogonal matrices.

$$C^T = C. \tag{3.9.7}$$

Requiring that J commute with S gives the restrictions

$$B^T = -B \tag{3.9.8}$$

$$C = A. \tag{3.9.9}$$

Thus, the most general S^c is of the form

$$S^c = \begin{pmatrix} A & B \\ -B & A \end{pmatrix} \tag{3.9.10}$$

with the restrictions (9.6) and (9.8). Correspondingly, JS^c is of the form

$$JS^c = \begin{pmatrix} -B & A \\ -A & -B \end{pmatrix}. \tag{3.9.11}$$

Let W be the unitary and (complex) symplectic matrix

$$W = \frac{1}{\sqrt{2}} \begin{pmatrix} I & iI \\ iI & I \end{pmatrix}. \tag{3.9.12}$$

Here each block in W is $n \times n$. Then it is easily verified that the similarity transformation produced by W brings matrices of the form JS^c to block diagonal form. From (9.11) and (9.12) we find the result

$$W^{-1}(JS^c)W = \begin{pmatrix} -B + iA & 0 \\ 0 & -B - iA \end{pmatrix}. \tag{3.9.13}$$

Here each block is again $n \times n$. Now observe that matrices of the form $-B + iA$ with A and B real and obeying (9.6) and (9.8) span the space of all $n \times n$ anti-Hermitian matrices. Consequently, upon exponentiation, matrices of the form $-B+iA$ generate the unitary group $U(n)$. Correspondingly, the matrices $-B-iA$ generate the complex conjugate representation for which we employ the abusive notation $\overline{U}(n)$. Therefore, the Lie algebra spanned by the matrices JS^c is reducible, and is a variant of $u(n)$, the Lie algebra of $U(n)$.

To see how this works in more detail, exponentiate both sides of (9.13) to get the result

$$\exp[W^{-1}(JS^c)W] = \begin{pmatrix} \exp(-B + iA) & 0 \\ 0 & \exp(-B - iA) \end{pmatrix}. \tag{3.9.14}$$

Define a matrix v by the rule

$$v = \exp(-B + iA). \tag{3.9.15}$$

As described earlier, v is unitary as a result of (9.6) and (9.8),

$$v^\dagger = v^{-1}. \tag{3.9.16}$$

Also, any unitary matrix can be written in the form (9.15). Next, observe that the left side of (9.14) can be written in the form

$$\exp[W^{-1}(JS^c)W] = W^{-1}\exp(JS^c)W = W^{-1}MW. \tag{3.9.17}$$

Finally, solving (9.17) and (9.14) for M gives the result

$$M(v) = W \begin{pmatrix} v & 0 \\ 0 & \overline{v} \end{pmatrix} W^{-1}. \tag{3.9.18}$$

Suppose m is an arbitrary $n \times n$ matrix with possibly complex entries. Define an associated $2n \times 2n$ matrix $M(m)$ by the rule

$$M(m) = W \begin{pmatrix} m & 0 \\ 0 & \overline{m} \end{pmatrix} W^{-1}. \tag{3.9.19}$$

Then it is easily verified that there are the relations

$$M(I_n) = I_{2n}, \tag{3.9.20}$$

$$M(m_1 m_2) = M(m_1)M(m_2), \tag{3.9.21}$$

$$M(m^{-1}) = M^{-1}(m), \tag{3.9.22}$$

$$M^{\dagger}(m) = M(m^{\dagger}). \tag{3.9.23}$$

Here I_n denotes the $n \times n$ identity matrix. Also, if (9.19) is multiplied out explicitly, we find the result

$$M(m) = \begin{pmatrix} \mathrm{Re}(m) & \mathrm{Im}(m) \\ -\mathrm{Im}(m) & \mathrm{Re}(m) \end{pmatrix}. \tag{3.9.24}$$

It follows that $M(m)$ is real for any m. Consequently, we also have the relation

$$M^T(m) = M^{\dagger}(m) = M(m^{\dagger}). \tag{3.9.25}$$

Use of (9.24) for the case $m = iI_n$ gives the result

$$M(iI_n) = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix} = J. \tag{3.9.26}$$

[Note that the matrix (iI_n) is unitary. Note also that (9.26) is consistent with J providing an almost complex structure. See Exercise 2.6.] Suppose we compute M^TJM. By using (9.21), (9.25), and (9.26), we find the result

$$\begin{aligned} M^T(m)JM(m) &= M(m^{\dagger})M(iI_n)M(m) = M[m^{\dagger}(iI_n)m] \\ &= M[m^{\dagger}m(iI_n)] = M(m^{\dagger}m)M(iI_n) \\ &= M(m^{\dagger}m)J. \end{aligned} \tag{3.9.27}$$

However, from (9.24) we also have the result

$$M(m^{\dagger}m) = \begin{pmatrix} \mathrm{Re}(m^{\dagger}m) & \mathrm{Im}(m^{\dagger}m) \\ -\mathrm{Im}(m^{\dagger}m) & \mathrm{Re}(m^{\dagger}m) \end{pmatrix}. \tag{3.9.28}$$

Consequently, inspection of (9.27) and (9.28) shows that a necessary and sufficient condition for $M(m)$ to be symplectic is that m be unitary,

$$m^{\dagger}m = I. \tag{3.9.29}$$

Also, if m is unitary, then use of (9.25) and (9.22) gives the result

$$M^T(m) = M^\dagger(m) = M(m^\dagger) = M(m^{-1}) = M^{-1}(m). \tag{3.9.30}$$

Thus $M(m)$ is also orthogonal if m is unitary.

Conversely, suppose M is a real symplectic matrix that is also orthogonal. Write M in the form

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \tag{3.9.31}$$

where the matrices a, b, c, and d are real and $n \times n$. Impose on M the condition (8.7), which is equivalent to M being both sympletic and orthogonal. See (1.9). Doing so gives the results

$$c = -b \quad , \quad d = a. \tag{3.9.32}$$

Consequently, a real symplectic orthogonal M must be of the form

$$M = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}. \tag{3.9.33}$$

Next, define the $n \times n$ matrix m by the relation

$$m = a + ib. \tag{3.9.34}$$

As a result of (9.33) and (9.34), M can be written in the form

$$M = M(m) \tag{3.9.35}$$

with $M(m)$ defined by (9.24). Finally, as has been seen, use of the symplectic condition for M implies that m is unitary, so (9.30) also holds. Thus, we conclude that (9.24), (9.34), and (9.35) give a one-to-one correspondence between $2n \times 2n$ real symplectic orthogonal matrices and $n \times n$ unitary matrices. Moreover, the relations (9.20) through (9.22) show that this correspondence is an isomorphism. More precisely, the set of $2n \times 2n$ real symplectic orthogonal matrices forms a group that is the representation $U(n) \oplus \overline{U}(n)$ of $U(n)$. We will sometimes refer to these matrices, which form the subgroup we have called H, as the $U(n)$ subgroup of $Sp(2n, \mathbb{R})$.

At this point we can also provide another proof of the fact that symplectic matrices must have determinant $+1$. For simplicity, we will restrict our discussion to the case of real symplectic matrices. Suppose M is written in the polar form (8.3). Then we know from (8.8) and (8.9) that both factors P and O are symplectic and hence, according to (1.8), must satisfy the relations

$$\det P = \pm 1, \tag{3.9.36}$$

$$\det O = \pm 1. \tag{3.9.37}$$

But since P is real positive definite and symmetric, its eigenvalues must be real and positive, and hence its determinant must be positive. Thus we must have the relation

$$\det P = +1. \tag{3.9.38}$$

Next consider the matrix O, which is symplectic and real orthogonal. According to (9.35) and (9.19), O can be written in the form

$$O = W \begin{pmatrix} m & 0 \\ 0 & \overline{m} \end{pmatrix} W^{-1}. \tag{3.9.39}$$

Now take the determinant of both sides of (9.39). Doing so gives the result

$$\begin{aligned} \det(O) &= [\det(W)][\det(m)][\det(\overline{m})][\det(W^{-1})] \\ &= [\det(m)][\det(\overline{m})] = |\det(m)|^2 \geq 0. \end{aligned} \tag{3.9.40}$$

Comparison of (9.37) and (9.40) gives the result

$$\det O = +1. \tag{3.9.41}$$

Note that (9.40) and (9.41) are consistent with the fact that $|\det(m)|^2 = 1$ for any unitary matrix m. And from (8.3), (9.38), and (9.41) we conclude that

$$\det M = +1. \tag{3.9.42}$$

Finally it remains to be shown, as promised, that a real symplectic orthogonal matrix M can be written in the form (8.10) with F real and satisfying (8.11) and (8.13). As has been seen, such an M can be written in the form (9.19) with m unitary. Since m is unitary, there exist real matrices A and B satisfying (9.6) and (9.8) such that m can be written in the form

$$m = \exp(-B + iA). \tag{3.9.43}$$

Correspondingly, using (9.43) and (9.19), M can be written in the form

$$M = W \begin{pmatrix} \exp(-B + iA) & 0 \\ 0 & \exp(-B - iA) \end{pmatrix} W^{-1}. \tag{3.9.44}$$

However, the right side of (9.44) can be manipulated to take the form

$$W \begin{pmatrix} \exp(-B + iA) & 0 \\ 0 & \exp(-B - iA) \end{pmatrix} W^{-1} = W \exp(H) W^{-1} = \exp(WHW^{-1}), \tag{3.9.45}$$

where here H is a matrix defined by the relation

$$H = \begin{pmatrix} -B + iA & 0 \\ 0 & -B - iA \end{pmatrix}. \tag{3.9.46}$$

Define a matrix F by the relation

$$F = WHW^{-1}. \tag{3.9.47}$$

Then, use of (9.44) through (9.47) gives the result

$$M = \exp(F). \tag{3.9.48}$$

Also, explicit calculation using (9.46), (9.47), and (9.12) gives the result

$$F = \begin{pmatrix} -B & A \\ -A & -B \end{pmatrix}. \tag{3.9.49}$$

It is readily verified from (9.6) and (9.8) that F satisfies (8.11) and (8.13). Finally, if this F is used in (8.15) to solve for S^c, one finds the result (9.10).

We close this section with one last observation. Consider the $n \times n$ diagonal unitary matrix v given by the relation

$$v(\phi_1, \phi_2, \cdots \phi_n) = \begin{pmatrix} \exp(i\phi_1) & & & \\ & \exp(i\phi_2) & & \\ & & \ddots & \\ & & & \exp(i\phi_n) \end{pmatrix}. \tag{3.9.50}$$

Let $V(\phi_1, \phi_2, \cdots \phi_n)$ be the associated real symplectic and orthogonal matrix given by the relation

$$V = M(v). \tag{3.9.51}$$

Explicit calculation gives the result

$$V = \begin{pmatrix} \mathrm{Re}(v) & \mathrm{Im}(v) \\ -\mathrm{Im}(v) & \mathrm{Re}(v) \end{pmatrix} = \begin{pmatrix} C & S \\ -S & C \end{pmatrix}. \tag{3.9.52}$$

Here C and S are $n \times n$ diagonal matrices given by the relations

$$C = \begin{pmatrix} \cos(\phi_1) & & & \\ & \cos(\phi_2) & & \\ & & \ddots & \\ & & & \cos(\phi_n) \end{pmatrix}, \tag{3.9.53}$$

$$S = \begin{pmatrix} \sin(\phi_1) & & & \\ & \sin(\phi_2) & & \\ & & \ddots & \\ & & & \sin(\phi_n) \end{pmatrix}. \tag{3.9.54}$$

Let us seek to write V in exponential form. Since V belongs to the $U(n)$ subgroup of $Sp(2n, \mathbb{R})$, there must be a matrix $\hat{S}^c$ such that

$$V = \exp(J\hat{S}^c). \tag{3.9.55}$$

From Exercise 3.8.14 we know that

$$V = \exp(J\hat{S}^c) = \cos(\hat{S}^c) + J\sin(\hat{S}^c). \tag{3.9.56}$$

But we also see from (9.52) that there is the relation

$$V = \begin{pmatrix} C & S \\ -S & C \end{pmatrix} = \begin{pmatrix} C & 0 \\ 0 & C \end{pmatrix} + \begin{pmatrix} 0 & S \\ -S & 0 \end{pmatrix} = \begin{pmatrix} C & 0 \\ 0 & C \end{pmatrix} + J\begin{pmatrix} S & 0 \\ 0 & S \end{pmatrix}. \tag{3.9.57}$$

Upon comparing (9.56) and (9.57) we find that

$$\cos(\hat{S}^c) = \begin{pmatrix} C & 0 \\ 0 & C \end{pmatrix} \tag{3.9.58}$$

and

$$\sin(\hat{S}^c) = \begin{pmatrix} S & 0 \\ 0 & S \end{pmatrix}. \tag{3.9.59}$$

It follows that

$$\hat{S}^c = \begin{pmatrix} \phi & 0 \\ 0 & \phi \end{pmatrix} \tag{3.9.60}$$

where ϕ is the diagonal matrix

$$\phi = \begin{pmatrix} \phi_1 & & & \\ & \phi_2 & & \\ & & \ddots & \\ & & & \phi_n \end{pmatrix}. \tag{3.9.61}$$

Evidently incrementing any of the angles ϕ_ℓ by 2π brings V (or v) back to itself. Thus these elements form an *n-torus* within $Sp(2n,\mathbb{R})$. An n-torus is the topological product of n circles, has dimension n, and will be denoted by the symbol T^n. The n-torus $V(\phi_1, \phi_2, \cdots \phi_n)$, with each ϕ_ℓ ranging over $[0, 2\pi]$, is called a *maximal* torus within $Sp(2n, R)$ because there is no torus within $Sp(2n,\mathbb{R})$ having a dimension larger than n.

By construction, V is symplectic with with respect to the J given by (1.1). Let us find the corresponding V' that is symplectic with respect to the J' given by (2.10). According to (2.15), it is given by the relation

$$V' = PVP^T \tag{3.9.62}$$

where, here, P is the permutation matrix of Section 3.2. Note that, since P is orthogonal, V' will also be orthogonal. It is easily verified that carrying out the calculation (9.62) gives the result

$$V'(\phi_1, \phi_2, \cdots \phi_n) = N(\phi_1, \phi_2, \cdots \phi_n) \tag{3.9.63}$$

where N is given by (8.85). Observe that the normal form N given by (8.85), or by (5.60) and (5.61), is orthogonal and real symplectic for the J' given by (2.10). Moreover, from the work of Section 3.5, we know that any real symplectic M with all eigenvalues distinct and on the unit circle is conjugate to such an N by a real symplectic similarity transformation. See also Exercises (8.9) and (8.12). We conclude that all these matrices are related to the maximal n-torus $V(\phi_1, \phi_2, \cdots \phi_n)$.

Exercises

3.9.1. Verify the relations (9.1) through (9.3).

3.9.2. Consider the matrix M written in the form (8.26). Show that it can also be written in the form

$$M = \exp(JS^c)\exp(JS^{a'}). \tag{3.9.64}$$

Find the matrix $S^{a'}$.
Answer: $S^{a'} = [\exp(-JS^c)]S^a[\exp(JS^c)]$. Show that $S^{a'}$ is symmetric and anticommutes with J.

3.9.3. Verify that the requirement that J commute with S does indeed give the restrictions (9.8) and (9.9).

3.9.4. Verify that W as given by (9.12) is unitary. That is, $W^\dagger W = I$. Show also that W is (complex) symplectic. That is, show that W belongs to $Sp(2n, C)$.

3.9.5. Verify (9.13).

3.9.6. Verify (9.17).

3.9.7. Verify (9.20) through (9.23).

3.9.8. Verify (9.24).

3.9.9. Verify (9.25) and (9.30).

3.9.10. Find the dimension of the Lie algebra generated by all $2n \times 2n$ matrices of the form JS^c. Verify that this dimension is the same as that of $u(n)$. See Exercise 7.27. Find the dimension of the vector space spanned by all $2n \times 2n$ matrices of the form JS^a. You should have found the dimensions n^2 and $(n^2 + n)$, respectively. Verify, in accord with (8.29), that their sum is $\dim sp(2n)$ as given by (7.42).

3.9.11. Verify (9.49) starting with (9.43) and (9.12). Verify that F satisfies (8.11) and (8.13).

3.9.12. Use the methods of this section to show that all (real) symplectic matrices of the form $\exp(JS^c)$, i.e. all real symplectic orthogonal matrices, can be brought to the normal form (5.60) and (5.61) even if the eigenvalues are not necessarily distinct. In addition, show that the transforming matrix A can be taken to be both real symplectic and orthogonal. Hint: Use the fact that any unitary matrix can be brought to diagonal form by a unitary similarity transformation.

3.9.13. Suppose M is real orthogonal and symplectic with respect to the J of (1.1). Show that then M' as given by (2.15) is real orthogonal and symplectic with respect to the J' of (2.10), and vice versa.

3.9.14. Show that J belongs to the $U(n)$ subgroup of $Sp(2n, \mathbb{R})$, and also commutes with all matrices in $U(n)$.

3.9.15. Verify the relations (9.36) through (9.42).

3.9.16. Was the condition $\det M = +1$ used to derive (9.48) and (9.49)? Show that (9.48), (9.49), (9.8), and (7.104) imply the relation $\det(M) = +1$.

3.9.17. Verify (9.63).

3.9.18. Refer to Exercise 2.6. Given the real $2n$-vector z in (2.18), let $w(z)$ denote the complex n-vector given by (2.19). Suppose m is a (possibly complex) $n \times n$ matrix. Let m act on w to get the result

$$\begin{aligned} mw &= [\mathrm{Re}(m) + i\mathrm{Im}(m)][x + iy] \\ &= [\mathrm{Re}(m)x - \mathrm{Im}(m)y] + i[\mathrm{Im}(m)x + \mathrm{Re}(m)y]. \end{aligned} \tag{3.9.65}$$

Define a $2n \times 2n$ real matrix $N(m)$ by the rule

$$N(m) = \begin{pmatrix} \mathrm{Re}(m) & -\mathrm{Im}(m) \\ \mathrm{Im}(m) & \mathrm{Re}(m) \end{pmatrix}. \tag{3.9.66}$$

Prove the relations

$$mw(z) = w(N(m)z), \tag{3.9.67}$$

$$(mw, mw') = (Nz, Nz') + i(Nz, JNz'). \tag{3.9.68}$$

Suppose m is unitary. Show that N is then both orthogonal and symplectic. Refer to (9.23). Show that

$$N(m) = M(\overline{m}), \tag{3.9.69}$$

where an overbar denotes the operation of complex conjugation. Show that if m is unitary, then so is $\overline{m}$.

3.9.19. The purpose of this exercise is to understand more about the correspondence relation (9.19). Consider the set of all matrices $g \in GL(2n, \mathbb{R})$. Next consider the subset of such matrices that also commute with J. Show that these matrices form a subgroup H of $GL(2n, \mathbb{R})$. But, what is this subgroup H?

Write g in the block form

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \tag{3.9.70}$$

where the matrices a, b, c, and d are real and $n \times n$. Show that there are the results

$$Jg = \begin{pmatrix} c & d \\ -a & -b \end{pmatrix}, \tag{3.9.71}$$

and

$$gJ = \begin{pmatrix} -b & a \\ -d & c \end{pmatrix}. \tag{3.9.72}$$

Show that requiring that g commute with J yields the restrictions

$$c = -b \tag{3.9.73}$$

and

$$d = a. \tag{3.9.74}$$

Thus, g is of the form

$$g = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}. \tag{3.9.75}$$

Show that the dimension of H is $2n^2$.

Next define matrices A and B by the rules

$$A = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}, \tag{3.9.76}$$

and

$$B = \begin{pmatrix} b & 0 \\ 0 & b \end{pmatrix}. \tag{3.9.77}$$

Verify that both A and B commute with J. Show that there is also the relation

$$JB = \begin{pmatrix} 0 & b \\ -b & 0 \end{pmatrix}. \tag{3.9.78}$$

Therefore, we may also write

$$g = A + JB. \tag{3.9.79}$$

Suppose g_1 and g_2 are two matrices that commute with J and we use the representation (9.79) to write

$$g_k = A_k + JB_k. \tag{3.9.80}$$

Then, recalling that the A_k and B_k commute with J and that $J^2 = -I$, show that there is the product relation

$$g_1g_2 = (A_1A_2 - B_1B_2) + J(A_1B_2 + B_1A_2). \tag{3.9.81}$$

We see that, in (9.80) and (9.81), the matrix J plays a role analogous to the imaginary number i. Recall Exercise 2.6 that dealt with almost complex structure.

This analogy can be made explicit using the machinery of this section. An arbitrary $n \times n$ matrix m with possibly complex entries can be written in the form (9.34) where a and b are real $n \times n$ matrices. Let us multiply two such matrices together. Show that so doing gives the result

$$m_1m_2 = (a_1a_2 - b_1b_2) + i(a_1b_2 + b_1a_2). \tag{3.9.82}$$

Note the resemblance between the pairs (9.79), (9.34) and (9.81), (9.82). To pursue the analogy further, verify that there is the relation

$$g = M(m). \tag{3.9.83}$$

Next take the determinant of both sides of (9.83). Show that doing so gives the result

$$\begin{aligned} \det(g) &= [\det(W)][\det(m)][\det(\overline{m})][\det(W^{-1})] \\ &= [\det(m)][\det(\overline{m})] = |\det(m)|^2 \geq 0. \end{aligned} \tag{3.9.84}$$

Matrices of the form (9.34) constitute the group $GL(n, \mathbb{C})$ provided we add the condition

$$\det(m) \neq 0. \tag{3.9.85}$$

In view of (9.20) through (9.22), you have shown that the set of matrices $g \in GL(2n, \mathbb{R}, +)$ that also commute with J constitutes a group that is the representation $GL(n, \mathbb{C}) \oplus \overline{GL(n, \mathbb{C})}$ of $GL(n, \mathbb{C})$; and (9.19) is the relation that provides the isomorphism between them. Note, as a sanity check, that the dimension of $GL(n, \mathbb{C})$ is $2n^2$, which you have already shown is also the dimension of H.

Suppose we impose the further condition

$$\det(m) = 1. \tag{3.9.86}$$

Show that then

$$\det(g) = 1. \tag{3.9.87}$$

Matrices of the form (9.34) subject to the further condition (9.86) constitute the group $SL(n, \mathbb{C})$. In view of (9.20) through (9.22) and (9.87), you have shown that the set of matrices $g \in SL(2n, \mathbb{R})$ that also commute with J constitutes a group that is isomorphic to $SL(n, \mathbb{C})$. More precisely, the set of matrices $g \in SL(2n, \mathbb{R})$ that also commute with J constitutes a group that is the representation $SL(n, \mathbb{C}) \oplus \overline{SL(n, \mathbb{C})}$ of $SL(n, \mathbb{C})$.

3.9.20. This exercise explores some further properties of the matrix W given by (9.12). To begin, review Exercise 9.4. Next, show that from (9.19) and (9.26) that there is the relation

$$WJW^{-1} = \begin{pmatrix} iI & 0 \\ 0 & -iI \end{pmatrix}. \tag{3.9.88}$$

Thus, W provides a similarity transformation that diagonalizes J.[43]

Suppose a vector w is defined by the rule

$$w = Wz. \tag{3.9.89}$$

Show that if z is given by (1.7.9), then w has the entries

$$w = (1/\sqrt{2})(q_1 + ip_1, \cdots, q_n + ip_n; iq_1 + p_1, \cdots, iq_n + p_n). \tag{3.9.90}$$

Suppose instead a vector w is defined by the rule

$$w = W^{-1}z. \tag{3.9.91}$$

Show that then

$$w = (1/\sqrt{2})(q_1 - ip_1, \cdots, q_n - ip_n; -iq_1 + p_1, \cdots, -iq_n + p_n). \tag{3.9.92}$$

Let S^a be the symmetric matrix defined by the rule

$$S^a = \begin{pmatrix} -I & 0 \\ 0 & I \end{pmatrix}. \tag{3.9.93}$$

[43] In Chapter 27 it will be found that the Lie transformation realization of W diagonalizes all the Lie operators $: (p_j^2 + q_j^2)/2 :$.

Verify that the matrix JS^a is given by the relation

$$JS^a = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}. \tag{3.9.94}$$

Verify, as the notation indicates, that S^a anticommutes with J. Let $U(\theta)$ be the matrix defined by the relation

$$U(\theta) = \exp(i\theta JS^a). \tag{3.9.95}$$

Verify that U is (complex) symplectic because S^a is symmetric and U is unitary because JS^a is Hermitian. Verify that there is the relation

$$(JS^a)^2 = I. \tag{3.9.96}$$

Use this relation to sum the series implied by (9.95) to find the relation

$$U(\theta) = I\cos(\theta) + iJS^a\sin(\theta) = \begin{pmatrix} \cos(\theta) & i\sin(\theta) \\ i\sin(\theta) & \cos(\theta) \end{pmatrix}. \tag{3.9.97}$$

Show that there is the relation

$$U(\pi/4) = W. \tag{3.9.98}$$

Verify that there is the relation

$$W^4 = U(\pi) = -I. \tag{3.9.99}$$

Suppose w is defined by the rule

$$w = U(\theta)z. \tag{3.9.100}$$

Show that

$$w = (w_1, \cdots, w_n; w_{n+1}, \cdots, w_{2n}) \tag{3.9.101}$$

with

$$w_a = q_a\cos(\theta) + ip_a\sin(\theta) \quad \text{for} \quad a = 1, n \tag{3.9.102}$$

and

$$w_{n+a} = iq_a\sin(\theta) + p_a\cos(\theta) \quad \text{for} \quad a = 1, n. \tag{3.9.103}$$

3.10 Other Subgroup Structure

Consider symplectic matrices of the form (3.9) through (3.11). We have seen that they generate all symplectic matrices. We will now see that, when taken individually, they generate subgroups.

Consider first matrices of the form (3.9). If M and M' are two such matrices, we find the multiplication rule

$$M'M = \begin{pmatrix} I & B' \\ 0 & I \end{pmatrix}\begin{pmatrix} I & B \\ 0 & I \end{pmatrix} = \begin{pmatrix} I & B' + B \\ 0 & I \end{pmatrix}. \tag{3.10.1}$$

It follows, if the matrices B are taken to be arbitrary $n\times n$ matrices, then matrices of the form (3.9) comprise a group. Moreover, (10.1) shows that the elements of this group commute. That is, the group is Abelian. (See Exercise 7.5 for the definition of *Abelian*). Further thought reveals that this group is isomorphic to the translation group in n^2 dimensions. Finally, if B' and B satisfy (3.12), so does their sum $B'+B$. We conclude that symplectic matrices of the form (3.9) comprise a subgroup of the symplectic group. Moreover, this subgroup is isomorphic to the translation group in $n(n+1)/2$ dimensions.

Suppose B satisfies (3.12). Then the matrix S defined by the equation

$$S = \begin{pmatrix} 0 & 0 \\ 0 & B \end{pmatrix} \tag{3.10.2}$$

is symmetric and satisfies the relation

$$JS = \begin{pmatrix} 0 & B \\ 0 & 0 \end{pmatrix}. \tag{3.10.3}$$

Furthermore, the matrix JS is *nilpotent*. That is, JS satisfies the relation

$$(JS)^2 = \begin{pmatrix} 0 & B \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & B \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = 0. \tag{3.10.4}$$

Consequently, the exponential of JS is given by the simple relation

$$\exp(JS) = I + JS = \begin{pmatrix} I & B \\ 0 & I \end{pmatrix} = M. \tag{3.10.5}$$

We conclude that symplectic matrices of the form (3.9) can be written in the exponential form (10.5) with S given by (10.2).

Similar statements can be made about matrices of the form (3.10). They also form an Abelian subgroup. They can be written in the form

$$M = \exp(JS) \tag{3.10.6}$$

with S given by the relation

$$S = \begin{pmatrix} -C & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.10.7}$$

Moreover, matrices of the form (3.10) are *conjugate* to matrices of the form (3.9) under the action of J. Compute the matrix $J^{-1}MJ$ with M given by (3.9). Matrix multiplication gives the result

$$J^{-1}MJ = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}\begin{pmatrix} I & B \\ 0 & I \end{pmatrix}\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ -B & I \end{pmatrix}. \tag{3.10.8}$$

Consider matrices of the form (3.11). Let M and M' be two such matrices. We find the multiplication rule

$$M'M = \begin{pmatrix} A' & 0 \\ 0 & D' \end{pmatrix}\begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix} = \begin{pmatrix} A'A & 0 \\ 0 & D'D \end{pmatrix}. \tag{3.10.9}$$

Also, if D and D' satisfy (3.13), we have the result

$$D'D = [(A')^T]^{-1}[(A)^T]^{-1} = [(A'A)^T]^{-1}. \tag{3.10.10}$$

Consequently, matrices of the form (3.11) also form a subgroup. Note that the condition (3.13) places no restrictions on the matrices A save that they be invertible. Also, once A is given, D is completely specified by (3.13). This observation, when combined with the multiplication rule (10.9), shows that the subgroup is isomorphic to $GL(n, \mathbb{R})$, the *general linear* group of $n \times n$ invertible matrices over the real field.

Suppose the (real) matrix A is sufficiently near the identity. Then there is real matrix a such that A can be written in the form

$$A = \exp(a). \tag{3.10.11}$$

From (3.13) we find that D can be written in the form

$$D = \exp(-a^T). \tag{3.10.12}$$

Let S be the symmetric matrix defined by the relation

$$S = \begin{pmatrix} 0 & a^T \\ a & 0 \end{pmatrix}. \tag{3.10.13}$$

Then the matrix JS is given by the relation

$$JS = \begin{pmatrix} a & 0 \\ 0 & -a^T \end{pmatrix}. \tag{3.10.14}$$

Evidently, matrices M of the form (3.11) with A sufficiently near the identity can be written as

$$M = \exp(JS) \tag{3.10.15}$$

with S given by (10.13).

Let M be a symplectic matrix of the form

$$M = \begin{pmatrix} A & B \\ 0 & D \end{pmatrix}, \tag{3.10.16}$$

and let M' be another such matrix. Then we find the multiplication rule

$$M'M = \begin{pmatrix} A' & B' \\ 0 & D' \end{pmatrix}\begin{pmatrix} A & B \\ 0 & D \end{pmatrix} = \begin{pmatrix} A'A & A'B + B'D \\ 0 & D'D \end{pmatrix}. \tag{3.10.17}$$

We conclude that such matrices also form a subgroup. This subgroup is the *semi-direct* product of the subgroups of matrices of the forms (3.11) and (3.9). Note that as far as the subgroup of matrices of the form (3.11) is concerned, the multiplication rule (10.17) is the same as (10.9). That is, the diagonal blocks of (10.9) and (10.17) are the same. However, the upper right block of (10.17) is not the same as that of (10.1), but instead involves A' and D. We see that the subgroup of matrices of the form (3.9) is *transformed* under the action of the subgroup of matrices of the form (3.11). For this reason, the subgroup product is said to be semi-direct rather than simply *direct*. Sometimes it is convenient to use the matrix identity

$$\begin{pmatrix} A' & 0 \\ 0 & D' \end{pmatrix} \begin{pmatrix} I & B' \\ 0 & I \end{pmatrix} = \begin{pmatrix} A' & A'B' \\ 0 & D' \end{pmatrix}. \tag{3.10.18}$$

[Observe that the right side of (10.18) has the desired subgroup block form (10.16), and the matrices on the left have the subgroup block forms (3.11) and (3.9).] When this done, the only conditions that need to be enforced to ensure symplecticity are of the forms (3.12) and (3.13).

In a similar fashion it can be shown that symplectic matrices of the form

$$M = \begin{pmatrix} A & 0 \\ C & D \end{pmatrix} \tag{3.10.19}$$

also constitute a subgroup. This subgroup is the semi-direct product of the subgroups of matrices of the forms (3.11) and (3.10). For this subgroup it is sometimes convenient to use the matrix identity

$$\begin{pmatrix} A' & 0 \\ 0 & D' \end{pmatrix} \begin{pmatrix} I & 0 \\ C' & I \end{pmatrix} = \begin{pmatrix} A' & 0 \\ D'C' & D' \end{pmatrix}. \tag{3.10.20}$$

Exercises

3.10.1. Strictly speaking, (10.1) shows only that the set of matrices of the form (3.9) is closed under multiplication. Show that the other requirements for a (sub)group are also satisfied. See Section 3.6. Verify that matrices of the form (3.10) also constitute a subgroup. Verify (10.6) through (10.8).

3.10.2. Verify the relations (10.2) through (10.5).

3.10.3. Verify the relations (10.6) through (10.8). Also verify that the requirements for a subgroup are met. [See Exercise (10.1) above.]

3.10.4. Verify the relations (10.9) through (10.15). Also verify that the requirements for a subgroup are met. [See Exercise (10.1) above.]

3.10.5. Verify that symplectic matrices of the form (10.16) constitute a subgroup. [See Exercise (10.1) above.] Also verify that symplectic matrices of the form (10.19) constitute a subgroup.

3.11 Other Factorizations/Decompositions

Sections 3.3.1 and 3.10 demonstrated that usually a symplectic matrix can be written as a product of three symplectic matrices of the form (3.9) through (3.11), and a product of six such factors always suffices. Section 3.8 showed that any symplectic matrix has a polar decomposition, and hence can be written as a product of two symplectic matrices in the form (8.26). The purpose of this section is to describe other possible factorizations/decompositions of symplectic matrices that may be of subsequent use.

3.12 Cayley Representation of Symplectic Matrices

In Sections 3.7 and 3.8 we saw that there is a connection between symplectic matrices and symmetric matrices, namely the relations (7.36) and (8.26). In this section we will find another connection, and in Section 5.13 we will see that this connection is but one of a whole family of such connections.[44] The connection to be described here is based on the *Cayley* representation/transformation.[45] It is a matrix generalization of the hyperbolic function identity

$$\exp(z) = \cosh(z) + \sinh(z) = [1 + \tanh(z/2)]/[1 - \tanh(z/2)].$$

Let M be a (real) symplectic matrix sufficiently near the identity. Then, according to (7.36), M can be written in the form

$$M = \exp(JS) \tag{3.12.1}$$

with S real and symmetric. Now watch closely. By algebraic manipulation involving properties of the exponential function, we may write the following chain of relations:

$$\begin{aligned} M &= \exp(JS) = [\exp(\frac{1}{2}JS)][\exp(-\frac{1}{2}JS)]^{-1} \\ &= [\cosh(\frac{1}{2}JS) + \sinh(\frac{1}{2}JS)][\cosh(\frac{1}{2}JS) - \sinh(\frac{1}{2}JS)]^{-1} \\ &= [I + \tanh(JS/2)][I - \tanh(JS/2)]^{-1}. \end{aligned} \tag{3.12.2}$$

Next, define a matrix W by the equation

$$W = -J\tanh(JS/2). \tag{3.12.3}$$

Then, we also have the relation

$$JW = \tanh(JS/2). \tag{3.12.4}$$

Consequently, using (12.2) and (12.4), M can be written in the form

$$M = (I + JW)(I - JW)^{-1} = (I - JW)^{-1}(I + JW). \tag{3.12.5}$$

We will call this form the *Cayley* representation of M.[46]

The alert reader will have observed that, in going from (12.1) to (12.5), no use was made of the symplectic condition. We now show that M being *symplectic* implies that W is *symmetric*, and vice versa,

$$W = W^T \Leftrightarrow M^T JM = J. \tag{3.12.6}$$

[44] And, in Section 6.7, we will see that there is an analogous connection between symplectic maps and gradient maps.

[45] Arthur Cayley (1821-1895), in his 1858 "A Memoir on the Theory of Matrices", was the first to define matrices abstractly and to describe general matrix algebra including matrix inversion.

[46] The nomenclature *Cayley transform* or *Cayley trivialization* is also used in the literature.

First, suppose W is symmetric. Then taking the transpose of (12.5) gives the representation

$$\begin{aligned} M^T &= [(I - JW)^T]^{-1}[(I + JW)^T] \\ &= (I + WJ)^{-1}(I - WJ). \end{aligned} \tag{3.12.7}$$

Next use the representations (12.5) and (12.7) to compute the quantity $M^T JM$. Doing so gives the result

$$M^T JM = (I + WJ)^{-1}(I - WJ)J(I + JW)(I - JW)^{-1}. \tag{3.12.8}$$

Insert judicious factors of $I = J^{-1}J$ into part of (12.8) to get the simplification

$$\begin{aligned} J(I + JW)(I - JW)^{-1} &= J(I + JW)J^{-1}J(I - JW)^{-1}J^{-1}J \\ &= (I + WJ)(I - WJ)^{-1}J. \end{aligned} \tag{3.12.9}$$

Here use has been made of (1.3). Correspondingly, (12.8) now simplifies to the form

$$M^T JM = (I + WJ)^{-1}(I - WJ)(I + WJ)(I - WJ)^{-1}J. \tag{3.12.10}$$

Observe that the second and third factors in the right side of (12.10) commute. Thus, we also have the relation

$$\begin{aligned} M^T JM &= (I + WJ)^{-1}(I + WJ)(I - WJ)(I - WJ)^{-1}J \\ &= J, \end{aligned} \tag{3.12.11}$$

which is what we wanted to prove.

Conversely, suppose that M is symplectic. Solve (12.5) for the quantity JW to get the relation

$$JW = (M + I)^{-1}(M - I) = (M - I)(M + I)^{-1}. \tag{3.12.12}$$

Now take the transpose of (12.12) to get the result

$$- W^T J = (M^T - I)(M^T + I)^{-1}. \tag{3.12.13}$$

The symplectic condition can be written in the form

$$M^T = JM^{-1}J^{-1}. \tag{3.12.14}$$

See (1.9). Consequently, (12.13) can also be written in the form

$$\begin{aligned} -W^T J &= (JM^{-1}J^{-1} - I)(JM^{-1}J^{-1} + I)^{-1} \\ &= J(M^{-1} - I)(M^{-1} + I)^{-1}J^{-1} \\ &= J(I - M)M^{-1}[(I + M)M^{-1}]^{-1}J^{-1} \\ &= J(I - M)(I + M)^{-1}J^{-1} \\ &= J(-JW)J^{-1} = -WJ. \end{aligned} \tag{3.12.15}$$

It follows from (12.15) that W is symmetric,

$$W^T = W. \tag{3.12.16}$$

Now consider matrices of the form JW. We know they they are Hamiltonian, that is, they belong to the Lie algebra $sp(2n,\mathbb{R})$ [or, more generally, $sp(2n,\mathbb{C})$] if W is symmetric. Since we have seen that W is symmetric if M is symplectic, we conclude that for M sufficiently near the identity matrix and for JW sufficiently near the zero matrix there is the relation

$$M \in Sp(2n,\mathbb{R}) \Leftrightarrow JW \in sp(2n,\mathbb{R}), \tag{3.12.17}$$

or, more generally,

$$M \in Sp(2n,\mathbb{C}) \Leftrightarrow JW \in sp(2n,\mathbb{C}). \tag{3.12.18}$$

Thus, near the identity in group space and near the origin in Lie-algebra space, the Cayley representation, like the exponential map, provides a local bijection between group elements and Lie-algebra elements.

Again we note that the symplectic condition as expressed by (1.2) is a set of quadratic relations, and the use of the Cayley representation converts these quadratic relations into the simple linear relations (12.16).

We also need to make an important observation. It is easily checked that $-I$ is a symplectic matrix. However, $(M+I)$ is singular for $M=-I$. Indeed, $(M+I)$ is singular whenever M has -1 as an eigenvalue. It follows that JW does not exist in these cases. Consequently, unlike the two-exponentials product representation (8.26), the Cayley representation is not global.

We note for future use that (12.12) can be solved for W to give the relation

$$W = (-JM + J)(M+I)^{-1}. \tag{3.12.19}$$

Finally, we observe that (12.5) and the inverse relation (12.19) stand on their own without any need of the motivational assumption (12.1).

We close this section by noting that there are also Cayley representations for all the so-called *quadratic* matrix groups including orthogonal, unitary, and Lorentz transformation matrices. See Exercises 12.5 and 12.6.

Exercises

3.12.1. Show that (12.3) and (12.4) have the expansions

$$\begin{aligned} W &= -J\tanh(JS/2) = -J[(JS/2) - (1/3)(JS/2)^3 + (2/15)(JS/2)^5 - \cdots] \\ &= S/2 - SJSJS/24 + SJSJSJSJS/240 - \cdots . \end{aligned} \tag{3.12.20}$$

$$\begin{aligned} JW &= \tanh(JS/2) = (JS/2) - (1/3)(JS/2)^3 + (2/15)(JS/2)^5 - \cdots \\ &= JS/2 - (JS)^3/24 + (JS)^5/240 - \cdots . \end{aligned} \tag{3.12.21}$$

Show directly from (12.20) that W is symmetric if S is. Show from (12.21) that the matrices JW and JS commute,

$$\{JW, JS\} = 0. \tag{3.12.22}$$

Show that (12.20) and (12.21) can be inverted to give the relations

$$JS/2 = \tanh^{-1}(JW) = [JW + (1/3)(JW)^3 + (1/5)(JW)^5 + \cdots], \tag{3.12.23}$$

$$JS = 2\tanh^{-1}(JW) = 2[JW + (1/3)(JW)^3 + (1/5)(JW)^5 + \cdots], \tag{3.12.24}$$

$$\begin{aligned} S &= -2J\tanh^{-1}(JW) = -2J[JW + (1/3)(JW)^3 + (1/5)(JW)^5 + \cdots] \\ &= 2W + (2/3)WJWJW + (2/5)WJWJWJWJW + \cdots . \end{aligned} \tag{3.12.25}$$

Show from (12.25) that, conversely, S is symmetric if W is.

3.12.2. Derive (12.12) from (12.5).

3.12.3. Find the Cayley representation for the matrix N given by (5.60) and (5.61). That is, find the matrix W in this case. Show explicitly that the representation is not global, i.e., does not hold for all values of ϕ_ℓ.

3.12.4. Read Exercise 8.13. Let W and M be the matrices appearing in the Cayley relations (12.5) and (12.19). Define what we will call the Cayley quadratic form Q_c by the relation

$$Q_c(z) = (z, Wz). \tag{3.12.26}$$

Verify that W is of the form

$$W = Jg(M) \tag{3.12.27}$$

with

$$g(M) = (-M + I)(M + I)^{-1} \tag{3.12.28}$$

and that M commutes with $g(M)$. Show that Q_c is invariant under the action of M.

Show, for the case described in Exercise 8.14, that Q_c is given by the relation

$$Q_c(z) = \sum_{\ell=1}^{n}[\tan(\phi_\ell/2)](\hat{p}_\ell^2 + \hat{q}_\ell^2). \tag{3.12.29}$$

3.12.5. Section 3.12 described the Cayley representation of symplectic matrices. The purpose of this exercise is to explore Cayley representations for other kinds of matrices including orthogonal, unitary, and Lorentz transformation matrices. Let L be any fixed real nonsingular $m \times m$ matrix, and consider all $m \times m$ matrices M such that

$$M^T L M = L. \tag{3.12.30}$$

Show that

$$\det M = \pm 1, \tag{3.12.31}$$

and therefore all such matrices are invertible. Indeed, show from (12.30) that

$$M^{-1} = L^{-1} M^T L. \tag{3.12.32}$$

Note that, while matrix inversion is usually a computationally intensive task, all that is required in this case is the inversion of L, which can be done once and for all, the transposing of M, and two matrix multiplications.

Verify that all matrices that satisfy (12.30) form a group, call it G. Since the relation (12.30), is an algebraic one among the entries in M, G is an algebraic group. Indeed, since (12.30) is a quadratic relation, G is also sometimes called a *quadratic* group. Thus, for example, the orthogonal, symplectic, and Lorentz groups are quadratic groups.

Let $(*,*)$ denote the usual real inner product. Define an angular inner product, more accurately a bilinear form, $\langle *,* \rangle$ by the rule

$$\langle u, v\rangle = (u, Lv). \tag{3.12.33}$$

Verify that

$$\langle Mu, Mv\rangle = (Mu, LMv) = (u, M^T LMv) = (u, Lv) = \langle u, v\rangle. \tag{3.12.34}$$

That is, G preserves the bilinear form $\langle *,* \rangle$.

Show that G consists of two disconnected components comprised of elements with determinant $+1$ and elements with determinant -1. [Actually, it can happen, as is the case for $Sp(2n,\mathbb{R})$, that the component with determinant -1 is empty.] Show that the matrices $M \in G$ such that $\det M = 1$ form a subgroup, call it SG.

Consider matrices in SG that are sufficiently close to the identity so that they can be written in the exponential form

$$M = \exp(\epsilon A) \tag{3.12.35}$$

where ϵ is a sufficiently small parameter. Show, by equating powers of ϵ, that (12.30) and (12.34) require that A obey the relation

$$A^T L + LA = 0 \tag{3.12.36}$$

or, equivalently,

$$L^{-1}A^T L = -A. \tag{3.12.37}$$

Conversely, show that if A satisfies the relation (12.36), then any M given by (12.34) satisfies (12.30), and therefore belongs to G. Verify that matrices A that satisfy (12.36) form a Lie algebra, and therefore G is a Lie group. Show that (12.36) implies the relation

$$\text{tr } A = 0, \tag{3.12.38}$$

and therefore any M of the form (12.34) belongs to SG. Correspondingly, following our usual nomenclature, we may define sg to be the Lie algebra of all matrices A that satisfy (12.36).

Set $\epsilon = 1$ in (12.34). Following the logic of Section 3.12, show that matrices M sufficiently near the identity can be written in the form

$$M = (I + V)/(I - V) \tag{3.12.39}$$

where

$$V = \tanh(A/2) = (A/2) - (1/3)(A/2)^3 + (2/15)(A/2)^5 + \cdots \tag{3.12.40}$$

and

$$A = 2\tanh^{-1} V = 2[V + (1/3)V^3 + (1/5)V^5 + \cdots]. \tag{3.12.41}$$

Show that if A satisfies (12.36), then so does V, and vice versa,

$$L^{-1}A^TL = -A \Leftrightarrow L^{-1}V^TL = -V. \tag{3.12.42}$$

[Here it assumed that A is sufficiently small for the series (12.39) and its inverse relation (12.40) to be convergent.] We conclude that $V \in sg$.

Verify that (12.38) can be solved for V to yield the inverse relation

$$V = (M - I)/(M + I). \tag{3.12.43}$$

Verify that, for M sufficiently near the identity matrix and for V sufficiently near the zero matrix, there is the relation

$$M \in SG \Leftrightarrow V \in sg. \tag{3.12.44}$$

Consequently, for quadratic groups, (12.38) and (12.42) provide a mapping between SG and sg, which is a bijection between elements in SG sufficiently near the identity and elements in sg sufficiently near the origin.

Sometimes it is convenient to define a function that is a variant of the relation (12.38). Given any matrix X that does not have -1 as an eigenvalue, define a matrix function cay by the rule

$$\text{cay}(X) = (I - X)/(I + X). \tag{3.12.45}$$

With this definition, (12.38) becomes

$$M = \text{cay}(-V) \tag{3.12.46}$$

and (12.42) becomes

$$V = -\text{cay}(M). \tag{3.12.47}$$

Show that we may also write

$$\text{cay}(V) = M^{-1} \tag{3.12.48}$$

and

$$\text{cay}(M^{-1}) = V. \tag{3.12.49}$$

Note that $V \in sg$ implies that $-V \in sg$, and vice versa; and $M \in SG$ implies that $M^{-1} \in SG$, and vice versa. Therefore, in our context, the function cay also provides a bijection between elements in SG sufficiently near the identity and elements in sg sufficiently near the origin. A map or operator whose square is the identity is often called an *involution*. Show that the map cay is an *involution*. That is, show that

$$(\text{cay})^2(X) = \text{cay}[\text{cay}(X)] = X. \tag{3.12.50}$$

Verify that in the case of $SO(m)$, for which $L = I$, the matrices A and V are antisymmetric.

Scan Exercise 6.2.6. Be aware that in this exercise the symbol g is used to denote the metric tensor. Verify that the Lorentz group is a quadratic group, and therefore has a Cayley representation.

Reapply, with necessary modifications, the arguments made so far to the case of matrices M that satisfy the relation

$$M^{\dagger} L M = L. \tag{3.12.51}$$

Show that such matrices form a group G and that there is a Cayley representation that provides a map between G and its Lie algebra g. Apply your results to the case of $U(m)$, for which $L = I$, and show that in this case the matrices A and V are anti-Hermitian. That is, $A \in u(m)$ and $V \in u(m)$. Verify that in the quantum-mechanical theory of scattering, for which M is the unitary scattering matrix S, there is the relation

$$S = M = (I + iK)/(I - iK) \tag{3.12.52}$$

where the so called K matrix given by $K = -iV$ is Hermitian. Verify also the inverse relation

$$K = -i(S - I)/(S + I). \tag{3.12.53}$$

The relations (12.51) and (12.52) provide a map between unitary matrices S and Hermitian matrices K. We remark that if the scattering process is time symmetric, then it can be shown that K is real, and therefore also symmetric.

Are there familiar examples of groups for which there is no Cayley representation? It depends what one means by a Cayley representation. If one means that the Cayley relations are required to supply a bijection between the group and its Lie algebra, then there are are groups for which there is no Cayley representation in the sense that the Cayley relations do not provide a bijection between the group and its Lie algebra. The groups $SL(m, \mathbb{R})$ for $m > 2$ are examples. In working Exercise 7.25 you should have found that $sl(m, \mathbb{R})$, the Lie algebra of $SL(m, \mathbb{R})$, consists of all real $m \times m$ matrices A that are traceless. Consider the case of $sl(3, \mathbb{R})$ and the Lie algebraic element

$$A = \epsilon \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix} \tag{3.12.54}$$

where ϵ is small. Verify that V as given by (12.39) is not traceless and therefore does not belong to $sl(3, \mathbb{R})$. Thus, there is no Cayley representation for $SL(3, \mathbb{R})$. That is, although the relations (12.38) and (12.42) continue to hold, there are group elements $M \in SL(3, \mathbb{R})$, and arbitrarily near the identity, for which the corresponding V given by (12.42) is not in $sl(3, \mathbb{R})$.

Verify that $SL(3, \mathbb{R})$ is a subgroup of $SL(m, \mathbb{R})$ for $m > 3$ and therefore there is no Cayley representation for $SL(m, \mathbb{R})$ when $m > 2$. Is $SL(m, \mathbb{R})$ an algebraic group?

3.12.6. The aim of this exercise is to explore in some detail the use of Cayley parameterizations for the cases of $SO(3, \mathbb{R})$ and $SU(2)$.

Assume, for the case of $SO(3, \mathbb{R})$, that group elements are parameterized in *exponential* form by the relation

$$R_e(\boldsymbol{\lambda}) = \exp(\boldsymbol{\lambda} \cdot \boldsymbol{L}). \tag{3.12.55}$$

Show that they have an associated *Cayley* parameterization of the form

$$R_c(\boldsymbol{\mu}) = (I + \boldsymbol{\mu} \cdot \boldsymbol{L})/(I - \boldsymbol{\mu} \cdot \boldsymbol{L}). \tag{3.12.56}$$

Show that if $R_e(\boldsymbol{\lambda}) = R_c(\boldsymbol{\mu}) = R$, then the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are interconnected by the relations

$$\boldsymbol{\mu}\cdot\boldsymbol{L} = \tanh(\boldsymbol{\lambda}\cdot\boldsymbol{L}/2) \tag{3.12.57}$$

and

$$\boldsymbol{\lambda}\cdot\boldsymbol{L} = 2\tanh^{-1}(\boldsymbol{\mu}\cdot\boldsymbol{L}). \tag{3.12.58}$$

Also verify that (12.55) can be inverted to give the relation

$$\boldsymbol{\mu}\cdot\boldsymbol{L} = [R-I]/[R+I]. \tag{3.12.59}$$

Verify, for the case of $so(3,\mathbb{R})$, that there is the relation

$$(\boldsymbol{\nu}\cdot\boldsymbol{L})^3 = (i|\boldsymbol{\nu}|)^2\boldsymbol{\nu}\cdot\boldsymbol{L} \tag{3.12.60}$$

for any 3-vector $\boldsymbol{\nu}$. See (7.201). Use this relation to show that (12.56) and (12.57) can be rewritten in the forms

$$\boldsymbol{\mu}\cdot\boldsymbol{L} = (\boldsymbol{\lambda}\cdot\boldsymbol{L}/|\boldsymbol{\lambda}|)\tan(|\boldsymbol{\lambda}|/2) \tag{3.12.61}$$

and

$$\boldsymbol{\lambda}\cdot\boldsymbol{L} = 2(\boldsymbol{\mu}\cdot\boldsymbol{L}/|\boldsymbol{\mu}|)\tan^{-1}(|\boldsymbol{\mu}|). \tag{3.12.62}$$

Hint: Expand (12.56) and (12.57) in Taylor series, use (12.59) in these series, and then sum the transformed series to get the advertised results. Show it follows from (12.60) and (12.61) that

$$\boldsymbol{\mu} = (\boldsymbol{\lambda}/|\boldsymbol{\lambda}|)\tan(|\boldsymbol{\lambda}|/2) \tag{3.12.63}$$

and

$$\boldsymbol{\lambda} = 2(\boldsymbol{\mu}/|\boldsymbol{\mu}|)\tan^{-1}(|\boldsymbol{\mu}|). \tag{3.12.64}$$

Verify that both (12.62) and (12.63) imply the relation

$$|\boldsymbol{\mu}| = \tan(|\boldsymbol{\lambda}|/2). \tag{3.12.65}$$

Observe that (12.64) is singular when $|\boldsymbol{\lambda}| = \pi$. Verify that this singularity is to be expected because then R has -1 as an eigenvalue, from which it follows that the factor $[R+I]^{-1}$ in (12.58) is singular.

Assume, for the case of $SU(2)$, that group elements are parameterized in exponential form by the relation

$$u_e(\boldsymbol{\lambda}) = \exp(\boldsymbol{\lambda}\cdot\boldsymbol{K}). \tag{3.12.66}$$

In the notation of Exercise 12.5, group elements near the identity have an associated parameterization of the form

$$u_c = (I+V)/(I-V). \tag{3.12.67}$$

Show that setinc $u_e = u_c = u$ yields the result

$$V = \tanh(\boldsymbol{\lambda}\cdot\boldsymbol{K}/2). \tag{3.12.68}$$

In order for this parameterization to be a Cayley parameterization we must verify that $V \in su(2)$. Check, for the case of $su(2)$, that there is the relation

$$(\boldsymbol{\nu}\cdot\boldsymbol{K})^3 = (i|\boldsymbol{\nu}|/2)^2\boldsymbol{\nu}\cdot\boldsymbol{K} \tag{3.12.69}$$

for any 3-vector $\boldsymbol{\nu}$. See (7.187). Use this relation in the Taylor series for the right side of (12.67) to show that (12.67) can be rewritten in the form

$$V = (\boldsymbol{\lambda} \cdot \boldsymbol{K})(2/|\boldsymbol{\lambda}|)\tan(|\boldsymbol{\lambda}|/4), \tag{3.12.70}$$

from which it follows, in particular, that $V \in su(2)$, and $SU(2)$ has a Cayley parameterization.[47] Therefore, we may write (12.66) in the Cayley form

$$u_c(\boldsymbol{\mu}) = (I + \boldsymbol{\mu} \cdot \boldsymbol{K})/(I - \boldsymbol{\mu} \cdot \boldsymbol{K}) \tag{3.12.71}$$

with

$$\boldsymbol{\mu} \cdot \boldsymbol{K} = V = (\boldsymbol{\lambda} \cdot \boldsymbol{K})(2/|\boldsymbol{\lambda}|)\tan(|\boldsymbol{\lambda}|/4). \tag{3.12.72}$$

From (12.71) show that

$$\boldsymbol{\mu} = 2(\boldsymbol{\lambda}/|\boldsymbol{\lambda}|)\tan(|\boldsymbol{\lambda}|/4) \tag{3.12.73}$$

and

$$|\boldsymbol{\mu}| = 2\tan(|\boldsymbol{\lambda}|/4). \tag{3.12.74}$$

Show also that (12.70) can be inverted to give the relation

$$\boldsymbol{\mu} \cdot \boldsymbol{K} = [u - I]/[u + I]. \tag{3.12.75}$$

Note that (12.73) is singular when $|\boldsymbol{\lambda}| = 2\pi$. Verify that this singularity is to be expected because then $u = -I$, see (7.189), from which it follows that the factor $[u + I]^{-1}$ in (12.74) is singular. Finally, show that (12.72) and (12.73) can be solved for $\boldsymbol{\lambda}$ to give the inverse relation

$$\boldsymbol{\lambda} = 4(\boldsymbol{\mu}/|\boldsymbol{\mu}|)\tan^{-1}(|\boldsymbol{\mu}|/2). \tag{3.12.76}$$

Note that both (12.62) and (12.72) yield the relation $\boldsymbol{\mu} \simeq \boldsymbol{\lambda}/2$ for small $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. But they differ in higher order.

3.13 General Symplectic Forms, Darboux Transformations, Pfaffians, and Variant Symplectic Groups

3.13.1 General Symplectic Forms

According to Exercise 2.7, the symplectic group consists of all linear transformations that preserve the fundamental symplectic 2-form (2.3). The matrix J in this 2-form has the property that it is real, antisymmetric, and nonsingular. We will now see that there is an endless supply of matrices K with this property; and we will call each (w, Kz) a *generalized* symplectic 2-form.

First we assert that such a matrix must be $2n \times 2n$ for some choice of n. For suppose K is $m \times m$ with m odd. Then we find that

$$\det K = \det K^T = \det(-K) = (-1)^m \det K = -\det K, \tag{3.13.1}$$

[47] We remark that although $U(n)$ has a Cayley parameterization, $SU(n)$ does not when $n > 2$.

from which it follows that $\det K = 0$ and therefore K is singular, contrary to one of our stipulations about K.

Next, let N be any matrix in $GL(2n, \mathbb{R})$. Define an associated matrix K by the rule

$$K = NJN^T. \tag{3.13.2}$$

That is, K and J are congruent under the action of N. Evidently K is real. We also find by direct calculation that

$$K^T = (NJN^T)^T = NJ^TN^T = -NJN^T = -K. \tag{3.13.3}$$

Moreover,

$$\det K = (\det N)(\det J)(\det N^T) = (\det N)^2 > 0. \tag{3.13.4}$$

Therefore K is nonsingular.

The converse is also true. Given any real, antisymmetric, and nonsingular $2n \times 2n$ matrix K, there is a matrix $N \in GL(2n, \mathbb{R})$ such that (13.2) holds.

We begin the demonstration of this claim by showing that there is a set of (real) basis vectors v^1, v^2, $\cdots$, v^{2n} such that

$$(v^i, Kv^j) = J'_{ij}, \tag{3.13.5}$$

where J' is the matrix given by (2.10). The construction of the v^i is very similar to that used for Darboux symplectification. Let w^1, $\cdots$, w^{2n} be any set of $2n$ real and linearly independent vectors. For convenience, they might be taken to be the unit vectors e^1, $\cdots$, e^{2n} given by (6.4). Now follow this algorithm:

1. Define v^1 by the simple rule

$$v^1 = w^1. \tag{3.13.6}$$

2. Starting with w^2, search through the w^j with $j \geq 2$ to find the first j, call it k, with the property

$$(v^1, Kw^j) \neq 0. \tag{3.13.7}$$

[Better yet, if one is working numerically and therefore only to finite precision, select j so that $|(v^1, Kw^j)|$ is maximized. The analogous choices should also be made in steps 6, 10, etc. below.] Renumber the vectors $w^2 \cdots w^{2n}$ so that w^k becomes w^2.

3. Define v^2 by the rule

$$v^2 = w^2/[(v^1, Kw^2)]. \tag{3.13.8}$$

We then have the result

$$(v^1, Kv^2) = 1 = J'_{12}. \tag{3.13.9}$$

And, since K is antisymmetric, at this stage we have the result

$$(v^i, Kv^j) = J'_{ij} \text{ for } i, j = 1 \text{ to } 2. \tag{3.13.10}$$

4. Using the remaining vectors $w^3 \cdots w^{2n}$, define new vectors ${}^1w^j$ for $j \geq 3$ by the rule

$$ {}^1w^j = w^j + (v^2, Kw^j)v^1 - (v^1, Kw^j)v^2. \tag{3.13.11}$$

   As a result of this rule there are the relations

$$(v^i, K\ {}^1w^j) = 0 \text{ for } i = 1, 2 \text{ and } j = 3, 4, \cdots 2n. \tag{3.13.12}$$

5. Define v^3 by the rule

$$v^3 = {}^1w^3. \tag{3.13.13}$$

6. Starting with ${}^1w^4$, search through the ${}^1w^j$ with $j \geq 4$ to find the first j, call it k, with the property

$$(v^3, K\ {}^1w^j) \neq 0. \tag{3.13.14}$$

   Renumber the vectors ${}^1w^4 \cdots {}^1w^{2n}$ so that ${}^1w^k$ becomes ${}^1w^4$.

7. Define v^4 by the rule

$$v^4 = {}^1w^4/[(v^3, K\ {}^1w^4)]. \tag{3.13.15}$$

   At this stage we have the results

$$(v^i, Kv^j) = J'_{ij} \text{ for } i, j = 1 \text{ to } 4. \tag{3.13.16}$$

8. Using the remaining vectors ${}^1w^5 \cdots {}^1w^{2n}$, define new vectors ${}^2w^j$ for $j \geq 5$ by the rule

$$ {}^2w^j = {}^1w^j + (v^4, K\ {}^1w^j)v^3 - (v^3, K\ {}^1w^j)v^4. \tag{3.13.17}$$

   Now we have the relations

$$(v^i, K\ {}^2w^j) = 0 \text{ for } i = 1 \text{ to } 4 \text{ and } j = 5, 6, \cdots 2n. \tag{3.13.18}$$

9. Define v^5 by the rule

$$v^5 = {}^2w^5. \tag{3.13.19}$$

10. Starting with ${}^2w^6$, search through the ${}^2w^j$ with $j \geq 6$ to find the first j, call it k, with the property

$$(v^5, K\ {}^2w^j) \neq 0. \tag{3.13.20}$$

   Renumber the vectors ${}^2w^6 \cdots {}^2\ w^{2n}$ so that ${}^2w^k$ becomes ${}^2w^6$.

11. Define v^6 by the rule

$$v^6 = {}^2w^6/[(v^5, K\ {}^2w^6)]. \tag{3.13.21}$$

   At this stage we have the results

$$(v^i, Kv^j) = J'_{ij} \text{ for } i, j = 1 \text{ to } 6. \tag{3.13.22}$$

12. Proceed with the obvious extension of the above process to construct $v^7, v^8, \cdots v^{2n-2}$. Then at the last stage we have

$$v^{2n-1} = {}^m w^{2n-1}, \tag{3.13.23}$$

$$v^{2n} = {}^m w^{2n}/[(v^{2n-1}, K\ {}^m w^{2n})], \tag{3.13.24}$$

with

$$m = n - 1. \tag{3.13.25}$$

As was the case with Darboux symplectification, how does one know that the required vectors ${}^m w^k$ described in steps 2, 6, 10, etc. exist? And how does one know that the vectors $v^3, v^5, \cdots v^{2n-1}$ given in steps 5, 9, etc. are nonzero? Again difficulties do not arise because the w^i are assumed to be linearly independent and K is assumed to be invertible. See Exercise 13.1.

Next, let e^j denote the unit column vector with 1 in its jth entry and zeroes elsewhere. See (6.4). Then (13.5) can be written in the form

$$(v^i, Kv^j) = (e^i, J'e^j). \tag{3.13.26}$$

Define a linear transformation L by the rule

$$v^j = Le^j. \tag{3.13.27}$$

It has the matrix elements

$$L_{ij} = (e^i, Le^j) = (e^i, v^j). \tag{3.13.28}$$

Upon inserting (13.27) into (13.5) we find the relation

$$J'_{ij} = (e^i, J'e^j) = (Le^i, KLe^j) = (e^i, L^T KLe^j), \tag{3.13.29}$$

which is equivalent to the matrix relation

$$J' = L^T KL. \tag{3.13.30}$$

We also observe that L is invertible. Indeed, taking the determinant of both sides of (13.30) yields the relation

$$\det J' = (\det L^T)(\det K)(\det L) = (\det K)(\det L)^2, \tag{3.13.31}$$

from which we find the result

$$(\det L)^2 = (\det J')/(\det K) = 1/(\det K) \neq 0 \text{ or } \infty \tag{3.13.32}$$

since K is assumed to be invertible.

We are almost done. Since L is invertible, we may also write (13.30) in the form

$$K = (L^{-1})^T J' L^{-1}. \tag{3.13.33}$$

Now make use of (2.14) to find the result

$$K = (L^{-1})^T PJP^T L^{-1}. \tag{3.13.34}$$

Finally, define N by the rule

$$N = (L^{-1})^T P. \tag{3.13.35}$$

This N is evidently real and in $GL(2n, \mathbb{R})$, and direct calculation shows that it has the desired property

$$NJN^T = [(L^{-1})^T P]J[(L^{-1})^T P]^T = (L^{-1})^T PJP^T L^{-1} = K. \tag{3.13.36}$$

3.13.2 Darboux Transformations

We have found that K and J are congruent under the action of the intertwining transformation N. An intertwining congruency transformation, such as N above, that relates two different antisymmetric matrices is sometimes called a *Darboux* transformation because he was the first to study such transformations in the context of Classical Mechanics, and N can be called a Darboux matrix.[48] Darboux transformations and matrices will be essential for the work of Sections 5.13 and 6.7 and Chapter 34. We note in passing that the relations (2.14), (6.118), and (6.119) are Darboux relations.

Suppose K and $\hat{K}$ are two symplectic 2-form matrices of the same dimension. Then we know that there are Darboux matrices N and $\hat{N}$ that connect them to the J of this same dimension by the relations (13.2) and

$$\hat{K} = \hat{N}J\hat{N}^T. \tag{3.13.37}$$

Upon combining (13.2) and (13.37), we see that

$$\hat{K} = (\hat{N}N^{-1})K(\hat{N}N^{-1})^T. \tag{3.13.38}$$

Thus, $\hat{K}$ and K are connected by the Darboux matrix $(\hat{N}N^{-1})$.

Given K, what can be said about the N that satisfy (13.2)? Suppose N' is another matrix that satisfies (13.2),

$$K = N'J(N')^T. \tag{3.13.39}$$

Combining (13.2) and (13.39) yields the relation

$$NJN^T = N'J(N')^T, \tag{3.13.40}$$

from which we conclude that

$$[N^{-1}N']J[N^{-1}N']^T = J. \tag{3.13.41}$$

Therefore, if we make the definition

$$M = N^{-1}N', \tag{3.13.42}$$

we see that M is a symplectic matrix. Moreover, (13.42) can be rewritten in the form

$$N' = NM. \tag{3.13.43}$$

That is, N' and N are related by multiplication on the right by a symplectic matrix. Finally, suppose that M is any symplectic matrix, and use (13.43) to define N'. Then we find the result

$$N'J(N')^T = NMJM^TN^T = NJN^T = K. \tag{3.13.44}$$

Thus, all Darboux matrices (for any fixed K) are related by multiplication on the right by symplectic matrices, and this symplectic matrix can be any symplectic matrix.[49] It follows

[48] The reader is warned that the words *Darboux transformation* are also employed, with a different meaning, in the context of differential equations.

[49] Note that (13.43) can be rewritten in the form $N = N'M^{-1}$. We know that M^{-1} is symplectic if M is. Thus N and N' are also related by multiplication on the right by a symplectic matrix.

that the dimensionality of the space of Darboux matrices (for any fixed K) is the same as that of $Sp(2n, \mathbb{R})$, namely $n(2n + 1)$.

Matrices N' and N in $GL(2n, \mathbb{R})$ that are related by an equation of the form (13.43) are said to be in the same (left) *coset* of $GL(2n, \mathbb{R})$ relative to the subgroup $Sp(2n, \mathbb{R})$. The collection of these cosets is denoted by the symbols $GL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$. See Section 5.12 for a discussion of cosets. We conclude that the coset space $GL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$ is in one-to-one correspondence with the set of all symplectic 2-forms on $2n$-dimensional space,

$$GL(2n, \mathbb{R})/Sp(2n, \mathbb{R}) \leftrightarrow \{K \mid K \text{ is real, } 2n \times 2n, \text{ antisymmetric, and nonsingular}\}. \tag{3.13.45}$$

Observe, as a sanity check, that the dimension of $GL(2n, \mathbb{R})$ is $(2n)^2$, the dimension of $Sp(2n, \mathbb{R})$ is $n(2n + 1)$, and the dimension of the space of all real $2n \times 2n$ antisymmetric matrices is $(1/2)[(2n)^2 - 2n]$. But there is the relation

$$(1/2)[(2n)^2 - 2n] = (2n)^2 - n(2n + 1), \tag{3.13.46}$$

which verifies that the dimensionality count works out properly. In Section 5.12 we will learn that the set of all symplectic 2-forms on $2n$-dimensional space constitutes a *homogeneous* space under the action of $GL(2n, \mathbb{R})$.

Can we restrict our attention to Darboux matrices N that have unit determinant so that $N \in SL(2n, \mathbb{R})$? Then the 2-form matrices K will also have unit determinant, which we might like. The answer is *no*. Consider the 2×2 case and suppose

$$K = -J. \tag{3.13.47}$$

According to (1.7) there is the relation

$$NJN^T = [\det(N^T)]J = [\det(N)]J. \tag{3.13.48}$$

Thus in this case, for (13.2) and (13.47) to hold, we must have the relation

$$\det N = -1. \tag{3.13.49}$$

Note also that if we instead impose the condition

$$\det N = \pm 1, \tag{3.13.50}$$

then, according to (13.4), we will still have the result

$$\det K = 1. \tag{3.13.51}$$

There is another interesting feature of Darboux transformations. Let us use the representation (13.2) to compute K^2. Doing so gives the result

$$K^2 = NJN^TNJN^T = NJ(N^TN)JN^T \tag{3.13.52}$$

Suppose N is orthogonal. Then we find that

$$K^2 = NJ^2N^T = -NN^T = -I. \tag{3.13.53}$$

Thus symplectic 2-form matrices K that are related to J by orthogonal Darboux transformations are those that are most analogous to J. We have already seen an instance of this fact in the case of the symplectic form matrix J' given by (2.10). Since orthogonal matrices from a group, we see from (13.38) that symplectic-form matrices that are related to J by orthogonal Darboux matrices are also related to each other by orthogonal Darboux matrices.

Suppose, conversely, that

$$K^2 = -I. \tag{3.13.54}$$

Then we find from (13.52) and (13.54) that

$$(N^T N)J(N^T N)^T = J, \tag{3.13.55}$$

from which it follows that M defined by

$$M = N^T N \tag{3.13.56}$$

is a symplectic matrix. Also, we see from (13.56) that M is symmetric and positive definite. Therefore we know, from the work of Section 3.8, that there is a unique symmetric matrix S^a such that

$$N^T N = \exp(JS^a). \tag{3.13.57}$$

Suppose we make a polar decomposition for N^T by writing

$$N^T = PO. \tag{3.13.58}$$

See Section 4.2 for information about polar decomposition. Then we find that

$$N^T N = P^2, \tag{3.13.59}$$

from which we conclude that

$$P = \exp(JS^a/2) \tag{3.13.60}$$

and

$$N^T = \exp(JS^a/2)O. \tag{3.13.61}$$

Taking the transpose of both sides of (13.61) gives the result

$$N = O' \exp(JS^a/2) \tag{3.13.62}$$

where $O' = O^T$ is also an orthogonal matrix. This is the most general form for N when (13.54) holds. All such N belong to cosets $GL(2n,\mathbb{R})/Sp(2n,\mathbb{R})$ that contain an orthogonal matrix. In the case that N is orthogonal, we see that $S^a = 0$.

Finally, suppose we replace O' in (13.62) by O'' where

$$O'' = O' \exp(JS^c). \tag{3.13.63}$$

We know that all matrices of the form $\exp(JS^c)$ are orthogonal and therefore O'' is also orthogonal. They are also symplectic, and form a subgroup H of the symplectic group. Recall the work of Section 3.9. Therefore N' defined by

$$N' = O'' \exp(JS^a/2) = O'[\exp(JS^c)\exp(JS^a/2)] \tag{3.13.64}$$

produces the same K as N does when used in (13.2). We conclude that what matters in (13.62) is the coset $O(2n,\mathbb{R})/H$ to which O' belongs.

3.13.3 Symplectic Forms and Pfaffians

Let A be a $2n \times 2n$ antisymmetric matrix. The *Pfaffian* of A, denoted by Pf(A), is a certain polynomial of degree n in the entries of A (with real coefficients). For our present purposes we need not know all about Pfaffians, but only that the they have certain remarkable properties.

The first of these is that

$$\det A = [\text{Pf}(A)]^2. \tag{3.13.65}$$

From (13.65) we see that any real nonsingular antisymmetric matrix must have a positive determinant. This result can also be proved without the use of Pfaffians. See Exercise 13.2.

Other remarkable Pfaffian properties are given by the relations

$$\text{Pf}(NAN^T) = [\det(N)]\text{Pf}(A), \tag{3.13.66}$$

$$\text{Pf}(\lambda A) = \lambda^n \text{Pf}(A), \tag{3.13.67}$$

$$\text{Pf}(J') = 1. \tag{3.13.68}$$

From (2.14), (13.66), and (13.68) we deduce the relation

$$\text{Pf}(J) = (-1)^{n(n-1)/2}. \tag{3.13.69}$$

As special cases of (13.54) we find the results

$$\text{Pf}(J) = 1, -1, -1 \tag{3.13.70}$$

for $n = 1, 2, 3$, respectively.

Upon employing (13.66) in (13.2) and using (13.69), we find the result

$$\text{Pf}(K) = [\det(N)](-1)^{n(n-1)/2}. \tag{3.13.71}$$

We see that symplectic forms can be classified according to the signs of their Pfaffians. Suppose K and K' are two symplectic forms. Then, from (2.38), we know that they are related by an equation of the form

$$K' = MKM^T \tag{3.13.72}$$

with $M \in GL(2n, \mathbb{R})$. If their Pfaffians have the same sign, then $M \in GL(2n, \mathbb{R}, +)$. Here $GL(2n, \mathbb{R}, +)$ denotes the set of real $2n \times 2n$ matrices with *positive* determinant. Such matrices evidently form a subgroup of $GL(2n, \mathbb{R})$. If the Pfaffians of K and K' have different signs, then $M \in GL(2n, \mathbb{R}, -)$. Here $GL(2n, \mathbb{R}, -)$ denotes the set of real $2n \times 2n$ matrices with *negative* determinant. They evidently do not form a subgroup of $GL(2n, \mathbb{R})$, but rather are in a disconnected piece of $GL(2n, \mathbb{R})$ that does not contain the identity matrix I. Given any element $F \in GL(2n, \mathbb{R}, -)$, all elements of $GL(2n, \mathbb{R}, -)$ can be obtained by multiplying F (either on the left or right) by all elements of $GL(2n, \mathbb{R}, +)$.

3.13.4 Variant Symplectic Groups

Consider the general symplectic 2-form (w, Kz), and suppose that R is a real matrix that preserves this 2-form. Then it follows that R must satisfy the generalized symplectic relation

$$R^T K R = K. \tag{3.13.73}$$

It is easily verified that all such matrices form a group. One might wonder if this group is something new or is merely $Sp(2n, \mathbb{R})$ in disguise. We will see that the latter is true. It follows that the group $Sp(2n, \mathbb{R})$ is as general as might be desired.

Suppose we employ (13.2) in (13.73). Doing so gives the relation

$$R^T N J N^T R = N J N^T \tag{3.13.74}$$

from which it follows that

$$[N^{-1} R^T N] J [N^{-1} R^T N]^T = J. \tag{3.13.75}$$

We conclude that the M now defined by the relation

$$M^T = N^{-1} R^T N \tag{3.13.76}$$

is a symplectic matrix. Upon solving (13.76) for R we find the result

$$R = N^T M (N^T)^{-1}. \tag{3.13.77}$$

Thus, we see that the group of matrices R is related to the group $Sp(2n, \mathbb{R})$ simply by the similarity transformation (13.77).

Exercises

3.13.1. Review Exercise 6.12. Show that the steps 1 through 12 in Section 3.13.1 can always be executed. Alternatively, verify by induction on n that the construction of the desired v^j is always possible.

3.13.2. Verify that any real nonsingular antisymmetric matrix must have a positive determinant. Hint: Use (13.4).

3.13.3. In the 2×2 case verify that using

$$N = B_3 \tag{3.13.78}$$

in (13.2), with B_3 given by (7.61), yields (13.47). Note also that (13.49) holds in this case as it should.

3.13.4. Verify that (13.52) and (13.54) together imply (13.55).

3.13.5. Verify that the matrices R that satisfy (13.73) form a group.

3.13.6. Take the Pfaffian of both sides of (1.10) or (1.2), and use (13.66) and (13.69), to show that symplectic matrices always have determinant +1.

3.13.7. Let N_2 be the matrix

$$N_2 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{3.13.79}$$

Verify that

$$N_2 J_2 N_2^T = -J_2, \tag{3.13.80}$$

and therefore N_2 is a Darboux matrix relating J_2 and $-J_2$. Recall (2.11). Use these results to show that N' defined by

$$N' = \begin{pmatrix} N_2 & & & \\ & N_2 & & \\ & & \ddots & \\ & & & N_2 \end{pmatrix} \tag{3.13.81}$$

has the property

$$N'J'(N')^T = -J'. \tag{3.13.82}$$

Therefore N' is a Darboux matrix relating J' and $-J'$. Define the matrix N by the rule

$$N = P^T N' P. \tag{3.13.83}$$

Verify the relation

$$NJN^T = P^T N' P J P^T (N')^T P = P^T N' J' (N')^T P = -P^T J' P = -J, \tag{3.13.84}$$

which demonstrates that N is a Darboux matrix relating J and $-J$. Verify that

$$\det N = \det(P^T N' P) = \det N' = (-1)^n. \tag{3.13.85}$$

Show that in fact N has the simple block form

$$N = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} \tag{3.13.86}$$

so that (13.84) and (13.85) follow immediately.

3.13.8. Suppose a real $2n \times 2n$ matrix M satisfies the condition

$$M^T J M = -J. \tag{3.13.87}$$

Such a matrix is said to be *antisymplectic*. (In Section 31.1 it will be shown that antisymplectic matrices arise in the study of *reversal* symmetry.) Show that if M is antisymplectic, then

$$\det M = \pm 1, \tag{3.13.88}$$

and therefore M is invertible. Show that if M is antisymplectic, then so are $-M$, M^T, and M^{-1}. Show that the product of two antisymplectic matrices is symplectic, and the product (in either order) of a symplectic and an antisymplectic matrix is antisymplectic. Thus, antisymplectic matrices do not form a group. For example, the identity matrix is symplectic, but not antisymplectic. Show, when taken together, that symplectic and antisymplectic matrices do form a group. This group does not seem to have have a name, but might be called the *complete* symplectic group.

Since, as you have proved, (13.87) implies the relation

$$MJM^T = -J, \tag{3.13.89}$$

it follows that M is a Darboux matrix connecting J and $-J$. Take the Pfaffian of both sides of (13.89) and use (13.67) to conclude that

$$\text{Pf}(MJM^T) = \text{Pf}(-J) = (-1)^n \text{Pf}(J). \tag{3.13.90}$$

But, by (13.66), we also have the relation

$$\text{Pf}(MJM^T) = [\det(M)]\text{Pf}(J). \tag{3.13.91}$$

Upon comparing (13.90) and (13.91) you have shown that, in fact,

$$\det M = (-1)^n. \tag{3.13.92}$$

Let N be the Darboux matrix given by (13.83) or (13.86) so that

$$NJN^T = -J. \tag{3.13.93}$$

Evidently N is antisymplectic. Show that any antisymplectic M can be written in the form

$$M = LN = NL' \tag{3.13.94}$$

where L and L' are symplectic. Show that (13.92) also follows from (13.94) and (13.85). Show that the set of antisymmetric matrices is *connected.* (See Section 5.9.1.) Show that what we have called the complete symplectic group consists of two disconnected pieces in $GL(2n, \mathbb{R})$, each of which itself is connected.

As in (1.7.9), write

$$z = (q_1, \cdots, q_n; p_1, \cdots, p_n). \tag{3.13.95}$$

Define $\bar{z}$ by the rule

$$\bar{z} = Nz. \tag{3.13.96}$$

Show that

$$\bar{z} = (q_1, \cdots, q_n; -p_1, \cdots, -p_n) \tag{3.13.97}$$

so that N leaves the q_j in peace and changes the signs of all the p_j. Verify that this same result holds when the N' given by (13.81) is used, for which z has the form

$$z = (q_1, p_1, q_2, p_2, \cdots, q_n, p_n). \tag{3.13.98}$$

Bibliography

Matrix Theory

[1] R.E. Bellman, *Introduction to Matrix Analysis*, Second Edition, Society for Industrial and Applied Mathematics (1997).

[2] G. Hadley, *Linear Algebra*, Addison-Wesley (1961).

[3] P. Lax, *Linear Algebra and its Applications*, Second Edition, John Wiley (2007).

[4] J.N. Franklin, *Matrix Theory*, Dover (2000).

[5] F.R. Gantmacher, *The Theory of Matrices, Vols. One and Two*, Chelsea (1959).

[6] F.R. Gantmacher and M. G. Krein, *Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems*, AMS Chelsea (2002).

[7] F. Zhang, *Matrix Theory*, Springer Verlag (1999).

[8] C. Cullen, *Matrices and Linear Transformations*, Second Edition, Dover (1990).

[9] M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Dover (1992).

[10] A. E. Fekete, *Real Linear Algebra*, Marcel Dekker (1985).

Linear Stability Analysis and Stability Diagrams

[11] J.E. Howard and R.S. MacKay, “Linear Stability of Symplectic Maps”, *J. Math. Phys.* **28**, 1036 (1987).

[12] J.E. Howard and R.S. MacKay, “Calculation of Linear Stability Boundaries for Equilibria of Hamiltonian Systems”, *Phys. Let.* A **122**, 331 (1987).

[13] J.E. Howard, *Celestial Mechanics and Dynamical Astronomy* **48**, 267 (1990).

Normal Forms

[14] J. Moser, *Comm. Pure and Appl. Math.* **11**, 81 (1958).

[15] A. Weinstein, *Bull. Am. Math. Soc.* **77**, 814 (1971).

[16] J. Williamson, *Am. J. Math.* **58**, 141 (1936); **59**, 599 (1937).

[17] A. Wintner, *Ann. di. Mat.* **13**, 105 (1934).

[18] N. Burgoyne and R. Cushman, *Celes. Mech.* **8**, 435 (1974).

[19] N. Burgoyne and R. Cushman, "Normal forms in linear Hamiltonian systems", published in the *1976 Ames Research Center* (NASA) *conference on geometric control theory*, C. Martin and R. Hermann, eds. Math. Sci. Press (Brookline, Mass., 1977).

[20] A.J. Laub and K. Meyer, *Celes. Mech* **9**, 213 (1974).

[21] R. Abraham and J.E. Marsden, *Foundations of Mechanics*, American Mathematical Society (2008).

[22] V.I. Arnold, *Mathematical Methods of Classical Mechanics*, Second Edition, Springer-Verlag (1989).

[23] M. Moshinsky and P. Winternitz, *J. Math Phys.* **21**, 1667 (1980).

[24] A.J. Dragt *et al.*, *Phys. Rev. A* **45**, 2572 (1992).

[25] S-N. Chow, C. Li, and D. Wang, *Normal Forms and Bifurcation of Planar Vector Fields*, Cambridge University Press (1994).

[26] R. Churchill and M. Kummer, "A Unified Approach to Linear and Nonlinear Normal Forms for Hamiltonian systems", *J. Symbolic Computation* **27**, p. 49, (1999).

[27] J. Murdock, *Normal Forms and Unfoldings for Local Dynamical Systems*, Springer-Verlag (2003).

[28] H. Hofer and E. Zehnder, *Symplectic Invariants and Hamiltonian Dynamics*, Birkhäuser Verlag (1994).

[29] Y. Long, *Index Theory for Symplectic Paths with Applications*, Progress in Mathematics, Vol. 207, Birkhäuser Verlag (2002).

[30] K. Weierstrass, *Mathematische Werke* Band I: 233-246, Band II: 19-44, Nachtrag: 139-148, Berlin (1858).

Krein-Moser Theory and Periodic Linear Systems

[31] M. Krein, "A Generalization of some Investigations on Linear Differential Equations with Periodic Coefficients", *Doklady Akad. Nauk. SSSR N.S.*, Vol. 73, 445-448 (1950).

[32] M. Krein, "On the Application of an Algebraic Proposition in the Theory of Monodromy Matrices", *Uspekhi Math. Nauk.* **6**, 171-177 (1951).

[33] M. Krein, "On the Theory of Entire Matrix-Functions of Exponential Type", *Ukrainian Math. Journal* **3**, 164-173 (1951).

[34] M. Krein, "On Some Maximum and Minimum Problems for Characteristic Numbers and Liapunov Stability Zones", *Prikl. Math. Mekh.* **15**, 323-348 (1951).

[35] M. Krein, "On Criteria for Stability and Boundedness of Solutions of Periodic Canonical Systems, *Prikl. Math. Mekh.* **19**, 641-680 (1955).

[36] M. Krein and G. Lyubarski, "On Analytical Properties of Multipliers of Periodic Canonical Differential Systems of Positive Type, *Izv. Ak. Nauk. SSSR* **26**, 542-572 (1962).

[37] J. Moser, "New Aspects in the Theory of Stability of Hamiltonian Systems", *Communications on Pure and Applied Mathematics*, Vol. XI, 81-114 (1958).

[38] I. M. Gelfand and L. D. Lidskii, "On the structure of stability of linear Hamiltonian systems of differential equations with periodic coefficients", *Uspekhi Mat Nauk* **10**, *AMS Translation* **2** 8 pp. 143-181 (1958).

[39] N. Erugin, *Linear Systems of Ordinary Differential Equations with Periodic and Quasi-Periodic Coefficients*, Academic Press (1966).

[40] V. Yakubovich and V. Starzhinskii, *Linear Differential Equations with Periodic Coefficients*, Vols. 1 and 2, Wiley (1975).

[41] F. M. Arscott, *Periodic Differential Equations: An Introduction to Mathieu, Lamé, and Allied Functions*, MacMillan (1964).

[42] I. Gohberg, P. Lancaster, and L. Rodman, *Matrices and Indefinite Scalar Products*, Birkhäuser Verlag (1983).

[43] I. Ekeland, *Convexity Methods in Hamiltonian Mechanics*, Springer-Verlag (1990).

[44] J. Meiss, *Differential Dynamical Systems*, Section 9.11, SIAM (2007).

[45] E. Forest, *Beam Dynamics, a New Attitude and Framework*, Section 4.5.1, Harwood (1998)

[46] M. Kuwamura and E. Yanagida, "Krein's formula for indefinite multipliers in linear periodic Hamiltonian systems", *J. Differential Equations* **230**, 446-464 (2006).

[47] R. Cordeiro and R. Vieira Martins, "Krein Stability in the Disturbed Two-Body Problem", *Chaos, Resonance, and Collective Dynamical Phenomena in the Solar System*, F. Ferraz-Mello, Ed., pp. 369-374, International Astronomical Union (1992).

[48] T. Bridges and J. Furter, *Singularity Theory and Equivariant Symplectic Maps*, Springer-Verlag (1993).

[49] A. Abbondandolo, *Morse theory for Hamiltonian systems*, Chapman & Hall/CRC (2001).

Vector and Matrix Norms

[50] L. Collatz, *Functional Analysis and Numerical Mathematics*, Section 9, Academic Press (1966).

[51] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series, and Products*, Section 15, Academic Press (1980).

Linear Algebra, Polar Decomposition, Orthogonalization, and Symplectic Bases

[52] N. Jacobson, *Lectures in Abstract Algebra*, Vol. II - *Linear Algebra*, D. Van Nostrand (1953). See Section 10, beginning on page 159, for a discussion of symplectic forms. See page 188 for a description of polar decomposition. See also the book of F.R. Gantmacher cited in reference 5 above.

[53] P.R. Halmos, *Finite-Dimensional Vector Spaces*, D. Van Nostrand (1958).

[54] P.R. Halmos, *Linear Algebra Problem Book*, Mathematical Association of America (1995).

[55] V. Moretti, *Multi-Linear Algebra, Tensors and Spinors in Mathematical Physics*. See the Web site http://www.science.unitn.it/~moretti/tensori.pdf.

[56] N. Higham, *Functions of Matrices: Theory and Computation*, SIAM (2008).

[57] J.B. Keller, “Closest Unitary, Orthogonal and Hermitian Operators to a Given Operator”, *Mathematics Magazine* **48**, p. 192 (1975).

[58] H.C. Schweinler and E.P. Wigner, *J. Math. Phys.* **11**, p. 1693 (1970).

[59] R. Simon, S. Chaturvedi, and V. Srinivasan, *J. Math. Phys.* **40**, p. 3632 (1999).

[60] P. Libermann and C-M. Marle, *Symplectic Geometry and Analytical Mechanics*, D. Reidel (1987).

[61] A. Baker, *Matrix Groups: An Introduction to Lie Group Theory*, Springer (2006). See Section 8.8 for Darboux symplectification.

Exponential Representations and Logarithms

See also the Matrix Theory and Polar Decomposition references at the beginning of this bibliography.

[62] J. Williamson, *Am. J. Math* **61**, 897 (1939).

[63] V. Yakubovich and V. Starzhinskii, *Linear Differential Equations with Periodic Coefficients*, Vols. 1 and 2, Wiley (1975).

[64] Y. Sibuya, “Note on Real Matrices and Linear Dynamical Systems with Periodic Coefficients”, *J. Math. Anal. Appl.* **1**, 363-372 (1960).

[65] K. Meyer, G. Hall, and D. Offin, *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, Second Edition, Springer (2009).

[66] W. Culver, "On the Existence and Uniqueness of the Real Logarithm of a Matrix", *Proceedings of the American Mathematical Society* **17**, p. 1146 (1966).

[67] J. Gallier, "Logarithms and Square Roots of Real Matrices", available on the Web at ScholarlyCommons (2008): `http://repository.upenn.edu/cis_reports/876/`.

[68] N. Higham, *Functions of Matrices: Theory and Computation*, SIAM (2008).

[69] Zhong-Qi Ma, *Group Theory for Physicists*, World Scientific (2007).

[70] A. Dooley and N. Wildberger, "Harmonic analysis and the global exponential map for compact Lie groups", *Functional Analysis and Its Applications* **27**, p. 21 (1993).

Numerical Methods for Symplectic and Hamiltonian Matrices

[71] H. Fassbender, *Symplectic Methods for the Symplectic Eigenproblem*, Kluwer/Plenum (2000).

[72] P. Benner, R. Byers, and E. Barth, "Algorithm 800: Fortran 77 Subroutines for Computing the Eigenvalues of Hamiltonian Matrices I: The Square-Reduced Method", *ACM Transactions on Mathematical Software* **26**, p. 49-77 (2000).

[73] L. Dieci, "Considerations on computing real logarithms of matrices, Hamiltonian logarithms, and skew-symmetric logarithms", *Linear Algebra Appl.* **244**, p. 35-54 (1996).

[74] L. Dieci, "Real Hamiltonian logarithm of a symplectic matrix", *Linear Algebra Appl.* **281**, p. 227-246 (1998).

Lie Algebra/Group Theory

See also the Group/Lie Algebra Theory sections of the Bibliographies for Chapters 5 and 27.

[75] M. Hamermesh, *Group Theory and its Application to Physical Problems*, Addison-Wesley (1962).

[76] K. Tapp, *Matrix Groups for Undergraduates*, American Mathematical Society (2005). For a supplementary 10^{th} chapter, see the Web site `http://people.sju.edu/~ktapp/`.

[77] A.O. Barut and R. Raczka, *Theory of Group Representations and Applications*, World Scientific (1986).

[78] N. Bourbaki, *Lie Groups and Lie Algebras, Elements of Mathematics, Chapters 1-3*, Springer-Verlag (1989).

[79] T. Brocker and T.T. Dieck, *Representations of Compact Lie Groups*, Springer-Verlag (1985).

[80] D. Bump, *Lie Groups*, Springer (2004).

[81] R. Cahn, *Semi-Simple Lie Algebras and their Representations*, Dover (2006).

[82] J-Q. Chen, *Group Representation Theory for Physicists*, World Scientific (1989).

[83] W. Fulton and J. Harris, *Representation Theory, A First Course*, Corrected third printing, Springer-Verlag (1996).

[84] H. Georgi, *Lie Algebras in Particle Physics*, Perseus Books (1999).

[85] R. Goodman and N.R. Wallach, *Representations and Invariants of the Classical Groups*, Cambridge University Press (1998).

[86] R. Goodman and N.R. Wallach, *Symmetry, Representations, and Invariants*, Springer (2009).

[87] J.E. Humphreys, *Introduction to Lie Algebras and Representation Theory*, Springer-Verlag (1972).

[88] N. Jacobson, *Lectures in Abstract Algebra*, Vol. I - *Basic Concepts*, D. Van Nostrand (Princeton, 1951).

[89] N. Jacobson, *Lie Algebras*, Interscience Publishers (1962).

[90] A.W. Knapp, *Lie Groups Beyond an Introduction*, Second Edition, Birkhäuser (2005).

[91] A.W. Knapp, *Representation Theory of Semisimple Groups, An Overview Based on Examples*, Princeton (1986).

[92] C. Procesi, *Lie Groups, An Approach through Invariants and Representations*, Springer (2007).

[93] V.S. Varadarajan, Review of "Lie Groups, An Approach through Invariants and Representations, by C. Procesi", *Bulletin of the American Mathematical Society* **45**, p. 661 (2008).

[94] V.S. Varadarajan, *Lie Groups, Lie Algebras, and Their Representations*, Springer-Verlag (1984).

[95] D.H. Sattinger and O.L. Weaver, *Lie Groups and Algebras with Applications to Physics, Geometry, and Mechanics*, Springer-Verlag (1986).

[96] J.E. Campbell, *Introductory Treatise on Lie's Theory of Finite Continuous Transformation Groups*, Chelsea Publishing (1903 and 1966).

[97] H. Weyl, *The Classical Groups: Their Invariants and Representations*, Princeton University Press (1946).

[98] E.P. Wigner, *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra*, Academic Press (1959).

[99] B.G. Wybourne, *Classical Groups for Physicists*, John Wiley and Sons (1974).

[100] A. Baker, *Matrix Groups: An Introduction to Lie Group Theory*, Springer (2006).

[101] J.G.F. Belinfante and B. Kolman, *A Survey of Lie Groups and Lie Algebras with Applications and Computational Methods*, Society for Industrial and Applied Mathematics (1972).

[102] D. Montgomery and L. Zippin, *Topological Transformation Groups*, Interscience (1955).

[103] P. Tondeur, *Introduction to Lie Groups and Transformation Groups*, Lecture Notes in Mathematics **7**, Springer-Verlag (1965).

[104] J. Stillwell, *Naive Lie Theory*, Springer (2008).

[105] P. Szekeres, *A Course in Modern Mathematical Physics: Groups, Hilbert Space, and Differential Geometry*, Cambridge University Press (2005).

[106] M. Curtis, *Matrix Groups*, Second Edition, Springer (1984).

[107] R. Gilmore, *Lie Groups, Lie Algebras, and Some of Their Applications*, Dover (2006).

[108] Arvind, B. Dutta, N. Mukunda, and R. Simon, "The Real Symplectic Groups in Quantum Mechanics and Optics", arXiv:quant-ph/9509002v3 24 Nov 1995, (2008).

[109] P. Ramond, *Group Theory, A Physicist's Survey*, Cambridge University Press (2010).

[110] W. Miller, *Symmetry Groups and Their Applications*, Academic Press (1972).

[111] J. Gallier, *Geometric Methods and Applications: For Computer Science and Engineering*, Springer-Verlag (2001).

[112] P. Winternitz, Edit., *Group Theory and Numerical Analysis*, American Mathematical Society (2005).

[113] N. Ibragimov, *Transformation Groups Applied to Mathematical Physics*, D. Reidel (1985).

[114] L. P. Eisenhart, *Continuous Groups of Transformations*, Dover (1961).

[115] A. Henderson, *Representations of Lie algebras: An Introduction Through $g\ell(n)$*, Cambridge (2012).

[116] G. Bredon, *Introduction to Compact Transformation Groups*, Academic Press (1972).

[117] N. Ibragimov, *Transformation Groups and Lie Algebras*, World Scientific (2013).

[118] W. Pfeifer, *The Lie Algebras $su(n)$: An Introduction*, Birkhäuser Verlag (2003).

[119] P. Teodorescu and N.-A. Nicorovici, *Applications of the Theory of Groups in Mechanics and Physics*, Kluwer (2004).

[120] R. Carter, G. Segal, and I. Macdonald, *Lectures on Lie Groups and Lie Algebras*, Cambridge University Press (1995).

[121] S. Sternberg, *Lie Algebras*, (2004). See the Web site http://www.math.harvard.edu/~shlomo/docs/lie_algebras.pdf.

[122] P. Cvitanović, *Group Theory: Birdtracks, Lie's, and Exceptional Groups*, Princeton University Press (2008).

[123] J. Arthur, *The Endoscopic Classification of Representations: Orthogonal and Symplectic Groups*, American Mathematical Society (2013).

[124] J. Talman, *Special Functions: A Group Theoretic Approach Based on Lectures by E. Wigner*, Benjamin (1968).

[125] N. Vilenkin, *Special Functions and the Theory of Group Representations*, American Mathematical Society (1968).

[126] F. Iachello, *Lie Algebras and Applications*, 2nd edition, Springer (2015).

[127] A. Zee, *Group Theory in a Nutshell for Physicists*, Princeton University Press (2016).

[128] B. Hall, *Lie Groups, Lie Algebras, and Representations: an Elementary Introduction*, 2nd edition, Springer (2015).

[129] A. Baker, *Matrix Groups: An Introduction to Lie Group Theory*, Springer (2002).

[130] K. Erdmann and M. Wildon, *Inroduction to Lie Algebras*, Springer (2011).

[131] H. Pollatsek, *Lie Groups: A Problem-Oriented Introduction via Matrix Groups*, Mathematical Association of America (2009).

[132] P. Woit, *Quantum Theory, Groups and Representations: An Introduction*, Springer (2017). See also the Web site http://www.math.columbia.edu/~woit/QM/qmbook.pdf.

[133] A. Kirillov Jr., *An Introduction to Lie Groups and Lie Algebras*, Cambridge University Press (2017).

[134] Adam Marsh, *Mathematics for Physics: An Illustrated Handbook*, World Scientific (2018).

[135] Gregory W. Moore, *Applied Group Theory*, http://www.physics.rutgers.edu/~gmoore/618Spring2018/GroupTheory-Spring2018.html (2018).

[136] J. D. Vergados, *Group and Representation Theory*, World Scientific (2017).

[137] S. Garibaldi,"E_8, The Most Exceptional Group", *Bulletin of the American Mathematical Society*, Volume 53, Number 4, (October 2016).

Pfaffians

[138] R. Vein and P. Dale *Determinants and Their Applications in Mathematical Physics*, Springer (1999).

[139] Google “Pfaffian” and look at the *Wikipedia* and *PlanetMath* entries.

Chapter 4

Matrix Exponentiation and Symplectification

Matrix Exponentiation

We have learned in Section 3.8 and elsewhere that we need to compute matrices of the form $\exp(JS)$. That is, we need to *exponentiate* the matrix JS. Sometimes, as will be seen in later chapters, this exponentiation can be done analytically. However in many cases numerical methods are required. When numerical methods are used, it is desirable that these methods be fast and accurate. No completely satisfactory method is known for this purpose, but one of the better methods available will be described in Section 4.1. This description begins with the problem of computing the ordinary exponential function, and then moves on to the computation of the matrix exponential function.

Matrix Symplectification

When numerical methods are used for evaluating $\exp(JS)$, it is often desirable that the result be *symplectic* to machine precision even if the result is not *accurate* to machine precision. One approach is to employ some procedure that takes a matrix that is nearly symplectic and produces a *nearby* matrix that is exactly symplectic. We will refer to such a procedure as *matrix symplectification*. There are several circumstances in which matrix symplectification may be useful. Four come to mind:

First, as just described, numerical exponentiation of JS may lead to a result that is not as symplectic as desired. Second, suppose that over the course of a numerical calculation we have multiplied together several symplectic matrices. For example, we will learn in Chapter 8 [see (8.4.20)] that such multiplication is required if we wish to *concatenate* a large number of maps. Then the net matrix result may not be exactly symplectic due to round-off error. Although we cannot recover an exact result, we can at least produce a result that is exactly symplectic (to machine precision) and also near the exact result.

Third, we will see in Section 9.3 that in the treatment of translations it is necessary to evaluate linear transformations of the form $\exp(: k_2 :)$ where k_2 arises solely from nonlinear feed-down effects. Since all calculations are carried out within the quotient algebra L^0/L^ℓ, k_2 in this case is only known up to some order in the size of the translation, and it seems pointless to evaluate $\exp(:k_2:)$ to any order higher than what is known for k_2. However, we may very well wish to have a result that is exactly symplectic. The meaning of the new notation just employed and the concepts alluded to will becomes evident in Chapter 9. Suffice it to say that there are cases where JS, while being exactly Hamiltonian, is yet only known approximately. In these cases we are content with a correspondingly approximate (but, we hope, rapidly computable) result for $\exp(JS)$ which is, nevertheless, exactly symplectic.

Fourth, there are occasions in which we may wish to factor a map into symplectic and nonsymplectic parts. See Section 29.1. The first step in this process is to factor, in some standard way, a matrix into symplectic and nonsymplectic parts.

Section 4.2 provides an initial background by describing the completely understood subject of orthogonal polar decomposition. Then Sections 4.3 and 4.4 provide a theoretical background for the more complicated subject of matrix symplectification and symplectic polar decomposition. They also give information concerning how the symplectic group lies within the general linear group. This information is useful when one considers non-Hamiltonian perturbations of Hamiltonian dynamics. Again see Section 29.1. Finally, Sections 4.5 through 4.8 describe four known methods for matrix symplectification.

4.1 Exponentiation by Scaling and Squaring

4.1.1 The Ordinary Exponential Function

The ordinary exponential function $\exp(z)$, where z is a complex variable, is defined by the Taylor series

$$\exp(z) = \sum_{\ell=0}^{\infty} z^{\ell}/\ell!. \tag{4.1.1}$$

This series converges everywhere, but is useful for computation only for small z. Consider computing, for example, $\exp(20)$. For $z = 20$, we find the numerical result

$$(20)^{60}/60! = 1.4 \times 10^{-4}. \tag{4.1.2}$$

Consequently, when $z = 20$, at least 60 terms must be retained in (1.1) to even begin to get convergence. And at this stage the convergence is still quite slow since the ratio of successive terms is only about

$$(20)/(60) = 1/3. \tag{4.1.3}$$

Finally, if we want to compute $\exp(-20)$ using the Taylor series, we would have to use very high precision arithmetic to take into account the high degree of cancellation that in this case must occur between very large terms.

There is a better way to compute the exponential function based on the observation that it satisfies the functional *scaling* equation

$$\exp(z) = [\exp(z/m)]^m. \tag{4.1.4}$$

Suppose we set m to an integer power of 2,

$$m = 2^n. \tag{4.1.5}$$

Then the right side of (1.4) can be calculated by n successive *squarings*,

$$\exp(z) = \{\exp[z/(2^n)]\}^{2^n} = \{\cdots\{\{\exp[z/(2^n)]\}^2\}^2\cdots\}^2 \text{ (n squarings)}. \tag{4.1.6}$$

Next we observe that if $[z/(2^n)]$ is small enough, the quantity $\exp[z/(2^n)]$ can be computed to good accuracy using a Taylor series truncated at relatively low order,

$$\exp[z/(2^n)] \sim \sum_0^N [z/(2^n)]^\ell/\ell!. \tag{4.1.7}$$

Let tNexp(z) denote the *truncated* exponential function defined by the relation

$$\text{tNexp}(z) = \sum_0^N z^\ell/\ell!. \tag{4.1.8}$$

Suppose we define *my* exponential function by the rule

$$\text{myexp}(z) = \{\cdots\{\{\text{tNexp}[z/(2^n)]\}^2\}^2\cdots\}^2 \text{ (n squarings)}. \tag{4.1.9}$$

Then we might hope that for a suitable value of n (which depends on z) we would have to good accuracy the relation

$$\exp(z) \sim \text{myexp}(z). \tag{4.1.10}$$

In fact, using Taylor's formula with remainder, we find the result

$$\text{myexp}(z) = \exp(z) - \exp(z) z [z/(2^n)]^N/(N+1)! + h.o.t. \tag{4.1.11}$$

where "*h.o.t.*" denotes still higher order error terms. Thus, the magnitude of the *relative* estimated error is given by the relation

$$\text{estimated error } \sim |z|[|z|/(2^n)]^N/(N+1)!. \tag{4.1.12}$$

We see that to achieve good accuracy what we must do is make N sufficiently large and $[z/(2^n)]$ sufficiently small that the error term above is small. Given moderate values of z, this can be done with quite small values of N and n. We also observe that the required value of n only grows as $\log(|z|)$, and that for a given $|z|$ and modest N the accuracy increases very rapidly with increasing n. The tables below show results for $N = 6$ and 9, $-20 < z < 20$, and n selected so that $|[z/(2^n)]| < (1/10)$. The error is also shown, and is consistent with the estimates (1.11) and (1.12). Note that (with $N = 9$) at most 16 $(9 - 1 + 8 = 16)$ multiplications are required to achieve full (64 bit) machine precision. Indeed, the errors listed in Table 1.2 fluctuate in sign, and are mostly the result of working with only 64 bit arithmetic.

Table 4.1.1: $N = 6$; scaling n values chosen to make $|[z/(2^n)]| < (1/10)$.

z	n	myexp(z)	error	relative error
-20	8	0.206E-08	0.199E-17	0.966E-09
-19	8	0.560E-08	0.377E-17	0.672E-09
-18	8	0.152E-07	0.699E-17	0.459E-09
-17	8	0.414E-07	0.127E-16	0.307E-09
-16	8	0.113E-06	0.225E-16	0.200E-09
-15	8	0.306E-06	0.388E-16	0.127E-09
-14	8	0.832E-06	0.648E-16	0.779E-10
-13	8	0.226E-05	0.105E-15	0.463E-10
-12	7	0.614E-05	0.108E-13	0.175E-08
-11	7	0.167E-04	0.158E-13	0.948E-09
-10	7	0.454E-04	0.219E-13	0.483E-09
-9	7	0.123E-03	0.283E-13	0.229E-09
-8	7	0.335E-03	0.335E-13	0.999E-10
-7	7	0.912E-03	0.355E-13	0.390E-10
-6	6	0.248E-02	0.217E-11	0.877E-09
-5	6	0.674E-02	0.163E-11	0.242E-09
-4	6	0.183E-01	0.915E-12	0.500E-10
-3	5	0.498E-01	0.218E-10	0.439E-09
-2	5	0.135E+00	0.338E-11	0.250E-10
-1	4	0.368E+00	0.460E-11	0.125E-10
0	0	0.100E+01	0.000E+00	0.000E+00
1	4	0.272E+01	-0.304E-10	-0.112E-10
2	5	0.739E+01	-0.165E-09	-0.224E-10
3	5	0.201E+02	-0.748E-08	-0.372E-09
4	6	0.546E+02	-0.245E-08	-0.448E-10
5	6	0.148E+03	-0.313E-07	-0.211E-09
6	6	0.403E+03	-0.300E-06	-0.745E-09
7	7	0.110E+04	-0.388E-07	-0.354E-10
8	7	0.298E+04	-0.267E-06	-0.896E-10
9	7	0.810E+04	-0.164E-05	-0.203E-09

Table 4.1.1 continued

z	n	myexp(z)	error	relative error
10	7	0.220E+05	-0.928E-05	-0.421E-09
11	7	0.599E+05	-0.488E-04	-0.815E-09
12	7	0.163E+06	-0.242E-03	-0.149E-08
13	8	0.442E+06	-0.187E-04	-0.423E-10
14	8	0.120E+07	-0.852E-04	-0.708E-10
15	8	0.327E+07	-0.374E-03	-0.114E-09
16	8	0.889E+07	-0.159E-02	-0.179E-09
17	8	0.242E+08	-0.659E-02	-0.273E-09
18	8	0.657E+08	-0.266E-01	-0.406E-09
19	8	0.178E+09	-0.105E+00	-0.590E-09
20	8	0.485E+09	-0.409E+00	-0.843E-09

Table 4.1.2: $N = 9$; scaling n values chosen to make $|[z/(2^n)]| < (1/10)$.

z	n	myexp(z)	error	relative error
-20	8	0.206E-08	-0.773E-22	-0.375E-13
-19	8	0.560E-08	0.108E-21	0.193E-13
-18	8	0.152E-07	-0.586E-21	-0.385E-13
-17	8	0.414E-07	-0.242E-20	-0.585E-13
-16	8	0.113E-06	0.132E-20	0.118E-13
-15	8	0.306E-06	-0.116E-20	-0.381E-14
-14	8	0.832E-06	-0.392E-20	-0.471E-14
-13	8	0.226E-05	0.775E-19	0.343E-13
-12	7	0.614E-05	0.110E-19	0.179E-14
-11	7	0.167E-04	0.146E-18	0.872E-14
-10	7	0.454E-04	-0.854E-18	-0.188E-13
-9	7	0.123E-03	-0.239E-17	-0.193E-13
-8	7	0.335E-03	0.195E-17	0.582E-14
-7	7	0.912E-03	-0.217E-17	-0.238E-14
-6	6	0.248E-02	0.217E-17	0.875E-15
-5	6	0.674E-02	-0.633E-16	-0.940E-14
-4	6	0.183E-01	0.555E-16	0.303E-14
-3	5	0.498E-01	0.208E-16	0.418E-15
-2	5	0.135E+00	0.194E-15	0.144E-14
-1	4	0.368E+00	0.278E-15	0.754E-15
0	0	0.100E+01	0.000E+00	0.000E+00

Table 4.1.2 continued

z	n	myexp(z)	error	relative error
1	4	0.272E+01	0.488E-14	0.180E-14
2	5	0.739E+01	0.258E-13	0.349E-14
3	5	0.201E+02	0.355E-13	0.177E-14
4	6	0.546E+02	0.384E-12	0.703E-14
5	6	0.148E+03	-0.568E-13	-0.383E-15
6	6	0.403E+03	0.142E-11	0.352E-14
7	7	0.110E+04	-0.432E-11	-0.394E-14
8	7	0.298E+04	0.414E-10	0.139E-13
9	7	0.810E+04	-0.236E-10	-0.292E-14
10	7	0.220E+05	-0.182E-10	-0.826E-15
11	7	0.599E+05	0.800E-10	0.134E-14
12	7	0.163E+06	0.114E-08	0.697E-14
13	8	0.442E+06	0.827E-08	0.187E-13
14	8	0.120E+07	-0.978E-08	-0.813E-14
15	8	0.327E+07	0.118E-06	0.362E-13
16	8	0.889E+07	0.248E-06	0.279E-13
17	8	0.242E+08	-0.110E-05	-0.455E-13
18	8	0.657E+08	-0.380E-06	-0.579E-14
19	8	0.178E+09	0.149E-04	0.833E-13
20	8	0.485E+09	-0.715E-06	-0.147E-14

4.1.2 The Matrix Exponential Function

So far we have been discussing the ordinary exponential function. The matrix exponential function (3.7.1) has similar properties. Again its Taylor series may be only very slowly convergent, and again scaling and squaring can be used to good advantage. Let s be a parameter and Z any $m \times m$ matrix. Consider the matrix function $F(s)$ defined by the equation

$$F(s) = \exp(-sZ). \tag{4.1.13}$$

The function F satisfies the relations

$$F(0) = I, \tag{4.1.14}$$

$$F(1) = \exp(-Z), \tag{4.1.15}$$

$$(d/ds)F(s) = -Z\exp(-sZ). \tag{4.1.16}$$

See Exercise 3.7.1. Integrate both sides of (1.16) to get the result

$$\int_0^1 (d/ds)F(s)ds = F(s)|_0^1 = \exp(-Z) - I. \tag{4.1.17}$$

By combining (1.16) and (1.17) we find the integral formula

$$\exp(-Z) - I = -Z \int_0^1 \exp(-sZ)ds. \tag{4.1.18}$$

Now multiply both sites of (1.18) by $\exp(Z)$ to get the result

$$\exp(Z) = I + Z\exp(Z) \int_0^1 \exp(-sZ)ds. \tag{4.1.19}$$

Integration by parts yields the general formula

$$\int_0^1 \exp(-sZ)s^n ds = [1/(n+1)]\exp(-Z) + [1/(n+1)]Z \int_0^1 \exp(-sZ)s^{n+1}ds. \tag{4.1.20}$$

Now use (1.20) repeatedly in (1.19) to get the truncated Taylor series with remainder result

$$\exp(Z) = \sum_{\ell=0}^{N} Z^\ell/\ell! + (Z^{N+1}/N!)\exp(Z) \int_0^1 \exp(-sZ)s^N ds. \tag{4.1.21}$$

As before, we define a truncated exponential function by the formula

$$\text{tNexp}(Z) = \sum_{\ell=0}^{N} Z^\ell/\ell!. \tag{4.1.22}$$

Then from (1.21) and (1.22) we get the result

$$\text{tNexp}(Z) = \exp(Z)[I - (Z^{N+1}/N!) \int_0^1 \exp(-sZ)s^N ds]. \tag{4.1.23}$$

In analogy to (1.9) we define myexp(Z) by the rule

$$\text{myexp}(Z) = \{\text{tNexp}[Z/(2^n)]\}^{2^n} = \{\cdots\{\{\text{tNexp}[z/(2^n)]\}^2\}^2\cdots\}^2 \text{ (n squarings)}. \tag{4.1.24}$$

Now scale and square both sides of (1.23). Upon combining (1.23) and (1.24) we find the final result

$$\text{myexp}(Z) = \exp(Z)\{I - (1/N!)[Z/(2^n)]^{N+1} \int_0^1 \exp[-sZ/(2^n)]s^N ds\}^{2^n}. \tag{4.1.25}$$

Suppose we decide to make the approximation

$$\exp(Z) \sim \text{myexp}(Z). \tag{4.1.26}$$

It is easily checked that the *relative* error made in doing so has an estimated *norm* given by the relation

$$\text{estimated error } = \parallel 2^n\{(1/N!)[Z/(2^n)]^{N+1} \int_0^1 \exp[-sZ/(2^n)]s^N ds\} \parallel . \tag{4.1.27}$$

By using the properties (3.7.10) through (3.7.13) for a norm the expression (1.27) can be simplified to the form

$$\text{estimated error } \sim \{[1/(N+1)!] \parallel Z \parallel \parallel Z/(2^n) \parallel^N \exp[u \parallel Z/(2^n) \parallel]\}, \qquad (4.1.28)$$

where u is a number in the range $0 < u < 1$.

For purposes of illustration, suppose we set $N = 10$ and select n so that $\parallel Z/(2^n) \parallel < (1/20)$. Then we find the estimates

$$\exp[u \parallel Z/(2^n) \parallel] < \exp(1/20) \sim 1.05, \qquad (4.1.29)$$

$$\parallel Z/(2^n) \parallel^N < (1/20)^{10} \sim 9.8 \times 10^{-14}, \qquad (4.1.30)$$

$$1/(N+1)! = 1/(11!) \sim 2.5 \times 10^{-8}. \qquad (4.1.31)$$

Correspondingly, the error estimate becomes

$$\text{estimated error } \sim (2.6 \times 10^{-21}) \parallel Z \parallel, \qquad (4.1.32)$$

which for reasonable values of $\parallel Z \parallel$ is well below round-off error for 64 bit arithmetic. We conclude that the error committed in using (1.26) can made quite small by using modest values for N and n. Consequently, the computation of $\exp(Z)$ by scaling and squaring can be both very fast and very accurate. See Exercise 1.2.

We close this section by remarking that there are alternatives to using the truncated exponential series (1.22) to evaluate the exponential of the scaled exponent. These alternatives, which include the use of Padé approximants, give even better numerical performance at the expense of more elaborate programming. For further detail, see the references at the end of this chapter.

Exercises

4.1.1. Verify (1.6). Verify (1.17) through (1.25). Verify (1.27) through (1.32).

4.1.2. Suppose n is selected so that

$$\parallel Z/(2^n) \parallel = \parallel Z \parallel /(2^n) < (1/20). \qquad (4.1.33)$$

Verify that n grows with increasing $\parallel Z \parallel$ like

$$n \sim [\log(20)]/\log(2) + [\log(\parallel Z \parallel)]/\log(2). \qquad (4.1.34)$$

Consequently, for reasonable values of $\parallel Z \parallel$, the number of squarings required to evaluate (1.24) is quite modest, and the computation of $\exp(Z)$ by scaling and squaring is both accurate and remarkably fast. Show that for a given $\parallel Z \parallel$ and N, the relative error (1.28) decreases *exponentially* with increasing n.

4.2 (Orthogonal) Polar Decomposition

4.2.1 Real Matrix Case

Consider the set of all *real* $n \times n$ matrices. This set obviously forms a Lie algebra, with the commutator as a Lie product, and this Lie algebra is $g\ell(n, \mathbb{R})$. Any matrix B in $g\ell(n, \mathbb{R})$ can be written in the form

$$B = S + A, \tag{4.2.1}$$

where S is (real) symmetric and A is (real) antisymmetric, and both are unique. Next we observe that the antisymmetric matrices form a Lie subalgebra by themselves,

$$\{A, A'\} = A'', \tag{4.2.2}$$

and this Lie algebra is $so(n, \mathbb{R})$. Finally, we observe that the remaining commutation rules for $g\ell(n, \mathbb{R})$ can be written in the form

$$\{A, S\} = S', \tag{4.2.3}$$

$$\{S, S'\} = A. \tag{4.2.4}$$

That is, the commutator of an antisymmetric and a symmetric matrix is a symmetric matrix, and the commutator of two symmetric matrices is an antisymmetric matrix.

If M is a matrix in $GL(n, \mathbb{R})$ sufficiently near the identity, it can be written in the exponential form

$$M = \exp(B). \tag{4.2.5}$$

See Section 3.7. Correspondingly, it can also be written in the form

$$M = \exp(S') \exp(A') \tag{4.2.6}$$

where S' is symmetric and A' is antisymmetric. Indeed, near the identity in $GL(n, \mathbb{R})$, which corresponds to being near the origin in $g\ell(n, \mathbb{R})$, one can in principle pass back and forth between the representations (2.5) and (2.6) by means of the BCH formula and an appropriate *Zassenhaus* formula. See Sections 3.7 and 8.8.

We observe that matrices of the form $\exp(S)$ are positive-definite symmetric, and matrices of the form $\exp(A)$ are orthogonal. See Exercise 2.2. Thus, any M sufficiently near the identity has the polar decomposition

$$M = PO \tag{4.2.7}$$

where P is positive-definite symmetric and O is orthogonal. To be more precise, we might call (2.7) an *orthogonal* polar decomposition to emphasize that the second factor in (2.7) is orthogonal.

So far we have examined matrices near the identity. In fact, the decomposition (2.7) can be made globally and is unique. It is easy to check that the matrix (MM^T) is positive symmetric, and consequently has a unique positive symmetric square root. See Exercise 2.3. Let us therefore define P by the rule

$$P = (MM^T)^{1/2}, \tag{4.2.8}$$

with the corresponding result

$$P^2 = MM^T. \tag{4.2.9}$$

Next assume M is invertible, in which case P is also invertible. Define a matrix O by the rule

$$O = P^{-1}M. \tag{4.2.10}$$

Calculation reveals that O is orthogonal,

$$\begin{aligned} OO^T &= (P^{-1}M)(P^{-1}M)^T = P^{-1}MM^T(P^{-1})^T \\ &= P^{-1}P^2(P^T)^{-1} = P^{-1}P^2P^{-1} = I. \end{aligned} \tag{4.2.11}$$

Thus (2.10) is equivalent to (2.7).

It can be shown that if M is invertible, then both P and O are unique, and P is positive-definite symmetric. See also Exercise 2.3. Moreover, the decomposition (2.7) is still possible if M is not invertible, and P in this case is still unique. However, O is no longer uniquely defined.

We close this section by noting that there is another way of looking at the decomposition (2.7) that deserves emphasis. We have been dealing with the group $GL(n, \mathbb{R})$ and its subgroup $O(n, \mathbb{R})$. Form the coset space $GL(n, \mathbb{R})/O(n, \mathbb{R})$ consisting of the left cosets of $GL(n, \mathbb{R})$ with respect to $O(n, \mathbb{R})$. See Section 5.12 for a detailed description of cosets. Equation (2.7) indicates that the elements of this coset space can be labeled by positive-definite symmetric matrices P. Moreover, any positive-definite symmetric matrix P can be written in the form

$$P = \exp(S) \tag{4.2.12}$$

where S is symmetric, and conversely. Symmetric matrices that are $n \times n$ form, in turn, a linear vector space whose dimension m is given by the relation

$$m = \dim(S) = (1/2)n(n + 1). \tag{4.2.13}$$

It follows that matrices P of the form (2.12) have the topology of E^m, m-dimensional Euclidean space, with m given by (2.13). Correspondingly, $GL(n, \mathbb{R})$ has the topology of $E^m \times O(n, \mathbb{R})$.

4.2.2 Complex Matrix Case

Finally we remark that there is an analogous result for *complex* matrices. In that case a factorization of the form (2.7) still holds but now M is complex, P is Hermitian and positive definite, and O is unitary. This result is also called a polar decomposition. See Exercise 2.5.

Exercises

4.2.1. The purpose of this exercise is to verify that the decomposition (2.1) is unique. To begin, define matrices S and A by the explicit formulas

$$S = (1/2)(B + B^T), \tag{4.2.14}$$

$$A = (1/2)(B - B^T). \tag{4.2.15}$$

Verify that S is symmetric, A is antisymmetric, and (2.1) is satisfied. Next assume that there are symmetric and antisymmetric matrices S' and A' such that

$$B = S' + A'. \tag{4.2.16}$$

Verify that there must be the relations

$$S' = S, \tag{4.2.17}$$

$$A' = A. \tag{4.2.18}$$

4.2.2. Verify the commutation rules (2.2) through (2.4).

4.2.3. This exercise examines some of the properties of $\exp(A)$ and $\exp(S)$ where A and S are arbitrary real antisymmetric and symmetric matrices, respectively.

a) Define a matrix O by the rule

$$O = \exp(A). \tag{4.2.19}$$

Show that

$$O^T = \exp(A^T) = \exp(-A), \tag{4.2.20}$$

and therefore

$$O^T O = O O^T = I. \tag{4.2.21}$$

Show, using (3.7.129), that

$$\det O = 1, \tag{4.2.22}$$

and therefore $O \in SO(n, \mathbb{R})$ if A is $n \times n$.

b) Define a matrix P by the rule

$$P = \exp(S). \tag{4.2.23}$$

Show that P is real and symmetric. Use (3.7.129) to prove that P is nonsingular. Since S is real and symmetric, show that there is a real orthogonal matrix O such that

$$S = ODO^{-1} \tag{4.2.24}$$

where D is diagonal and real. Show that $\exp(D)$ is diagonal, real, and positive definite. Show that

$$P = \exp(S) = O\exp(D)O^{-1} = O\exp(D)O^T. \tag{4.2.25}$$

Show, based on the representation (2.25), that P is positive definite.

There is also a more direct proof of this fact that does not involve matrix diagonalization. Show that the matrix $P^{1/2}$ defined by

$$P^{1/2} = \exp(S/2) \tag{4.2.26}$$

is real, symmetric, and nonsingular, and has the property

$$P^{1/2}P^{1/2} = P. \tag{4.2.27}$$

As a result, show that for any vector v there is the relation

$$(v, Pv) = (v, P^{1/2}P^{1/2}v) = (P^{1/2}v, P^{1/2}v) = ||P^{1/2}v||^2 \geq 0. \tag{4.2.28}$$

Finally, demonstrate that $(v, Pv) = 0$ implies that $v = 0$.

c) What about the converse? Suppose P is a real positive-definite symmetric matrix. Show that there is a real orthogonal matrix O such that

$$P = ODO^{-1} \tag{4.2.29}$$

where D is diagonal and has all positive entries on the diagonal. Define a symmetric matrix S' by the rule

$$S' = \log D \tag{4.2.30}$$

where $\log D$ is defined to be the diagonal matrix whose diagonal entries are the logarithms of the corresponding diagonal entries in D. Verify that, by this definition, S' is real and symmetric and has the feature

$$D = \exp(S'). \tag{4.2.31}$$

Define the matrix S by the rule

$$S = OS'O^{-1} = OS'O^T. \tag{4.2.32}$$

Verify that S is real and symmetric and has the property

$$P = \exp(S). \tag{4.2.33}$$

d) There is more that can be said about the relation between real symmetric matrices S and real positive-definite symmetric matrices P. According to (2.23), P is a real analytic function of S. That is, each entry (matrix element) of P is an analytic function of the various entries in S, and each entry is real when the entries in S are real. (See Section 35.2 for a discussion of analyticity in several complex variables.) This result follows because all powers of S are analytic functions of S and the exponential series converges in norm for all S.

We will see the converse is also true. Namely, S is a real analytic function of P. This fact is not obvious from the work so far because the construction of S as given by (2.32) involved O and D, and their analytic properties are not evident. Indeed, as described in Section 26.13, the eigenvalues of a matrix M (and eigenvalues of P are involved in the construction of both O and D) need not be analytic functions of the entries in M. What is required is an alternate procedure for constructing S in terms of P that is manifestly analytic. Begin with the matrix Q defined in terms of P by the rule

$$Q = [1/\text{tr}(P)]P. \tag{4.2.34}$$

Evidently $[1/\mathrm{tr}(P)]$ is a real analytic function of P as long as $\mathrm{tr}(P) \neq 0$, and correspondingly Q is also an analytic function of P under this same proviso. Moreover, since P is positive definite, all its eigenvalues λ_j are real and positive, and therefore

$$\mathrm{tr}(P) = (\sum \lambda_j) > 0. \tag{4.2.35}$$

The eigenvalues of Q, call them μ_j, are given by the relation

$$\mu_j = \lambda_j / (\sum \lambda_i), \tag{4.2.36}$$

and therefore are real and satisfy the conditions $0 < \mu_j < 1$. Next consider the matrix $I - Q$. Its eigenvalues, call them ν_j, satisfy the conditions $0 < \nu_j < 1$. It therefore follows that

$$||I - Q|| < 1 \tag{4.2.37}$$

when the spectral norm is used. See Section 3.7.1. Moreover, suppose P is not exactly real, symmetric, and positive definite, but is in some sufficiently small and possibly complex neighborhood of such a matrix. Show that (2.37) will continue to hold for such a P. [Use the norm property (3.7.12) and the fact that although the eigenvalues of M need not be analytic functions of M, they are continuous functions of M.] Now define S in terms of P by the rule

$$S = \{\log[\mathrm{tr}(P)]\}I - \sum_{k=1}^{\infty}(1/k)(I - Q)^k = \{\log[\mathrm{tr}(P)]\}I + \log(Q). \tag{4.2.38}$$

Evidently the infinite sum in (2.38) converges in norm because of (2.37), and therefore is an analytic function of P as long as P is in a sufficiently small neighborhood of a real positive-definite symmetric matrix. Also, the first term in (2.38) is an analytic function of P under the same proviso. Therefore S is an analytic function of P. Moreover, S is manifestly real and symmetric when P is real, positive definite, and symmetric. Finally, we find that

$$\begin{aligned} \exp(S) &= \exp\{\{\log[\mathrm{tr}(P)]\}I + \log(Q)\} \\ &= I\exp\{\log[\mathrm{tr}(P)]\}\exp[\log(Q)] = [\mathrm{tr}(P)]Q = P. \end{aligned} \tag{4.2.39}$$

4.2.4. This exercise further explores polar decomposition. Let M be any real matrix. Show that the matrix Q defined by

$$Q = MM^T \tag{4.2.40}$$

is positive symmetric. Show that it has a positive symmetric square root. Hint: Since Q is symmetric, show that there exists a real orthogonal matrix O' that *diagonalizes* it,

$$O'Q(O')^{-1} = D. \tag{4.2.41}$$

Show that the entries in D are real and positive or zero. Define $D^{1/2}$ to be a diagonal matrix with entries equal to the positive square roots of the corresponding entries in D. Now construct P by the rule

$$P = (O')^{-1}D^{1/2}O'. \tag{4.2.42}$$

Show that P is positive symmetric and satisfies

$$P^2 = Q. \tag{4.2.43}$$

Show that if M is invertible, then so is P, and P is then positive-definite symmetric.

A bit more can be said in the way of analyticity if M is invertible. Review Exercise 2.3. From the definition (2.40) show that Q is real, symmetric, and positive definite if M is real and invertible. Verify that Q is analytic in M. From part d of Exercise 2.3 we know there is a real symmetric matrix S such that

$$Q = \exp(S) \tag{4.2.44}$$

and S is analytic in Q, and hence also in M. Now define P by the rule

$$P = \exp(S/2). \tag{4.2.45}$$

Evidently P is real, symmetric, and positive definite. Verify that it also satisfies (2.43). Also, we see that P is analytic in Q, and hence also in M. Finally, if we make a polar decomposition of M, we see from (2.10) that O is also analytic in M. This follows because P^{-1} is analytic in M if P is.

4.2.5. The purpose of this exercise is to define and study polar decomposition for complex matrices. Review the work of Subsection 4.2.1. We will follow an analogous path in the complex case.

Consider the set of all possibly complex $n \times n$ matrices. This set obviously forms a Lie algebra, with the commutator as a Lie product, and this Lie algebra is $g\ell(n, \mathbb{C})$. Verify that any matrix B in $g\ell(n, \mathbb{C})$ can be written uniquely in the form

$$B = H + A \tag{4.2.46}$$

where

$$H = (B + B^\dagger)/2 \tag{4.2.47}$$

and

$$A = (B - B^\dagger)/2. \tag{4.2.48}$$

Show that H is *Hermitian*,

$$H^\dagger = H, \tag{4.2.49}$$

and A is *anti-Hermitian*,

$$A^\dagger = -A. \tag{4.2.50}$$

Next we observe that anti-Hermitian matrices form a Lie subalgebra by themselves,

$$\{A, A'\} = A'', \tag{4.2.51}$$

and this Lie algebra is $u(n)$. Finally, we observe that the remaining commutation rules for $g\ell(n, \mathbb{C})$ can be written in the form

$$\{A, H\} = H', \tag{4.2.52}$$

$$\{H, H'\} = A. \tag{4.2.53}$$

That is, the commutator of an anti-Hermitian and a Hermitian matrix is a Hermitian matrix, and the commutator of two Hermitian matrices is an anti-Hermitian matrix. Verify these claims.

If M is a matrix in $GL(n, \mathbb{R})$ sufficiently near the identity, it can be written in the exponential form

$$M = \exp(B). \tag{4.2.54}$$

See Section 3.7. Correspondingly, it can also be written in the form

$$M = \exp(H') \exp(A') \tag{4.2.55}$$

where H' is Hermitian and A' is anti-Hermitian. Indeed, near the identity in $GL(n, \mathbb{C})$, which corresponds to being near the origin in $g\ell(n, \mathbb{C})$, one can in principle pass back and forth between the representations (2.54) and (2.55) by means of the BCH formula and an appropriate *Zassenhaus* formula. See Sections 3.7 and 8.8.

Verify that matrices of the form $\exp(H)$ are positive-definite Hermitian (see Exercise 3.7.44), and matrices of the form $\exp(A)$ are unitary. Thus, any M sufficiently near the identity has the polar decomposition

$$M = PU \tag{4.2.56}$$

where $P = \exp(H')$ is positive-definite Hermitian and $U = \exp(A')$ is unitary. To be more precise, we might call (2.56) a *unitary* polar decomposition to emphasize that the second factor in (2.56) is unitary.

So far we have examined matrices near the identity. In fact, the decomposition (2.56) can be made globally and is unique. Assume, for a moment, the correctness of (2.56). Verify that then there is the relation

$$MM^\dagger = PUU^\dagger P^\dagger = PP^\dagger = P^2. \tag{4.2.57}$$

Verify that the matrix $(MM^\dagger)$ is positive Hermitian, and consequently has a unique positive Hermitian square root. Let us therefore *define* P by the rule

$$P = (MM^\dagger)^{1/2}, \tag{4.2.58}$$

with the corresponding result

$$P^2 = MM^\dagger. \tag{4.2.59}$$

Next assume M is invertible, in which case prove that P is also invertible. *Define* a matrix U by the rule

$$U = P^{-1}M. \tag{4.2.60}$$

Verify by calculation that U is unitary,

$$\begin{aligned} UU^\dagger &= (P^{-1}M)(P^{-1}M)^\dagger = P^{-1}MM^\dagger(P^{-1})^\dagger \\ &= P^{-1}P^2(P^\dagger)^{-1} = P^{-1}P^2P^{-1} = I. \end{aligned} \tag{4.2.61}$$

Thus the decomposition (2.56) can be made globally.

It can be shown that if M is invertible, then both P and U are unique, and P is positive-definite Hermitian. Moreover, the decomposition (2.56) is still possible if M is not invertible, and P in this case is still unique. However, U is no longer uniquely defined.

4.3 Symplectic Polar Decomposition

4.3.1 Introduction

Because we will be working with real symplectic matrices, consider the set of all real $2n \times 2n$ matrices. Since the matrix J is invertible, any matrix B in $g\ell(2n, \mathbb{R})$ can be written in the form

$$B = JS + JA. \tag{4.3.1}$$

We also know that matrices of the form JS constitute a Lie subalgebra,

$$\{JS, JS'\} = JS'', \tag{4.3.2}$$

and that this Lie algebra is $sp(2n, \mathbb{R})$. We next observe that the remaining matrices in $g\ell(2n, \mathbb{R})$ obey commutation rules of the form

$$\{JS, JA\} = JA', \tag{4.3.3}$$

$$\{JA, JA'\} = JS. \tag{4.3.4}$$

Finally, by arguments identical to those of the previous section, we conclude that if M is a matrix in $GL(2n, \mathbb{R})$ sufficiently near the identity, then it can be written in the form

$$M = \exp(JA') \exp(JS') \tag{4.3.5}$$

where S' is symmetric and A' is antisymmetric.

Any matrix R of the form

$$R = \exp(JS) \tag{4.3.6}$$

is symplectic, and therefore satisfies the relation

$$JR^T J^{-1} = R^{-1}. \tag{4.3.7}$$

See Sections 3.1 and 3.7. Let Q be any matrix of the form

$$Q = \exp(JA). \tag{4.3.8}$$

It is easily verified that Q satisfies the relation

$$JQ^T J^{-1} = Q. \tag{4.3.9}$$

We define a *J-symmetric* matrix to be *any* matrix Q that satisfies (3.9). See (3.1.12) and also note the similarity of (3.9) and (3.7.26). With these ideas in mind, we see from (3.5) that any M in $GL(2n, \mathbb{R})$ sufficiently near the identity has the decomposition

$$M = QR \tag{4.3.10}$$

where Q is J-symmetric and R is symplectic.

We call (3.10) a *symplectic* polar decomposition to emphasize that the second factor in (3.10) is symplectic. Of course, simply demanding that the second factor R in (3.10) be symplectic is not enough for a definition. We must also put requirements on Q because otherwise we could write $R = R'R''$, with R' and R'' symplectic, and then make the factorization $M = (QR')R''$ and claim R'' as the second factor. In what follows we will require that Q be J-symmetric. We might instead require that Q have a representation of the form (3.8). However, in all the cases for which we have been able to establish the existence of a symplectic polar decomposition, with Q being J-symmetric, we have then also been able to establish that Q has a representation of the form (3.8).

We have seen that orthogonal polar decomposition is possible globally and is unique. Is symplectic polar decomposition also possible globally and unique? To begin to answer these questions, we need to explore some of the properties of J-symmetric matrices. We will do so by proving a series of lemmas.

4.3.2 Properties of J-Symmetric Matrices

Lemma 3.1 Any matrix Q that is symmetric and commutes with J is J-symmetric. Evidently with these two assumptions about Q we have the result

$$JQ^TJ^{-1} = JQJ^{-1} = QJJ^{-1} = Q. \tag{4.3.11}$$

We conclude that all the matrices S^c are J-symmetric. See Section 3.8. In particular, the zero and identity matrices are J-symmetric.

Lemma 3.2 J-symmetric matrices form a linear vector space. We have already seen that the zero matrix is J-symmetric. Suppose Q_1, and Q_2 are J-symmetric. Let a_1 and a_2 be any two scalars. Then we find the result

$$J(a_1Q_1 + a_2Q_2)^TJ^{-1} = a_1JQ_1^TJ^{-1} + a_2JQ_2^TJ^{-1} = a_1Q_1 + a_2Q_2. \tag{4.3.12}$$

Lemma 3.3 Suppose Q is J-symmetric and has an inverse. Then Q^{-1} is J-symmetric:

$$J(Q^{-1})^TJ^{-1} = J(Q^T)^{-1}J^{-1} = (JQ^TJ^{-1})^{-1} = Q^{-1}. \tag{4.3.13}$$

Lemma 3.4 Suppose Q_1 and Q_2 are J-symmetric and commute. Then the product Q_1Q_2 is J-symmetric:

$$J(Q_1Q_2)^TJ^{-1} = J(Q_2Q_1)^TJ^{-1} = JQ_1^TQ_2^TJ^{-1} = JQ_1^TJ^{-1}JQ_2^TJ^{-1} = Q_1Q_2. \tag{4.3.14}$$

Lemma 3.5 If Q is J-symmetric, then so are all powers of Q including, if Q is invertible, all negative powers. This result follows from Lemmas 3.3 and 3.4. We also note that (by definition) $Q^0 = I$ and (by Lemma 3.1) I is J-symmetric.

Lemma 3.6 If Q is J-symmetric, then Q can be written in the form

$$Q = JA \tag{4.3.15}$$

where A is antisymmetric, and conversely. To see this, solve (3.15) for A to find the definition

$$A = J^TQ. \tag{4.3.16}$$

Then we find the result

$$A^T = (J^TQ)^T = Q^TJ = JJ^{-1}Q^TJ = JJQ^TJ^{-1} = JQ = -J^TQ = -A. \tag{4.3.17}$$

Conversely, if A is antisymmetric, we have from (3.15) the result

$$JQ^TJ^{-1} = J(JA)^TJ^{-1} = JA^TJ^TJ^{-1} = -JAJ^TJ^{-1} = JA = Q. \tag{4.3.18}$$

Lemma 3.7 If Q is J-symmetric and nonsingular, then

$$\det Q > 0. \tag{4.3.19}$$

Moreover, if M is nonsingular and has the symplectic polar decomposition (3.10), then

$$\det M > 0. \tag{4.3.20}$$

Thus, M is orientation preserving. To verify the first claim, take the determinant of both sides of (3.15) to find the relation

$$\det Q = (\det J)(\det A) = \det A. \tag{4.3.21}$$

We see that Q being nonsingular implies that A is nonsingular. But, according to Exercise 3.12.2, it follows that $\det A > 0$ and therefore (3.19) holds. To verify the second claim, take the determinant of both sides of (3.10) to find the relations

$$\det M = (\det Q)(\det R) = \det Q. \tag{4.3.22}$$

We see that Q is nonsingular if M is nonsingular. Therefore, if M is nonsingular and has a symplectic polar decomposition, (3.20) follows from the first claim.

Lemma 3.8 A J-symmetric matrix remains J-symmetric under the action of any symplectic similarity transformation. In other words, if Q is J-symmetric, if R is symplectic, and if we define the *transformed* matrix Q^{tr} by the rule

$$Q^{\text{tr}} = R^{-1}QR, \tag{4.3.23}$$

then Q^{tr} is J-symmetric. To check this claim, we carry out the computation

$$\begin{aligned} J(Q^{\text{tr}})^TJ^{-1} &= J(R^{-1}QR)^TJ^{-1} = JR^TQ^T(R^{-1})^TJ^{-1} \\ &= JR^TJ^{-1}JQ^TJ^{-1}J(R^{-1})^TJ^{-1} = R^{-1}QR = Q^{\text{tr}}. \end{aligned} \tag{4.3.24}$$

Note that if Q is written in the form (3.15), and Q^{tr} is written in the form

$$Q^{\text{tr}} = JA^{\text{tr}}, \tag{4.3.25}$$

then A and A^{tr} are related by the equation

$$A^{\text{tr}} = J^TQ^{\text{tr}} = J^TR^{-1}QR = J^TR^{-1}JAR = JR^{-1}J^{-1}AR = R^TAR. \tag{4.3.26}$$

Also note that if the matrix M has a symplectic polar decomposition, then so does the matrix $R'MR''$ where R' and R'' are any two symplectic matrices. To see this, use (3.10) to write

$$R'MR'' = R'QRR'' = R'Q(R')^{-1}R'RR'' = Q^{\rm tr}R''' \tag{4.3.27}$$

where now

$$Q^{\rm tr} = R'Q(R')^{-1} \tag{4.3.28}$$

and

$$R''' = R'RR''. \tag{4.3.29}$$

By the Lemma 3.8, $Q^{\rm tr}$ is J-symmetric. And, by the group property, R''' is symplectic. Therefore, the right side of (3.27) is a symplectic polar decomposition.

Conversely, suppose that a matrix M does not have a symplectic polar decomposition. (We will see in Subsection 4.3.5 that there are such matrices.) Again consider the matrix $R'MR''$ where R' and R'' are any two symplectic matrices. Then it is easy to verify, by *reductio ad absurdum*, that $R'MR''$ also does not have a symplectic polar decomposition. We conclude that the spaces of matrices that do and do not have symplectic polar decompositions are invariant under left and right translations/multiplications by elements in the symplectic group.

Lemma 3.9 Given any matrix M, form the matrix $N(M)$ by the rule

$$N(M) = MJM^TJ^T. \tag{4.3.30}$$

Then N is J-symmetric. To see this, simply compute. We find the result

$$\begin{aligned} JN^TJ^{-1} &= J(MJM^TJ^T)^TJ^{-1} = JJMJ^TM^TJ^{-1} \\ &= MJM^TJ^T = N. \end{aligned} \tag{4.3.31}$$

We remark that if M is symplectic, then $N(M) = I$. Also, suppose we take the determinant of both sides of (3.30). Doing so gives the result

$$\det N = (\det J)^2(\det M)^2 = (\det M)^2. \tag{4.3.32}$$

We see that if M is nonsingular, than so is N. Moreover, consistent with Lemma 3.7, N has a positive determinant.

Lemma 3.10 Suppose M is any matrix and R' and R'' are any two symplectic matrices. Then we have the relation

$$N(R'MR'') = R'N(M)(R')^{-1}. \tag{4.3.33}$$

Again we simply compute and use (3.7) to find the result

$$\begin{aligned} N(R'MR'') &= (R'MR'')J(R'MR'')^TJ^T \\ &= R'M[R''J(R'')^T]M^T(R')^TJ^T \\ &= R'MJM^TJ^TJ(R')^TJ^T \\ &= R'MJM^TJ^T(R')^{-1} = R'N(M)(R')^{-1}. \end{aligned} \tag{4.3.34}$$

As a special case of (3.33) we have the relation

$$N(MR'') = N(M), \tag{4.3.35}$$

which shows that $N(M)$ depends only on the coset $GL(2n,\mathbb{R})/Sp(2n,\mathbb{R})$ to which M belongs.

Lemma 3.11 If the matrix M is transformed by a symplectic similarity transformation, then so is the matrix $N(M)$. Suppose M is any matrix and R is a symplectic matrix. Define the transformed matrix M^{tr} by the rule

$$M^{\text{tr}} = R^{-1}MR. \tag{4.3.36}$$

Let us compute the matrix N^{tr} associated with M^{tr} by the rule (3.30). As a special case of (3.33) we have the result

$$N^{\text{tr}} = N(M^{\text{tr}}) = N(R^{-1}MR) = R^{-1}N(M)R. \tag{4.3.37}$$

4.3.3 Initial Result on Symplectic Polar Decomposition

With these lemmas in hand, we are prepared to say more about the possibility of achieving the factorization (3.10) for general matrices M. Suppose M is invertible, and suppose there exists a J-symmetric matrix Q such that

$$N(M) = Q^2 \tag{4.3.38}$$

with N defined by (3.30). Then M has the factorization (3.10) with R symplectic.

To prove this result, we first observe that Q is invertible: we know from (3.32) that N is invertible if M is, and from (3.38) we see that Q is invertible if N is. Next, since Q is invertible, we define R by the rule

$$R = Q^{-1}M. \tag{4.3.39}$$

Then the computation

$$\begin{aligned} RJR^T &= (Q^{-1}M)J(Q^{-1}M)^T = Q^{-1}MJM^T(Q^{-1})^T \\ &= Q^{-1}MJM^TJ^TJ(Q^{-1})^TJ^{-1}J = Q^{-1}NQ^{-1}J \\ &= Q^{-1}Q^2Q^{-1}J = J \end{aligned} \tag{4.3.40}$$

shows that R is symplectic. Conversely, suppose that Q and R in (3.10) are J-symmetric and symplectic, respectively. Then we find the result

$$\begin{aligned} N &= MJM^TJ^T = QRJ(QR)^TJ^T = QRJR^TQ^TJ^T \\ &= QJQ^TJ^{-1} = Q^2. \end{aligned} \tag{4.3.41}$$

We have learned that establishing the factorization (3.10) is equivalent to finding a J-symmetric matrix Q that satisfies (3.38).

Put another way, we seek a solution of the matrix equation

$$Q = [N(M)]^{1/2}, \tag{4.3.42}$$

provided a solution can be found, and further require that Q be J-symmetric. A standard general method for finding roots of a general matrix N is to first find, if possible, its logarithm.[1] Therefore, let us first try to compute $\log(N)$. From (3.7.2) we find the result

$$\log(N) = -\sum_{\ell=1}^{\infty}(I-N)^{\ell}/\ell. \tag{4.3.43}$$

Note that (by Lemmas 3.1, 3.2, 3.5, and 3.9) all the terms $[(I-N)^{\ell}/\ell]$ are J-symmetric matrices. Consequently, if the series (3.43) converges, then (by Lemma 3.2) $\log(N)$ will be a J-symmetric matrix. If the series does converge, let us define a matrix Q by the rule

$$Q = \exp[(1/2)\log(N)]. \tag{4.3.44}$$

The matrix Q will also be J-symmetric. [Apply to the series (3.7.1) arguments similar to those just made for $\log(N)$.] Moreover, Q will satisfy (3.38),

$$Q^2 = \{\exp[(1/2)\log(N)]\}^2 = \exp[\log(N)] = N. \tag{4.3.45}$$

Therefore we can achieve the factorization (3.10) if the series (3.43) converges.

The series (3.43) will converge if N is sufficiently near I. Specifically, the series will converge if $\parallel N-I \parallel< 1$ for some choice of matrix norm. But, according to the remark made in Lemma 3.9, $N = I$ if M is symplectic. Consequently, the series will converge if M is sufficiently near a symplectic matrix.

4.3.4 Extended Result on Symplectic Polar Decomposition

But still more can be said. Consider the matrix $N(\lambda M)$ which, according to (3.30), has the form

$$N(\lambda M) = (\lambda M)J(\lambda M)^T J^T = \lambda^2 MJM^TJ^T = \lambda^2 N(M) \tag{4.3.46}$$

where λ is any real scalar in the range $0 < \lambda < \infty$. Also, let us view the set of all $2n \times 2n$ matrices as a linear vector space. This space is shown schematically in Figure 3.1. There we have depicted the zero matrix as the origin, and have also displayed the identity matrix I. In addition we have depicted the various matrices $N(\lambda M)$ for fixed M as a ray (half line) emanating from the origin. These matrices do in fact lie on a ray because, according to (3.46), they are the λ^2 multiple of a fixed matrix. Finally, we have depicted the unit ball about the identity I. It is the set of matrices C that satisfy the requirement

$$\parallel C-I \parallel< 1 \tag{4.3.47}$$

for some choice of matrix norm. In drawing the unit ball about I we have assumed that the norm has the property $\parallel I \parallel= 1$ so that the zero matrix lies on the ball's surface. We are now ready to state an extended result in the form of a theorem.

[1] Recall from Exercise 2.3 that there are special methods for finding matrix roots if the matrix is positive symmetric.

Theorem 3.1 Suppose that for some value λ_0 the matrix $N(\lambda_0 M)$ lies *within* the unit ball around I. (This is the situation depicted in Figure 3.1.) Then the matrix M is invertible, has positive determinant, and has the symplectic polar decomposition (3.10).

Proof: Consider the matrix M_0 defined by the relation

$$M_0 = \lambda_0 M. \tag{4.3.48}$$

According to (3.46) the matrix N_0 associated with M_0 is given by the relation

$$N_0 = N(\lambda_0 M) = \lambda_0^2 N(M) = \lambda_0^2 N. \tag{4.3.49}$$

By hypothesis, we have the relation

$$\| N_0 - I \| = \| N(\lambda_0 M) - I \| < 1. \tag{4.3.50}$$

It follows that the series (3.43) for $\log(N_0)$ converges, and we can define a J-symmetric matrix Q_0 by the rule

$$Q_0 = \exp[(1/2)\log(N_0)]. \tag{4.3.51}$$

Moreover, according to Lemma 3.2, the matrix Q defined by

$$Q = (1/\lambda_0)Q_0 \tag{4.3.52}$$

is also J-symmetric. By (3.49), (3.51), and (3.52), it satisfies the relation

$$Q^2 = (1/\lambda_0)^2 Q_0^2 = (1/\lambda_0)^2 N_0 = N. \tag{4.3.53}$$

Consequently, M has the symplectic polar decomposition (3.10) with Q given by (3.52).

We also note that Q can be written in exponential form: We already know that $[(1/2)\log(N_0)]$ is J-symmetric. Consequently, according to Lemma 3.6, there exists an antisymmetric matrix A_0 such that

$$(1/2)\log(N_0) = JA_0, \tag{4.3.54}$$

and (3.51) can therefore be written in the form

$$Q_0 = \exp(JA_0). \tag{4.3.55}$$

Now use (3.52) and (3.55) to write Q in the form

$$\begin{aligned} Q &= (1/\lambda_0)Q_0 = \exp\{-[\log(\lambda_0)]I\}\exp(JA_0) \\ &= \exp\{JA_0 - [\log(\lambda_0)]I\} = \exp(JA) \end{aligned} \tag{4.3.56}$$

with A given by the relation

$$A = A_0 + [\log(\lambda_0)]J. \tag{4.3.57}$$

Finally, it follows from (3.56) and (3.7.129) that $\det Q > 0$ and hence, from (3.22), $\det M > 0$.

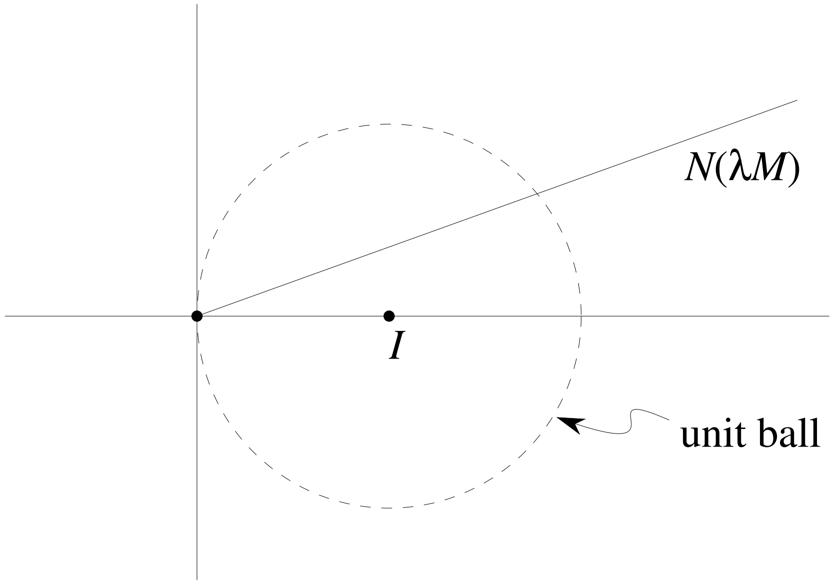


Figure 4.3.1: Schematic depiction of matrix space showing the zero matrix, the identity matrix I, the ray $N(\lambda M)$, and the unit ball around the identity matrix.

One might wonder if the condition of Theorem 3.1 is necessary. We can easily see that it is for the case of $GL(2,\mathbb{R})$. However, it is not necessary for some examples in the case of $GL(4,\mathbb{R})$, and presumably not for some examples in any $GL(2n,\mathbb{R})$ with $n \geq 2$. See Exercise 3.19.

In the $GL(2,\mathbb{R})$ case, for any 2×2 matrix M, we have the result

$$N(\lambda M) = \lambda^2 MJM^TJ^T = \lambda^2[\det(M)]I. \tag{4.3.58}$$

(See Exercise 3.1.2.) Thus we have the relation

$$\| N(\lambda M) - I \| = \| [\lambda^2 \det(M) - 1]I \| = |\lambda^2 \det(M) - 1| \; \| I \| \; . \tag{4.3.59}$$

Evidently, if $\det(M) > 0$, we can find a λ such that the right side of (3.59) is less than 1. Also, we can write M in the form (3.10) with Q and R given by the relation

$$Q = +[\det(M)]^{1/2}I, \tag{4.3.60}$$

$$R = +\{1/[\det(M)]^{1/2}\}M. \tag{4.3.61}$$

On the other hand, if $\det(M) < 0$, no choice of (real) λ will make the right side of (3.59) less than 1. This is consistent with Lemma 3.7 which states that $\det M > 0$ is a necessary condition for M to have a symplectic polar decompostion.

We also observe that we could replace the + signs in (3.60) and (3.61) by − signs and also obtain a (different) symplectic polar decomposition. Exercise 3.13 shows that the use of the Theorem 3.1 procedure, which is always possible in the 2×2 case when $\det M > 0$, produces the + signs choice.

Let us summarize our results in the language of cosets. (Again, see Section 5.12, if necessary, for a detailed discussion of cosets.) We have been dealing with the group $GL(2n,\mathbb{R})$ and its subgroup $Sp(2n,\mathbb{R})$. Form the coset space $GL(2n,\mathbb{R})/Sp(2n,\mathbb{R})$ consisting of the left cosets of $GL(2n,\mathbb{R})$ with respect to $Sp(2n,\mathbb{R})$. Equations (3.10), (3.56), and (3.57) indicate that the left cosets $GL(2n,\mathbb{R})/Sp(2n,\mathbb{R})$, for those M which satisfy the requirement of Theorem 3.1, can be put into one-to-one correspondence with $2n \times 2n$ (real) nonsingular antisymmetric matrices A. Such matrices form a linear vector space whose dimension m is given by the relation

$$m = \dim(A) = n(2n-1). \tag{4.3.62}$$

Thus, the portion of $GL(2n,\mathbb{R})$ that satisfies the requirement of Theorem 3.1 has the topology of $E^m\times$ $Sp(2n,\mathbb{R})$ with m given by (3.62).

4.3.5 Symplectic Polar Decomposition Not Globally Possible

We have already seen that symplectic polar decomposition is not possible for M in the cosets with $\det M < 0$. Are there other cosets as well for which symplectic decomposition is impossible? We will see that there are. Therefore symplectic polar decomposition is not possible globally even with the restriction $\det M > 0$.

Consider, as a possible 4×4 counter example, the diagonal matrix M given by

$$M = \begin{pmatrix} \mu_1 & 0 & 0 & 0 \\ 0 & \mu_2 & 0 & 0 \\ 0 & 0 & \nu_1 & 0 \\ 0 & 0 & 0 & \nu_2 \end{pmatrix} \tag{4.3.63}$$

where the μ_j and ν_j are real and nonzero. Its determinant satisfies the condition

$$\det(M) = \mu_1\mu_2\nu_1\nu_2. \tag{4.3.64}$$

Take for J the matrix

$$J = \begin{pmatrix} J_2 & 0 \\ 0 & J_2 \end{pmatrix}. \tag{4.3.65}$$

See (3.2.10). Then we find, using (3.30) and the results of Exercise 3.1.2, the relations

$$N(M) = \begin{pmatrix} \mu_1\mu_2 & 0 & 0 & 0 \\ 0 & \mu_1\mu_2 & 0 & 0 \\ 0 & 0 & \nu_1\nu_2 & 0 \\ 0 & 0 & 0 & \nu_1\nu_2 \end{pmatrix} \tag{4.3.66}$$

and

$$N(\lambda M) = \begin{pmatrix} \lambda^2\mu_1\mu_2 & 0 & 0 & 0 \\ 0 & \lambda^2\mu_1\mu_2 & 0 & 0 \\ 0 & 0 & \lambda^2\nu_1\nu_2 & 0 \\ 0 & 0 & 0 & \lambda^2\nu_1\nu_2 \end{pmatrix}. \tag{4.3.67}$$

It follows that

$$N(\lambda M)-I=-\left(\begin{array}{cccc} 1-\lambda^2\mu_1\mu_2 & 0 & 0 & 0 \\ 0 & 1-\lambda^2\mu_1\mu_2 & 0 & 0 \\ 0 & 0 & 1-\lambda^2\nu_1\nu_2 & 0 \\ 0 & 0 & 0 & 1-\lambda^2\nu_1\nu_2 \end{array}\right). \tag{4.3.68}$$

Therefore, using the spectral norm, which is the strongest, we find the result

$$||N(\lambda M)-I|| = \max[|(1-\lambda^2\mu_1\mu_2)|,\ |(1-\lambda^2\nu_1\nu_2)|]. \tag{4.3.69}$$

We see that the ray $N(\lambda M)$ does not pass through the interior of the unit ball about the identity if

$$\mu_1\mu_2 < 0 \ \text{ or } \ \nu_1\nu_2 < 0. \tag{4.3.70}$$

Consequently the series (3.43) used to construct $\log(N)$, and hence Q, might be expected to diverge in these cases.[2]

We will show that, in fact, for these cases there is no J-symmetric matrix Q such that

$$Q^2 = N(\lambda M). \tag{4.3.71}$$

Suppose that such a Q exists. By Lemma 3.6 there is an antisymmetric matrix A such that $Q = JA$. Write A in the 2×2 block form

$$A=\left(\begin{array}{cc} a & b \\ c & d \end{array}\right). \tag{4.3.72}$$

Then we find for Q the result

$$Q = JA = \left(\begin{array}{cc} J_2 & 0 \\ 0 & J_2 \end{array}\right)\left(\begin{array}{cc} a & b \\ c & d \end{array}\right)=\left(\begin{array}{cc} J_2a & J_2b \\ J_2c & J_2d \end{array}\right). \tag{4.3.73}$$

And for Q^2 we find the result

$$Q^2=\left(\begin{array}{cc} J_2a & J_2b \\ J_2c & J_2d \end{array}\right)\left(\begin{array}{cc} J_2a & J_2b \\ J_2c & J_2d \end{array}\right)=\left(\begin{array}{cc} (J_2a)^2+J_2bJ_2c & J_2aJ_2b+J_2bJ_2d \\ J_2cJ_2a+J_2dJ_2c & J_2cJ_2b+(J_2d)^2 \end{array}\right). \tag{4.3.74}$$

Let us work out the properties of the entries in Q^2. Since A is antisymmetric, the matrices a and d are antisymmetric and therefore have the form

$$a=\left(\begin{array}{cc} 0 & \alpha \\ -\alpha & 0 \end{array}\right), \tag{4.3.75}$$

$$d=\left(\begin{array}{cc} 0 & \delta \\ -\delta & 0 \end{array}\right). \tag{4.3.76}$$

Consequently, we have the relations

$$J_2a=-\alpha I, \tag{4.3.77}$$

[2] In fact, in this case because of the diagonal form of $[N(\lambda M)-I]$, it easily verified that the series (3.43) does diverge.

$$J_2 d = -\delta I, \tag{4.3.78}$$

from which it follows that

$$(J_2 a)^2 = \alpha^2 I, \tag{4.3.79}$$

$$(J_2 d)^2 = \delta^2 I. \tag{4.3.80}$$

Next we find, again using the results of Exercise 3.1.2, that

$$J_2 b J_2 c = -J_2 b J_2 b^T = \det(b) I \tag{4.3.81}$$

and

$$J_2 c J_2 b = -J_2 b^T J_2 b = \det(b) I. \tag{4.3.82}$$

Here we have used the relation

$$c = -b^T, \tag{4.3.83}$$

which also follows from the fact that A is antisymmetric. Finally, we have the results

$$J_2 a J_2 b + J_2 b J_2 d = -(\alpha + \delta) J_2 b, \tag{4.3.84}$$

$$J_2 c J_2 a + J_2 d J_2 c = -(\alpha + \delta) J_2 c = (\alpha + \delta) J_2 b^T. \tag{4.3.85}$$

Now require that (3.71) hold. So doing yields the relations

$$\alpha^2 + \det(b) = \lambda^2 \mu_1 \mu_2, \tag{4.3.86}$$

$$\delta^2 + \det(b) = \lambda^2 \nu_1 \nu_2, \tag{4.3.87}$$

$$-(\alpha + \delta) J_2 b = 0, \tag{4.3.88}$$

$$(\alpha + \delta) J_2 b^T = 0. \tag{4.3.89}$$

Note that the relations (3.88) and (3.89) are equivalent, and yield the two possibilities

$$\alpha = -\delta \tag{4.3.90}$$

or

$$J_2 b = 0 \text{ which implies } b = 0. \tag{4.3.91}$$

If (3.90) holds, the relations (3.86) and (3.87) become

$$\alpha^2 + \det(b) = \lambda^2 \mu_1 \mu_2, \tag{4.3.92}$$

$$\alpha^2 + \det(b) = \lambda^2 \nu_1 \nu_2, \tag{4.3.93}$$

and they are contradictory if $\mu_1 \mu_2 \neq \nu_1 \nu_2$. If (3.91) holds, the relations (3.86) and (3.87) become

$$\alpha^2 = \lambda^2 \mu_1 \mu_2, \tag{4.3.94}$$

$$\delta^2 = \lambda^2 \nu_1 \nu_2, \tag{4.3.95}$$

and at least one of them is an impossibility in the cases (3.70).

We conclude that no J-symmetric matrix Q exists that satisfies (3.71) when M is of the form (3.63) and (3.70) holds. Therefore, symplectic polar decomposition for such M is impossible. The same is true for any matrices M' that are in the same cosets as such M. Finally we note from (3.64) that if both the cases (3.70) hold, then it is still possible to have $\det(M) > 0$.

From Theorem 3.1 we know that a *sufficient* condition for M to have a symplectic polar decomposition is that the ray $N(\lambda M)$ intersect the unit ball about I. From the example of this section one might be tempted to conjecture that this intersection condition is also a *necessary* condition. However, as already mentioned earlier, Exercise 3.19 shows that the intersection condition is not necessary for a particular $GL(4, \mathbb{R})$ example.

4.3.6 Uniqueness of Symplectic Polar Decomposition

There remains the question of uniqueness. Suppose that M has the two symplectic polar decompositions

$$M = QR \tag{4.3.96}$$

and

$$M = Q'R'. \tag{4.3.97}$$

Then we see that

$$Q' = QR(R')^{-1}. \tag{4.3.98}$$

But, since symplectic matrices form a group, the matrix $R(R')^{-1}$ is symplectic, and therefore Q' and Q are in the same $GL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$ coset. By Lemma 3.2, $-Q$ is a J-symplectic matrix if $+Q$ is, and they are related under multiplication by the symplectic matrix $-I$, and are therefore in the same coset. Thus, if symplectic polar decomposition is possible at all, there are always at least two possibilities. In the case of Theorem 3.1 we imposed the unit ball condition of Figure 3.1, and were able to make the choice specified by (3.51) and (3.52). In the 2×2 case this choice dictates the + sign in (3.60). We will see that an analogous choice can be made in the general case.

As in Theorem 3.1, let Q_0 be the matrix associated with $N_0 = N(\lambda_0 M)$. Write the matrix identity

$$-2I = (-Q_0 - I) + (Q_0 - I) \tag{4.3.99}$$

and use the triangle inequality (3.7.12) to deduce the inequality

$$|| - 2I|| \leq || - Q_0 - I|| + ||Q_0 - I||. \tag{4.3.100}$$

With an appropriate norm, such as the spectral norm, we have the relation

$$||2I|| = 2, \tag{4.3.101}$$

and we conclude from (3.100) that

$$|| - Q_0 - I|| \geq 2 - ||Q_0 - I||. \tag{4.3.102}$$

We will now seek an estimate for the quantity $||Q_0 - I||$.

Consider the function $g(x)$ defined by the equation

$$(1 - x)^{1/2} = 1 - g(x). \tag{4.3.103}$$

It is readily verified that this function has the expansion

$$g(x) = \sum_{\ell=1}^{\infty} d_\ell x^\ell = x/2 + x^2/8 + \cdots \tag{4.3.104}$$

where all the coefficients d_ℓ are *positive*. Moreover, $g(x)$ satisfies the inequality

$$(x/2) \leq g(x) \leq x \text{ for } x \in [0,1]. \tag{4.3.105}$$

Instead of the method of Theorem 3.1, let us use a more direct (but equivalent) way of defining Q_0 by writing

$$Q_0 = (N_0)^{1/2} = [I - (I - N_0)]^{1/2} = I - g(I - N_0). \tag{4.3.106}$$

As in the proof of Theorem 3.1, let us also make the assumption that

$$||N_0 - I|| < 1. \tag{4.3.107}$$

Under this hypothesis the relation (3.106) yields the inequality

$$\begin{aligned} ||Q_0 - I|| &= || - g(I - N_0)|| = ||g(I - N_0)|| \\ &= \| \sum_{\ell=1}^{\infty} d_\ell (I - N_0)^\ell \| \leq \sum_{\ell=1}^{\infty} d_\ell \parallel I - N_0 \parallel^\ell \\ &= g(||I - N_0||) = g(||N_0 - I||) \leq ||N_0 - I|| < 1. \end{aligned} \tag{4.3.108}$$

Here we have made use of the positivity of the d_ℓ and the relation (3.105).

Finally, combine (3.102) and (3.108) to get the result

$$|| - Q_0 - I|| > 1. \tag{4.3.109}$$

We see from (3.108) that the use of (3.106) or, equivalently, the use of the method of Theorem 3.1, produces a Q_0 that is inside the unit ball shown in Figure 3.1. And, correspondingly, (3.109) shows that $-Q_0$ is outside this unit ball. Thus, the method of Theorem 3.1 assures that the J-symplectic factor Q_0 is as close to the identity I as possible.

4.3.7 Concluding Summary

Let us summarize what has been learned. We have seen that symplectic polar decomposition is possible and unique if M is sufficiently near the symplectic group so that $N(M)$ lies within the unit ball about I. We have extended this result to show that symplectic polar decomposition is possible and unique if the ray $N(\lambda M)$ passes through the unit ball about I. Also, we have found a family of counter examples that show that symplectic polar decomposition is not possible globally. Naturally, for any counter example, the ray $N(\lambda M)$ cannot pass through the unit ball about I. However, as illustrated in Exercise 3.21, there are examples where the ray $N(\lambda M)$ does not pass through the unit ball about I and symplectic polar decomposition is still possible and unique. See also Exercises 3.22 through 3.24 for further examples of when symplectic polar decomposition is and is not possible.

Exercises

4.3.1. Verify the commutation rules (3.2) through (3.4).

4.3.2. Suppose M is a $2n \times 2n$ matrix near the identity. Then M can be written in the form

$$M = \exp[\epsilon(JS + JA)] \tag{4.3.110}$$

where ϵ is a small parameter. Show that M can also be written in the form

$$M = \exp(\epsilon JA')\exp(\epsilon JS'), \tag{4.3.111}$$

and determine the first few terms in A' and S' when expressed as a power series expansion in ϵ.

4.3.3. Show that any matrix of the form (3.8) satisfies (3.9), and hence is J-symmetric.

4.3.4. Review Exercise 3.1.9. Employing a slightly different notation for the symplectic transpose, define the matrix M' by the rule

$$M' = M^S = JM^T J^{-1}. \tag{4.3.112}$$

Show that any matrix of the form JA is symmetric under this priming operation,

$$(JA)' = JA, \tag{4.3.113}$$

and any matrix of the form JS is antisymmetric,

$$(JS)' = -JS. \tag{4.3.114}$$

Thus, verify that (3.1) is a decomposition of B into symmetric and antisymmetric parts with respect to the symplectic transpose operation.

4.3.5. Verify the calculations associated with Lemmas 3.1 through 3.11.

4.3.6. Refer to Lemma 3.1. Show that any two of the following three properties implies the third: (i) symmetric, (ii) commutes with J, (iii) J-symmetric.

4.3.7. Suppose M_1 and M_2 are two commuting matrices, and suppose M_2 is invertible. Verify that M_1 and M_2^{-1} also commute. Show that the set of all commuting J-symmetric matrices in $GL(2n, \mathbb{R})$ forms a group.

4.3.8. Review Lemma 3.6. Show that if A is any antisymmetric matrix, there exists another antisymmetric matrix A' such that

$$JA = A'J. \tag{4.3.115}$$

4.3.9. Given any factorization of the form (3.10), use Lemma 3.8 to show that M also has the factorization

$$M = QR = RQ^{\text{tr}}. \tag{4.3.116}$$

If Q is of the form (3.8), find the A^{tr} associated with Q^{tr}.

4.3.10. If M is symplectic, verify that N as given by (3.30) satisfies $N = I$.

4.3.11. If you have not already done so in Exercise 3.5, verify (3.40) and (3.41).

4.3.12. Verify the steps in the proof of Theorem 3.1.

4.3.13. Verify (3.58) and (3.59). Show that

$$\parallel I \parallel \geq 1 \tag{4.3.117}$$

for any choice of norm. Hint: Apply (3.7.10) and (3.7.13) to $\parallel I^2 \parallel$. Show that if $\det(M) < 0$, no choice of λ will make the right side of (3.59) less than 1. Show that applying the method of Theorem 3.1 in the 2×2 case produces the symplectic polar decomposition given by (3.60) and (3.61).

4.3.14. Suppose the matrices M and M' belong to the same $GL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$ coset. Show that they then have the same determinant. Is the converse true?

4.3.15. Verify (3.66).

4.3.16. Suppose that, for the matrix M given by (3.63), there are the conditions $\mu_1\mu_2 > 0$ and $\nu_1\nu_2 > 0$. Show that in this case a J-symmetric solution to (6.71) is given by the relation

$$Q = [N(\lambda M)]^{1/2} = \lambda \begin{pmatrix} [\mu_1\mu_2]^{1/2} & 0 & 0 & 0 \\ 0 & [\mu_1\mu_2]^{1/2} & 0 & 0 \\ 0 & 0 & [\nu_1\nu_2]^{1/2} & 0 \\ 0 & 0 & 0 & [\nu_1\nu_2]^{1/2} \end{pmatrix}. \tag{4.3.118}$$

Instead, suppose one or both of the conditions (3.70) holds. Then, from (3.118), one might surmise that (3.71) has only imaginary solutions. This is not the case. Show, for example if both conditions (3.70) hold, then (3.71) has the *real* solution

$$Q = \lambda \begin{pmatrix} [-\mu_1\mu_2]^{1/2} J_2 & 0 \\ 0 & [-\nu_1\nu_2]^{1/2} J_2 \end{pmatrix}. \tag{4.3.119}$$

However note that this solution Q is *not* J-symmetric. Consequently, there is no contradiction with the results of Section 4.3.5.

4.3.17. Graph the function $g(x)$ given by (3.103). Verify all claims made for $g(x)$. Determine the coefficients d_ℓ and the domain of convergence of this series. Verify (3.106) and (3.108). Show that Q_0 as defined by (3.106) is J-symmetric. Show that use of (3.106) gives the same result as that of Theorem 3.1.

4.3.18. We know that $N(M)$ is J-symmetric and therefore, by Lemma 3.6, there is an antisymmetric matrix A' such that

$$N(M) = JA'. \tag{4.3.120}$$

By the same lemma, if Q is J-symmetric there is an antisymmetric A such that (3.15) holds. Now suppose that there is a Q of the form (3.15) such that (3.38) is satisfied. Show, using the representations (3.15) and (3.120), that there is the relation

$$JA' = JAJA, \tag{4.3.121}$$

from which it follows that

$$A' = AJA. \tag{4.3.122}$$

Since A is antisymmetric, (3.122) can also be written in the form

$$-A' = AJA^T. \tag{4.3.123}$$

Recall the work of Section 3.12. We see that if there exists a J-symmetric Q such that (3.38) holds, then the antisymmetric matrix A associated with this Q is *also* a Darboux matrix that transforms J to $-A'$.

4.3.19. In Section 4.3.5 we studied the space of all *diagonal* matrices in $GL(4,\mathbb{R})$ to determine which of them had symplectic polar decompositions. Ideally we would like to do the same for all matrices in $GL(4,\mathbb{R})$, but this seems to be a formidable task because $GL(4,\mathbb{R})$ is 16 dimensional.[3] We know that in principle it is sufficient to examine the coset space $GL(4,\mathbb{R})/Sp(4,\mathbb{R})$, which is 6 dimensional. However, the parameterization of the coset space $GL(2n,\mathbb{R})/Sp(2n,\mathbb{R})$ is complicated. See Appendix P. As a simpler task, this exercise begins to examine the subset of matrices $SO(4,\mathbb{R}) \subset GL(4,\mathbb{R})$. This set (also 6 dimensional), while incomplete in the sense of not embracing all cosets, is easier to study. The work of this exercise and the next will show that every element in $SO(4,\mathbb{R})$ has a symplectic polar decomposition. It will also provide information about $SO(4,\mathbb{R})$ that will be valuable for subsequent use.

The Lie algebra $so(4,\mathbb{R})$ consists of all real 4×4 antisymmetric matrices A. As with the case of symmetric matrices S, it is convenient to decompose A into matrices A^a and A^c that *anticommute* and *commute* with J, respectively,

$$A = A^a + A^c. \tag{4.3.124}$$

Show that

$$A^a = (1/2)(A - JAJ^{-1}) \tag{4.3.125}$$

and

$$A^c = (1/2)(A + JAJ^{-1}). \tag{4.3.126}$$

Verify that the Lie algebra formed by the set of antisymetric matrices has the property

$$\{A^c, (A^c)'\} = (A^c)'', \tag{4.3.127}$$

$$\{A^c, A^a\} = (A^a)', \tag{4.3.128}$$

$$\{A^a, (A^a)'\} = A^c. \tag{4.3.129}$$

[3] In fact, we would like to do the same for $GL(2n,\mathbb{R})$ for all n; but at least $GL(4,\mathbb{R})$ is the first nontrivial case.

If A is sufficiently small, the BCH and Zassenhaus series converge, and we may achieve the factorization

$$\exp(A) = \exp(A^a + A^c) = \exp[(A^a)'] \exp[(A^c)']. \tag{4.3.130}$$

Show that any element of the form $\exp(A^a)$ is J-symmetric, and any element of the form $\exp(A^c)$ is symplectic. The relation (3.130) shows that any $SO(4,\mathbb{R})$ element suficiently near the identity has a symplectic polar decomposition. This is to be expected because we already know that *every* matrix sufficiently near the identity has a symplectic polar decomposition.

The most general 4×4 antisymmetric matrix A can be written in the form

$$A = \begin{pmatrix} 0 & \alpha & \beta & \gamma \\ -\alpha & 0 & \delta & \epsilon \\ -\beta & -\delta & 0 & \zeta \\ -\gamma & -\epsilon & -\zeta & 0 \end{pmatrix}. \tag{4.3.131}$$

Using the form of J given by (3.65), show that

$$A^c = (1/2) \begin{pmatrix} 0 & 2\alpha & \beta+\epsilon & \gamma-\delta \\ -2\alpha & 0 & -\gamma+\delta & \beta+\epsilon \\ -\beta-\epsilon & \gamma-\delta & 0 & 2\zeta \\ -\gamma+\delta & -\beta-\epsilon & -2\zeta & 0 \end{pmatrix} \tag{4.3.132}$$

and

$$A^a = (1/2) \begin{pmatrix} 0 & 0 & \beta-\epsilon & \gamma+\delta \\ 0 & 0 & \gamma+\delta & -\beta+\epsilon \\ -\beta+\epsilon & -\gamma-\delta & 0 & 0 \\ -\gamma-\delta & \beta-\epsilon & 0 & 0 \end{pmatrix}. \tag{4.3.133}$$

Evidently the space of matrices of the form A^c is 4 dimensional, and the space of matrices of the form A^a is 2 dimensional. Let us seek a convenient basis for each.

Begin with the A^c. Evidently matrices of the form JS^c are antisymmetric and commute with J. Verify that there is the one-to-one correspondence

$$JS^c \leftrightarrow A^c. \tag{4.3.134}$$

We already know that the matrices JS^c are associated with the $u(2)$ part of $sp(4,\mathbb{R})$. Looking ahead, a convenient basis for these matrices, in the case that J is of the form (3.1.1), will be found in Exercise 5.7.8. They are the matrices B^0 through B^3 given in (5.7.44). If we can find their counterparts for the case that J is given by (3.65), then we will have found a convenient basis for the A^c. This is easily done. Review Section 3.2. Show that in the 4×4 case the matrix P of (3.2.5) is given by the relation

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{4.3.135}$$

Evidently P is both symmetric and orthogonal. Show that the desired basis for the A^c can be taken to be the matrices C^j defined by the rule

$$C^j = PB^jP. \tag{4.3.136}$$

Verify that the matrices C^j are given by the relations

$$C^0 = PB^0P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}, \tag{4.3.137}$$

$$C^1 = PB^1P = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}, \tag{4.3.138}$$

$$C^2 = PB^2P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}, \tag{4.3.139}$$

$$C^3 = PB^3P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \tag{4.3.140}$$

and are of the form (3.132). Verify that they satisfy the commutation rules

$$\{C^0, C^j\} = 0, \quad j = 0, 1, 2, 3; \tag{4.3.141}$$

$$\{C^1, C^2\} = -2C^3, \tag{4.3.142}$$

$$\{C^2, C^3\} = -2C^1, \tag{4.3.143}$$

$$\{C^3, C^1\} = -2C^2. \tag{4.3.144}$$

Next find a basis for the A^a. By looking at (3.133), show that a convenient choice is

$$E^1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix} \tag{4.3.145}$$

and

$$E^2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{4.3.146}$$

Show that they, together with the C^j, obey the commutation relations

$$\{E^1, E^2\} = 2C^0, \tag{4.3.147}$$

$$\{C^0, E^1\} = 2E^2, \tag{4.3.148}$$

$$\{C^0, E^2\} = -2E^1, \tag{4.3.149}$$

$$\{C^j, E^1\} = \{C^j, E^2\} = 0, \quad j = 1, 2, 3. \tag{4.3.150}$$

After a bit of algebraic experimentation (and in anticipation of Exercise 11.1.6), one finds that it is convenient to relabel and renormalize the basis just found by making the definitions

$$G^1 = -(1/2)E^1 = (1/2)\begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

$$G^2 = -(1/2)E^2 = (1/2)\begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix},$$

$$G^3 = (1/2)C^0 = (1/2)\begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}. \tag{4.3.151}$$

$$H^1 = (1/2)C^3 = (1/2)\begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$H^2 = (1/2)C^2 = (1/2)\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix},$$

$$H^3 = (1/2)C^1 = (1/2)\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}. \tag{4.3.152}$$

Show that the G^j and H^k satisfy the pleasing commutation rules

$$\{G^1, G^2\} = G^3, \tag{4.3.153}$$

$$\{G^2, G^3\} = G^1, \tag{4.3.154}$$

$$\{G^3, G^1\} = G^2, \tag{4.3.155}$$

$$\{H^1, H^2\} = H^3, \tag{4.3.156}$$

$$\{H^2, H^3\} = H^1, \tag{4.3.157}$$

$$\{H^3, H^1\} = H^2, \tag{4.3.158}$$

$$\{G^j, H^k\} = 0 \ \text{ for } \ j, k = 1, 2, 3; \tag{4.3.159}$$

and the anticommutation relations

$$\{G^j, G^k\}_+ = \{H^j, H^k\}_+ = -(1/2)\delta_{jk} I. \tag{4.3.160}$$

Show also that there are the relations

$$(G^1)^2 + (G^2)^2 + (G^3)^2 = (H^1)^2 + (H^2)^2 + (H^3)^2 = -(3/4)I. \tag{4.3.161}$$

You have verified, as advertised in the discussion associated with Table 3.7.1, that the Lie algebra $so(4, \mathbb{R})$ is the direct sum of two mutually commuting $su(2)$ Lie algebras. [Strictly speaking, based only on the commutation rules, we cannot tell at this stage whether it is $su(2)$ or $so(3, \mathbb{R})$ that we have found. In the next exercise you will verify that it is indeed $su(2)$.] Note that all the matrices G^j and H^k are real and antisymmetric, and form a basis for the 6-dimensional set of antisymmetric 4×4 matrices. Verify that G^1 through G^3 are linear combinations of pair-wise commuting generators for rotations in the (1,4 and 2,3), (1,3 and 2,4), and (1,2 and 3,4) planes, respectively. Verify that H^1 through H^3 are also linear combinations of pair-wise commuting generators for rotations in the (1,2 and 3,4), (1,3 and 2,4), and (1,4 and 2,3) planes, respectively. Verify that, given a four-element set, there are three ways of forming pairs of disjoint two-element subsets. This combinatorial fact lies behind the possible construction of the three G^j and the three H^k.

4.3.20. Review Exercise 3.19 above. It set up the machinery for a study of $SO(4, \mathbb{R})$. The purpose of this exercise is to show that *all* elements of $SO(4, \mathbb{R})$ have symplectic polar decompositions. In so doing we will also learn more about $so(4, \mathbb{R})$ and the two mutually commuting $su(2)$ Lie algebras within it.

Introduce the notation

$$\boldsymbol{G} = (G^1, G^2, G^3), \ \ \boldsymbol{H} = (H^1, H^2, H^3). \tag{4.3.162}$$

Also introduce the vectors

$$\boldsymbol{s} = (s^1, s^2, s^3), \ \ \boldsymbol{t} = (t^1, t^2, t^3). \tag{4.3.163}$$

Finally employ the notation

$$\boldsymbol{s} \cdot \boldsymbol{G} = s_1 G^1 + s_2 G^2 + s_3 G^3, \ \text{ etc.} \tag{4.3.164}$$

We know that the G^j and H^k form a basis for the Lie algebra $so(4, \mathbb{R})$. It follows from Section 3.8.1 that the most general element in $SO(4, \mathbb{R})$ can be written in the form

$$O(\boldsymbol{s}, \boldsymbol{t}) = \exp(\boldsymbol{s} \cdot \boldsymbol{G} + \boldsymbol{t} \cdot \boldsymbol{H}). \tag{4.3.165}$$

Now, since the G^j and H^k commute, we may also write

$$O(\boldsymbol{s}, \boldsymbol{t}) = \exp(\boldsymbol{s} \cdot \boldsymbol{G}) \exp(\boldsymbol{t} \cdot \boldsymbol{H}). \tag{4.3.166}$$

We observe that the factor $\exp(\boldsymbol{t}\cdot\boldsymbol{H})$ is an element in $Sp(4,\mathbb{R})$. Therefore, in order to achieve a symplectic polar decomposition for $O(\boldsymbol{s},\boldsymbol{t})$, we only need to achieve a symplectic polar decomposition for $\exp(\boldsymbol{s}\cdot\boldsymbol{G})$.

At this point let us pause to explore more of the properties of the G^j and H^k. Review Exercise 8.7.12. Verify that the G^j and H^j satisfy the *same* multiplication rules as the K^j. But, unlike some of the K^j, they are purely real. Verify in particular that there are the relations

$$(\boldsymbol{s}\cdot\boldsymbol{G})^2=-(1/4)(\boldsymbol{s}\cdot\boldsymbol{s})I \tag{4.3.167}$$

and

$$(\boldsymbol{t}\cdot\boldsymbol{H})^2=-(1/4)(\boldsymbol{t}\cdot\boldsymbol{t})I. \tag{4.3.168}$$

Use these relations to show that there are the explicit results

$$\exp(\boldsymbol{s}\cdot\boldsymbol{G})=I\cos(s/2)+(\boldsymbol{s}\cdot\boldsymbol{G})(2/s)\sin(s/2), \tag{4.3.169}$$

$$\exp(\boldsymbol{t}\cdot\boldsymbol{H})=I\cos(t/2)+(\boldsymbol{t}\cdot\boldsymbol{H})(2/t)\sin(t/2) \tag{4.3.170}$$

where

$$s=(\boldsymbol{s}\cdot\boldsymbol{s})^{1/2},\ \ t=(\boldsymbol{t}\cdot\boldsymbol{t})^{1/2}. \tag{4.3.171}$$

It follows that the G^j and H^k generate bona fide realizations of the group $SU(2)$ rather than the group $SO(3,\mathbb{R})$. See Exercise 3.7.30 for the distinction. Note also the coefficient (3/4) occurring in (3.161) is the same as that in (3.7.203) for $su(2)$, and not that in (3.7.204) for $so(3,\mathbb{R})$.

After this pleasant interruption, let us return to the main discussion. Write $\exp(\boldsymbol{s}\cdot\boldsymbol{G})$ in the Euler angle form

$$\exp(\boldsymbol{s}\cdot\boldsymbol{G})=\exp(\phi G_3)\exp(\theta G_2)\exp(\psi G_3). \tag{4.3.172}$$

Then we may also write

$$\exp(\boldsymbol{s}\cdot\boldsymbol{G})=\{\exp(\phi G_3)\exp(\theta G_2)\exp(-\phi G_3)\}\{\exp[(\phi+\psi)G_3]\}. \tag{4.3.173}$$

We know that $\exp(\theta G_2)$ is J-symmetric and $\exp(\phi G_3)$ is symplectic. Therefore, by Lemma 3.8, the first curly-bracketed factor in (3.173) is J-symmetric. Also, the second curly-bracketed factor in (3.173) is symplectic. Consequently, we have achieved the desired symplectic polar decomposition.

We end this exercise with a few more observations about $SO(4,\mathbb{R})$. Let us write (3.166) in the form

$$O(U,V)=UV \tag{4.3.174}$$

where

$$U(\boldsymbol{s})=\exp(\boldsymbol{s}\cdot\boldsymbol{G}) \tag{4.3.175}$$

and

$$V\boldsymbol{t})=\exp(\boldsymbol{t}\cdot\boldsymbol{H}). \tag{4.3.176}$$

Evidently the matrices U and V form two separate subgroups of $SO(4,\mathbb{R})$ and each of these subgroups has the same topology as $SU(2)$. From (3.169) we see that

$$U(\boldsymbol{s})=-I \text{ when } s=2\pi, \tag{4.3.177}$$

with an analogous result for V. Show that it follows that if U is a matrix of the form (3.175), then so is $-U$. Verify an analogous result for V. We also conclude from (3.174) that

$$O(-U,-V) = O(U,V). \tag{4.3.178}$$

Therefore, (3.174) provides a two-to-one homomorphism of $SU(2) \otimes SU(2)$ onto $SO(4,\mathbb{R})$.

4.3.21. The two previous exercises showed that all elements in $SO(4,\mathbb{R})$ have symplectic polar decompositions. This exercise examines a particular one-parameter subgroup of $SO(4,\mathbb{R})$. Since it is a subgroup of $SO(4,\mathbb{R})$, all its elements must have symplectic polar decompositons. We will apply the methods of Theorem 3.1 to matrices in this subgroup; and in so doing we will discover that the conditions of Theorem 3.1, while sufficient, are not necessary.

Consider a rotation by angle θ in the q_1, q_2 plane. It has the effect

$$q_1' = q_1 c + q_2 s, \tag{4.3.179}$$

$$q_2' = -q_1 s + q_2 c, \tag{4.3.180}$$

$$p_1' = p_1, \tag{4.3.181}$$

$$p_2' = p_2 \tag{4.3.182}$$

where

$$c = \cos\theta \tag{4.3.183}$$

and

$$s = \sin\theta. \tag{4.3.184}$$

Show that in the (q_1, p_1, q_2, p_2) basis this rotation is represented by the matrix $O(\theta)$ given by the relation

$$O(\theta) = \begin{pmatrix} c & 0 & s & 0 \\ 0 & 1 & 0 & 0 \\ -s & 0 & c & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{4.3.185}$$

Seek to write $O(\theta)$ in exponential form. For small θ, and through terms of degree one, show that (3.185) has the expansion

$$O(\theta) = I + \theta A \tag{4.3.186}$$

with

$$A = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{4.3.187}$$

Verify that

$$O(\theta) = \exp(\theta A). \tag{4.3.188}$$

Next, using (3.125) and (3.126), verify that A has the decomposition (3.124) with

$$A^a = (1/2)E^2 = -G^2 \tag{4.3.189}$$

and

$$A^c = (1/2)C^2 = H^2. \tag{4.3.190}$$

Observe from (3.159) that in this case A^a and A^c commute. Verify, therefore, that there is the relation

$$O(\theta) = \exp(\theta A) = \exp[\theta(A^a + A^c)] = \exp(\theta A^a)\exp(\theta A^c). \tag{4.3.191}$$

For future use show that there are the explicit matrix results

$$\exp(\theta A^a) = \exp[(\theta/2)E^2] = I\cos(\theta/2) + E^2\sin(\theta/2) = \begin{pmatrix} c' & 0 & s' & 0 \\ 0 & c' & 0 & -s' \\ -s' & 0 & c' & 0 \\ 0 & s' & 0 & c' \end{pmatrix} \tag{4.3.192}$$

and

$$\exp(\theta A^c) = \exp[(\theta/2)C^2] = I\cos(\theta/2) + C^2\sin(\theta/2) = \begin{pmatrix} c' & 0 & s' & 0 \\ 0 & c' & 0 & s' \\ -s' & 0 & c' & 0 \\ 0 & -s' & 0 & c' \end{pmatrix} \tag{4.3.193}$$

where

$$c' = \cos(\theta/2), \tag{4.3.194}$$

and

$$s' = \sin(\theta/2). \tag{4.3.195}$$

Verify, by explicit matrix multiplication, that (3.191) holds.

Define matrices $Q(\theta)$ and $R(\theta)$ by the rules

$$Q(\theta) = \exp(\theta A^a) = \exp[(\theta/2)E^2] \tag{4.3.196}$$

and

$$R(\theta) = \exp(\theta A^c) = \exp[(\theta/2)C^2] \tag{4.3.197}$$

so that (3.191) can be written in the form

$$O = QR. \tag{4.3.198}$$

Verify that Q is J-symmetric and R is symplectic. You have shown, as expected, that $O(\theta)$ has a symplectic polar decomposition for *all* θ. Verify that $Q(\theta)$ can be written in the form

$$Q(\theta) = \exp[JA'(\theta)] \tag{4.3.199}$$

Find $A'(\theta)$ explicitly and verify that it is real and antisymmetric.

Suppose now that we are just given the matrix M, with

$$M = O(\theta), \tag{4.3.200}$$

and we attempt to find a symplectic polar decomposition for M using the method of Theorem 3.1. Show that

$$N(\lambda M) = \lambda^2 \begin{pmatrix} c & 0 & s & 0 \\ 0 & c & 0 & -s \\ -s & 0 & c & 0 \\ 0 & s & 0 & c \end{pmatrix}. \tag{4.3.201}$$

As a sanity check, verify that

$$N(M) = Q^2 \tag{4.3.202}$$

using (3.196) and the explicit matrix results (3.192) and (3.201).

The next task is to compute the spectral norm of $[N(\lambda M) - I]$. Define the matrix V by the rule

$$V = N(\lambda M) - I = \begin{pmatrix} e & 0 & \lambda^2 s & 0 \\ 0 & e & 0 & -\lambda^2 s \\ -\lambda^2 s & 0 & e & 0 \\ 0 & \lambda^2 s & 0 & e \end{pmatrix} \tag{4.3.203}$$

where

$$e = \lambda^2 c - 1. \tag{4.3.204}$$

Verify that

$$\begin{aligned} V^T V &= \begin{pmatrix} e & 0 & -\lambda^2 s & 0 \\ 0 & e & 0 & \lambda^2 s \\ \lambda^2 s & 0 & e & 0 \\ 0 & -\lambda^2 s & 0 & e \end{pmatrix} \begin{pmatrix} e & 0 & \lambda^2 s & 0 \\ 0 & e & 0 & -\lambda^2 s \\ -\lambda^2 s & 0 & e & 0 \\ 0 & \lambda^2 s & 0 & e \end{pmatrix} \\ &= \begin{pmatrix} e^2+\lambda^4 s^2 & 0 & 0 & 0 \\ 0 & e^2+\lambda^4 s^2 & 0 & 0 \\ 0 & 0 & e^2+\lambda^4 s^2 & 0 \\ 0 & 0 & 0 & e^2+\lambda^4 s^2 \end{pmatrix}. \end{aligned} \tag{4.3.205}$$

Evidently $V^T V$ is diagonal and has the repeated eigenvalue $(e^2 + \lambda^4 s^2)$. Therefore, the matrix $[N(\lambda M) - I]$ has the spectral norm

$$\begin{aligned} ||N(\lambda M) - I|| &= (e^2 + \lambda^4 s^2)^{1/2} = [(\lambda^2 c - 1)^2 + \lambda^4 s^2]^{1/2} \\ &= (1 - 2\lambda^2 c + \lambda^4 c^2 + \lambda^4 s^2)^{1/2} = [1 + \lambda^4 - 2\lambda^2 c]^{1/2}. \end{aligned} \tag{4.3.206}$$

We see that when $c > 0$ there is a $\lambda > 0$ such that $||N(\lambda M) - I|| < 1$. That is, the ray $\lambda^2 N(M)$ passes through the unit ball about I. However, this is not true when $c \leq 0$. That is,

$$||N(\lambda M) - I|| > 1 \text{ when } \lambda > 0 \text{ and } c \leq 0. \tag{4.3.207}$$

We conclude from (3.206) that when $\theta \in (-\pi/2, \pi/2)$ there is a $\lambda > 0$ such that there is the inequality $||N(\lambda M) - I|| < 1$. That is, the ray $\lambda^2 N(M)$ passes through the unit ball about I. However, this is not true for $\theta \notin (-\pi/2, \pi/2)$. That is,

$$||N(\lambda M) - I|| > 1 \text{ when } \lambda > 0 \text{ and } \theta \notin (-\pi/2, \pi/2). \tag{4.3.208}$$

But we know that symplectic polar decomposition is possible for $M = O(\theta)$ for all θ. We have discovered examples where symplectic polar decomposition is possible but the ray $\lambda^2 N(M)$ does not pass through the unit ball about I.

Finally, to study uniqueness, let us compute the spectral norm of $[Q(\theta) - I]$. Similar to what was done in the case of $N(\lambda M)$, now write

$$T = Q - I = \exp(\theta A^a) - I = \begin{pmatrix} e & 0 & s' & 0 \\ 0 & e & 0 & -s' \\ -s' & 0 & e & 0 \\ 0 & s' & 0 & e \end{pmatrix} \tag{4.3.209}$$

where now

$$e = c' - 1. \tag{4.3.210}$$

Here we have used (3.196) and (3.192). Verify that in this case

$$\begin{aligned} V^T V &= \begin{pmatrix} e & 0 & -s' & 0 \\ 0 & e & 0 & s' \\ s' & 0 & e & 0 \\ 0 & -s' & 0 & e \end{pmatrix} \begin{pmatrix} e & 0 & s' & 0 \\ 0 & e & 0 & -s' \\ -s' & 0 & e & 0 \\ 0 & s' & 0 & e \end{pmatrix} \\ &= \begin{pmatrix} e^2 + (s')^2 & 0 & 0 & 0 \\ 0 & e^2 + (s')^2 & 0 & 0 \\ 0 & 0 & e^2 + (s')^2 & 0 \\ 0 & 0 & 0 & e^2 + (s')^2 \end{pmatrix}. \end{aligned} \tag{4.3.211}$$

Evidently $V^T V$ is diagonal and has the repeated eigenvalue $[e^2 + (s')^2]$. Therefore, $(Q - I)$ has the spectral norm

$$\begin{aligned} ||Q - I|| &= [e^2 + (s')^2]^{1/2} = [(c' - 1)^2 + (s')^2]^{1/2} \\ &= [1 - 2c' + (c')^2 + (s')^2]^{1/2} = (2 - 2c')^{1/2}. \end{aligned} \tag{4.3.212}$$

We see that Q lies outside the unit ball about I when $c' < 1/2$. This occurs when $|\theta/2| > 60°$ and therefore $|\theta| > 120°$. So, for $|\theta| < 120°$, there is a symplectic polar decomposition that is unique. But the ray $\lambda^2 N(M)$ lies outside the unit circle about I when $\theta \in (90°, 120°)$ or $\theta \in (-120°, -90°)$. Thus we have found situations where symplectic polar decomposition is possible and unique, but for which the ray $\lambda^2 N(M)$ lies outside the unit circle about I. Finally, we note that $Q(\pm 2\pi) = -I$.

4.3.22. Section 4.3.6 showed that 4×4 diagonal matrices do not have symplectic polar decompositions when $\mu_1\mu_2 < 0$ and $\nu_1\nu_2 < 0$ (and $\mu_1\mu_2 \neq \nu_1\nu_2$). Exercise 3.16 showed that they do when when $\mu_1\mu_2 > 0$ and $\nu_1\nu_2 > 0$. Note that in both cases $\det(M) > 0$. One might wonder if these two cases are joined by a continuous path in $GL(4, \mathbb{R})$. You are to show that they are. Thus there must be some point along the path where symplectic polar decomposition becomes impossible.

Let D be the diagonal matrix given by (3.63), and assume $\mu_1\mu_2 > 0$ and $\nu_1\nu_2 > 0$ so that D has a symplectic polar decomposition. Show that any matrix M sufficiently close to

D must also have a symplectic polar decomposition. Define a continuous family of matrices $M(\theta)$ by the rule

$$M(\theta) = O(\theta)D \tag{4.3.213}$$

where $O(\theta)$ is the matrix given by (3.185). Verify that

$$M(\theta) \in GL(4, \mathbb{R}) \ \text{ for } \ \text{all} \ \ \theta. \tag{4.3.214}$$

Show that

$$M(0) = D = \begin{pmatrix} \mu_1 & 0 & 0 & 0 \\ 0 & \mu_2 & 0 & 0 \\ 0 & 0 & \nu_1 & 0 \\ 0 & 0 & 0 & \nu_2 \end{pmatrix} \tag{4.3.215}$$

and

$$M(\pi) = \begin{pmatrix} -\mu_1 & 0 & 0 & 0 \\ 0 & \mu_2 & 0 & 0 \\ 0 & 0 & -\nu_1 & 0 \\ 0 & 0 & 0 & \nu_2 \end{pmatrix}. \tag{4.3.216}$$

Thus, the continuous path (3.213) joins the two cases. It would be interesting to know where on the path $M(\theta)$ ceases to have a symplectic polar decomposition.

4.3.23. Let H be the subgroup consisting of all elements $g \in GL(2n, R, +)$ such that

$$\{g, J\} = 0. \tag{4.3.217}$$

According Exercise 3.9.18, H is isomorphic to $GL(n, C)$. We know that symplectic polar decomposition is possible for all elements $g \in H$ that are sufficiently near the identity. Is symplectic polar decomposition possible for all elements $g \in H$?

Show that (3.217) implies the commutation relation

$$\{g^T, J\} = 0. \tag{4.3.218}$$

That is, if g commutes with J, so does g^T, and vice versa. Use this result to show that

$$N(g) = gJg^TJ^T = gg^TJJ^T = gg^T. \tag{4.3.219}$$

Does N have a J-symmetric square root? Use the correspondence relation (3.9.19) to obtain the representation

$$g = M(m). \tag{4.3.220}$$

Show, using (3.9.25) and (3.9.21), that there is the result

$$N(g) = gg^T = M(m)M^T(m) = M(m)M(m^\dagger) = M(mm^\dagger). \tag{4.3.221}$$

Verify that $mm^\dagger$ is Hermitian and positive definite. Verify that there is a unitary matrix u such that

$$mm^\dagger = udu^\dagger \tag{4.3.222}$$

where d is diagonal with real positive entries. Define $d^{1/2}$ to be the diagonal matrix whose entries are the positive square roots of the corresponding entries in d. Use this definition to write the relation

$$mm^\dagger = udu^\dagger = ud^{1/2}d^{1/2}u^\dagger = ud^{1/2}u^\dagger ud^{1/2}u^\dagger = (ud^{1/2}u^\dagger)^2. \tag{4.3.223}$$

Show that

$$N(g) = Q^2 \tag{4.3.224}$$

where

$$Q = M(ud^{1/2}u^\dagger). \tag{4.3.225}$$

Is Q J-symmetric? Verify that

$$Q^T = M[(ud^{1/2}u^\dagger)^\dagger] = M(ud^{1/2}u^\dagger) = Q. \tag{4.3.226}$$

Also, verify that

$$\begin{aligned} JQJ^{-1} &= M(iI_n)M(ud^{1/2}u^\dagger)M(-iI_n) \\ &= M[(iI_n)(ud^{1/2}u^\dagger)(-iI_n)] = M(ud^{1/2}u^\dagger) = Q. \end{aligned} \tag{4.3.227}$$

You have shown that Q is J-symmetric, and therefore all elements $g \in H$ have a symplectic polar decomposition.

Consider the ray $\lambda^2 N(g)$. Does it intersect the unit ball around I? Show that

$$\begin{aligned} \lambda^2 N(g) - I &= \lambda^2 M(udu^\dagger) - I = M(u)[\lambda^2 M(d) - I]M(u^\dagger) \\ &= M(u)WD(\lambda)W^{-1}M(u^\dagger) \end{aligned} \tag{4.3.228}$$

where

$$D(\lambda) = \begin{pmatrix} \lambda^2 d - I_n & 0 \\ 0 & \lambda^2 d - I_n \end{pmatrix}. \tag{4.3.229}$$

Use the properties of a matrix norm to show that

$$||\lambda^2 N(g) - I|| \le ||M(u)||\ ||W||\ ||D||\ ||W^{-1}||\ ||M(u^\dagger)||. \tag{4.3.230}$$

Verify that when the spectral norm is used, there are the relations

$$||M(u)|| = ||W|| = ||W^{-1}|| = ||M(u^\dagger)|| = 1. \tag{4.3.231}$$

Consequently, verify that for this norm

$$||\lambda^2 N(g) - I|| \le ||D||. \tag{4.3.232}$$

We are ready for the final step. Since d is diagonal with all entries positive, show that there is a $\lambda_0 > 0$ such that

$$||D(\lambda_0)|| < 1. \tag{4.3.233}$$

Conclude that

$$||\lambda_0^2 N(g) - I|| < 1 \tag{4.3.234}$$

so that the ray $\lambda^2 N(g)$ does indeed intersect the unit ball around I.

4.3.24. Consider matrices M of the form (3.3.10) where C is an arbitrary $n \times n$ real matrix. Show that in this case

$$N(M) = \begin{pmatrix} I & 0 \\ (C - C^T) & I \end{pmatrix} \tag{4.3.235}$$

and therefore

$$[N(M)]^{1/2} = \begin{pmatrix} I & 0 \\ (C - C^T)/2 & I \end{pmatrix}. \tag{4.3.236}$$

Show that such matrices M have a symplectic polar decomposition of the form (3.10) with

$$Q = \begin{pmatrix} I & 0 \\ (C - C^T)/2 & I \end{pmatrix} \tag{4.3.237}$$

and

$$R = \begin{pmatrix} I & 0 \\ (C + C^T)/2 & I \end{pmatrix}. \tag{4.3.238}$$

Carry out an analogous demonstration for matrices of the form (3.3.9).

4.4 Finding the Closest Symplectic Matrix

4.4.1 Background

Let M be any $2n \times 2n$ matrix, and let N be the matrix associated with M by the rule (3.30). Since $N = I$ when M is symplectic, we may define a measure f of the *failure* of M to be symplectic by the rule

$$f = f(M) = \parallel N(M) - I \parallel . \tag{4.4.1}$$

Suppose f is small. Then M is nearly symplectic, and we might hope to find a matrix R that is both near M and exactly symplectic. One way to enforce this nearness condition would be to require the relation

$$\parallel M - R \parallel \sim f. \tag{4.4.2}$$

However, there is also another possibility. If R is close to M, then MR^{-1} is close to the identity I. Consequently, we could equally well require the relation

$$\parallel MR^{-1} - I \parallel \sim f. \tag{4.4.3}$$

Both (4.2) and (4.3) state the hope that if M fails to be symplectic by an amount f, then there should be a symplectic matrix R that is, so to speak, roughly within a distance f from M.

Given (4.1) with $f < 1$, we will show that there is a symplectic R that satisfies both (4.2) and (4.3). Such a matrix R is entitled to be called a *symplectification* of M. Our proof will be based on the results of the previous section. Recall the function $g(x)$ defined by (3.103). Similar to what was done before, use it to define Q by the rule

$$Q = (N)^{1/2} = [I - (I - N)]^{1/2} = I - \sum_{\ell=1}^{\infty} d_\ell (I - N)^\ell. \tag{4.4.4}$$

This series will converge if $f < 1$, and in that case we have the inequality

$$\| Q - I \| = \| \sum_{\ell=1}^{\infty} d_\ell (I - N)^\ell \| \leq \sum_{\ell=1}^{\infty} d_\ell \| I - N \|^\ell \leq \sum_{\ell=1}^{\infty} d_\ell f^\ell \leq f. \tag{4.4.5}$$

We may also define Q^{-1} by the series

$$Q^{-1} = [I - (I - Q)]^{-1} = \sum_{\ell=0}^{\infty} (I - Q)^\ell. \tag{4.4.6}$$

According to (4.5) this series also converges if $f < 1$. Since Q has been defined and is invertible, we may use (3.36) to define a symplectic matrix R and thereby achieve the symplectic polar decomposition (3.10).

Let us use this R and this decomposition to test the relations (4.2) and (4.3). For (4.2) we find the result

$$\| M - R \| = \| QR - R \| = \| (Q - I)R \| \leq \| Q - I \| \| R \| \leq \| R \| f. \tag{4.4.7}$$

Testing the relation (4.3) gives the result

$$\| MR^{-1} - I \| = \| QRR^{-1} - I \| = \| Q - I \| \leq f. \tag{4.4.8}$$

We have learned that if M is such that its failure f to be symplectic satisfies $f < 1$, then it has a symplectic polar decomposition and the factor R in this decomposition provides a symplectification that satisfies the nearness relations (4.7) and (4.8).

Suppose R' is a symplectic matrix that is sufficiently near R in the sense that

$$\| R' - R \| \leq f. \tag{4.4.9}$$

(Because symplectic matrices form a Lie group there are many such R'.) Then we find the result

$$\begin{aligned} \| M - R' \| = \| (M - R) + (R - R') \| \\ \leq \| M - R \| + \| R - R' \| \leq \| R \| f + f = (\| R \| + 1) f, \end{aligned} \tag{4.4.10}$$

and conclude that R' satisfies the nearness requirement (4.2) and hence is also an acceptable symplectification of M. Alternatively, suppose R' is a symplectic matrix that is sufficiently near R in the sense that

$$\| R(R')^{-1} - I \| \leq f. \tag{4.4.11}$$

Then we find the result

$$\begin{aligned} \| M(R')^{-1} - I \| &= \| MR^{-1}R(R')^{-1} - I \| = \| QR(R')^{-1} - I \| \\ &= \| (Q - I) + Q[R(R')^{-1} - I] \| \leq \| Q - I \| + \| Q[R(R')^{-1} - I \| \\ &\leq f + \| Q \| \| R(R')^{-1} - I \| \leq f + \| Q \| f = (\| Q \| + 1) f, \end{aligned} \tag{4.4.12}$$

and conclude that R' satisfies the nearness requirement (4.3) and hence is also an acceptable symplectification of M. We have learned that a matrix M, whose failure f to be symplectic satisfies $f < 1$ in some norm, has many acceptable symplectifications R' that satisfy (4.2) or (4.3) and, in particular, (4.10) or (4.12). Sections 4.5 through 4.8 describe four methods for finding such symplectifications.

Since there are many symplectifications R' that meet our requirements, we may wonder which one is actually *closest* to M. The discussion so far has made only rather general assumptions about the matrix norm $\| * \|$ employed to determine nearness. It has served only as a tool to establish the convergence of various series; but, of course, the quantities defined by these series, if they are defined at all, are independent of the choice of norm. Now, however, we have a more specific question than those discussed above: Imagine we are given a matrix M and we consider all symplectic matrices R'. Is there a closest symplectic matrix R_c that *minimizes* the quantity $\| M - R' \|$? Alternatively, is there a closest symplectic matrix R_c that minimizes the quantity $\| M(R')^{-1} - I \|$? The answers to these questions do depend on the choice of matrix norm.

4.4.2 Use of Euclidean Norm

Let us explore the question of determining the closest symplectic matrix using the nearness condition (4.2) and the Euclidean matrix norm. Consider the set of all (real) $2n \times 2n$ matrices. It obviously forms a linear vector space under the operations of scalar multiplication and matrix addition. Let A and B be any two vectors (matrices) in this space. We define an inner product between them by the rule

$$(A, B) = \text{ tr } (A^T B). \tag{4.4.13}$$

It is easily verified that this rule satisfies all the requirements for a positive-definite inner product. (See Exercise 4.3.) Let O' and O'' denote any $2n \times 2n$ orthogonal matrices. Then it can also be shown that the inner product (4.13) is invariant under the action of the orthogonal group in the sense that

$$(O'AO'', O'BO'') = (A, B). \tag{4.4.14}$$

Next, as in Exercise 3.7.1, we define the *Euclidean* norm $\| M \|_E$ of any matrix M by the rule

$$\| M \|_E = (M, M)^{1/2}. \tag{4.4.15}$$

This rule satisfies all the conditions (3.7.10) through (3.7.14) required for a norm. The Euclidean norm is not particularly powerful for establishing convergence in some circumstances because it gives for the $2n \times 2n$ identity matrix the result

$$\| I \|_E = (2n)^{1/2}. \tag{4.4.16}$$

By contrast there are more powerful norms, the maximum column sum norm (3.7.15) and spectral norm (3.7.17) for examples, that give the optimal result

$$\| I \| = 1. \tag{4.4.17}$$

However, as a consequence of (4.14), the Euclidean norm does have the convenient feature that

$$\| O'MO'' \|_E = \| M \|_E \tag{4.4.18}$$

where O' and O'' are any orthogonal matrices.

It can be shown that if M is any matrix, then the *orthogonal* matrix O that is closest to M, in the sense of minimizing $\| M - O \|_E$, is given by the orthogonal matrix appearing in the polar decomposition (2.7). The situation with regard to the closest symplectic matrix is more complicated. One might entertain the analogous conjecture that the *symplectic* matrix R that is closest to any symplectifiable M, in the sense of minimizing $\| M - R \|_E$, is given by the symplectic matrix appearing in the symplectic polar decomposition (3.10). However, this conjecture is wrong.

As a counter example in the 2 × 2 case, consider the matrix M given by the relation

$$M = \mu K. \tag{4.4.19}$$

Here K is a symplectic diagonal matrix of the form

$$K = \begin{pmatrix} k & 0 \\ 0 & k^{-1} \end{pmatrix}, \tag{4.4.20}$$

and we assume that $\det(M) > 0$ so that μ has the value

$$\mu = +[\det(M)]^{1/2}. \tag{4.4.21}$$

For this M we find from (3.61) the result

$$R = K. \tag{4.4.22}$$

Next, let X be the symplectic diagonal matrix

$$X(x) = \begin{pmatrix} x & 0 \\ 0 & x^{-1} \end{pmatrix}. \tag{4.4.23}$$

Let us examine whether there is a choice of x such that $\| M - X \|_E$ has a value smaller than $\| M - K \|_E$. To make such a study, consider the function $h(x)$ defined by the relation

$$h(x) = [\| M - X \|_E]^2. \tag{4.4.24}$$

Does h have a minimum at $x = k$? From the definitions (4.13) and (4.15) we find that

$$h(x) = \text{tr } [(M - X)^T (M - X)] = (\mu k - x)^2 + (\mu k^{-1} - x^{-1})^2. \tag{4.4.25}$$

Suppose we differentiate h with respect to x and evaluate the result at $x = k$. Doing so gives the result

$$h'(k) = 2(\mu - 1)(k^{-3} - k). \tag{4.4.26}$$

We see that in general $h'(k) \neq 0$. It follows that, for this example, setting $X = K = R$ does not minimize $\| M - X \|_E$.

One can also construct counter examples for which the symplectic matrix R produced by the symplectic polar decomposition of M does not give the symplectic matrix closest to M in the sense of minimizing $\| M(R')^{-1} - I \|_E$. We also remark that if the transpose operation in (4.13) is omitted, which produces a different inner product about to be discussed, these conclusions remain unchanged. See Exercise 4.5.

4.4.3 Geometric Interpretation of Symplectic Polar Decomposition

Although the symplectic matrix R produced by the symplectic polar decomposition of M does not necessarily give the symplectic matrix closest to M as defined by either the nearness condition (4.2) or (4.3) and the use of some inner product norm, one might still wonder if it has some other *geometric* interpretation. It does, but some further concepts need to be developed to show that this is the case. We note that the nearness condition (4.2) is related to the matrix operation of *addition* (actually, in this case, subtraction) while the nearness condition (4.3) is related to the operation of matrix *multiplication*. We will explore the use of nearness conditions related to *group* properties. Since group properties are based on the operation of matrix multiplication, these nearness conditions are similar in spirit to the condition (4.3).

Consider the group $GL(2n, \mathbb{R})$. Near the identity any matrix M can be written in the form

$$M = \exp(B) \tag{4.4.27}$$

where B, an arbitrary matrix of $g\ell(2n, \mathbb{R})$, has the decomposition (3.1). Let B_0, B_1, $\cdots$ be a set of basis vectors (matrices) for this space. There are $(2n)^2$ such matrices. For example, for the simplest case $n = 1$, a convenient basis is given by the choice

$$B_0 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{4.4.28}$$

$$B_1 = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}, \tag{4.4.29}$$

$$B_2 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{4.4.30}$$

$$B_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{4.4.31}$$

We note that the first 3 basis matrices, B_0 through B_2, are of the form JS [see (5.6.7), (5.6.13), and (5.6.14)], and the last is of the form JA with $A = -J$:

$$B_3 = J(-J) = I. \tag{4.4.32}$$

In terms of the basis provided by the B_ℓ, any matrix in $g\ell(2n, \mathbb{R})$ can be written in the form

$$B(b) = \sum_\ell b^\ell B_\ell, \tag{4.4.33}$$

where the b^ℓ are real, but otherwise arbitrary, coefficients. We may view the b^ℓ as the entries of a vector b in a $(2n)^2$ dimensional vector space. Correspondingly, we may view (4.27) and (4.33) as a mapping between points in this vector space and elements in the group $GL(2n, \mathbb{R})$ near the identity. Put another way, this mapping and the entries in b constitute a *coordinate patch* for $GL(2n, \mathbb{R})$ at the identity. Alternatively, we may regard the b^ℓ as the *components* of the vectors in the *tangent space* of $GL(2n,\mathbb{R})$ at the identity. Consequently, the vectors b are in and span the *cotangent* space of $GL(2n,\mathbb{R})$ at the identity.

Next, suppose b and b' are any two vectors. Let us introduce an inner product between them by the rule

$$(b,b') = \text{tr}[B(b)B(b')] = \text{tr}[\sum_{\ell\ell'} b^{\ell}(b')^{\ell'} B_{\ell}B_{\ell'}] = \sum_{\ell\ell'} b^{\ell}(b')^{\ell'}\text{tr}(B_{\ell}B_{\ell'}). \tag{4.4.34}$$

This inner product can be expressed in the form

$$(b,b') = \sum_{\ell\ell'} b^{\ell}(b')^{\ell'} g_{\ell\ell'} \tag{4.4.35}$$

where g is the *metric tensor*

$$g_{\ell\ell'} = \text{tr}(B_{\ell}B_{\ell'}). \tag{4.4.36}$$

In general, this metric tensor is *not* positive definite. For example, use of the basis given by (4.28) through (4.31) in the case $n = 1$ gives the result

$$g = \begin{pmatrix} -2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}. \tag{4.4.37}$$

Although the metric g is not positive definite, it does have two attractive features. The first feature is this: Suppose b is a vector such that $B(b)$ is a matrix of the form JS, and b' is a vector such that $B(b')$ is a matrix of the form JA. Then b and b' are orthogonal,

$$(b,b') = \text{tr}[B(b)B(b')] = \text{tr}(JSJA) = 0. \tag{4.4.38}$$

To verify this assertion, we make the observation

$$\text{tr}(JSJA) = \text{tr}[(JSJA)^T] = \text{tr}(A^TJ^TS^TJ^T) = \text{tr}(-AJSJ) = -\text{tr}(JSJA). \tag{4.4.39}$$

A description of the second attractive feature requires some background discussion. Suppose M^1 is some matrix, not necessarily near the identity, and we wish to examine matrices near M^1. To do so, we might consider all matrices ${}^LM^1$ of the form

$${}^LM^1(b) = M^1\exp[B(b)] \tag{4.4.40}$$

where $B(b)$ is again given by (4.33). The relation (4.40) provides a coordinate patch for $GL(2n,\mathbb{R})$ at the point M^1. Indeed, looking at (4.40), we may say that this coordinate patch is obtained by *translating* the coordinate patch at the identity to the point M^1. This translation may be called *left* translation since (4.40) involves multiplication by M^1 on the left. [Hence the notation ${}^LM^1$ in (4.40).] Alternatively, we may view the entries in b as the components of the vectors in the tangent space of $GL(2n,\mathbb{R})$ at the point M^1, and vectors in this tangent space are to be regarded as associated with those at the identity by the operation of left translation.

There is an obvious alternative to (4.40). Namely, we might equally well use *right* translation to examine matrices near M^1 by considering matrices ${}^RM^1$ of the form

$$ {}^RM^1(c) = \{\exp[B(c)]\}M^1. \tag{4.4.41}$$

Here the entries in c may again be viewed as the components of vectors in the tangent space of $GL(2n, \mathbb{R})$ at the point M^1, but vectors in this tangent space are now to be regarded as associated with those at the identity by the operation of right translation.

Suppose both (4.40) and (4.41) are used to represent the same element,

$$ {}^LM^1(b) = {}^RM^1(c). \tag{4.4.42}$$

Then, using (4.40) and (4.41), we find the equation

$$M^1 \exp[B(b)] = \{\exp[B(c)]\}M^1, \tag{4.4.43}$$

which is essentially a relation between b and c. Indeed, this relation may be rewritten in the form

$$\exp[B(c)] = M^1\{\exp[B(b)]\}(M^1)^{-1} = \exp\{M^1[B(b)](M^1)^{-1}\}, \tag{4.4.44}$$

from which we conclude that

$$B(c) = M^1[B(b)](M^1)^{-1}. \tag{4.4.45}$$

Also, since the matrices B_ℓ form a basis, we must have relations of the form

$$M^1B_\ell(M^1)^{-1} = \sum_{\ell'} d_{\ell\ell'}(M^1)B_{\ell'} \tag{4.4.46}$$

where the $d_{\ell\ell'}(M^1)$ are coefficients that depend on M^1. Correspondingly, we find the result

$$M^1[B(b)](M^1)^{-1} = \sum_\ell b^\ell M^1 B_\ell (M^1)^{-1} = \sum_{\ell\ell'} b^\ell d_{\ell\ell'} B_{\ell'} = \sum_{\ell'}(\sum_\ell b^\ell d_{\ell\ell'})B_{\ell'}, \tag{4.4.47}$$

from which we conclude that c is given in terms of b by the equation

$$c^{\ell'} = \sum_\ell b^\ell d_{\ell\ell'}. \tag{4.4.48}$$

Now a question arises: We have seen how (4.34) can be used to define an inner product between two vectors b and b' whose entries are the components for the two vectors in the tangent space at the identity. What can be done for pairs of tangent vectors at the general point M^1? These tangent-vector pairs can be associated with either of the coordinate pairs b, b' or c, c' depending on whether left or right translation is used. We will define their inner products to be the same as those for their counterparts at the origin,

$$(b, b')^L = \text{tr}[B(b)B(b')], \tag{4.4.49}$$

$$(c, c')^R = \text{tr}[B(c)B(c')], \tag{4.4.50}$$

where we have used the superscripts L and R to indicate that either left or right translation has been used to make a correspondence between the tangent space at the identity and the tangent space at the general point M^1.

However, we know that b and c are related by (4.45), and the same is true for b' and c'. Consequently, from (4.45), (4.49), and (4.50), we find the relation

$$\begin{aligned}(c,c')^R &= \mathrm{tr}[B(c)B(c')] = \mathrm{tr}\{M^1B(b)(M^1)^{-1}M^1B(b')(M^1)^{-1}\}\\ &= \mathrm{tr}\{M^1B(b)B(b')(M^1)^{-1}\} = \mathrm{tr}[B(b)B(b')] = (b,b')^L.\end{aligned} \tag{4.4.51}$$

We have learned that the inner product definition (4.34) has the feature that its extension from the identity to an arbitrary point M^1 is *independent* of whether left or right translations are used.

Now that an inner product has been defined on the tangent (and cotangent) spaces at any point in $GL(2n,\mathbb{R})$, we can discuss the *lengths* of paths in matrix space. Let M^0 and M^1 be any two matrices. Suppose they are joined by a path $M(\tau)$, where τ is a parameter lying in the range [0,1],

$$M(0) = M^0, \tag{4.4.52}$$

$$M(1) = M^1. \tag{4.4.53}$$

Consider the two nearby points $M(\tau)$ and $M(\tau{+}d\tau)$ on the path. Suppose we view $M(\tau{+}d\tau)$ as being related to $M(\tau)$ by right translation of elements near the identity. That is, we write a relation of the form

$$M(\tau+d\tau) = \exp[d\tau C(\tau)]M(\tau) = \{I + d\tau C(\tau) + O[(d\tau)^2]\}M(\tau). \tag{4.4.54}$$

At this juncture we note that (4.54) can be rewritten in the form

$$M(\tau+d\tau)M^{-1}(\tau) - I = d\tau C(\tau) + O[(d\tau)^2], \tag{4.4.55}$$

and we see that the left side of (4.55) is reminiscent of the nearness condition (4.3). Now make the Taylor expansion

$$M(\tau+d\tau) = M(\tau) + (dM/d\tau)d\tau + O[(d\tau)^2 = M(\tau) + \dot{M}(\tau)d\tau + O[(d\tau)^2]. \tag{4.4.56}$$

Upon comparing (4.54) and (4.56), we may solve for $C(\tau)$, which we may view as the tangent vector to the path at the point $M(\tau)$, to find the result

$$C(\tau) = \dot{M}(\tau)M^{-1}(\tau). \tag{4.4.57}$$

Let us define an *energy functional* $E[M]$ associated with any path $M(\tau)$ by the relation

$$E[M] = (1/2)\int_0^1 d\tau\ \mathrm{tr}[C(\tau)C(\tau)]. \tag{4.4.58}$$

[Note that this definition employs the inner product (4.34).] Then an *affine geodesic* in matrix space is defined to be a path ${}^{ag}M(\tau)$ that extremizes the energy functional. For a discussion of geodesics and affine geodesics, see Exercise 1.6.17.

Why should these definitions interest us? Suppose M^0 and M^1 are close in the sense that $[M^1(M^0)^{-1}]$ is near the identity. Then there exists a matrix B such that

$$M^1(M^0)^{-1} = \exp(B) \tag{4.4.59}$$

or

$$M^1 = \exp(B)M^0. \tag{4.4.60}$$

Consider the particular path $M(\tau)$ given by the rule

$$M(\tau) = \exp(\tau B)M^0. \tag{4.4.61}$$

Evidently this path satisfies (4.52) and (4.53), and therefore joins M^0 and M^1. Moreover, this path satisfies the differential equation

$$\dot{M}(\tau) = B[\exp(\tau B)]M^0 = BM(\tau), \tag{4.4.62}$$

and consequently by (4.57) has the *constant* tangent vector

$$C(\tau) = B. \tag{4.4.63}$$

Thus, in this sense, the path (4.61) in matrix space is the analog of a *straight line* in Euclidean space, which also has a constant tangent. But even more can be said about this analogy. The path (4.61) is also an affine geodesic! See Exercise 4.6.

Let us now apply these general considerations to the problem at hand. Suppose that M is a matrix that meets the condition of Theorem 3.1. Define a path $M(\tau)$ in matrix space by the rule

$$M(\tau) = \exp(\tau JA)R, \tag{4.4.64}$$

where R is the symplectic matrix in the factorization (3.10) and A is defined by (3.57). Evidently this path joins R, the symplectic factor of M, to M itself,

$$M(0) = R, \tag{4.4.65}$$

$$M(1) = M. \tag{4.4.66}$$

See Figure 4.1. Comparison of (4.61), (4.63), and (4.64) shows that this path has the constant tangent vector JA. Consider the point R at which the path $M(\tau)$ meets the group of symplectic matrices. We know that any vector in the tangent space of the group of symplectic matrices is of the form JS. We see from (4.39) that the path $M(\tau)$ is *perpendicular* to the subspace of symplectic matrices at the point R. Finally, we know that the path $M(\tau)$ is an affine geodesic. We conclude that R has the special geometric property that it is connected to M by a path that [in terms of the tangent-space metric (4.36)] is both an affine geodesic and is perpendicular to the subspace of symplectic matrices at the point R.

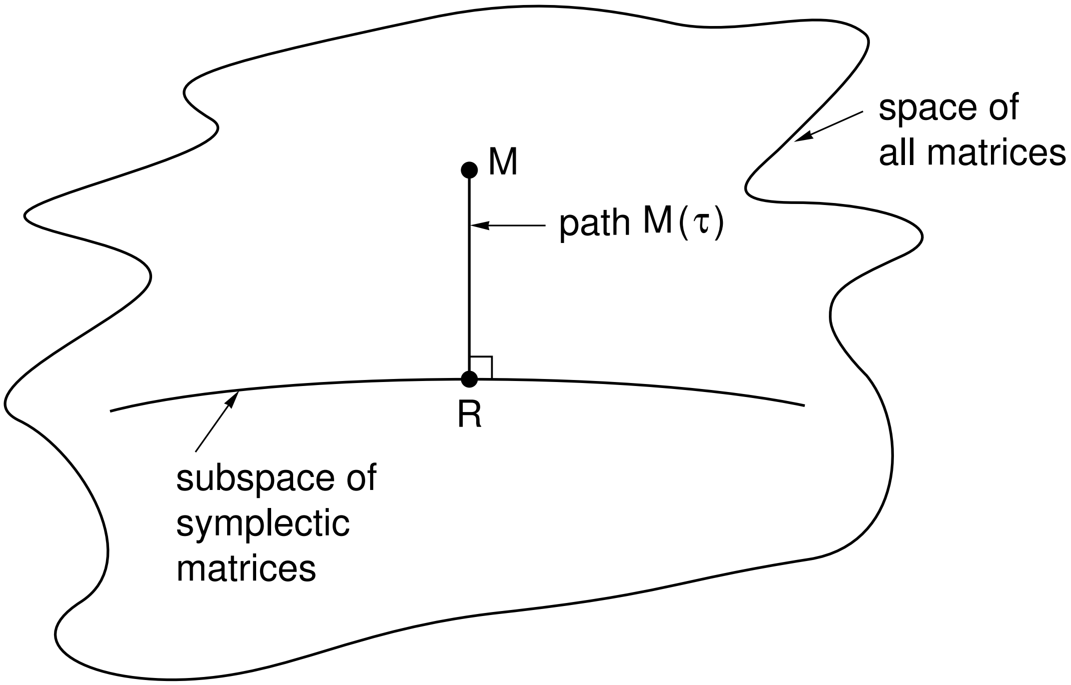


Figure 4.4.1: The matrices R and M are connected by a path that is both an affine geodesic and is perpendicular to the subspace of symplectic matrices at the point R.

Exercises

4.4.1. Verify that the rule (4.13) satisfies all the requirements for an inner product including the positive-definite conditions

$$(A, A) \geq 0, \tag{4.4.67}$$

$$(A, A) = 0 \Leftrightarrow A = 0. \tag{4.4.68}$$

4.4.2. Verify (4.14) through (4.18).

4.4.3. If you have not already worked Exercise 3.7.1, verify that (4.15) satisfies all the conditions (3.7.10) through (3.7.14) required for a matrix norm.

4.4.4. Verify (4.25) and (4.26).

4.4.5.

4.4.6. Suppose that, instead of using (4.57), which involves right translation, the tangent vector $C(\tau)$ is defined by the left-translation relation

$$C(\tau) = M^{-1}(\tau)\dot{M}(\tau). \tag{4.4.69}$$

Show that the value of the energy functional $E[M]$ given by (4.58) remains unchanged.

The remainder of this exercise is devoted to showing that $M(\tau)$ as given by (4.61) is an affine geodesic. To do so, we will need to evaluate $E[M]$ for paths near (4.61) and show that, through first order, $E[M]$ remains unchanged when small changes are made about the path (4.61). Parameterize paths near (4.61) by writing

$$M(\epsilon,\tau) = \exp[\epsilon F(\tau)]\exp(\tau B)M^0 = \exp[\epsilon F(\tau)]M(\tau), \tag{4.4.70}$$

where $F(\tau)$ is an arbitrary matrix function save for the end-point conditions

$$F(0) = F(1) = 0. \tag{4.4.71}$$

Evaluate $E(\epsilon)$ for paths of the form (4.70) and show that

$$(dE/d\epsilon)|_{\epsilon=0} = 0. \tag{4.4.72}$$

Hints: Using (4.57) and (4.70), show that

$$C(\tau) = C_0 + \epsilon C_1 + O(\epsilon^2) \tag{4.4.73}$$

where

$$C_0 = B, \tag{4.4.74}$$

$$C_1 = \dot{F}(\tau) + \{F(\tau), B\}. \tag{4.4.75}$$

Next show that

$$E = E_0 + \epsilon E_1 + O(\epsilon^2) \tag{4.4.76}$$

where

$$E_1 = \int_0^1 d\tau\ \mathrm{tr}(C_0C_1). \tag{4.4.77}$$

Finally, show that

$$E_1 = 0. \tag{4.4.78}$$

For extra credit, suppose B is of the form JA as in (4.64), $F(0)$ is of the form JS, and $F(1) = 0$. Show that (4.78) still holds in this case, and give a geometrical interpretation of this fact in terms of Figure 4.1.

4.4.7. Consider using the positive-definite inner product (4.13) instead of the indefinite inner product (4.34). See Exercise 4.1. Show that in this case matrix pairs of the form JS and JA are again orthogonal,

$$\mathrm{tr}[(JS)^TJA] = \ \mathrm{tr}(S^TJ^TJA) = \ \mathrm{tr}(SA) = 0. \tag{4.4.79}$$

Use (4.13) to define an energy functional by writing

$$E[M] = (1/2)\int_0^1 d\tau\ \mathrm{tr}[C^T(\tau)C(\tau)]. \tag{4.4.80}$$

Following the discussion in Exercise 4.6, show that in this case

$$E_1 = \int_0^1 d\tau\ \mathrm{tr}(C_0^TC_1) \tag{4.4.81}$$

with C_0 and C_1 given by (4.74) and (4.75) as before. Show that in this case E_1 has the value

$$E_1 = \int_0^1 d\tau \ \mathrm{tr}[\{B, B^T\}F], \tag{4.4.82}$$

and that the necessary and sufficient condition for E_1 to vanish for all F satisfying (4.71) is the requirement

$$\{B, B^T\} = 0. \tag{4.4.83}$$

For the case of the path (4.64), B is the matrix given by the relation

$$B = JA. \tag{4.4.84}$$

Verify that (4.83) holds in the case of $g\ell(2, \mathbb{R})$, but need not be true in the cases of $g\ell(4, \mathbb{R})$, $g\ell(6, \mathbb{R})$, etc.

4.4.8. Consider orthogonal polar decompositions of the form (2.7). Suppose M is invertible. Show that there exists a real symmetric matrix S such that M can be written in the form

$$M = \exp(S)O. \tag{4.4.85}$$

See Exercise 2.3. It follows that O and M can be joined by the path

$$M(\tau) = \exp(\tau S)O, \tag{4.4.86}$$

with constant tangent vector S. We know that the orthogonal matrices form a group, and that any vector in the tangent space of this group at any point in the group is of the form A where A is an antisymmetric matrix. Show that matrix pairs of the form S and A are orthogonal for both the inner products (4.13) and (4.34). Show that $M(\tau)$ as given by (4.86) is an affine geodesic for both the energy functionals (4.58) and (4.80). Show that the length of this affine geodesic is the same independent of whether (4.57) or (4.69) is used to define the tangent vector $C(\tau)$.

4.5 Symplectification Using Symplectic Polar Decomposition

We are now prepared to discuss various symplectification processes. The first uses symplectic polar decompostion. Closely related is an iterative procedure. If successful, it produces the same result as symplectic polar decomposition. We will begin with a review of the process and properties of symplectic polar decomposition, and then proceed to describe how and when iteration may be used to obtain the same results.

4.5.1 Properties of Symplectification Using Symplectic Polar Decomposition

Let M denote any $2n \times 2n$ matrix. Consider the mapping $\mathcal{S}$ of the space of such matrices into itself defined by the rule

$$\mathcal{S}(M) = (MJM^TJ^T)^{-1/2}M = [N(M)]^{-1/2}M. \tag{4.5.1}$$

Here we have used (3.30). Moreover, to define $N^{-1/2}$ we will write

$$\begin{aligned} N^{-1/2} &= [I + (N - I)]^{-1/2} = I + \sum_{\ell=1}^{\infty} e_\ell (N - I)^\ell \\ &= I - (1/2)(N - I) + (3/8)(N - I)^2 - \cdots \end{aligned} \tag{4.5.2}$$

and assume that in some norm (4.1) is satisfied with $f < 1$ in order to ensure convergence. Of course, this assumption places some restrictions on the domain of $\mathcal{S}$.

Suppose R' and R'' are any two symplectic matrices. Then we find from (3.33) the result

$$[N(R'MR'') - I] = R'[N(M) - I](R')^{-1}, \tag{4.5.3}$$

and hence

$$[N(R'MR'') - I]^\ell = R'[N(M) - I]^\ell (R')^{-1}. \tag{4.5.4}$$

Consequently, since matrix multiplication and infinite summation can be interchanged, we find from (5.2) the result

$$[N(R'MR'')]^{-1/2} = R'[N(M)]^{-1/2}(R')^{-1}. \tag{4.5.5}$$

See Exercise 5.1. It follows from the definition (5.1) that $\mathcal{S}$ has the property

$$\begin{aligned} \mathcal{S}(R'MR'') &= R'[N(M)]^{-1/2}(R')^{-1}(R'MR'') \\ &= R'\mathcal{S}(M)R''. \end{aligned} \tag{4.5.6}$$

As a special case of (5.6) we have the result

$$\mathcal{S}(MR'') = \mathcal{S}(M)R''. \tag{4.5.7}$$

We note that this result can be proved directly without concern about the effect of interchanging the operations of matrix multiplication and infinite summation since we have as a special case of (3.33) the relation

$$N(MR'') = N(M). \tag{4.5.8}$$

Next suppose Q' is any (invertible) J-symmetric matrix. Then we find the result

$$\mathcal{S}(Q') = [Q'J(Q')^T J^T]^{-1/2} Q' = [(Q')^2]^{-1/2} Q' = I. \tag{4.5.9}$$

Finally, suppose M has the symplectic polar decomposition (3.10). Then, using (5.7) and (5.9), we find the result

$$\mathcal{S}(M) = \mathcal{S}(QR) = \mathcal{S}(Q)R = R. \tag{4.5.10}$$

Consequently, as one might expect from (3.38) and (3.39), the map $\mathcal{S}$ is a *symplectifying* map that sends M into the symplectic factor in its symplectic polar decomposition.

There are three properties of the symplectifying map $\mathcal{S}$ provided by symplectic polar decomposition that are worth noting. First, we have the result

$$\mathcal{S}(R) = R \tag{4.5.11}$$

for *any* symplectic matrix R. Thus, if M is already symplectic, the map $\mathcal{S}$ given by (5.1) leaves M in peace. The second property is that already stated in (5.6): Suppose the matrix M is given left and right symplectic translations by sending it to the matrix $(R'MR'')$, and this translated matrix is then symplectified using symplectic polar decomposition. Equation (5.6) states that the result is the same as that obtained by first symplectifying M and then giving the symplectified M the same translations. We may say that the the symplectification process provided by $\mathcal{S}$ is *invariant* under left and right *symplectic translations.* Finally, suppose M has the symplectic polar decomposition (3.10). Then we find for M^{-1} the result

$$M^{-1} = R^{-1}Q^{-1} = (R^{-1}Q^{-1}R)R^{-1} = \acute{Q}R^{-1}. \tag{4.5.12}$$

By Lemmas 3.5 and 3.8 the matrix $\acute{Q} = (R^{-1}Q^{-1}R)$ is J-symmetric. Therefore we have the result

$$\mathcal{S}(M^{-1}) = R^{-1} = [\mathcal{S}(M)]^{-1}. \tag{4.5.13}$$

We may say that the the symplectification process provided by $\mathcal{S}$ is also *invariant* under *inversion.*

4.5.2 Iteration

Suppose we define a map $\mathcal{S}_1$, related to $\mathcal{S}$, by retaining only the $\ell = 1$ term in the series appearing in (5.2). This map has the definition

$$\begin{aligned}\mathcal{S}_1(M) &= [I - (1/2)(N - I)]M = (1/2)[3I - N(M)]M \\ &= (1/2)(3I - MJM^TJ^T)M.\end{aligned} \tag{4.5.14}$$

It is readily verified that $\mathcal{S}_1$ also satisfies a relation of the form (5.6),

$$\mathcal{S}_1(R'MR'') = R'\mathcal{S}_1(M)R'', \tag{4.5.15}$$

and now, because the series (5.2) has been truncated, there is no concern about convergence.

Now let Q' be any J-symmetric matrix. Using (3.9) and (5.14), we find that

$$\mathcal{S}_1(Q') = (3/2)Q' - (Q')^3/2. \tag{4.5.16}$$

It follows from this result and Lemmas 3.2 and 3.5 that $\mathcal{S}_1$ maps the space of J-symmetric matrices into itself. In addition note that $Q' = I$ is a fixed pont of $\mathcal{S}_1$,

$$\mathcal{S}_1(I) = I. \tag{4.5.17}$$

Let us examine the nature of the fixed point $Q' = I$. To do so, write Q' in the form

$$Q' = I + W, \tag{4.5.18}$$

where W is "small". By Lemmas 3.1 and 3.2, $W = Q' - I$ is also J-symmetric if Q' is J-symmetric. Upon inserting the form (5.18) into (5.16), we find the result

$$\mathcal{S}_1(Q') = S_1(I + W) = I - (3/2)W^2 - (1/2)W^3. \tag{4.5.19}$$

At this point it is convenient to introduce the map $\mathcal{U}_1$ defined on J-symmetric matrices by the rule

$$\mathcal{U}_1(W) = S_1(I + W) - I. \tag{4.5.20}$$

From this definition it follows, by combining (5.20) with (5.19), that

$$\mathcal{U}_1(W) = -(3/2)W^2 - (1/2)W^3. \tag{4.5.21}$$

Evidently, $\mathcal{U}_1$ has the fixed point $W = 0$, which corresponds precisely to the fixed point $Q' = I$ of $\mathcal{S}_1$.

To exploit this correspondence, define *translation* maps $\mathcal{T}$ and $\mathcal{T}^{-1}$ by the rules

$$\mathcal{T}(W) = W + I, \tag{4.5.22}$$

$$\mathcal{T}^{-1}(W) = W - I. \tag{4.5.23}$$

We these definitions, we have the relations

$$\mathcal{U}_1 = \mathcal{T}^{-1}\mathcal{S}_1\mathcal{T}, \tag{4.5.24}$$

$$\mathcal{S}_1 = \mathcal{T}\mathcal{U}_1\mathcal{T}^{-1}. \tag{4.5.25}$$

From (5.25) we see that

$$\mathcal{S}_1^m = \mathcal{T}\mathcal{U}_1^m\mathcal{T}^{-1}, \tag{4.5.26}$$

and conclude that the behavior of $\mathcal{S}_1^m$ on J-symmetric matrices of the form $Q' = I + W$ is governed by the behavior of $\mathcal{U}_1^m$ on the matrices W. Moreover, since the right side of (5.21) is quadratic in W, we expect that $W = 0$ will be an *attractor* of $\mathcal{U}_1$, and correspondingly $Q' = I$ will be an *attractor* of $\mathcal{S}_1$.

An estimate of the basin of attraction of $\mathcal{U}_1$ can be obtained by requiring that

$$\| \mathcal{U}_1(W) \| = \| -(3/2)W^2 - (1/2)W^3 \| < \| W \| . \tag{4.5.27}$$

This condition is difficult to work with, and we will use instead a poorer estimate. Suppose we require that

$$[(3/2) \| W \|^2 + (1/2) \| W \|^3] < \| W \| . \tag{4.5.28}$$

By the properties (3.7.11) through (3.7.13) of a norm we will then have the result

$$\| -(3/2)W^2 - (1/2)W^3 \| < \| W \| . \tag{4.5.29}$$

Consequently, W that satisfy (5.28) will lie in the basin of $W = 0$. It is easily verified that (5.28) is equivalent to the condition

$$\| W \| < (-3/2 + (1/2)\sqrt{17}) \simeq (.56). \tag{4.5.30}$$

We conclude that if $\| W \| < .56$, then we have the result

$$\lim_{m\to\infty} \mathcal{U}_1^m(W) = 0. \tag{4.5.31}$$

That is, repeated application (iteration) of $\mathcal{U}_1$ will drive such W to 0. Moreover, in view of (5.21), once convergence gets underway it will be quadratic and therefore very rapid.

We next show that W as given by (5.18) satisfies the inequality

$$\parallel W \parallel \leq f, \tag{4.5.32}$$

where f is given by (4.1) and N is defined in terms of Q'. When $M = Q'$, we find from (3.27) the result

$$N(Q') = Q'J(Q')^T J^T = (Q')^2. \tag{4.5.33}$$

Consequently, following (4.4) and (4.5), we may write the relations

$$Q' = (N)^{1/2} = [I - (I - N)]^{1/2} = I - \sum_{\ell=1}^{\infty} d_\ell (I - N)^\ell, \tag{4.5.34}$$

$$\parallel W \parallel = \parallel Q' - I \parallel \leq f. \tag{4.5.35}$$

We conclude that if $f < (.56)$, then we again have the result (5.31). Correspondingly, we also have the result

$$\lim_{m\to\infty} \mathcal{S}_1^m(Q') = I. \tag{4.5.36}$$

We are now ready for the master stroke. Suppose M is some matrix whose failure f to be symplectic satisfies $f < (.56)$. Then since $f < 1$, we know that such a matrix has the symplectic polar decomposition (3.10), and that [according to (4.5)] the J-symmetric factor Q of M must satisfy the relation

$$\parallel Q - I \parallel \leq (.56). \tag{4.5.37}$$

Let us compute the matrices $\mathcal{S}_1^m(M)$ for successive values of m. From (3.10) and (5.15) we find the result

$$\mathcal{S}_1^m(M) = \mathcal{S}_1^m(QR) = \mathcal{S}_1^m(Q)R. \tag{4.5.38}$$

Now take the limit $m \to \infty$. In view of (5.36), doing so gives the result

$$\lim_{m\to\infty} \mathcal{S}_1^m(M) = R. \tag{4.5.39}$$

We see that repeated application (iteration) of $\mathcal{S}_1$ drives M to its symplectification R. Since $\mathcal{S}_1$ is simple to evaluate, see (5.14), and the convergence is very rapid, we conclude that this iterative method is well suited to numerical computation.

As an example of how well the iterative method works, consider the 2 × 2 case. In this case W as defined by (5.18) must be a multiple of the identity matrix so that we may write

$$W = wI \tag{4.5.40}$$

and

$$\mathcal{T}_1(W) = -[(3/2)w^2 + (1/2)w^3]I. \tag{4.5.41}$$

Exhibit 5.1 below shows successive values of w given by the recursion relation

$$w_{n+1} = -(3/2)(w_n)^2 - (1/2)(w_n)^3 \tag{4.5.42}$$

for various initial conditions w_0. Evidently the convergence is very rapid as expected.

Exhibit 4.5.1: Convergence of symplectification by iteration in the 2×2 case. Successive values of w_n for various initial conditions w_0.

```
 n   wn
 0   0.1000000000000000
 1  -1.5500000000000000E-02
 2  -3.5851306250000000E-04
 3  -1.9277438384305627E-07
 4  -5.5742941017167727E-14
 5  -4.6609132098651603E-27
 6   0.0000000000000000E+00

 0  -0.1000000000000000
 1  -1.4500000000000000E-02
 2  -3.1385068750000000E-04
 3  -1.4773792356625795E-07
 4  -3.2739739477203758E-14
 5  -1.6078358115527613E-27
 6   0.0000000000000000E+00

 0   0.6000000000000000
 1  -0.6480000000000000
 2  -0.4938071040000000
 3  -0.3055618747255026
 4  -0.1257872293112257
 5  -2.2738510956402836E-02
 6  -7.6968146227769021E-04
 7  -8.8838634673523115E-07
 8  -1.1838451010276436E-12
 9  -2.1022338348398979E-24
10   0.0000000000000000E+00

 0  -0.6000000000000000
 1  -0.4320000000000000
 2  -0.2396252160000000
 3  -7.9250697011794843E-02
 4  -9.1721356097234983E-03
 5  -1.2580629042388899E-04
 6  -2.3739838483241409E-08
 7  -8.4536989012593641E-16
 8  -1.0719753766973067E-30
 9   0.0000000000000000E+00
```

At this point at least two thoughts come to mind. First, it would be nice to have a procedure that would work whenever $f < 1$ rather than the condition $f < (.56)$, which is more restrictive. Of course, the map $\mathcal{S}$ does meet this requirement; but its use requires summing the infinite series (5.2), which may be only slowly convergent. It is easily verified that the series for $\mathcal{S}(M)$ and $\mathcal{S}(M^{\rm tr})$ have the same convergence properties when M and $M^{\rm tr}$ are related by a condition of the form (3.33) for some symplectic matrix R. Note that (3.33) defines an *equivalence relation.* (For the definition of an equivalence relation, see Exercise 5.12.7.) It follows that the convergence of the series for $\mathcal{S}(M)$ depends only on the equivalence class to which M belongs. The same is true for the convergence of the sequence $\mathcal{S}_1^m(M)$. From (5.15) we see that its behavior also depends only on the equivalence class to which M belongs. We note that we have proved that $f < (.56)$ is sufficient to ensure convergence of the sequence $\mathcal{S}_1^m(M)$. However, there may be equivalence classes for which it is not necessary. For example, in the $2n \times 2n$ case, one equivalence class consists of matrices Q' of the form (5.18) with W given by (5.40) with I now being the $2n \times 2n$ identity matrix. It can be shown that the fixed point $w = 0$ of the sequence (5.42) has a larger basin of attraction than that given by the condition $|w| < (.56)$. See Exercise 5.3. Indeed, examination of Exhibit 5.1 shows that convergence occurs when $|w| = (.60)$.

A second thought that comes to mind concerns the properties of maps $\mathcal{S}_k$ produced by discarding in the series (5.2) all terms beyond $\ell = k$. They have properties analogous to those of $\mathcal{S}_1$, and they can also be iterated to produce R. What would be their basins of attraction and their rates of convergence? See Exercise 5.4 for a discussion of the properties of $\mathcal{S}_2$.

Although these questions may be interesting, they do not seem to be of practical importance for problems encountered to date. That is, the condition $f < (.56)$ always seems to be well satisfied in practice whenever a symplectification is required. Consequently, for the present, we will not pursue these questions further.

Exercises

4.5.1. Verify the expansion (5.2) and compute the first few coeffcients e_ℓ. Show that the series $\sum e_\ell x^\ell$ has a radius of convergence of 1. Verify that the series (5.2) converges in norm when $f < 1$, and therefore verify (5.6).

4.5.2. Verify (5.14) through (5.26). Verify the steps that led from (5.27) to (5.30).

4.5.3. Consider the map $\mathcal{M}$ given by (5.42). Show that it has the four fixed points

$$w^f = -2, -1, 0, \pm\infty \tag{4.5.43}$$

where the points $\pm\infty$ are to be identified in a manner similar to the way that all points at infinity are identified by use of the Riemann sphere. Examine the stability of each. You should find that $w^f = -1$ is unstable, and the rest are stable. Show that $\bar{w}$ defined by the equation

$$\bar{w} = -1 + \sqrt{3} \approx .732 \tag{4.5.44}$$

is the positive root of the cubic equation

$$\bar{w}^3 + 3\bar{w}^2 = 2. \tag{4.5.45}$$

Show that the open interval $w \in (-1, \bar{w})$ is in the basis of attraction of the fixed point $w^f = 0$, and points just outside the interval are not. Show that $\mathcal{M}$ sends the two endpoints of the interval into the unstable fixed point $w^f = -1$. Make a numerical study of the w axis to see if there are any other points in the basin of attraction of $w^f = 0$. You should find, for example, that points near $w = -3$ are in the basin of $w^f = 0$. Color the w axis in three colors depending on whether a point on the axis is in the basin of $-2, 0$, or $\pm\infty$. As already illustrated in Figure 1.2.8, the basin of an attracting fixed point can have disjoint pieces.

4.5.4. Study the properties of $\mathcal{S}_2$.

4.5.5. Suppose x, p_x, y, p_y, t, p_t is a set of canonical coordinates as in Exercise 1.6.1. With this order of variables the J' of (3.2.10) should be used. In this context a matrix M is called *static* if it is of the form

$$M = \begin{pmatrix} * & * & * & * & 0 & * \\ * & * & * & * & 0 & * \\ * & * & * & * & 0 & * \\ * & * & * & * & 0 & * \\ * & * & * & * & 1 & * \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{4.5.46}$$

Here the entries denoted by * are arbitrary. Later, in Chapter 7, it will be evident that static symplectic matrices are related to Lie transformations generated by quadratic polynomials f_2 with the property $\partial f_2/\partial t = 0$. Show that if M is a static (but not necessarily symplectic) matrix, the result of symplectifying M by iteration is a static symplectic matrix. That is, symplectification by iteration preserves the property of being static.

4.5.6. Show that if Q is an invertible J-symmetric matrix, then so is Q^T. Review Section 4.5.1. Show that under the transposition operation the symplectifying map $\mathcal{S}$ also has the invariance property

$$\mathcal{S}(M^T) = [\mathcal{S}(M)]^T. \tag{4.5.47}$$

4.5.7. Section 4.5.2 studied symplectification by iteration. This exercise explores orthogonalization by iteration. Suppose M is any real matrix with nonzero determinant and we wish to make the polar decomposition (2.7). In analogy to the symplectic case, define a matrix function $N(M)$ by the rule

$$N(M) = MM^T. \tag{4.5.48}$$

Also, in analogy to (5.1), define a mapping $\mathcal{S}$ by the rule

$$\mathcal{S}(M) = [N(M)]^{-1/2}M. \tag{4.5.49}$$

Show, using (2.8) through (2.10), that

$$\mathcal{S}(M) = O. \tag{4.5.50}$$

Again compute $N^{-1/2}$ using (5.2). Define, in analogy to (5.14), the map $\mathcal{S}_1$ by the rule

$$\mathcal{S}_1(M) = (1/2)(3I - MM^T)M. \tag{4.5.51}$$

Show that

$$\mathcal{S}_1(O'MO'') = O'\mathcal{S}_1(M)O'' \tag{4.5.52}$$

where O' and O'' are any two orthogonal matrices. Show that $\mathcal{S}_1$ maps the set of symmetric matrices into itself, and that

$$\mathcal{S}_1(I) = I \tag{4.5.53}$$

so that $M = I$ is a fixed point of $\mathcal{S}_1$. Show that, on the set of symmetric matrices, this fixed point is an attractor of $\mathcal{S}_1$. Show that if $N(M)$ is sufficiently near I, then

$$\lim_{m\to\infty} \mathcal{S}_1^m(M) = O \tag{4.5.54}$$

where O is the orthogonal factor appearing in the polar decomposition (2.7).

4.6 Modified Darboux Symplectification

Suppose one is given a matrix M whose determinant is nonzero. The columns of M may be regarded as vectors m^1, m^2, m^3, $\cdots$, and the condition $\det(M) \neq 0$ is equivalent to the statement that the vectors m^j are linearly independent. Given a set of linearly independent vectors m^j, there is the Darboux process for constructing an associated set of symplectic vectors r^j. Finally, the vectors r^j may be viewed as the columns of a matrix R, and this matrix will be symplectic. Thus, given any nonsingular matrix M, there is a procedure for constructing a corresponding symplectic matrix R. Moreover, if M itself is nearly symplectic, then R will be near M. Indeed, if M happens to be symplectic, then R will coincide with M. See Sections 3.6.3 and 3.6.5. In this section we will describe what we will call *modified* Darboux symplectification, and will examine how close R is to M if M is nearly symplectic.

Let M be a $2n \times 2n$ matrix. Rather than using (4.1), we will describe the *failure* of M to be symplectic in terms of an antisymmetric matrix F defined by the relation

$$F = M^T JM - J. \tag{4.6.1}$$

From (4.1) and (6.1) we have the result

$$\begin{aligned} \parallel F \parallel &= \parallel (M^T JMJ^T - I)J \parallel \leq \parallel M^T JMJ^T - I \parallel \parallel J \parallel \\ &\leq \parallel M^T JMJ^T - I \parallel = \parallel M^T J(M^T)^T J^T - I \parallel = f(M^T). \end{aligned} \tag{4.6.2}$$

Here we have assumed that the matrix norm employed has the property

$$\parallel J \parallel = 1, \tag{4.6.3}$$

which is true for the maximum column sum norm (3.7.15) and the spectral norm (3.7.17). If the norm also has the property (3.7.97), which we shall also assume, then the matrix elements of F are bounded by the relation

$$|F_{jk}| \leq f(M^T). \tag{4.6.4}$$

Suppose we view M as a collection of column vectors m^1, m^2, $\cdots$ m^{2n}. Let m_i^j denote the ith component of the jth such vector. Then, following the usual matrix element labelling scheme, we have the relation

$$m_i^j = M_{ij}. \tag{4.6.5}$$

In terms of the vectors m^j, the relation (6.1) can be rewritten in the form

$$(M^T JM)_{jk} = (m^j, Jm^k) = J_{jk} + F_{jk}. \tag{4.6.6}$$

Correspondingly if R is a symplectic matrix and we view it as a collection of column vectors r^j, then the symplectic condition (3.1.2) can be written in the form

$$(r^j, Jr^k) = J_{jk}. \tag{4.6.7}$$

Assume we are given an M for which $f(M^T)$ is sufficiently small. From M we extract the vectors m^j using (6.5). Our task is to use these m^j, which obey (6.6), to construct a set of vectors r^j that obey (6.7). Moreover, this construction is to be made in such a way that the corresponding symplectic matrix R is near M in the sense that

$$\parallel M - R \parallel \sim f(M^T). \tag{4.6.8}$$

We will construct the vectors r^j two at a time, beginning with r^1 and r^2. To simplify our presentation, we will use a J matrix of the form (3.2.10). For this choice we have the relation

$$J_{12} = 1, \tag{4.6.9}$$

and (6.6) gives the result

$$(m^1, Jm^2) = 1 + F_{12}. \tag{4.6.10}$$

According to (6.4) and (6.10), if $f(M^T)$ is sufficiently small, the quantity (m^1, Jm^2) will be positive and hence will have a positive square root γ_{12},

$$\gamma_{12} = +[(m^1, Jm^2)]^{1/2}. \tag{4.6.11}$$

We can therefore define “normalized” vectors r^1 and r^2 by the rules

$$r^1 = m^1/\gamma_{12}, \tag{4.6.12}$$

$$r^2 = m^2/\gamma_{12}. \tag{4.6.13}$$

Note that by (6.4) and (6.10), γ_{12} will be near 1 if $f(M^T)$ is sufficiently small. Correspondingly r^1 and r^2 will be near m^1 and m^2, respectively. By construction, these vectors satisfy the relation

$$(r^1, Jr^2) = 1 = J_{12}, \tag{4.6.14}$$

as required by (6.7). Also, because J is antisymmetric, we automatically get from (6.14) the relations

$$(r^j, Jr^k) = J_{jk} \text{ when } j = 1, 2 \text{ and } k = 1, 2, \tag{4.6.15}$$

as is also required by (6.7).

Next we construct the vectors r^3 and r^4. We begin by defining intermediate vectors s^3 and s^4 according to the rule

$$s^3 = m^3 + \alpha_{31}r^1 + \alpha_{32}r^2, \tag{4.6.16}$$

$$s^4 = m^4 + \alpha_{41}r^1 + \alpha_{42}r^2, \tag{4.6.17}$$

where the α's are coefficients still to be determined. According to (6.7) we must have the relations

$$(r^j, Jr^k) = 0 \text{ when } j = 1, 2 \text{ and } k = 3, 4. \tag{4.6.18}$$

Let us therefore require the relations

$$(r^j, Js^k) = 0 \text{ when } j = 1, 2 \text{ and } k = 3, 4. \tag{4.6.19}$$

Doing so determines the values of the coefficients α:

$$\alpha_{31} = (r^2, Jm^3), \tag{4.6.20}$$

$$\alpha_{32} = -(r^1, Jm^3), \tag{4.6.21}$$

$$\alpha_{41} = (r^2, Jm^4), \tag{4.6.22}$$

$$\alpha_{42} = -(r^1, Jm^4). \tag{4.6.23}$$

If $f(M^T)$ is sufficiently small then, according to (6.4), (6.6), (6.11) through (6.13), and (6.20) through (6.23), all the α's are of order $f(M^T)$. It also follows that the quantity (s^3, Js^4) will be positive and consequently will have the positive square root

$$\gamma_{34} = +[(s^3, Js^4)]^{1/2}. \tag{4.6.24}$$

Finally, we define the normalized vectors r^3 and r^4 by the rules

$$r^3 = s^3/\gamma_{34}, \tag{4.6.25}$$

$$r^4 = s^4/\gamma_{34}. \tag{4.6.26}$$

Upon reflection we see that we have now constructed four vectors r^1 through r^4 that are, respectively, near m^1 through m^4 if $f(M^T)$ is small; and these vectors satisfy the relations

$$(r^j, Jr^k) = J_{jk} \text{ when } j, k = 1, 2, 3, 4. \tag{4.6.27}$$

Moreover the general pattern is now clear. We see that the construction can be continued to include r^5 and r^6 (and still more r's if we are dealing with more than a 6-dimensional phase space). We simply write the analogs of (6.16) and (6.17), for example

$$s^5 = m^5 + \alpha_{51}r^1 + \alpha_{52}r^2 + \alpha_{53}r^3 + \alpha_{54}r^4, \tag{4.6.28}$$

$$s^6 = m^6 + \alpha_{61}r^1 + \alpha_{62}r^2 + \alpha_{63}r^3 + \alpha_{64}r^4, \tag{4.6.29}$$

determine the α's, and then normalize the results. Finally, we may view all the r^j we have constructed in this manner as the columns of a matrix R. This matrix will be symplectic and will be close to M in the sense of satisfying (6.8).

There is one last nuisance to be resolved. All our estimates have involved the quantity $f(M^T)$ whereas it would be more pleasant to work with $f(M)$. This defect can be overcome by using the modified Darboux procedure just described to symplectify the matrix M^T instead of M. Call the resulting symplectic matrix R'. Using (3.7.51) and (6.8) we will then have the result

$$|(M^T)_{jk} - R'_{jk}| \sim f(M). \tag{4.6.30}$$

Finally we define R, which is to be the symplectification of M, by writing

$$R = (R')^T. \tag{4.6.31}$$

Combining (6.30) and (6.31) then gives the desired result

$$|M_{jk} - R_{jk}| \sim f(M). \tag{4.6.32}$$

Exercises

4.6.1. Show that F as defined by (6.1) is antisymmetric.

4.6.2. Refer to Exercise 5.4. Show that modified Darboux symplectification also preserves the property of being static.

4.7 Exponential and Cayley Symplectifications

Both the exponential and Cayley representations of a matrix provide additional methods for matrix symplectification. We will first describe the use of the exponential representation. Subsequently we will consider the use of the Cayley representation, which is based on the exponential representation.

4.7.1 Exponential Symplectification

As before, let M be a (real) $2n \times 2n$ matrix. Consider, in matrix space, the ray λM where λ lies in the range $0 < \lambda < \infty$. Suppose that for some value λ_0 the matrix $\lambda_0 M$ lies *within* the unit ball about I. [The geometric picture for this situation is similar to that of Figure 4.1 except that the ray $N(\lambda M)$ is replaced by the ray λM.] Then M can be written in the exponential form

$$M = \exp(B) \tag{4.7.1}$$

where B is a real matrix. The proof for this assertion is straightforward: Since by hypothesis $\lambda_0 M$ lies within the unit ball about I, the series of the form (3.43) for $\log(\lambda_0 M)$ converges, and we may write

$$\lambda_0 M = \exp[\log(\lambda_0 M)]. \tag{4.7.2}$$

It follows that M can be written in the form

$$M = [(\lambda_0)^{-1} I][\lambda_0 M] = \exp[-I \log(\lambda_0)] \exp[\log(\lambda_0 M)] = \exp(B) \tag{4.7.3}$$

where B is defined by the equation

$$B = \log(\lambda_0 M) - I\log(\lambda_0). \tag{4.7.4}$$

It is now a simple matter to find a symplectification R for M. Without loss of generality, the matrix B can be written in the form (3.1) where S and A are uniquely defined. We simply take R to be the symplectic matrix given by the relation

$$R = \exp(JS). \tag{4.7.5}$$

4.7.2 Cayley Symplectification

The symplectification provided by (7.5) has the defect that it requires the summation of the infinite exponential series. Although this problem can be overcome by the method of Section 4.1, it is worthwhile to explore other possibilities. Suppose M can be written in the exponential form (7.1). Then we may write the relations

$$\begin{aligned} M &= \exp(B) = [\exp(B/2)]/[\exp(-B/2)] \\ &= [\cosh(B/2) + \sinh(B/2)]/[\cosh(B/2) - \sinh(B/2)] \\ &= [I + \tanh(B/2)]/[I - \tanh(B/2)]. \end{aligned} \tag{4.7.6}$$

Define a matrix T by the equation

$$T = \tanh(B/2). \tag{4.7.7}$$

With the aid of T, M as given by (7.6) has the Cayley representation

$$M = (I + T)(I - T)^{-1} = (I - T)^{-1}(I + T). \tag{4.7.8}$$

The relation (7.8) can be solved for T to given the result

$$T = (M + I)^{-1}(M - I) = (M - I)(M + I)^{-1}. \tag{4.7.9}$$

Now view (7.9) as the *definition* of T in terms of M. That is, this definition can be made without any reference to B. Define the matrix V by the equation

$$V = J^{-1}T. \tag{4.7.10}$$

We know that V will be symmetric if M is symplectic, and vice versa. See Section 3.11. Consequently, V will be nearly symmetric if M is nearly symplectic. Let us define a symmetric matrix W by taking the symmetric part of V,

$$W = (V + V^T)/2. \tag{4.7.11}$$

Then we may define a symplectic matrix R by writing

$$R = (I + JW)(I - JW)^{-1} = (I - JW)^{-1}(I + JW), \tag{4.7.12}$$

and R will be a symplecification of M that we will call the *Cayley* symplectification. Note that while the evaluation of (7.5) requires the summation of an infinite series, the evaluation of (7.9) and (7.12) requires only matrix inversion.

Let us view R, the result of this Cayley symplectification process applied to M, as the outcome of a Cayley symplectifying map $\mathcal{S}_C$ applied to M,

$$R = \mathcal{S}_C(M). \tag{4.7.13}$$

Then it is easily verified that Cayley symplectification has the feature

$$\mathcal{S}_C(M^{-1}) = [\mathcal{S}_C(M)]^{-1}. \tag{4.7.14}$$

That is, Cayley symplectification, like symplectic polar decomposition symplectification, is invariant under inversion. See (5.13). Moreover, suppose $\acute{R}$ is any symplectic matrix. Then it can be shown that Cayley symplectification has the feature

$$\mathcal{S}_C(\acute{R}M\acute{R}^{-1}) = \acute{R}[\mathcal{S}_C(M)]\acute{R}^{-1}. \tag{4.7.15}$$

We may say that Cayley symplectification is invariant under symplectic similarity transformation. This property, although weaker than and a special case of the symplectic translational invariance described by (5.6), is still significant.

4.7.3 Cayley Symplectification Near the Identity

Cayley symplectification is particularly useful near the identity. Consider the problem of evaluating $\exp(\epsilon JS)$ where ϵ is small and S is symmetric and may itself have the form of a power series in ϵ beginning with constant terms. As discussed at the beginning of this chapter, such is the problem in evaluating linear transformations of the form $\exp(: k_2 :)$ where k_2 arises solely from nonlinear feed-down effects. See Chapter 9. According to (7.6) we have the result

$$R = \exp(\epsilon JS) = [I + \tanh(\epsilon JS/2)][I - \tanh(\epsilon JS/2)]^{-1}. \tag{4.7.16}$$

The hyperbolic tangent function has the Taylor expansion

$$\begin{aligned} \tanh(\epsilon JS/2) &= \sum_{\ell=1}^{\infty} a_\ell(\epsilon JS/2)^\ell = (\epsilon JS/2) - (1/3)(\epsilon JS/2)^3 + (2/15)(\epsilon JS/2)^5 \\ &- (17/315)(\epsilon JS/2)^7 + (62/2835)(\epsilon JS/2)^9 - \cdots . \end{aligned} \tag{4.7.17}$$

Note that the coefficients a_ℓ vanish for even ℓ.[4] Suppose we *truncate* the series (7.17) by omitting terms beyond $\ell = k$, and use this truncated series to define a matrix W_t by the relation

$$W_t = J^{-1}\sum_{\ell=1}^{k} a_\ell(\epsilon JS/2)^\ell. \tag{4.7.18}$$

[4] It is tempting to regard (7.16) through (7.18) as a diagonal Padé approximate to the exponential function. However, it is not. For example, the 3,3 diagonal Padé approximate (approximation through cubic terms in the numerator and denominator) for the exponential function has different coefficients. In particular, it contains both even and odd powers: $\exp(z) \simeq (1 + z/2 + z^2/10 + z^3/120)/(1 - z/2 + z^2/10 - z^3/120)$.

Let us use W_t to define the matrix R_a, which will be an *approximation* to the matrix R, by the equation

$$R_a = (I + JW_t)(I - JW_t)^{-1} = (I - JW_t)^{-1}(I + JW_t). \tag{4.7.19}$$

It is easily verified that W_t is a symmetric matrix, and hence R_a will be symplectic. Moreover, R_a will be near to R in the sense of satisfying relations of the form

$$\parallel R - R_a \parallel \sim \epsilon^{k+2}, \tag{4.7.20}$$

$$\parallel R(R_a)^{-1} - I \parallel \sim \epsilon^{k+2}. \tag{4.7.21}$$

We conclude that the use of (7.18) and (7.19) is well suited to the calculation of $\exp(\epsilon JS)$ where S, although symmetric, is only known through some power in some smallness parameter ϵ. Correspondingly, in the language of and as will be needed for Chapter 9, this method is well suited to the calculation of $\exp(: k_2 :)$ when k_2 itself is only known through some power in some smallness parameter ϵ.

Exercises

4.7.1. Show that the two factors in (7.8) commute as indicated. Show the same for the two factors in (7.9).

4.7.2. Verify the invariance properties (7.14) and (7.15).

4.7.3. Show that W_t as given by (7.18) is symmetric.

4.7.4. Verify the estimates (7.20) and (7.21).

4.8 Generating Function Symplectification

It is well known that canonical transformations (symplectic maps as defined in Section 6.1) can be produced by the method of mixed-variable generating functions, often referred to as F_1 through F_4. The generating functions are called *mixed* because they involve both "old" and "new" variables. In this section we will outline how quadratic mixed-variable generating functions can be used to symplectify matrices. See section 6.5 for a more extensive discussion of the mixed-variable generating functions F_1 through F_4.[5]

Since the method of generating functions does not treat coordinate and momentum variables on a common footing, it is convenient to introduce the notation

$$z = (q_1 \cdots q_n, p_1 \cdots p_n), \tag{4.8.1}$$

[5] We remark that the adjective *generating* often occurs in an "infinitesimal" context" in the sense that one says that Lie algebras generate Lie groups or Hamiltonians generate symplectic maps. That is, generation involves some sort of "exponentiation/integration" process. By contrast, in the case of mixed-variable generating functions, results are immediate with no need to pass from the infinitesimal to the finite. Still, there is no free lunch. The complexity of exponentiation/integration is replaced by the complexity of making initially implicit relations explicit.

$$Z = (Q_1 \cdots Q_n, P_1 \cdots P_n). \tag{4.8.2}$$

Let R be a symplectic matrix that maps z to Z according to the rule

$$Z = Rz. \tag{4.8.3}$$

Then, under certain conditions, the transformation (8.3) can be produced by a mixed-variable generating function.

For example, let us attempt to use a generating function of the second kind, $F_2(q, P)$. Its use gives the implicit equations

$$p_\ell = \partial F_2/\partial q_\ell, \tag{4.8.4}$$

$$Q_\ell = \partial F_2/\partial P_\ell. \tag{4.8.5}$$

In view of (8.4) and (8.5), and since the relation (8.3) is linear, we will consider a quadratic generating function. The most general such function (of the second kind) can be written in the form

$$F_2(q, P) = (1/2)\sum_{i,j} \alpha_{ij} q_i q_j + \sum_{i,j} \beta_{ij} q_i P_j + (1/2)\sum_{i,j} \delta_{ij} P_i P_j, \tag{4.8.6}$$

where the matrices α and δ are symmetric,

$$\alpha^T = \alpha, \tag{4.8.7}$$

$$\delta^T = \delta, \tag{4.8.8}$$

and the matrix β is arbitrary. (Soon, however, we will require that β be invertible. Also, here the matrix δ is not to be confused with the Kronecker delta.)

Applying the rules (8.4) and (8.5) to this F_2 gives the set of implicit equations

$$p = \alpha q + \beta P, \tag{4.8.9}$$

$$Q = \beta^T q + \delta P. \tag{4.8.10}$$

These equations may be made explicit to give the relations

$$P = -\beta^{-1}\alpha q + \beta^{-1} p, \tag{4.8.11}$$

$$Q = (\beta^T - \delta\beta^{-1}\alpha) q + \delta\beta^{-1} p. \tag{4.8.12}$$

(Here we have assumed that β is invertible.) Suppose R is written in the $n \times n$ block form

$$R = \begin{pmatrix} A & B \\ C & D \end{pmatrix}. \tag{4.8.13}$$

Then comparison of (8.3) with (8.11) and (8.12) gives the relations

$$A = \beta^T - \delta\beta^{-1}\alpha, \tag{4.8.14}$$

$$B = \delta\beta^{-1}, \tag{4.8.15}$$

$$C = -\beta^{-1}\alpha, \tag{4.8.16}$$

$$D = \beta^{-1}. \tag{4.8.17}$$

These relations may be solved for the matrices α, β, and δ to give the results

$$\alpha = -D^{-1}C, \tag{4.8.18}$$

$$\beta = D^{-1}, \tag{4.8.19}$$

$$\delta = BD^{-1}. \tag{4.8.20}$$

We conclude that a necessary condition for (8.3) to be produced by an $F_2(q, P)$ is that the matrix D be invertible. Moreover, it is easily checked that the matrices A through D given by (8.14) through (8.17) satisfy the symplectic conditions (3.3.3) through (3.3.5). Consequently, both the necessary and sufficient condition for the linear symplectic transformation (8.3) to be produced by the F_2 defined in (8.6) is that the D matrix associated with R be invertible.

We momentarily interrupt our discussion to observe for future use that the relations (8.13) through (8.17), which relate R to the matrices α, β, and δ, can be written in a more compact form. Let W be the *symmetric* matrix defined by the equation

$$W = \begin{pmatrix} \alpha & \beta \\ \beta^T & \delta \end{pmatrix}, \tag{4.8.21}$$

and define matrices E through H by the rules

$$E = \begin{pmatrix} 0 & 0 \\ I & 0 \end{pmatrix}, \tag{4.8.22}$$

$$F = \begin{pmatrix} 0 & I \\ 0 & 0 \end{pmatrix}, \tag{4.8.23}$$

$$G = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}, \tag{4.8.24}$$

$$H = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}. \tag{4.8.25}$$

Here, as with R, all blocks in W and in the matrices E through H are $n \times n$. With these definitions, it can be verified that R can be written in terms of W in the compact form

$$R = (FW + G)(EW + H)^{-1}. \tag{4.8.26}$$

Equation (8.26) is an example of symplectic and symmetric matrices being related by a Möbius transformation. See Section 5.13 for further discussion of this topic.

To continue our discussion of symplectification, suppose M is an arbitrary $2n \times 2n$ matrix written in the $n \times n$ block form

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}. \tag{4.8.27}$$

Let us seek to symplectify M. First we make the restriction that d is invertible, and define a matrix β by the rule

$$\beta = d^{-1}. \tag{4.8.28}$$

Next, following (8.18), we form the matrix $(-d^{-1}c)$ and define a matrix α by taking its symmetric part,

$$\alpha = -[(d^{-1}c) + (d^{-1}c)^T]/2. \tag{4.8.29}$$

Also, following (8.20), we form the matrix (bd^{-1}) and define a matrix δ by taking its symmetric part,

$$\delta = [(bd^{-1}) + (bd^{-1})^T]/2. \tag{4.8.30}$$

Finally, from the α, β, and δ matrices just defined, we construct the matrices A through D given by (8.14) through (8.17). In so doing we have constructed a *symplectic* matrix R of the form (8.13), and this matrix may be taken to be a symplectification of M.

There are also the generating functions $F_1(q,Q)$, $F_3(p,Q)$, and $F_4(p,P)$. They too can be used for symplectification in ways analogous to that described for F_2. We close this section by noting that there are nearly symplectic matrices that cannot be symplectified by using any of the generating functions F_1 through F_4. For example the matrix R given by

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \tag{4.8.31}$$

is symplectic, but cannot be produced by any of the generating functions $F_1(q,Q)$ through $F_4(p,P)$. Correspondingly, there are nonsymplectic matrices M near R that cannot be symplectified by use of the generating functions $F_1(q,Q)$ through $F_4(p,P)$. However, there are other mixed-variable generating functions that can be used. See Section 6.7.4.

Exercises

4.8.1. Verify the relations (8.9) through (8.20). Show that the symplectic conditions (3.3.3) through (3.3.5) are satisfied.

4.8.2. Verify (8.26). Hint: Along the way you will have to verify the relation

$$(EW + H)^{-1} = \begin{pmatrix} I & 0 \\ -\beta^{-1}\alpha & \beta^{-1} \end{pmatrix}. \tag{4.8.32}$$

4.8.3. Referring to (8.27) through (8.30), work out explicitly the relations giving the matrices A through D in terms of the matrices a through d. Show that R coincides with M if M is symplectic, and is near M if M is nearly symplectic.

4.8.4. Verify that the matrix R given by (8.31) is symplectic. Verify that the matrices A through D that compose R as in (8.13) have the properties

$$\det(A) = \det(B) = \det(C) = \det(D) = 0, \tag{4.8.33}$$

and therefore fail to have inverses.

Bibliography

Taylor Series with Remainder

[1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover (1972). Also available on the Web by Googling "abramowitz and stegun 1972".

[2] F. Olver, D. Lozier, R. Boisvert, and C. Clark, Editors, *NIST Handbook of Mathematical Functions*, Cambridge (2010). See also the Web site http://dlmf.nist.gov/.

Matrix Exponentiation

[3] C. Moler and C. Van Loan, "Nineteen Dubious Ways to Compute the Exponential of a Matrix", *SIAM Review* **20**, 801 (1978); "Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later", *SIAM Review* **45**, 3 (2003).

[4] C. Kenney and A. Laub, "A Schur-Fréchet Algorithm for Computing the Logarithm and Exponential of a Matrix", to appear in *SIAM Journal of Matrix Analysis and Applications* (1998).

[5] N. Higham, *Functions of Matrices: Theory and Computation*, SIAM (2008).

[6] E. Celledoni and A. Iserles, "Methods for the approximation of the matrix exponential in a Lie-algebraic setting", arXiv:math/9904122v1 (2008).

(Orthogonal) Polar Decomposition

[7] F.R. Gantmacher, *The Theory of Matrices*, Vols. One and Two, Chelsea (1959).

[8] N. Jacobson, *Lectures in Abstract Algebra, Vol. II – Linear Algebra*, p. 188, D. Van Nostrand (Princeton, 1953).

[9] J.B. Keller, "Closest Unitary, Orthogonal and Hermitian Operators to a Given Operator", *Mathematics Magazine* **48**, p. 192 (1975).

[10] N. J. Higham, D. S. Mackey, N. Mackey, and F. Tisseur, "Computing the Polar Decomposition and the Matrix Sign Decomposition in Matrix Groups", *SIAM J. of Matrix Analysis and Appl.*, v.25, No. 4, pp. 1178 - 1192 (2004).

[11] N. Higham, *Functions of Matrices: Theory and Computation*, SIAM (2008).

Analyticity

[12] The treatment of analyticity in part d of Exercise 2.2 was suggested by R. Goodman.

Symplectification by Iteration

[13] M. Furman, "Simple Method to Symplectify Matrices", SSC Central Design Group Report SSC-TM-4001 (1985).

Modified Darboux Symplectification

[14] F. Neri invented what we have called modified Darboux symplectification, and incorporated it in the code MaryLie circa 1986.

Generating Function Symplectification

[15] For a description of the mixed-variable generating functions F_1 through F_4, see H. Goldstein, *Classical Mechanics*, Addison-Wesley (1980). See also H.D. Block, *J. of Mathematics and Physics* **32**, p. 207 (1953-54).

[16] For a description of how mixed-variable generating functions can be used to symplectify matrices arising in the context of Accelerator Physics, see D.R. Douglas, "Interpolation of Off-Energy Matrices: Symplectic and Otherwise (A Comparison of Methods)", SSC Central Design Group Report SSC-TM-4003 (1985).

Chapter 5

Preliminary Lie Concepts for Classical Mechanics and Related Delights

In this chapter we will begin a study of the Lie algebraic structure of Classical Mechanics. We will learn about Lie operators and Lie transformations, and how they can be used to represent the symplectic group. We will see that the symplectic group is related to the quaternion field just as the orthogonal group and the unitary group are related to the real and complex fields. We will also find that there is a close connection between symplectic and symmetric matrices. Along the way we will learn something about Cartan's method for understanding the nature of simple Lie algebras and their representations, Clebsch-Gordan series, the topology of $Sp(2n)$, Siegel and homogeneous spaces, Möbius transformations, and Lagrangian planes.

5.1 Properties of the Poisson Bracket

The Poisson bracket has already been defined in Section 1.7. The purpose of this section is to review its properties. Suppose that f and g are any two functions of the variables q, p, t. We recall that the *Poisson bracket* of f and g, denoted by the symbol $[f, g]$, is defined by the equation

$$[f,g] = \sum_i [(\partial f/\partial q_i)(\partial g/\partial p_i) - (\partial f/\partial p_i)(\partial g/\partial q_i)]. \tag{5.1.1}$$

We also recall from Section 1.7 that it is convenient to introduce the $2n$ variables $(z_1 \ldots z_{2n})$ by the rule

$$z = (z_1 \cdots z_n, z_{n+1} \cdots z_{2n}) = (q_1 \cdots q_n, p_1 \cdots p_n). \tag{5.1.2}$$

When this is done, the Poisson bracket in terms of the variables z, t can be written more compactly in the forms

$$[f,g] = \sum_{a,b} (\partial f/\partial z_a) J_{ab} (\partial g/\partial z_b), \tag{5.1.3}$$

$$[f,g] = (\partial_z f, J\partial_z g). \tag{5.1.4}$$

Here J is the fundamental $2n \times 2n$ matrix given by (1.7.11) or (3.1.1), and used in defining symplectic matrices. Note that the Poisson bracket symbol $[,]$ is the same as that used earlier for a commutator. This is somewhat awkward, but unfortunately there are not always enough convenient symbols to go around.

We also saw in Section 1.7 that the Poisson bracket has several obvious properties. These are again listed below along with one less obvious property, the Jacobi identity. You are instructed to verify it in Exercise 1.3.

1. Distributive property
$$[(af+bg),h]=a[f,h]+b[g,h] \tag{5.1.5}$$
for arbitrary constants a, b.

2. Antisymmetry condition,
$$[f,g]=-[g,f]. \tag{5.1.6}$$

3. Jacobi identity,
$$[f,[g,h]]+[g,[h,f]]+[h,[f,g]]=0. \tag{5.1.7}$$

4. Derivation with respect to ordinary multiplication,
$$[f,gh]=[f,g]h+g[f,h]. \tag{5.1.8}$$

Now the stage is set for a subtle conclusion. Observe that the set of all functions of the variables q, p, t or z, t forms a *linear vector space*. That is, any linear combination of two such functions is again such a function. Thus, we have the first ingredient for a Lie algebraic structure. Now define the Lie product of any two functions to be the Poisson bracket (1.1). Equations (1.5) and (1.6) show that conditions 1 through 4 for a Lie algebra are satisfied. See Section 3.7. And (1.7) shows that condition 5 is satisfied. Consequently, the set of functions of the variables q, p, t or z, t forms a Lie algebra! This Lie algebra will be called the Poisson bracket Lie algebra of dynamical variables. It is evidently infinite dimensional since the set of all functions on phase space is infinite dimensional.

Exercises

5.1.1. If you have not already done so, work out Exercises 1.7.1 through 1.7.4.

5.1.2. Determine the dimensionality of the Poisson bracket Lie algebra of dynamical variables.
Answer: The set of functions of q, p, t or z, t is an infinite dimensional vector space.

5.1.3. Verify the Jacobi identity (1.7). Hint: Use the relation (1.3).

5.1.4. Verify the relation
$$[f,g]=-\sum_{a,b}[f,z_a][z_a,z_b][z_b,g]. \tag{5.1.9}$$

5.2 Equations, Constants, and Integrals of Motion

It has already been shown in Section 1.7 that any dynamical variable $f(z,t)$ of a dynamical system governed by a Hamiltonian H obeys the equation of motion

$$df/dt = \partial f/\partial t + [f, H]. \tag{5.2.1}$$

A special case of this relation is the fact that the dynamical variables z_a obey the equations of motion

$$\dot{z}_a = (J\partial_z H)_a, \tag{5.2.2}$$

or, in more compact vector notation,

$$\dot{z} = J\partial_z H. \tag{5.2.3}$$

A dynamical variable f is called a *constant of motion* if its total time derivative vanishes. In view of (2.1), a constant of motion satisfies the equation

$$\partial f/\partial t + [f, H] = 0. \tag{5.2.4}$$

It can be shown in general that any Hamiltonian dynamical system with n degrees of freedom has $2n$ functionally independent constants of motion. See Exercise 2.4.

Suppose that a constant of motion f does not explicitly depend on the time t,

$$\partial f/\partial t = 0. \tag{5.2.5}$$

A constant of motion that does not explicitly depend on the time will be called an *integral of motion*. By definition, an integral of motion is a constant of motion, but a constant of motion is not an integral of motion if it has explicit time dependence. Evidently, an integral of motion obeys the equation

$$[f, H] = 0. \tag{5.2.6}$$

The question of the existence of integrals of motion is quite complicated. Observe that if $f(z)$ is an integral of motion, then any given trajectory must remain for all time on a general hypersurface in phase space defined by an equation of the form

$$f(z) = \text{constant}. \tag{5.2.7}$$

If there are several functionally independent integrals of motion, then the general trajectory is further restricted to lie in the intersection of several hypersurfaces for all time. Thus, the greater the number of integrals, the more that can be said about the behavior of a dynamical system.

Consider a time-independent Hamiltonian $H(z)$. A point z^c in phase space for which the vector $\partial_z H$ is zero is called a *critical point*. Evidently, according to (2.2) or (2.3), a critical point is some kind of equilibrium point. Now suppose some small region R of phase space contains *no* critical points. Then it can be shown that, provided R is small enough, the dynamical system described by the Hamiltonian $H(z)$ has $2n-1$ functionally independent integrals of motion in the region R. Furthermore, n of these integrals can be arranged to be in *involution.* (Two functions f and g are said to be in involution if their Poisson bracket $[f, g]$ is zero.)[1] See Exercises 2.5 and 2.6.

The result just stated is of limited use unless all trajectories starting in R happen to remain in R. In general, and contrary to the impression given by most textbooks, most dynamical Hamiltonian systems do not have global integrals of motion. If a time-independent Hamiltonian dynamical system with n degrees of freedom has n functionally independent global integrals of motion in involution, the system is said to be *completely integrable.* In general, only the soluble problems found in textbooks fall into this category. Most Hamiltonian dynamical systems, including the majority encountered in real life, are not completely integrable and are therefore sufficiently complicated to be in some sense insoluble. In particular, the behavior of most Hamiltonian systems is sufficiently complicated that the trajectories are not generally confined to lie on hypersurfaces in phase space.

Exercises

5.2.1. Verify (2.2).

5.2.2. Suppose that the Hamiltonian H for a dynamical system does not depend explicitly on the time t. Show that then H is an integral of motion.

5.2.3. Suppose that the dynamical variables f and g are constants of motion. Verify *Poisson's* theorem, which states that the quantity $[f, g]$ is then also a constant of motion. Suppose that f and g are integrals of motion. Show that $[f, g]$ is then also an integral of motion. Hint: Use the Jacobi identity.

5.2.4. Suppose that $f_1, f_2, \cdots, f_n$ are n constants of motion. Let c be any function of the f_j,

$$c = c(f_1, f_2, \cdots, f_n). \tag{5.2.8}$$

Show that c is then also a constant of motion. Suppose that $f_1, f_2, \cdots, f_n$ are n integrals of motion, and that c is again defined as above. Show that c is then also an integral of motion.

5.2.5. Let t^i denote some *initial* time. Given t and $z(t)$, we can always integrate the equations of motion backward (or forward) in time to the time t^i to find the initial conditions z^i. The result of this process will generally depend on t and $z(t)$. Thus, we obtain $2n$ functions $z^i_a(z, t)$. Show that these functions are functionally independent and are constants of motion. Carry out this construction explicitly for the case of the one-dimensional simple harmonic oscillator.

5.2.6. Problem on constructing local integrals of motion (Hamiltonian flow-box or straightening-out theorem).

[1] It is a confusing fact that the term *involution* has multiple meanings. It can also refer to a map or operator whose square is the identity. See, for example, Exercise 3.12.5.

5.3 Lie Operators

Let $f(z,t)$ be any function of the phase-space variables z and perhaps the time t. Associated with each f is a *Lie operator* that we denote by the symbol $: f :$. The Lie operator $: f :$ is a *differential* operator defined by the rule

$$: f : \overset{\text{def}}{=} \sum_i (\partial f/\partial q_i)(\partial/\partial p_i) - (\partial f/\partial p_i)(\partial/\partial q_i). \tag{5.3.1}$$

In particular, if $: f :$ acts on any phase-space function g, one finds the result

$$: f : g = \sum_i (\partial f/\partial q_i)(\partial g/\partial p_i) - (\partial f/\partial p_i)(\partial g/\partial q_i) = [f, g]. \tag{5.3.2}$$

Thus, one may heuristically view a Lie operator as a Poisson bracket waiting to happen. Note that in view of (1.3), the defining relation (3.1) can also be written in the form

$$: f := \sum_{a,b} (\partial f/\partial z_a) J_{ab} (\partial/\partial z_b). \tag{5.3.3}$$

We also remark that in the Mathematics literature the Lie operator $: f :$ is sometimes referred to as $ad(f)$ where ad is shorthand for *adjoint*. Note the similarity of the relations (3.7.71) and (3.2). See also the discussion in Section 8.1. We use the notation $: f :$ instead of $ad(f)$ because it facilitates the writing of complicated expressions.

Powers of $: f :$ can be defined by repeated application, which amounts to taking repeated Poisson brackets. For example, $: f :^2$ is defined by the relation

$$: f :^2 g =: f :: f : g =: f : [f, g] = [f, [f, g]]. \tag{5.3.4}$$

Finally, $: f :$ to the zero power is defined to be the identity operator,

$$: f :^0 = \mathcal{I} \Leftrightarrow : f :^0 g = g. \tag{5.3.5}$$

We note that Lie operators, as well as their powers, are linear operators because of (1.5) and (1.6)

As result of (1.5), the sum of two Lie operators is again a Lie operator. Specifically, one finds the relation

$$a : f : + b : g :=: (af + bg) : \tag{5.3.6}$$

for any two scalars a, b and any two functions f, g. Therefore, the set of Lie operators forms a linear vector space.

A Lie operator is also a *derivation* with respect to the operation of ordinary multiplication. That is, a Lie operator satisfies the product rule analogous to that for differentiation: Let g and h be any two functions. Then, according to (1.8), $: f :$ obeys the rule

$$: f : (gh) = (: f : g)h + g(: f : h). \tag{5.3.7}$$

In addition to being a derivation with respect to ordinary multiplication, a Lie operator is also a derivation with respect to Poisson bracket multiplication. Suppose g and h are any two functions. Then the Jacobi identity (1.7) can be written in the form

$$[f, [g, h]] = [[f, g], h] + [g, [f, h]]. \tag{5.3.8}$$

or equivalently, using Lie operator notation,

$$:f:[g,h] = [:f:g,h] + [g,:f:h]. \tag{5.3.9}$$

Since the set of Lie operators forms a linear vector space, it is of interest to inquire whether the vector space can be given a multiplication rule that will convert it into a Lie algebra. The answer is yes, as is nearly obvious, since Lie operators are linear operators and linear operators are quite similar to matrices. The Lie product of two Lie operators $:f:$ and $:g:$ is simply taken to be their commutator. Denoting the Lie product of two Lie operators by the symbol $\{:f:,:g:\}$, the Lie product is defined by the rule

$$\{:f:,:g:\} =:f::g: - :g::f:. \tag{5.3.10}$$

See Exercise (3.5). Note that there are now two Lie algebras that have to be kept in mind. First, there is the Lie algebra of functions of z,t with the Lie product defined to be the Poisson bracket. Second, there is the Lie algebra of Lie operators with the Lie product defined to be the commutator.

One point, however, has been overlooked. Namely, is the right side of (3.10) a Lie operator? To answer this question, it is useful to view the Jacobi identity (1.7) for Poisson brackets from yet another perspective. For any function h, the Jacobi identity can be written in the form

$$[f,[g,h]] - [g,[f,h]] = [[f,g],h]. \tag{5.3.11}$$

However, using Lie operator notation, this same equation can be written in the form

$$:f::g:h - :g::f:h =:[f,g]:h, \tag{5.3.12}$$

or more compactly, using (3.10),

$$\{:f:,:g:\}h =:[f,g]:h. \tag{5.3.13}$$

But, since h is an arbitrary function, (3.13) can also be viewed as the operator identity

$$\{:f:,:g:\} =:[f,g]:. \tag{5.3.14}$$

Evidently, the commutator of two Lie operators $:f:$ and $:g:$ is again a Lie operator, and is in fact the Lie operator associated with the function $[f,g]$.

Put another way, (3.14) shows that there is a close connection between the Lie algebra of functions and the Lie algebra of Lie operators. Specifically, the Lie product (commutator) of two Lie operators is the Lie operator of the Lie product (Poisson bracket) of the two associated functions. Mathematicians have a word for such a situation. They would say that the two Lie algebras are *homomorphic*. To see that this relation between the two Lie algebras is a homomorphism and not an isomorphism, suppose two Lie operators $:f:$ and $:g:$ are equal,

$$:f:=:g:. \tag{5.3.15}$$

Then from (3.15) we can only deduce the relation

$$f = g + c, \tag{5.3.16}$$

where c is an arbitrary constant. That is, as is obvious from the definition (3.1), the Lie operator associated with any constant is identically zero.

We close this subsection by noting that what we have called a Lie operator is actually a special case of a more general object. Let x denote a collection of N variables $x_1, x_2, \cdots x_N$. Also, let $\boldsymbol{g} = (g_1, g_2, \cdots g_N)$ be a collection of N functions of x and perhaps the time t. The Lie operator $\mathcal{L}_{\boldsymbol{g}}$ associated with the collection of functions $g_b(x, t)$ is defined to be the differential operator given by the rule

$$\mathcal{L}_{\boldsymbol{g}} = \sum_{b=1}^{N} g_b(x, t)(\partial/\partial x_b). \tag{5.3.17}$$

The relation (3.17) is the general definition of a Lie operator. It is also sometimes called a *vector field*. With the introduction of the notation $\boldsymbol{\partial} = (\partial/\partial x_1, \partial/\partial x_2, \cdots \partial/\partial x_N)$, it is often convenient to write $\mathcal{L}_{\boldsymbol{g}}$ in the suggestive form

$$\mathcal{L}_{\boldsymbol{g}} = \boldsymbol{g} \cdot \boldsymbol{\partial}. \tag{5.3.18}$$

There is an intimate connection between vector fields and ordinary differential equations. Consider the set of first-order differential equations

$$\dot{x}_a = g_a(x, t). \tag{5.3.19}$$

Then, using (3.17), this set can also be written in the form

$$\dot{x}_a = \mathcal{L}_{\boldsymbol{g}} x_a. \tag{5.3.20}$$

Also, let h be any function of x and perhaps the time t. Then, by the chain rule, the time derivative of h along a trajectory is given by the relation

$$dh/dt = \partial h/\partial t + \sum_b (\partial h/\partial x_b)\dot{x}_b = \partial h/\partial t + \sum_b g_b(\partial h/\partial x_b) = \partial h/\partial t + \mathcal{L}_{\boldsymbol{g}} h. \tag{5.3.21}$$

Upon comparison of (3.17) with (3.3), we see that we have assumed $N = 2n$ and

$$g_b(z, t) = \sum_a (\partial f/\partial z_a) J_{ab}. \tag{5.3.22}$$

For future reference we note that (3.22) can also be written in the form

$$g_b(z, t) = [f, z_b] = [z_b, (-f)]. \tag{5.3.23}$$

We conclude that, in the case of interest for Hamiltonian systems, the collection of functions g_b arises from a *single* function f according to the relation (3.22). Thus, to be more precise, what we have called and will continue to call a Lie operator could better be called a *Hamiltonian* Lie operator or a *Hamiltonian* vector field. Non-Hamiltonian vector fields are of use for describing dissipative effects including, in the field of accelerator physics, synchrotron radiation effects and electron and ionization cooling. Our primary attention will be focused on Hamiltonian Lie operators. However, where applicable, we will also present results for general Lie operators. General polynomial vector fields, both Hamiltonian and non-Hamiltonian, are treated and classified in Chapter 27.

Finally, in the case $N = 2n$, define quantities η_c by the rule

$$\eta_c = \sum_b J_{bc} g_b. \tag{5.3.24}$$

If the g_b arise from a single function f as in (3.22), we find the result

$$\begin{aligned} \eta_c &= \sum_{ab} J_{bc} J_{ab} (\partial f/\partial z_a) = \sum_{ab} J_{ab} J_{bc} (\partial f/\partial z_a) \\ &= \sum_a (J^2)_{ac} (\partial f/\partial z_a) = -\partial f/\partial z_c. \end{aligned} \tag{5.3.25}$$

It follows from (3.25) that the collection of functions η_c then has the property

$$\partial \eta_c/\partial z_d - \partial \eta_d/\partial z_c = -\partial^2 f/\partial z_d \partial z_c + \partial^2 f/\partial z_c \partial z_d = 0. \tag{5.3.26}$$

Evidently (3.26) is a necessary condition for a vector field to be Hamiltonian. In Section 6.4 we will see that it is also sufficient.

It is easily verified that that the set of all vector fields in N variables forms a Lie algebra with the commutator taken as the Lie product (see Exercise 3.8), and (in the even dimensional case) the set of Hamiltonian vector fields forms a Lie subalgebra of the Lie algebra of all vector fields. In subsequent chapters we will learn that the set of all vector fields is the Lie algebra of the group of all diffeomorphisms, and the set of all Hamiltonian vector fields is the Lie algebra of the subgroup of all symplectic maps.

Exercises

5.3.1. Starting from (3.7), show that $: f :^n$ obeys the *Leibniz* rule

$$: f :^n (gh) = \sum_{m=0}^{n} \left(\begin{array}{c} n \\ m \end{array} \right) (: f :^m g)(: f :^{n-m} h), \tag{5.3.27}$$

where $\left(\begin{array}{c} n \\ m \end{array} \right)$ is the binomial coefficient defined by

$$\left(\begin{array}{c} n \\ m \end{array} \right) = \frac{n!}{(m!)(n-m)!}. \tag{5.3.28}$$

Suggestion: Use induction and the relations $\left(\begin{array}{c} n \\ n \end{array} \right) = 1, \left(\begin{array}{c} n \\ 0 \end{array} \right) = 1,$ and $\left(\begin{array}{c} n \\ m-1 \end{array} \right) + \left(\begin{array}{c} n \\ m \end{array} \right) = \left(\begin{array}{c} n+1 \\ m \end{array} \right).$

According to some, perhaps apocryphal, lore it took Leibniz seven years to discover this rule (in the context of how to differentiate a product of two functions). He and others first assumed that the derivative of a product would be the product of the derivatives (which, in fact, is the case for the chain rule that applies to functions of functions).

5.3.2. Verify (3.8) and (3.9).

5.3.3. State and verify the analog of the Leibniz rule of Exercise (3.1) for the case of $: f :^n [g, h]$.

5.3.4. Verify that the Lie product defined by (3.10) satisfies the properties 1 through 5 required to make the set of Lie operators into a Lie algebra. See Section 3.7.

5.3.5. Verify (3.11), (3.12), and (3.13).

5.3.6. Let h_0 be any constant function. Verify that

$$: h_0 := 0. \tag{5.3.29}$$

5.3.7. Let G be any function of the variables z. A set of differential equations of the form

$$\dot{z}_a = -\partial G/\partial z_a \tag{5.3.30}$$

is called a *gradient* system, and the corresponding vector field

$$\mathcal{L}_G = -\sum_a (\partial G/\partial z_a)(\partial/\partial z_a) \tag{5.3.31}$$

is called a gradient vector field. At this point is is interesting to contrast Hamiltonian systems, see (2.2), with gradient systems. Both are derived from master functions, H and G, respectively. But their behavior can be very different. Verify the relation

$$dG/dt = \mathcal{L}_G G = -\sum_a (\partial G/\partial z_a)^2 \leq 0. \tag{5.3.32}$$

It follows that, for a gradient system, points on a trajectory move away from maxima of G and toward minima of G. Compare the behavior of Hamiltonian and gradient systems near and at local extrema of H and G, respectively. What happens at and near saddle points?

5.3.8. Suppose $\mathcal{L}_{\boldsymbol{f}}$ and $\mathcal{L}_{\boldsymbol{g}}$ are any two vector fields. Show that their commutator is also a vector field. That is, given $\boldsymbol{f}$ and $\boldsymbol{g}$, show that there is a relation of the form

$$\{\mathcal{L}_{\boldsymbol{f}}, \mathcal{L}_{\boldsymbol{g}}\} = \mathcal{L}_{\boldsymbol{h}} \tag{5.3.33}$$

and find a formula for $\boldsymbol{h}$ in terms of $\boldsymbol{f}$ and $\boldsymbol{g}$. Show that, for vector fields, double commutators composed of three vector fields obey the Jacobi identity,

$$\{\mathcal{L}_{\boldsymbol{f}}, \{\mathcal{L}_{\boldsymbol{g}}, \mathcal{L}_{\boldsymbol{h}}\}\} + \{\mathcal{L}_{\boldsymbol{g}}, \{\mathcal{L}_{\boldsymbol{h}}, \mathcal{L}_{\boldsymbol{f}}\}\} + \{\mathcal{L}_{\boldsymbol{h}}, \{\mathcal{L}_{\boldsymbol{f}}, \mathcal{L}_{\boldsymbol{g}}\}\} = 0. \tag{5.3.34}$$

Show that the set of all vector fields forms an infinite-dimensional Lie algebra with the commutator playing the role of a Lie product.

5.3.9. Suppose some N-dimensional Lie algebra L has structure constants $c^{\gamma}_{\alpha\beta}$. Consider an N-dimensional Euclidean space with coordinates $x_1, x_2, \cdots x_N$. Define N vector fields $\mathcal{L}_\alpha$ by the rule

$$\mathcal{L}_\alpha = -\sum_{\beta\gamma} c^{\gamma}_{\alpha\beta} x_\beta \partial/\partial x_\gamma. \tag{5.3.35}$$

Show that these vector fields satisfy the commutation relations

$$\{\mathcal{L}_\alpha, \mathcal{L}_\beta\} = \sum_{\gamma} c^{\gamma}_{\alpha\beta} \mathcal{L}_\gamma, \tag{5.3.36}$$

and therefore provide a vector-field realization of L. Since the vector fields are manufactured from the structure constants, might this realization be related to the adjoint representation of L? Using (3.7.54), show that

$$\mathcal{L}_\alpha x_\beta = -\sum_{\gamma} (\hat{B}_\alpha)_{\beta\gamma} x_\gamma \tag{5.3.37}$$

or, more compactly,

$$\mathcal{L}_\alpha x = -\hat{B}_\alpha x. \tag{5.3.38}$$

Suggestion: First verify (3.38) and then (3.36).

5.3.10. Let $\mathcal{L}_{\boldsymbol{f}}$ be a vector field and suppose g and h are any two functions. In analogy to (3.7), prove the derivation property

$$\mathcal{L}_{\boldsymbol{f}}(gh) = (\mathcal{L}_{\boldsymbol{f}} g)h + g(\mathcal{L}_{\boldsymbol{f}} h). \tag{5.3.39}$$

Find the Leibniz rule for $(\mathcal{L}_{\boldsymbol{f}})^n$ analogous to (3.27).

5.4 Lie Transformations

5.4.1 Definition and Some Properties

Since powers of $:f:$ have been defined, it is also possible to deal with power series in $:f:$. Of particular importance is the power series $\exp(:f:)$. This particular object is called the *Lie transformation* associated with $:f:$ or f.[2] The Lie transformation is also a linear operator, and is formally defined as expected by the exponential series

$$e^{:f:} = \exp(:f:) = \sum_{n=0}^{\infty} :f:^n /n!. \tag{5.4.1}$$

In particular, the action of $\exp(: f :)$ on any function g is given by the rule

$$\exp(: f :)g = g + [f, g] + [f, [f, g]]/2! + \cdots . \tag{5.4.2}$$

[2] Some authors use the terms *Lie transformation* and *Lie series* interchangeably. We prefer to refer to any power series in $:f:$ as a Lie series, and to refer to the particular power series $\exp(:f:)$ as a Lie transformation.

The fact that $: f :$ is a derivation with respect to ordinary multiplication, see (3.7), implies that the Lie transformation $\exp(: f :)$ is an *isomorphism* with respect to ordinary multiplication. (This is another remarkable property of the exponential function!) That is, suppose g and h are any two functions. Then the Lie transformation $\exp(: f :)$ has the property

$$\exp(: f :)(gh) = [\exp(: f :)g][\exp(: f :)h]. \tag{5.4.3}$$

In words, (4.3) says that one can either let a Lie transformation act on the product of two functions, or act on each function separately and then take the product of the results. Both operations give the same net result.

The relation (4.3) may be proved as follows. First, use the definition (4.1) to get the result

$$\exp(: f :)(gh) = \sum_{n=0}^{\infty}(: f :^n /n!)(gh). \tag{5.4.4}$$

Next, use the Leibniz rule (3.27), which is a consequence of the derivation property (3.7), to get the result

$$\exp(: f :)(gh) = \sum_{n=0}^{\infty}(1/n!)\sum_{m=0}^{n}\left(\begin{array}{c} n \\ m \end{array}\right)(: f :^m g)(: f :^{n-m} h). \tag{5.4.5}$$

The binomial coefficients obey the relation

$$(1/n!)\left(\begin{array}{c} n \\ m \end{array}\right) = 1/[(m!)(n-m)!]. \tag{5.4.6}$$

Consequently, (4.5) can also be written in the form

$$\exp(: f :)(gh) = \sum_{n=0}^{\infty}\sum_{m=0}^{n}\{[: f :^m /m!]g\}\{[: f :^{n-m} /(n-m)!]h\}. \tag{5.4.7}$$

Observe that the double sum on the right side of (4.7) can be rearranged to give the result

$$\exp(: f :)(gh) = \sum_{m=0}^{\infty}\sum_{n=m}^{\infty}\{[: f :^m /m!]g\}\{[: f :^{n-m} /(n-m)!]h\}. \tag{5.4.8}$$

See Figure 4.1 and Exercise 4.8. Finally, let $\ell = n - m$ be a new summation index. Then (4.8) takes the final form

$$\begin{aligned} \exp(: f :)(gh) &= \sum_{m=0}^{\infty}[: f :^m /m!]g\sum_{\ell=0}^{\infty}[: f :^{\ell} /\ell!]h \\ &= [\exp(: f :)g][\exp(: f :)h]. \end{aligned} \tag{5.4.9}$$

The relation (4.3) may be extended to products of Lie transformations acting on products of functions. Let a and b be any two functions, and let $\exp(: f :)$ and $\exp(: g :)$ be any two Lie transformations. Then we have, by using (4.3) repeatedly, the result

$$\begin{aligned} \exp(: f :)\exp(: g :)(ab) &= \exp(: f :)\{[\exp(: g :)a][\exp(: g :)b]\} \\ &= [\exp(: f :)\exp(: g :)a][\exp(: f :)\exp(: g :)b]. \end{aligned} \tag{5.4.10}$$

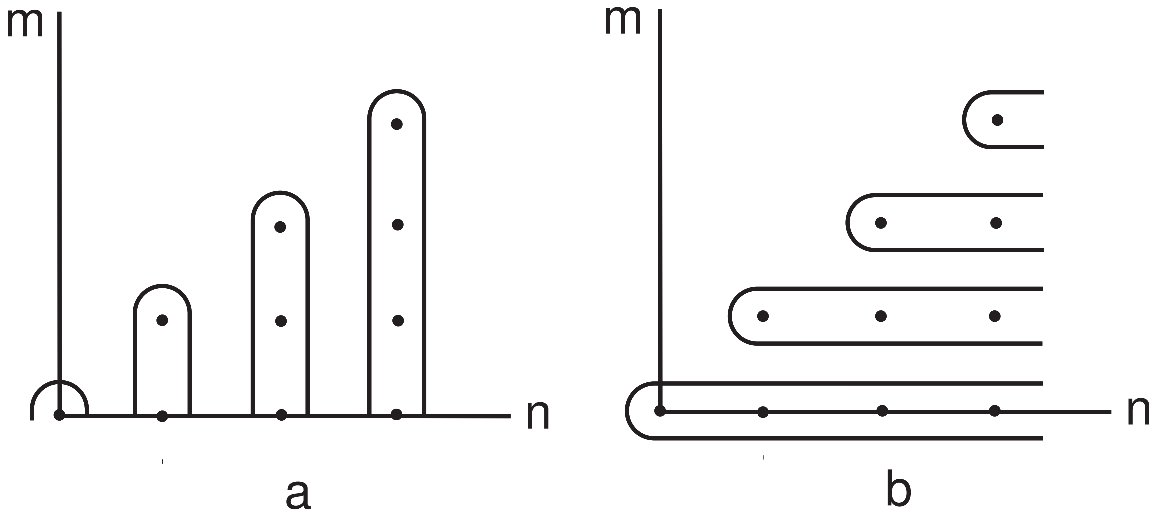


Figure 5.4.1: a) The summation points in m, n space for the sum (4.7) indicating that the inner sum is over m followed by a sum over n. b) The summation points for the sum (4.8) illustrating that the points are the same, but the inner sum is now over n followed by a sum over m.

Analogous results evidently hold for any number of Lie transformations and any number of functions.

The isomorphism property of $\exp(:f:)$ described by (4.3) often facilitates computations involving Lie transformations. Let the symbol z stand, as usual, for the collection of quantities $z_1 \cdots z_{2n}$. Similarly, let the symbol $\exp(:f:)z$ stand for the collection of quantities $\exp(:f:)z_1, \cdots \exp(:f:)z_{2n}$. Now let $g(z)$ be any function. Then it follows from (4.3) that

$$\exp(:f:)g(z) = g[\exp(:f:)z]. \tag{5.4.11}$$

That is, the action of a Lie transformation on a function is to perform a Lie transformation on its arguments.

To see the truth of (4.11), suppose first that g were a polynomial in the quantities $z_1 \cdots z_{2n}$. But a polynomial is just a sum of monomials of the form

$$z_1^{m_1} z_2^{m_2} \cdots z_{2n}^{m_{2n}}.$$

It follows from (4.9) that

$$\exp(:f:)z_1^{m_1} z_2^{m_2} \cdots z_{2n}^{m_{2n}} = [\exp(:f:)z_1]^{m_1} \cdots [\exp(:f:)z_{2n}]^{m_{2n}}. \tag{5.4.12}$$

Also, as mentioned earlier, $\exp(:f:)$ is a linear operator. Therefore a Lie transformation has the advertised property (4.11) when acting on polynomials. But, according to the *Weierstrass* approximation theorem, the set of monomials is dense in the complete set of functions on any bounded domain. Consequently, (4.11) holds in general by continuity.

As a consequence of (4.10), there is a result analogous to (4.11) for any product of Lie transformations acting on a function. For example, in the case of two Lie transformations $\exp(:f:)$ and $\exp(:g:)$ and a function h, we have the result

$$\exp(:f:)\exp(:g:)h(z) = h[\exp(:f:)\exp(:g:)z]. \tag{5.4.13}$$

Similar results hold for any number of Lie transformations. The proof of these results is similar to that just given for (4.11).

The last observation to be made is that since $:f:$ is also a derivation with respect to Poisson bracket multiplication, the Lie transformation $\exp(:f:)$ must also be an isomorphism with respect to Poisson bracket multiplication. That is, suppose g and h are any two functions. Then the Lie transformation $\exp(:f:)$ has the property

$$\exp(:f:)[g,h] = [\exp(:f:)g, \exp(:f:)h]. \tag{5.4.14}$$

This property will be essential for subsequent discussions of symplectic maps and charged-particle beam transport. Its proof is exactly analogous to that just given for the case of ordinary multiplication. Also, there are results analogous to (4.10) for products of Lie transformations acting on Poisson brackets. Suppose, for example, that a and b are any two functions. Then we have the result

$$\exp(:f:)\exp(:g:)[a,b] = [\exp(:f:)\exp(:g:)a, \exp(:f:)\exp(:g:)b], \tag{5.4.15}$$

and similar results hold for any number of Lie transformations.

5.4.2 Applications

Subsequent chapters and sections will be devoted to the use of Lie transformations for representing, manipulating, and analyzing symplectic maps. They can also be used to transform Hamiltonians to normal form. Indeed, this was their original use as envisioned by Hori and Deprit. Similarly, they can be used to transform vector and tensor fields. See the references listed at the end of this chapter.

Exercises

5.4.1. Let q and p be the phase-space coordinates for a system having one degree of freedom. Let f be the function

$$f = -\lambda p^2/2. \tag{5.4.16}$$

Show that

$$\begin{aligned} \exp(:f:)p &= p, \\ \exp(:f:)q &= q + \lambda p. \end{aligned} \tag{5.4.17}$$

Here λ is an arbitrary parameter.
Hint: Observe that the series (4.2) terminates in this case.

5.4.2. Repeat Exercise 4.1 for the case $f = \lambda q^2/2$.

5.4.3. Repeat Exercise 4.1 for the case $f = \lambda q^3/3$.

5.4.4. Repeat Exercise 4.1 for the case $f = -\lambda pq$. Now you must sum an infinite series.
Answer:

$$\begin{aligned} \exp(: f :)q &= (e^{\lambda})q, \\ \exp(: f :)p &= (e^{-\lambda})p. \end{aligned} \tag{5.4.18}$$

5.4.5. Repeat Exercise 4.1 for the case $f = -\lambda(p^2 + q^2)/2$.
Answer:

$$\begin{aligned} \exp(: f :)q &= q\cos\lambda + p\sin\lambda, \\ \exp(: f :)p &= -q\sin\lambda + p\cos\lambda. \end{aligned} \tag{5.4.19}$$

5.4.6. Repeat Exercise 4.1 for the case $f = -\lambda(p^2 - q^2)/2$.
Answer:

$$\begin{aligned} \exp(: f :)q &= q\cosh\lambda + p\sinh\lambda, \\ \exp(: f :)p &= q\sinh\lambda + p\cosh\lambda. \end{aligned} \tag{5.4.20}$$

5.4.7. Repeat Exercise 4.1 for the case $f = \lambda qp^2$.
Answer:

$$\begin{aligned} \exp(: f :)q &= q(1 - \lambda p)^2, \\ \exp(: f :)p &= p/(1 - \lambda p). \end{aligned} \tag{5.4.21}$$

See the end of Section 1.4.

5.4.8. Verify (4.3) for the case $f = \lambda q^2$ and $g = h = p$.

5.4.9. Verify the rearrangement required to go from (4.7) to (4.8). Hint: Mark out, in m, n space, the lattice of points that are summed over in (4.7). Show that the same points are summed over in (4.8). See Figure 4.1.

5.4.10. Prove (4.13).

5.4.11. Derive (4.14) from the definition (4.1) and the results of Exercise 3.3.

5.4.12. Prove (4.15).

5.4.13. Let c be any constant. Verify the result

$$\exp(: f :)c = c. \tag{5.4.22}$$

5.4.14. Let $\mathcal{L}_{\boldsymbol{f}}$ be a general vector field. In analogy to (4.1) define an associated Lie transformation by the rule

$$\exp(\mathcal{L}_{\boldsymbol{f}}) = \sum_{n=0}^{\infty} (\mathcal{L}_{\boldsymbol{f}})^n/n!. \tag{5.4.23}$$

Show, in analogy to (4.3), that this Lie transformation is also an isomorphism with respect to function multiplication,

$$\exp(\mathcal{L}_{\boldsymbol{f}})(gh) = [\exp(\mathcal{L}_{\boldsymbol{f}})g][\exp(\mathcal{L}_{\boldsymbol{f}})h]. \tag{5.4.24}$$

5.5 Realization of the $sp(2n,\mathbb{R})$ Lie Algebra

According to Exercise (1.2), the Poisson bracket Lie algebra of dynamical variables is infinite dimensional. The purpose of this section is to show that, for a $2n$-dimensional phase space, the Poisson bracket Lie algebra of dynamical variables contains $sp(2n,\mathbb{R})$ as a subalgebra.

Suppose f and g are homogeneous polynomials of degree 2 in the variables z. Then, inspection of (1.3) indicates that their Poisson bracket $[f,g]$ is also a homogeneous polynomial of degree two. We conclude that second-degree polynomials form a subalgebra of the Poisson bracket Lie algebra of all functions. In fact, calculation shows that this subalgebra is a realization of $sp(2n,\mathbb{R})$.

To verify this assertion, suppose that f and g are any two homogeneous second-degree polynomials in the variables z. They can be written in the form

$$f = (1/2)\sum_{a,b} S^f_{ab} z_a z_b = (1/2)(z, S^f z), \tag{5.5.1}$$

$$g = (1/2)\sum_{c,d} S^g_{cd} z_c z_d = (1/2)(z, S^g z), \tag{5.5.2}$$

where S^f and S^g are real *symmetric* matrices. Evidently, there is a one-to-one correspondence between homogeneous second-degree polynomials and symmetric matrices. We will indicate a one-to-one correspondence by the symbol $\leftrightarrow$. Since J is invertible, there is also an associated one-to-one correspondence between homogeneous second degree polynomials and matrices of the form JS. Indeed, the relations (5.1) and (5.2) can be written also in the form

$$f \leftrightarrow JS^f \Leftrightarrow f = (1/2)(Jz, JS^f z), \tag{5.5.3}$$

$$g \leftrightarrow JS^g \Leftrightarrow g = (1/2)(Jz, JS^g z). \tag{5.5.4}$$

Recall (3.1.6). Here the symbol $\Leftrightarrow$ denotes logical implication in both directions.

Now use the representations (5.1) and (5.2) to compute the Poisson bracket $[f,g]$. This calculation is facilitated by the relation

$$\begin{aligned} [z_a z_b, z_c z_d] &= z_a z_c J_{bd} + z_a z_d J_{bc} + z_b z_c J_{ad} + z_b z_d J_{ac} \\ &= \sum_{ef} (\delta_{ae}\delta_{cf} J_{bd} + \delta_{ae}\delta_{df} J_{bc} + \delta_{be}\delta_{cf} J_{ad} + \delta_{be}\delta_{df} J_{ac}) z_e z_f. \end{aligned} \tag{5.5.5}$$

[Note that in this realization the structure constants are related to the entries of J and the Kronecker delta. This result is not completely surprising because J also enters the definition of the Poisson bracket. Recall (1.3).] We find from (5.1), (5.2), and (5.5) the result

$$[f,g] = (z, S^f J S^g z) = (Jz, JS^f JS^g z). \tag{5.5.6}$$

Similarly, the Poisson bracket $[g,f]$ can be evaluated to give the result

$$[g,f] = (Jz, JS^g JS^f z). \tag{5.5.7}$$

Subtract (5.7) from (5.6) and use the antisymmetry condition (1.6). Doing so gives the result

$$[f,g] = (1/2)(Jz, \{JS^f, JS^g\} z). \tag{5.5.8}$$

Here the notation $\{,\}$ indicates the matrix commutator,

$$\{JS^f, JS^g\} = JS^f JS^g - JS^g JS^f. \tag{5.5.9}$$

Suppose the second-degree polynomial h is defined by the relation

$$h = [f, g]. \tag{5.5.10}$$

Then, comparison of (5.8) and (5.10) shows that h can be written also in the form

$$h = (1/2)(Jz, JS^h z), \tag{5.5.11}$$

where the matrix JS^h is defined by the relation

$$JS^h = \{JS^f, JS^g\}. \tag{5.5.12}$$

Observe that (5.10) is a Lie-algebraic relation in the Poisson bracket Lie algebra of second-degree polynomials, and (5.12) is a Lie-algebraic relation in $sp(2n)$. Thus we have the logical implication

$$h = [f, g] \Leftrightarrow JS^h = \{JS^f, JS^g\}. \tag{5.5.13}$$

What we have just shown is that these two Lie algebras are isomorphic under the one-to-one correspondence given by (5.3), (5.4), and (5.11).

In the next three sections we will study the problem of finding suitable bases for the Lie algebras $sp(2), sp(4)$, and $sp(6)$ when special attention is given to their $u(1), u(2)$, and $u(3)$ subalgebras, respectively. We close this section by finding a basis for $sp(2n)$ when special attention is given to the subgroups described in Section (3.10).

We have already studied the basis for $sp(2n)$ consisting of the monomials $z_a z_b$ and found that they satisfy the Poisson bracket rules (5.5). Another possible basis can be found by decomposing these monomials into those associated with the subgroups constituted by matrices of the form (3.3.9), (3.3.10), and (3.3.11), respectively. Consider first the subgroup associated with matrices of the form (3.3.9). In this case, S is of the form (3.10.2). Correspondingly, the polynomials f given by (5.1) are linear combinations of the monomials $p_j p_k$. They satisfy the Poisson bracket relations

$$[p_j p_k, p_\ell p_m] = 0. \tag{5.5.14}$$

The vanishing of all Lie products for elements of this subalgebra is expected since the associated subgroup is Abelian.

Consider next the subgroup associated with matrices of the form (3.3.10). In this case S is of the form (3.10.7), and the polynomials f given by (5.1) are linear combinations of the monomials $q_j q_k$. They satisfy the Poisson bracket relations

$$[q_j q_k, q_\ell q_m] = 0. \tag{5.5.15}$$

Again all Poisson brackets for this Lie subalgebra vanish since the associated subgroup is also Abelian.

Finally consider the subgroup associated with matrices of the form (3.3.11). In this case S is of the form (3.10.13). Correspondingly, the polynomials f given by (5.1) are linear combinations of the monomials $q_j p_k$. They satisfy the Poisson bracket relations

$$[q_j p_k, q_\ell p_m] = \delta_{jm} q_\ell p_k - \delta_{k\ell} q_j p_m. \tag{5.5.16}$$

Since the right side of (5.16) is again of the form $q_j p_k$, these monomials constitute a Lie subalgebra as expected. This subalgebra is the Lie algebra $g\ell(n,\mathbb{R})$, the Lie algebra of the group $GL(n,\mathbb{R})$.

It remains to compute the Poisson brackets of the monomials $p_j p_k, q_j q_k$, and $q_j p_k$ with each other. We find the results

$$[q_j p_k, p_\ell p_m] = \delta_{j\ell} p_k p_m + \delta_{jm} p_k p_\ell, \tag{5.5.17}$$

$$[q_j p_k, q_\ell q_m] = -\delta_{k\ell} q_j q_m - \delta_{km} q_j q_\ell, \tag{5.5.18}$$

$$[q_j q_k, p_\ell p_m] = \delta_{j\ell} q_k p_m + \delta_{jm} q_k p_\ell + \delta_{k\ell} q_j p_m + \delta_{km} q_j p_\ell. \tag{5.5.19}$$

Note that (5.17) indicates that the Lie algebra formed by the monomials $p_\ell p_m$ is transformed under the action of the Lie algebra formed by the monomials $q_j p_k$. Also, (5.16) and (5.17) together indicate that the set of monomials $q_j p_k$ and $p_\ell p_m$, when combined in linear combinations, still form a Lie subalgebra. This is the subalgebra associated with the subgroup of matrices of the form (3.10.16). The fact that the monomials $p_\ell p_m$ transform under the action of the monomials $q_j p_k$ is a consequence of the fact that the subgroup of matrices (3.10.16) is a *semidirect* product of the subgroups of matrices (3.3.11) and (3.3.9). Similarly, the relations (5.16) and (5.18) indicate that the monomials $q_j p_k$ and $q_\ell q_m$ span a Lie subalgebra associated with the subgroup of matrices of the form (3.10.19), and this subgroup is a semidirect product of the subgroups of matrices (3.3.11) and (3.3.10).

Exercises

5.5.1. Verify (5.5).

5.5.2. Verify (5.6), (5.7), and (5.8).

5.5.3. Verify the following Poisson bracket relation:

$$[z_a z_b, z_c] = z_a J_{bc} + z_b J_{ac}. \tag{5.5.20}$$

Suppose f is given by (5.1). Show that the matrix JS^f can be computed from f by the relation

$$: f : z_c = [f, z_c] = -(JS^f z)_c = -\sum_d (JS^f)_{cd} z_d. \tag{5.5.21}$$

5.5.4. Find the dimensions of the three Lie subalgebras spanned by the monomials of the form $p_j p_k$, monomials of the form $q_j q_k$, and monomials of the form $q_j p_k$, respectively.

5.5.5. Find the dimension of the Lie subalgebra spanned by the monomials of the form $q_j p_k$ plus monomials of the form $p_\ell p_m$. Find the dimension of the Lie subalgebra spanned by the monomials of the form $q_j p_k$ plus monomials of the form $q_\ell q_m$.

5.5.6. Show that the monomials $q_j q_k$ and $p_\ell p_m$ generate the full Lie algebra of $sp(2n)$ in the sense that taking suitable Poisson brackets of them produces all possible monomials $z_a z_b$.

5.6 Basis for $sp(2,\mathbb{R})$

The symplectic Lie algebras of primary interest for accelerator applications are $sp(2,\mathbb{R})$, $sp(4,\mathbb{R})$, and $sp(6,\mathbb{R})$.[3] The purpose of this and the next two sections is to discuss suitable bases for these Lie algebras when special attention is given to their unitary subalgebras $u(1), u(2)$, and $u(3)$. What we will be finding are the defining or fundamental representations of $sp(2,\mathbb{R})$, $sp(4,\mathbb{R})$, and $sp(6,\mathbb{R})$. See Section 3.7.6. We remark that these are the lowest dimensional representations that are faithful, in the sense of being isomorphic, to the underlying abstract Lie algebra.

One way to specify a basis is to select suitable matrices of the form JS. In Section (5.5) we learned that there is an isomorphism between the Poisson bracket Lie algebra of quadratic polynomials and the commutator Lie algebra of the matrices JS. Therefore, another way to specify a basis for $sp(2n,\mathbb{R})$ is to select suitable second-degree polynomials. We will mostly choose this second approach because of its convenience for later use. However, some of the calculations employed in selecting suitable polynomials will involve the associated matrices. Moreover, as indicated by (5.1) through (5.4), the associated matrices can easily be constructed from a knowledge of the associated second degree polynomials, and vice versa.

Because (as discussed in Section 3.9 and Exercise 3.9.1) matrices of the form JS^c form a Lie algebra in their own right, it is convenient to find the polynomials associated with these matrices first. By making use of the results of Sections 3.9 and 5.5, these polynomials can be arranged to give a realization of the Lie algebra $u(n)$. Then, when this is done, the polynomials associated with matrices of the form JS^a can be selected in a suitable manner. In particular, these polynomials can be selected in such a way that they have convenient transformation properties under the action of $u(n)$.

We begin with the case of $sp(2,\mathbb{R})$. In the 2×2 realization of $sp(2,\mathbb{R})$, the most general symmetric matrix S is of the form

$$S = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix}, \tag{5.6.1}$$

and J is simply the matrix

$$J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{5.6.2}$$

Requiring that J commute with S gives the restrictions

$$\beta = 0, \quad \gamma = \alpha. \tag{5.6.3}$$

Consequently, the most general S^c in the 2×2 case is just a multiple of the identity,

$$S^c = \alpha I, \tag{5.6.4}$$

and JS^c is simply a multiple of J,

$$JS^c = \alpha J. \tag{5.6.5}$$

[3] In addition, the Lie algebra $sp(8,\mathbb{R})$ is useful for the treatment of errors. See Section 9.4.

Let b^0 be the polynomial associated with S^c by a relation of the form (5.1). Set $\alpha = 1$ so that $S^c = I$, in which case b^0 is given by the relation

$$b^0 = (1/2)(z_1^2 + z_2^2) = (1/2)(q^2 + p^2). \tag{5.6.6}$$

Let B^0 denote the associated matrix of the form JS^c. Then, according to (6.5), B^0 is given by the relation

$$B^0 = JS^c = JI = J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = i\sigma^2. \tag{5.6.7}$$

[Here, and in (6.13) and (6.14), we have also referenced a Pauli matrix σ^α. See Exercise 3.7.31. This referencing will be useful later.] We observe that Exercise 3.7.23 shows that $u(1)$ is one dimensional. The fact that in the 2×2 case we have found only one linearly independent matrix of the form JS^c is consistent with this observation.

Next study matrices S^a that anticommute with J. Requiring that J anticommute with the S of (6.1) gives only the restriction

$$\alpha = -\gamma. \tag{5.6.8}$$

Consequently, S^a is of the general form

$$S^a = \gamma \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} + \beta \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{5.6.9}$$

and JS^a is of the general form

$$JS^a = \gamma \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + \beta \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{5.6.10}$$

Suppose we set $\gamma = 1$ and $\beta = 0$ in (6.9). Let f be the polynomial corresponding to this choice for S^a. It is given by the relation

$$f = (1/2)(-z_1^2 + z_2^2) = (1/2)(-q^2 + p^2). \tag{5.6.11}$$

Alternatively, suppose we set $\gamma = 0$ and $\beta = 1$ in (6.9). Let g be the polynomial corresponding to this choice for S^a. It is given by the relation

$$g = z_1 z_2 = qp. \tag{5.6.12}$$

[Again see (5.1).] Let F and G be the matrices associated with f and g. According to (6.10), F and G are given by the relations

$$F = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \sigma^1, \tag{5.6.13}$$

$$G = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \sigma^3. \tag{5.6.14}$$

It is readily verified that the polynomials b^0, f, and g obey the Poisson bracket rules

$$[b^0, f] = 2g, \tag{5.6.15}$$

$$[b^0, g] = -2f, \tag{5.6.16}$$

$$[f, g] = -2b^0. \tag{5.6.17}$$

Correspondingly, the matrices B^0, F, and G obey the analogous commutation rules,

$$\{B^0, F\} = 2G, \tag{5.6.18}$$

$$\{B^0, G\} = -2F, \tag{5.6.19}$$

$$\{F, G\} = -2B^0. \tag{5.6.20}$$

This in one version of the commutation rules for $sp(2, \mathbb{R})$. All others can be obtained by making the transformations (3.7.56) with a real invertible matrix T. Note that all the matrices B^0, F, and G are of the form JS with S real and symmetric. They therefore belong to $sp(2, \mathbb{R})$. Also, B^0 is real anti-Hermitian/antisymmetric, and therefore generates elements in $SO(2, \mathbb{R})$ upon exponentiation. By contrast, F and G are real Hermitian/symmetric and generate noncompact subgroups upon exponentiation.

Exercises

5.6.1. Verify that the requirement that J commute with S does indeed give the restrictions (6.3).

5.6.2. Verify that the requirement that J anticommute with S gives the restriction (6.8).

5.6.3. Verify that b^0, f, and g are associated with B^0, F, and G by relations of the form (5.3).

5.6.4. Verify the Lie algebraic relations (6.15) through (6.20).

5.6.5. Let g be the 2×2 matrix

$$g = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{5.6.21}$$

Let $U(1, 1)$ be the set of all complex 2×2 matrices that satisfy the relation

$$U^\dagger g U = g. \tag{5.6.22}$$

Show that $U(1, 1)$ is a group. Let $SU(1, 1)$ be the subset of matrices in $U(1, 1)$ that have unit determinant. Show that $SU(1, 1)$ is a group. Find the corresponding Lie algebras $u(1, 1)$ and $su(1, 1)$. Show that $su(1, 1)$ and $sp(2)$ are equivalent over the complex field.

5.6.6. Let g be the 3×3 matrix

$$g = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}. \tag{5.6.23}$$

Let $O(2,1)$ be the set of all real 3×3 matrices that satisfy the relation

$$O^T g O = g. \tag{5.6.24}$$

Show that $O(2,1)$ is a group. Let $SO(2,1)$ be the subset of matrices in $O(2,1)$ that have unit determinant. Show that $SO(2,1)$ is a group. Find the corresponding Lie algebra $so(2,1)$. Show that $so(2,1)$ and $sp(2)$ are equivalent over the complex field.

5.6.7. This exercise studies polar decomposition for two interesting symplectic matrices, call them L and M, defined by the equations

$$L = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \tag{5.6.25}$$

$$M = -L = \begin{pmatrix} -1 & -1 \\ 0 & -1 \end{pmatrix}. \tag{5.6.26}$$

Verify that both L and M are indeed symplectic. The matrix M is interesting because we know from Exercise 3.7.12 that it cannot be written in single exponential form. By contrast, verify that L can be written in the form

$$L = \exp(JS) \tag{5.6.27}$$

with

$$JS = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \tag{5.6.28}$$

where S is the symmetric matrix

$$S = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}. \tag{5.6.29}$$

Let us first work on finding the polar decomposition for L. That is, according to Subsection 3.8.2, we wish to write L in the form

$$L = PO. \tag{5.6.30}$$

Verify that from the properties of P and O it follows that

$$LL^T = POO^T P^T = P^2. \tag{5.6.31}$$

Show that the matrix (LL^T) is real positive-definite symmetric since L is real symplectic. Next show that (LL^T) has a unique real positive-definite symmetric square root. Thus, P is determined by the equation

$$P = (LL^T)^{1/2}. \tag{5.6.32}$$

Show for the problem at hand that

$$LL^T = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \tag{5.6.33}$$

and P is given by the relation

$$P = \frac{1}{\sqrt{5}} \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}. \tag{5.6.34}$$

Observe that P is symplectic and symmetric as desired. Moreover, let us check that P as given by (6.34) is indeed positive definite. Let v be a two-component real vector given by

$$v = \{v_1, v_2\}. \tag{5.6.35}$$

Verify that

$$(v, Pv) = (1/5)(3v_1^2 + 2v_1v_2 + 2v_2^2). \tag{5.6.36}$$

The *discriminant* D of a binary quadratic form

$$av_1^2 + bv_1v_2 + cv_2^2 \tag{5.6.37}$$

is defined by the relation

$$D = 4ac - b^2. \tag{5.6.38}$$

From the theory of binary quadratic forms it is known that such a form is positive definite if $D > 0$. Verify that for the form (6.36)

$$D = (1/25)(24 - 4) > 0, \tag{5.6.39}$$

and therefore P is positive definite.

Now that P is known, O is given by (4.2.10). Verify that

$$O = P^{-1}L = -JP^TJL = -JPJL = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 & 1 \\ -1 & 2 \end{pmatrix}. \tag{5.6.40}$$

Here we have used the fact that P is symplectic to compute its inverse. Verify that O is symplectic and orthogonal as desired. Finally, (6.30) holds because of the construction (6.40).

Let us now turn our attention to finding the polar decomposition for M, which we seek to write as

$$M = P'O'. \tag{5.6.41}$$

Show that

$$P' = P \tag{5.6.42}$$

and

$$O' = -O. \tag{5.6.43}$$

The last items we might wonder about for L and M are the matrices S^a and S^c and the associated polynomials f_2^a and f_2^c. The computation of these items is the task of Exercise 7.6.14.

5.7 Basis for $sp(4,\mathbb{R})$

The case of $sp(4,\mathbb{R})$ is somewhat more complicated. We again begin with the $u(n)$ Lie algebra, in this case $u(2)$. Any matrix v in $U(2)$ can be written in the form

$$v = e^{i\tau} \tag{5.7.1}$$

where τ is some linear combination (with real coefficients) of the Hermitian matrices σ^0, σ^1, σ^2, and σ^3, which will be specified shortly. Comparison of (7.1) with (3.9.15) gives the relation

$$\tau = A + iB. \tag{5.7.2}$$

It is convenient to select σ^0 to be the 2×2 identity matrix, and to require that the remaining σ^j be the *Pauli* matrices,

$$\sigma^0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \;\; , \;\; \sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

$$\sigma^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \;\; , \;\; \sigma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{5.7.3}$$

(See Exercise 3.7.31.) Note that the Pauli matrices are all Hermitian and, save for σ^0, are traceless. Then successively setting $\tau = \sigma^j$, with $j = 0, 1, 2, 3$, and using (7.2) specifies four pairs of A, B matrices. Correspondingly, according to (3.9.10), these four pairs of A, B matrices specify four matrices of the form S^c. Finally, using the correspondence (5.1), these four S^c matrices specify four second-degree polynomials. Call these polynomials b^0, b^1, b^2, and b^3, respectively. Carrying out the required calculations gives the results

$$b^0 = (1/2)(z_1^2 + z_2^2 + z_3^2 + z_4^2) = (1/2)(q_1^2 + p_1^2 + q_2^2 + p_2^2),$$

$$\begin{aligned} b^1 &= z_1 z_2 + z_3 z_4 = q_1 q_2 + p_1 p_2, \\ b^2 &= -z_1 z_4 + z_2 z_3 = -q_1 p_2 + q_2 p_1, \\ b^3 &= (1/2)(z_1^2 - z_2^2 + z_3^2 - z_4^2) = (1/2)(q_1^2 + p_1^2 - q_2^2 - p_2^2). \end{aligned} \tag{5.7.4}$$

It is readily verified that the polynomials b^0 through b^3 obey the Poisson bracket rules

$$[b^0, b^j] = 0 \;\; , \;\; j = 0, 1, 2, 3; \tag{5.7.5}$$

$$\begin{aligned} [b^1, b^2] &= -2b^3, \\ [b^2, b^3] &= -2b^1, \\ [b^3, b^1] &= -2b^2. \end{aligned} \tag{5.7.6}$$

These are the rules for the Lie algebra $u(2)$. Observe also that the relations (7.6) are a variant of the rules for the Lie algebra $su(2)$. That is, the rules (7.6) can be written in the form

$$[b^j, b^k] = -2 \sum_{\ell} \epsilon_{jk\ell} b^{\ell}, \tag{5.7.7}$$

where $\epsilon_{jk\ell}$ is the *Levi-Civita* tensor.

The reader is probably aware that the treatment of angular momentum in quantum mechanics, which essentially amounts to a study of the representations of $su(2)$, is facilitated by the introduction of *raising* and *lowering ladder* operators $J_\pm$ as well as the *diagonal* operator J_z. For our purposes it is convenient to employ the analogous polynomials $r(\pm)$ and c defined by the relations

$$r(\pm) = (i/2)(b^1 \pm ib^2) = (i/2)(q_1 \pm ip_1)(q_2 \mp ip_2), \tag{5.7.8}$$

$$c = (-i/\sqrt{2})b^3 = (-i/\sqrt{8})(q_1^2 + p_1^2 - q_2^2 - p_2^2). \tag{5.7.9}$$

They obey the Poisson bracket rules

$$[c, r(\pm)] = \pm(\sqrt{2})r(\pm), \tag{5.7.10}$$

$$[r(+), r(-)] = (\sqrt{2})c. \tag{5.7.11}$$

We note that these rules can also be written in the form

$$: c : r(\pm) = \pm(\sqrt{2})r(\pm), \tag{5.7.12}$$

$$: r(\pm) : c = \mp(\sqrt{2})r(\pm). \tag{5.7.13}$$

$$: r(+) : r(-) = (\sqrt{2})c, \tag{5.7.14}$$

We now turn to the problem of determining the matrices S^a that anticommute with J. As described earlier, the most general real symmetric S can be written in the form (3.9.1) subject to the conditions (3.9.2). Requiring that S^a anticommute with J gives the further restrictions

$$B^T = B, \tag{5.7.15}$$

$$C = -A. \tag{5.7.16}$$

Consequently, the most general S^a is of the form

$$S^a = \begin{pmatrix} A & B \\ B & -A \end{pmatrix}, \tag{5.7.17}$$

with both A and B real and symmetric,

$$A^T = A \quad , \quad B^T = B. \tag{5.7.18}$$

In the 4×4 case of $sp(4)$, both A and B are 2×2. Thus, since they are symmetric, they can be written in the form

$$A = a\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + b\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + c\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{5.7.19}$$

$$B = d\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + e\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + f\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{5.7.20}$$

where the coefficients a through f are abitrary. It follows that in the 4×4 case the vector space spanned by matrices of the form JS^a is six dimensional. Since the Lie algebra generated by matrices of the form JS^c is four dimensional as has already been seen, the dimension of the complete Lie algebra generated by both the matrices JS^c and JS^a is 4+6=10 dimensional, in accord with (3.7.35) for $n = 4$.

The last step is to make a suitable choice of six second-degree polynomials corresponding to six different choices for the matrices S^a. We have found it convenient to first introduce 3 complex polynomials $h^{\pm}$, h^0 by the rules

$$h^+ = -(1/2)(q_1 + ip_1)^2, \tag{5.7.21}$$

$$h^0 = -(1/\sqrt{2})(q_1 + ip_1)(q_2 + ip_2), \tag{5.7.22}$$

$$h^- = (1/2)(q_2 + ip_2)^2. \tag{5.7.23}$$

Under the action of $r(\pm)$ and c they are transformed according to the rules

$$: c : h^{\pm} = \pm(\sqrt{2})h^{\pm}, \tag{5.7.24}$$

$$: c : h^0 = 0, \tag{5.7.25}$$

$$: r(+) : h^- = (\sqrt{2})h^0, \tag{5.7.26}$$

$$: r(+) : h^0 = -(\sqrt{2})h^+. \tag{5.7.27}$$

Note that these rules are analogous to the relations (7.12) through (7.14). Consequently, we may view the 3 objects $h^{\pm}$ and h^0 as the components of a “spin” 1 (vector) object in a spherical basis. [Strictly speaking, we should invent a special terminology for this and related situations. Perhaps we should talk about *unitary* spin 1 to emphasize the fact that the spin we are referring to is with respect to an $SU(2)$ group, and is not an angular momentum spin related to some rotation group.] Finally, under the action of b^0, they are transformed according to the rules

$$: b^0 : h^{\pm} = 2ih^{\pm}, \tag{5.7.28}$$

$$: b^0 : h^0 = 2ih^0. \tag{5.7.29}$$

We now form 6 real polynomials f^j and g^j by taking suitable linear combinations of real and imaginary parts of $h^{\pm}$ and h^0,

$$f^3 = -\sqrt{2}\ \mathrm{Re}\ (h^0) = q_1q_2 - p_1p_2, \tag{5.7.30}$$

$$\begin{aligned} f^1 &= -(1/2)[b^2, f^3] = (1/2)(p_1^2 - q_1^2 - p_2^2 + q_2^2), \\ f^2 &= -(1/2)[b^3, f^1] = -q_1p_1 - q_2p_2, \\ g^3 &= -\sqrt{2}\ \mathrm{Im}\ (h^0) = q_1p_2 + q_2p_1, \qquad (5.7.31) \\ g^1 &= -(1/2)[b^2, g^3] = -q_1p_1 + q_2p_2, \\ g^2 &= -(1/2)[b^3, g^1] = (1/2)(-p_1^2 + q_1^2 - p_2^2 + q_2^2). \end{aligned}$$

The f’s and g’s have been selected in such a way that they obey the Lie algebraic rules

$$[b^j, f^k] = -2\sum_{\ell} \epsilon_{jk\ell} f^{\ell}, \tag{5.7.32}$$

$$[b^j, g^k] = -2\sum_\ell \epsilon_{jk\ell} g^\ell. \tag{5.7.33}$$

That is, they behave like the Cartesian components of spin 1 objects under the action of $su(2)$. Under the action of b^0, the f's and g's are transformed into each other,

$$[b^0, f^j] = -2g^j, \tag{5.7.34}$$

$$[b^0, g^j] = 2f^j.$$

Finally, the Poisson brackets of the f's and g's with each other are given by the relations

$$[f^j, f^k] = 2\sum_\ell \epsilon_{jk\ell} b^\ell, \tag{5.7.35}$$

$$[g^j, g^k] = 2\sum_\ell \epsilon_{jk\ell} b^\ell, \tag{5.7.36}$$

$$[f^j, g^k] = 2\delta_{jk} b^0. \tag{5.7.37}$$

Note that according to the relations (7.32) through (7.34), the f's and g's are transformed among each other under the action of $u(2)$; and the right sides of (7.35) through (7.37) are elements of $u(2)$. This result is in accord with Exercise (3.9.1).

When taken all together, the rules (7.5) through (7.7) and (7.32) through (7.37) specify the Lie algebra $sp(4)$.

Exercises

5.7.1. Carry out the calculations that produce the results (7.4).

5.7.2. Verify the Poisson bracket relations (7.5), (7.6), and (7.7). Show that the Pauli matrices satisfy the analogous commutation rules

$$\{i\sigma^0, i\sigma^j\} = 0, \tag{5.7.38}$$

$$\{i\sigma^j, i\sigma^k\} = -2\sum_\ell \epsilon_{jk\ell}(i\sigma^\ell) \text{ or } \{\sigma^j, \sigma^k\} = 2i\sum_\ell \epsilon_{jk\ell}\sigma^\ell \Leftrightarrow \{\sigma^1, \sigma^2\} = 2i\sigma^3, \text{etc.} \tag{5.7.39}$$

5.7.3. Verify that S^a is of the form (7.10) subject to the conditions (7.11).

5.7.4. Show that the f's and g's given by (7.30) and (7.31) do indeed correspond to matrices of the form S^a, and find these matrices.

5.7.5. Verify the Poisson bracket rules (7.32) through (7.37).

5.7.6. Consider the complex conjugates of the polynomials $h^\pm$ and h^0. Show that they are also transformed among each other as a spin 1 object under the action of $u(2)$. Thus, as already evidenced by the existence of the f^j and g^j, there are *two* spin 1 objects in $sp(4)$ corresponding to the 6 independent matrices of the form JS^a.

5.7.7. Review Exercise 7.2 above. The purpose of this exercise is to further explore properties of the Pauli matrices. Show that they obey the multiplication rules

$$\sigma^j\sigma^k = \delta_{jk}\sigma^0 + i\sum_\ell \epsilon_{jk\ell}\sigma^\ell \text{ for } j,k,\ell = 1,2,3 \Leftrightarrow (\sigma^j)^2 = I \text{ and } \sigma^1\sigma^2 = i\sigma^3, \text{ etc.} \tag{5.7.40}$$

Show, as a special case of (7.40), that they obey the *anticommutation* rules

$$\{\sigma^j,\sigma^k\}_+ = \sigma^j\sigma^k + \sigma^k\sigma^j = 2\delta_{jk}\sigma^0 \quad ; \quad j,k = 1,2,3. \tag{5.7.41}$$

In particular, they anticommute ($\sigma^j\sigma^k = -\sigma^k\sigma^j$) when $j \neq k$.

Show that the Pauli matrices σ^j for $j = 1,2,3$ span the vector space of 2×2 *traceless* Hermitian matrices, and obey the relations

$$\text{tr}(\sigma^j\sigma^k) = 2\delta_{jk} \quad ; \quad j,k = 1,2,3. \tag{5.7.42}$$

Show that there are the additional trace relations

$$\text{tr}(\sigma^j\{\sigma^k,\sigma^\ell\}_+) = 0, \tag{5.7.43}$$

$$\text{tr}(\sigma^j\{\sigma^k,\sigma^\ell\}) = 4i\epsilon_{jk\ell} = -4i(L^j)_{k\ell}, \tag{5.7.44}$$

$$\text{tr}(\sigma^j\sigma^k\sigma^\ell) = 2i\epsilon_{jk\ell} = -2i(L^j)_{k\ell}. \tag{5.7.45}$$

Recall (3.7.182).

Let $\boldsymbol{a}$ be a three-component vector with entries (a_1,a_2,a_3). Introduce the notation

$$\boldsymbol{a}\cdot\boldsymbol{\sigma} = \sum_{j=1}^{3} a_j\sigma^j. \tag{5.7.46}$$

Verify that

$$\boldsymbol{a}\cdot\boldsymbol{\sigma} = \begin{pmatrix} a_3 & a_1 - ia_2 \\ a_1 + ia_2 & -a_3 \end{pmatrix}. \tag{5.7.47}$$

Verify that

$$\det(\boldsymbol{a}\cdot\boldsymbol{\sigma}) = -\boldsymbol{a}\cdot\boldsymbol{a}. \tag{5.7.48}$$

Show that there are the multiplication, commutation, and anticommutation relations

$$(\boldsymbol{a}\cdot\boldsymbol{\sigma})(\boldsymbol{b}\cdot\boldsymbol{\sigma}) = (\boldsymbol{a}\cdot\boldsymbol{b})\sigma^0 + i(\boldsymbol{a}\times\boldsymbol{b})\cdot\boldsymbol{\sigma}, \tag{5.7.49}$$

$$\{(\boldsymbol{a}\cdot\boldsymbol{\sigma}),(\boldsymbol{b}\cdot\boldsymbol{\sigma})\} = 2i(\boldsymbol{a}\times\boldsymbol{b})\cdot\boldsymbol{\sigma}, \tag{5.7.50}$$

$$\{(\boldsymbol{a}\cdot\boldsymbol{\sigma}),(\boldsymbol{b}\cdot\boldsymbol{\sigma})\}_+ = 2(\boldsymbol{a}\cdot\boldsymbol{b})\sigma^0. \tag{5.7.51}$$

Show that the Pauli matrices and σ^0 span the vector space of 2×2 Hermitian matrices, and obey the relations

$$\text{tr}(\sigma^j\sigma^k) = 2\delta_{jk} \quad ; \quad j,k = 0,1,2,3. \tag{5.7.52}$$

5.7.8. The relation (5.3) associates a matrix JS with every quadratic polynomial. Find the matrices B^i (for $i = 0, 1, 2, 3$) associated with the polynomials b^i. Find the matrices F^j and G^j (for $j = 1, 2, 3$) associated with the polynomials f^j and g^j. Use (5.21) if you wish. The B^i, F^j, and G^j provide a basis for the 4×4 matrix representation of $sp(4)$. Find their commutation rules.
Answer:

$$B^0 = J = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} , \quad B^1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}, \tag{5.7.53}$$

$$B^2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix} , \quad B^3 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix},$$

$$F^1 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} , \quad F^2 = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$F^3 = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix} , \quad G^1 = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix},$$

$$G^2 = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} , \quad G^3 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{pmatrix}.$$

Note that these generators/matrices are given for the case that J has the form (3.1.1). In the case that J' as given by (3.2.10) is employed, one may find the related generators by means of the permutation matrix P given by (3.2.18).

5.7.9. Consider the matrices $i\sigma^j$ that obey the $su(2)$ commutation rules (7.39). Also, review Exercise 3.7.36. Form the associated hatted representation given by (3.7.218). Verify that this representation is equivalent to the original representation using the matrix

$$E = i\sigma^2 = J. \tag{5.7.54}$$

Form the associated checked representation given by (3.7.219). Show that in this case the result is the same as using the hatting operation (3.7.218). Consider the matrices B^1, B^2, B^3 given by (7.47). Verify that, as expected, they also provide a representation of $su(2)$. Show that these matrices are unaffected by either of the hatting or checking operations (3.7.218) and (3.7.219).

5.7.10. Review Exercise 3.7.36. Consider the representation of $sp(4,\mathbb{R})$ provided by the matrices (7.47). Since they are real, they are unaffected by the checking' operation (3.7.219). Find the hatted representation given by (3.7.218). Verify that, as expected, this representation is equivalent to the original representation using

$$E = J. \tag{5.7.55}$$

5.7.11. Review Exercise 4.3.19. Let z' denote the collection of phase-space variables with the ordering

$$z' = (q_1, p_1, q_2, p_2). \tag{5.7.56}$$

The purpose of this exercise is to find a relation between the polynomials b^j and the matrices C^j. Show that there is the relation

$$: b^j : z'_c = -\sum_d C^j_{cd} z'_d. \tag{5.7.57}$$

5.7.12. Consider the linear transformation on 4-dimensional phase space given by the rules

$$\bar{q}_1 = q_2, \tag{5.7.58}$$

$$\bar{q}_2 = q_1, \tag{5.7.59}$$

$$\bar{p}_1 = p_2, \tag{5.7.60}$$

$$\bar{p}_2 = p_1. \tag{5.7.61}$$

Verify that it is described by the matrix

$$R = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{5.7.62}$$

Evidently R interchanges the q_1, p_1 and q_2, p_2 planes. Verify that this transformation is symplectic and also that R is orthogonal. Thus, R is in the $U(2)$ subgroup of $Sp(4,\mathbb{R})$ and must be expressible in the form

$$R = \exp(JS^c). \tag{5.7.63}$$

Your task is to find S^c. Verify that R can be written in the form

$$R = M(v) \tag{5.7.64}$$

as described in Section 3.9 and show that

$$v = \sigma^1. \tag{5.7.65}$$

Using (3.7.159), show that σ^1 satisfies the relation

$$\sigma^1 = \exp[(i\pi/2)(\sigma^1 - \sigma^0)]. \tag{5.7.66}$$

Use this result to find S^c.

5.8 Basis for $sp(6,\mathbb{R})$

The case of $sp(6,\mathbb{R})$ is even more complicated, yet the procedure will still be the same. Again we will begin with the unitary Lie algebra, in this case $u(3)$, corresponding to matrices of the form JS^c. Then we will select a basis for matrices of the form JS^a (or, equivalently, a basis for the corresponding second degree polynomials) in such a way that these matrices (polynomials) have convenient transformation properties under the action of $u(3)$ or $su(3)$. This second step will require some discussion of representations of $su(3)$. Fortunately for us $su(3)$ has been well studied, initially by mathematicians, and subsequently by physicists because of its applications to Elementary Particle and Nuclear Physics and the Three-Body problem.

5.8.1 $U(3)$ Preliminaries

Any matrix in $U(3)$ can be written in the form (7.1) where τ is some linear combination (with real coefficients) of $3^2 = 9$ Hermitian matrices $\lambda^0, \lambda^1, \cdots \lambda^8$ that will be listed below. Once these matrices are specified, use of (7.2) in turn specifies 9 pairs of A, B matrices, which in turn according to (3.9.10) specifies 9 matrices of the form S^c. Finally, using the correspondence (5.1), these 9 S^c matrices specify 9 second-degree polynomials.

We select λ^0 to be the 3×3 identity matrix, and require that the remaining λ^j be the *Gell-Mann* (1929-2019) matrices,

$$\lambda^0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad , \quad \lambda^1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad , \quad \lambda^2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{5.8.1}$$

$$\lambda^3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad , \quad \lambda^4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad , \quad \lambda^5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix},$$

$$\lambda^6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad , \quad \lambda^7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix} \quad , \quad \lambda^8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

Note that all the λ matrices are Hermitian and the λ^j (for $j > 0$) are traceless. They satisfy the commutation rules

$$\{i\lambda^0, i\lambda^j\} = 0; \tag{5.8.2}$$

$$\{i\lambda^j, i\lambda^k\} = -2 \sum_{\ell} f_{jk\ell}(i\lambda^\ell) \text{ for } j, k, \ell \neq 0. \tag{5.8.3}$$

Taken together, the rules (8.2) and (8.3) are the commutation rules for $u(3)$. The rules (8.3), for which $j, k, \ell = 1, 2, \cdots 8$, are the commutation rules for $su(3)$.[4] The coefficients $f_{jk\ell}$ (up to a multiplicative constant) are the *structure constants* of $su(3)$. They are real and

[4] We remark that quantum physicists prefer, when possible, to work with Hermitian matrices because in Quantum Mechanics observables are associated with Hermitian operators. (See Exercise 3.7.43.) Therefore judicious factors of i are employed in definitions like (7.3) and (8.1) to achieve this end. In so doing, mathematically extraneous factors of i appear elsewhere. However, in writing (7.38), (7.39), (8.2), and (8.3),

antisymmetric under the interchange of any two (adjacent) indices. Thus most of them are zero. See Exercise 8.2. The table below lists some of them. All the rest are zero, or can be obtained from those listed by permutation of indices and use of the antisymmetry property.

Table 5.8.1: Structure Constants of $su(3)$.

$jk\ell$	$f_{jk\ell}$	$jk\ell$	$f_{jk\ell}$	$jk\ell$	$f_{jk\ell}$
123	1	246	1/2	367	$-1/2$
147	1/2	257	1/2	458	$\sqrt{3}/2$
156	$-1/2$	345	1/2	678	$\sqrt{3}/2$

Before going on, we remark that the Gell-Mann matrices also satisfy the *anticommutation* rules

$$\{\lambda^j, \lambda^k\}_+ = \lambda^j\lambda^k + \lambda^k\lambda^j = (4/3)\delta_{jk}\lambda^0 + 2\sum_\ell d_{jk\ell}\lambda^\ell. \tag{5.8.4}$$

Here the coefficients $d_{jk\ell}$, called the *symmetric coupling coefficients*, are *symmetric* under the interchange of any two indices. See Exercise 8.4. Table 8.2 below lists some of them. All the rest are zero, or can be gotten from those listed by permutation of indices and use of the symmetry property.

Table 5.8.2: Symmetric Coupling Coefficients of $su(3)$.

$jk\ell$	$d_{jk\ell}$	$jk\ell$	$d_{jk\ell}$	$jk\ell$	$d_{jk\ell}$	$jk\ell$	$d_{jk\ell}$
118	$1/\sqrt{3}$	247	$-1/2$	355	1/2	558	$-\sqrt{3}/6$
146	1/2	256	1/2	366	$-1/2$	668	$-\sqrt{3}/6$
157	1/2	338	$1/\sqrt{3}$	377	$-1/2$	778	$-\sqrt{3}/6$
228	$1/\sqrt{3}$	344	1/2	448	$-\sqrt{3}/6$	888	$-1/\sqrt{3}$

5.8.2 Polynomials for $u(3)$

Now successively set $\tau = \lambda^j$ with $j = 0, 1, \cdots 8$, and compute the corresponding second-degree polynomials $b^0, b^1, \cdots b^8$. Doing so gives the results

$$b^0 = (1/2)(q_1^2 + p_1^2 + q_2^2 + p_2^2 + q_3^2 + p_3^2), \tag{5.8.5}$$

$$b^1 = q_1q_2 + p_1p_2,$$

$$b^2 = -q_1p_2 + q_2p_1,$$

we have compensated for this mischief by explicitly displaying i factors in the commutation rules. From a Lie algebraic perspective, the natural basis for any $su(n)$ Lie algebra consists of anti-Hermitian matrices. Thus, for a mathematician, the natural basis for $su(3)$ consists of the matrices $i\lambda^j$ with $j = 1, 2 \cdots 8$.

$$b^3 = (1/2)(q_1^2 + p_1^2 - q_2^2 - p_2^2),$$

$$b^4 = q_1 q_3 + p_1 p_3,$$

$$b^5 = -q_1 p_3 + q_3 p_1,$$

$$b^6 = q_2 q_3 + p_2 p_3,$$

$$b^7 = -q_2 p_3 + q_3 p_2,$$

$$b^8 = (1/\sqrt{12})(q_1^2 + p_1^2 + q_2^2 + p_2^2 - 2q_3^2 - 2p_3^2).$$

It is readily verified that the polynomials b^0 through b^8 obey the Poisson bracket rules

$$[b^0, b^j] = 0 \text{ for } j = 0, 1, \cdots 8; \tag{5.8.6}$$

$$[b^j, b^k] = -2\sum_{\ell} f_{jk\ell} b^\ell \text{ for } j, k, \ell = 1, \cdots 8. \tag{5.8.7}$$

Taken together, these are the rules for the Lie algebra $u(3)$. By themselves, the relations (8.7) are the rules for the Lie algebra $su(3)$.

5.8.3 Plan for the Remaining Polynomials

We next turn to the problem of finding the second-degree polynomials corresponding to the matrices JS^a. As was the case for $sp(4)$, the most general S^a is of the form (7.17) with the matrices A, B subject to the symmetry conditions (7.18). In the 6×6 case of $sp(6)$, both A and B are 3×3. Since there are 6 linearly independent 3×3 symmetric matrices, the space spanned by matrices of the form JS^a is $2 \times 6 = 12$ dimensional. This is as it should be since $9 + 12 = 21$, the dimension of $sp(6)$. What we wish to do is select 12 second-degree polynomials corresponding to the matrices JS^a in such a way that these polynomials have convenient transformation properties under $su(3)$. To do so will require some discussion of what is called the *Cartan basis* for $su(3)$ and of the representations of $su(3)$.

5.8.4 Cartan Basis for $su(3)$

As already mentioned in the previous section, a study of the representations of $su(2)$ is facilitated by the introduction of *raising* and *lowering ladder* operators $J_\pm$ as well as the *diagonal* operator J_z. As discovered by *Killing* and *Cartan*, the same is true for $su(3)$ and all *simple* Lie algebras.[5] In the case of $su(3)$ there are 6 ladder elements that play roles analogous to $J_\pm$; and there are 2 commuting elements that play roles analogous to J_z [for this reason $su(3)$ is said to be of *rank* 2].

Abstractly speaking, a Lie algebra is any set of elements with the properties (3.7.43) through (3.7.45) and (3.7.48) and (3.7.49). For many purposes (including illustrative purposes) it is convenient to work with concrete matrix or differential operator *representations* of Lie algebras. In the matrix case, the Lie algebra consists of linear operators acting on a (usually finite-dimensional) vector space, and the Lie product is matrix commutation. In

[5] Recall that a Lie algebra is called simple if it has no ideals. See Section 8.9. For the use of ladder operators in the case of the symplectic Lie algebras, see Chapter 27.

the differential operator case, the Lie algebra consists of linear differential operators acting on a function space, and the Lie product is differential operator commutation. [As indicated, both these kinds of realizations (*representations*) of Lie algebras and their associated Lie groups are *linear*. There are also *nonlinear* realizations of groups as illustrated, for example, in Section 5.11.]

In the case of $su(3)$, as might be imagined, the smallest matrix representation is realized in terms of 3×3 matrices. For our purposes it is again convenient to employ the Gell-Mann matrices. [However, just as in the case of $su(2)$ for which the smallest matrix representation is realized in terms of the 2×2 Pauli matrices but there are also representations in terms of larger $(2j+1) \times (2j+1)$ matrices, so too there are also representations of $su(3)$ in terms of larger matrices.] We will therefore begin by illustrating for $su(3)$ how the 2 commuting elements and 6 ladder elements are set up in the 3×3 case.

Call the commuting elements C^1 and C^2. In the 3×3 case they are defined by the relations

$$C^1 = (1/\sqrt{2})\lambda^3 \quad , \quad C^2 = (1/\sqrt{2})\lambda^8. \tag{5.8.8}$$

It is easily checked that they do indeed commute,

$$\{C^1, C^2\} = 0. \tag{5.8.9}$$

The 6 ladder elements are conveniently labelled by 3 two-component vectors and their negatives, collectively called *root vectors*. Let $\boldsymbol{e}^1$ and $\boldsymbol{e}^2$ be orthogonal unit vectors. Define three vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ by the relations

$$\begin{aligned} \boldsymbol{\alpha} &= (\sqrt{2})\boldsymbol{e}^1, \qquad (5.8.10)\\ \boldsymbol{\beta} &= (1/\sqrt{2})\boldsymbol{e}^1 + (\sqrt{6}/2)\boldsymbol{e}^2,\\ \boldsymbol{\gamma} &= -(1/\sqrt{2})\boldsymbol{e}^1 + (\sqrt{6}/2)\boldsymbol{e}^2. \end{aligned}$$

Figure 8.1 shows these vectors and their negatives in what is called a *root diagram*. (Note that all the root vectors have length $\sqrt{2}$, and the angle between any two successive root vectors as one goes around the root diagram is 60 degrees.) We denote the ladder elements by $R(\boldsymbol{\mu})$ where $\boldsymbol{\mu}$ is one of the root vectors, i.e. one of the vectors (8.10) or their negatives. The ladder elements are defined by the relations

$$\begin{aligned} R(\pm\boldsymbol{\alpha}) &= (-1/2)(\lambda^1 \pm i\lambda^2), \qquad (5.8.11)\\ R(\pm\boldsymbol{\beta}) &= (-1/2)(\lambda^4 \pm i\lambda^5),\\ R(\pm\boldsymbol{\gamma}) &= (-1/2)(\lambda^6 \pm i\lambda^7). \end{aligned}$$

The choice of elements given by the relations (8.8) and (8.11) is called the *Cartan basis* for $su(3)$. The commuting elements C^j are referred to as the *Cartan subalgebra*, and the *rank* of a simple Lie algebra is the dimension of its Cartan subalgebra.[6] Inspection of (8.8), (8.11), and the Gell-Mann matrices (8.1) reveals that all the C^j and $R(\boldsymbol{\mu})$ are *real* matrices. This is a general feature of the Cartan basis for simple Lie algebras. See, for example, Chapter 27

[6] To be true to history, the Cartan subalgebra could better be called the *Killing* subalgebra since it was he who first recognized and employed it. We also remark that Killing discovered Lie algebras independently of Lie.

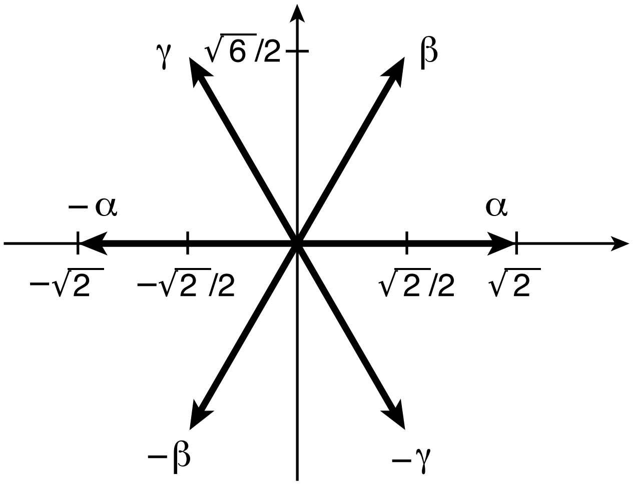


Figure 5.8.1: Root diagram showing the root vectors for $su(3)$.

for the case of $sp(2n, \mathbb{R})$. We also note that matrices of the form $\exp(i\theta_1 C^1 + i\theta_2 C^2)$ produce a *torus*, indeed a 2-torus, in $SU(3)$. See (8.1), (8.5), and Section 3.9. Moreover, this torus has largest dimension for any torus in $SU(3)$. Thus, exponentiating the Cartan subalgebra produces a maximal torus.

We remark that in the Lie algebraic mathematics literature it is customary to denote the elements of the Cartan subalgebra by the symbols H^j rather than our C^j, and the ladder elements by $E(\boldsymbol{\mu})$ rather than our $R(\boldsymbol{\mu})$. We have departed from this common notation because of our desire to generally reserve the symbol H for Hamiltonians and to employ the symbol E for other purposes.

The virtue of the Cartan basis is that the commutation rules take a particularly illuminating form. The commutator of C^j with $R(\boldsymbol{\mu})$ is

$$\{C^j, R(\boldsymbol{\mu})\} = (\boldsymbol{e}^j \cdot \boldsymbol{\mu})R(\boldsymbol{\mu}). \tag{5.8.12}$$

The C^j thus serve to establish the coordinate system for the root vectors.[7] The commutators between pairs of R's, $R(\boldsymbol{\mu})$ and $R(\boldsymbol{\nu})$, are of two types. If the root vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are equal and opposite, the commutator is given by the relation

$$\{R(\boldsymbol{\mu}), R(-\boldsymbol{\mu})\} = \sum_j (\boldsymbol{e}^j \cdot \boldsymbol{\mu})C^j. \tag{5.8.13}$$

If the sum of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ is again a root vector, the commutator takes the form

$$\{R(\boldsymbol{\mu}), R(\boldsymbol{\nu})\} = N(\boldsymbol{\mu}, \boldsymbol{\nu})R(\boldsymbol{\mu} + \boldsymbol{\nu}). \tag{5.8.14}$$

All other commutators vanish. Here $N(\boldsymbol{\mu},\, \boldsymbol{\nu})$ is a numerical factor equal to ±1. The positive N's are $N(\boldsymbol{\alpha},\, -\boldsymbol{\beta})$, $N(\boldsymbol{\gamma},\, \boldsymbol{\alpha})$, $N(-\boldsymbol{\beta},\, \boldsymbol{\gamma})$, $N(\boldsymbol{\beta},\, -\boldsymbol{\alpha})$, $N(-\boldsymbol{\alpha},\, -\boldsymbol{\gamma})$, and N(-$\boldsymbol{\gamma}$, $\boldsymbol{\beta}$).

[7] Note that 8.12 can also be written in the form (ad $C^j)R(\boldsymbol{\mu}) = (\boldsymbol{e}^j \cdot \boldsymbol{\mu})R(\boldsymbol{\mu})$. Thus, the components of the root vectors are the eigenvalues of the linear operators (ad C^j).

5.8.5 Representations of $su(3)$: Cartan's Approach

Suppose the C^j and $R(\boldsymbol{\mu})$ are *any* set of matrices obeying the commutation rules (8.9) and (8.12) through (8.14). Suppose further that a scalar product can be set up in the underlying vector space in such a way that the C^j are Hermitian. See Section 7.3. Let $|\boldsymbol{w}\rangle = |w_1 w_2\rangle$ denote an eigenvector of the C^j with the property

$$C^j|w_1w_2\rangle = w_j|w_1w_2\rangle \text{ or } C^j|\boldsymbol{w}\rangle = (\boldsymbol{e}^j \cdot \boldsymbol{w})|\boldsymbol{w}\rangle. \tag{5.8.15}$$

Since the C^j are Hermitian, the w_j are real. It is convenient, as shown, to treat them together as the components of a single labeling vector denoted as $\boldsymbol{w}$ and called a *weight*. Consider the vector $R(\boldsymbol{\mu})|\boldsymbol{w}\rangle$. From the commutation rules (8.12) we have the relation

$$\begin{aligned} C^jR(\boldsymbol{\mu})|\boldsymbol{w}\rangle &= R(\boldsymbol{\mu})C^j|\boldsymbol{w}\rangle + (\boldsymbol{e}^j\cdot\boldsymbol{\mu})R(\boldsymbol{\mu})|\boldsymbol{w}\rangle \\ &= R(\boldsymbol{\mu})w_j|\boldsymbol{w}\rangle + (\boldsymbol{e}^j\cdot\boldsymbol{\mu})R(\boldsymbol{\mu})|\boldsymbol{w}\rangle \\ &= [\boldsymbol{e}^j\cdot(\boldsymbol{w}+\boldsymbol{\mu})]R(\boldsymbol{\mu})|\boldsymbol{w}\rangle. \end{aligned} \tag{5.8.16}$$

It follows that if $R(\boldsymbol{\mu})|\boldsymbol{w}\rangle$ is different from zero, then it is an eigenvector of the C^j with weight $\boldsymbol{w}+\boldsymbol{\mu}$. Consequently, from a single weight we can produce a whole set of weights. The set of weight vectors can be *ordered* by means of the following definitions:

1. A vector is *positive* if its first nonvanishing component is positive.

2. A vector $\boldsymbol{w}$ is *higher* than the vector $\boldsymbol{w}'$ if $\boldsymbol{w}-\boldsymbol{w}'$ is positive.

We can now state the fundamental theorems of Cartan concerning representations:

1. In any irreducible representation, there is an eigenvector with highest weight, and this eigenvector is unique, i.e., non-degenerate.

2. Two irreducible representations are equivalent if they have the same highest weight.

3. Every highest weight $\boldsymbol{w}^h$ is a linear combination, with non-negative integer coefficients, of what are called *fundamental* weights. For a rank ℓ Lie algebra there are ℓ such fundamental weights. Thus, for a rank ℓ Lie algebra, each irreducible representation is (uniquely) specified by an ℓ-tuple of non-negative integers.

For example in the case of $su(3)$, which is of rank 2, the two fundamental weights $\boldsymbol{\phi}^1$ and $\boldsymbol{\phi}^2$ are given by the relations

$$\boldsymbol{\phi}^1 = (1/\sqrt{2})\boldsymbol{e}^1 + (1/\sqrt{6})\boldsymbol{e}^2, \tag{5.8.17}$$

$$\boldsymbol{\phi}^2 = (1/\sqrt{2})\boldsymbol{e}^1 - (1/\sqrt{6})\boldsymbol{e}^2. \tag{5.8.18}$$

These fundamental weights are shown in Figure 8.2 along with the $su(3)$ root vectors. Consequently, for $su(3)$, every highest weight $\boldsymbol{w}^h$ is of the form

$$\boldsymbol{w}^h = m\boldsymbol{\phi}^1 + n\boldsymbol{\phi}^2 = m[(1/\sqrt{2})\boldsymbol{e}^1 + (1/\sqrt{6})\boldsymbol{e}^2] + n[(1/\sqrt{2})\boldsymbol{e}^1 - (1/\sqrt{6})\boldsymbol{e}^2], \tag{5.8.19}$$

where m and n are arbitrary non-negative integers.

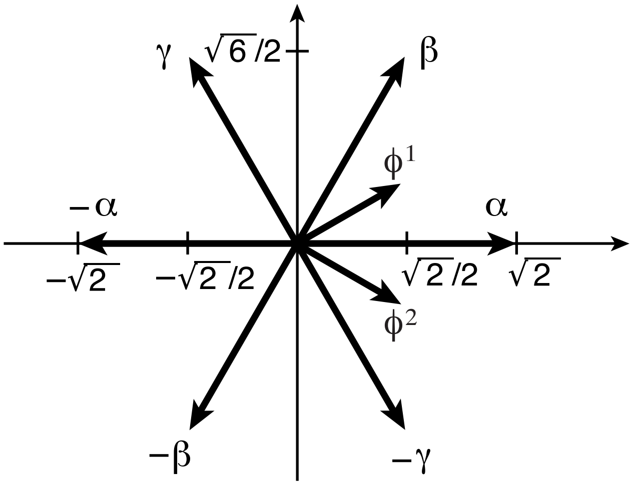


Figure 5.8.2: Fundamental weights $\boldsymbol{\phi}^1$ and $\boldsymbol{\phi}^2$ for $su(3)$. The root vectors are also shown.

Taken together, Cartan's theorems show that an irreducible representation of $su(3)$ is completely characterized by the two non-negative integers m and n. We denote this representation by $\Gamma(m,n)$. It can be shown that the conjugate representation is given by $\Gamma(n,m)$. That is,

$$\overline{\Gamma}(m,n) = \Gamma(n,m). \tag{5.8.20}$$

For discussion and examples see Exercises 3.7.36, 8.29, and 8.30. We also note for future use that the dimension of the representation $\Gamma(m,n)$ is given by the relation

$$\dim \Gamma(m,n) = (m+1)(n+1)(m+n+2)/2. \tag{5.8.21}$$

For quick reference, the dimensions of the first few representations are listed in Table 8.3 below. Note that, as expected, $\Gamma(m,n)$ and $\Gamma(n,m)$ have the same dimension. Finally, for simplicity and where no ambiguity is involved, we sometimes refer to a representation by its dimension. That is, in view of (8.20) and (8.21), we use the shorthand notation $1 = \Gamma(0,0)$, $3 = \Gamma(1,0)$, $\bar{3} = \Gamma(0,1)$, $6 = \Gamma(2,0)$, $\bar{6} = \Gamma(0,2)$, $8 = \Gamma(1,1)$, etc. Note however that $\Gamma(2,1)$ and $\Gamma(4,0)$ as well as their conjugates all have dimension 15.

Table 5.8.3: Dimensions of Representations of $su(3)$.

m	n	$\dim\Gamma(m,n)$	m	n	$\dim\Gamma(m,n)$
0	0	1	4	0	15
1	0	3	0	4	15
0	1	3	3	1	24
2	0	6	1	3	24
0	2	6	2	2	27
1	1	8	5	0	21
3	0	10	0	5	21
0	3	10	4	1	35
2	1	15	1	4	35
1	2	15	3	2	42
			2	3	42

5.8.6 Weight Diagrams for the First Few $su(3)$ Representations

We begin this subsection with the preparatory remark that the overall normalization of the root vectors $\boldsymbol{\mu}$ (and, correspondingly, that of the related fundamental weights) is arbitrary. We have chosen a normalization that facilitates comparison of the root vectors for $su(3)$, $sp(2)$, $sp(4)$, and $sp(6)$. See Chapter 27. For the normalization we have adopted, the $su(3)$ root vectors obey the relation

$$\sum_{\mu}(\boldsymbol{e}^i\cdot\boldsymbol{\mu})(\boldsymbol{\mu}\cdot\boldsymbol{e}^j)=6\delta_{ij}. \tag{5.8.22}$$

To continue our discussion, consider the representation $\Gamma(0,0)$. According to (8.19) its highest weight is the vector zero, and according to (8.21) this representation is one dimensional. Thus, $\Gamma(0,0)$ has only one weight vector. Figure 8.3 displays this vector in what is called a *weight diagram.* Since $\Gamma(0,0)$ is one dimensional, and as described earlier, it is often referred to by its dimension, 1.

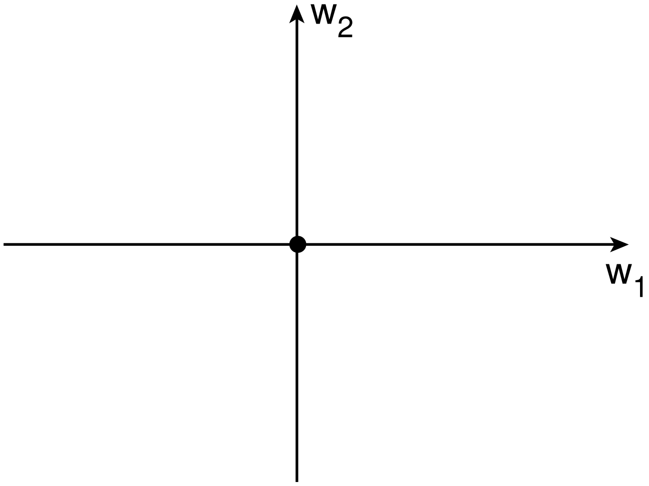


Figure 5.8.3: Weight diagram for the representation $1 = \Gamma(0, 0)$.

Consider the representation $\Gamma(1, 0)$. The highest weight $\boldsymbol{w}^h$ for this representation is shown in Figure 8.4. Also shown are all other weights obtained from $\boldsymbol{w}^h$ by adding and subtracting integer multiples of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$. We observe that there are 3 different weights. Correspondingly, in accord with (8.21), $\Gamma(1, 0)$ is a 3-dimensional representation. It is often referred to by its dimension, 3.

Next consider the representation $\Gamma(0, 1)$. Its weights are shown in Figure 8.5. Evidently this representation is also 3 dimensional. In view of (8.20) and (8.21), it is often referred to as $\overline{3}$. From (8.19) we find that the highest weights for the representations 3 and $\overline{3}$ are given by the relations

$$\boldsymbol{w}^h(3) = (1/\sqrt{2})\boldsymbol{e}^1 + (1/\sqrt{6})\boldsymbol{e}^2, \tag{5.8.23}$$

$$\boldsymbol{w}^h(\overline{3}) = (1/\sqrt{2})\boldsymbol{e}^1 - (1/\sqrt{6})\boldsymbol{e}^2. \tag{5.8.24}$$

Note also that all the weights for $\bar{3}$ are related to those for 3 by the operation of reflection across the w_1 axis. This is a general result. The weights for $\bar{\Gamma}(m, n)$ are related to those of $\Gamma(m, n)$ by reflection across the w_1 axis. It is a consequence of (8.20) and the fact that the fundamental weights $\boldsymbol{\phi}^1$ and $\boldsymbol{\phi}^2$ are interchanged by reflection across the $\boldsymbol{e}^1$ axis. See (8.17), (8.18), and Figure 8.2.

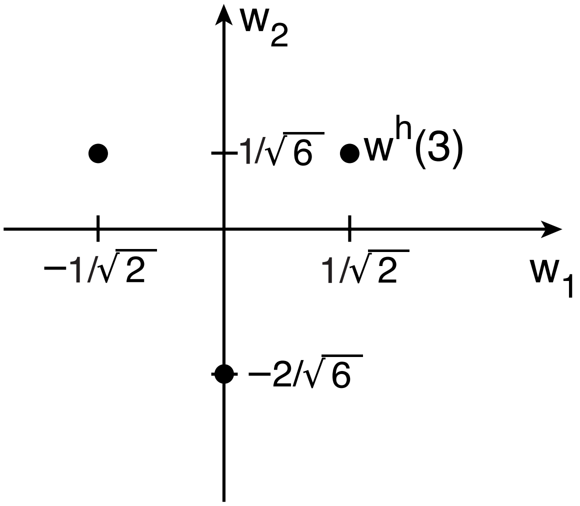


Figure 5.8.4: Weight diagram for the representation $3 = \Gamma(1,0)$.

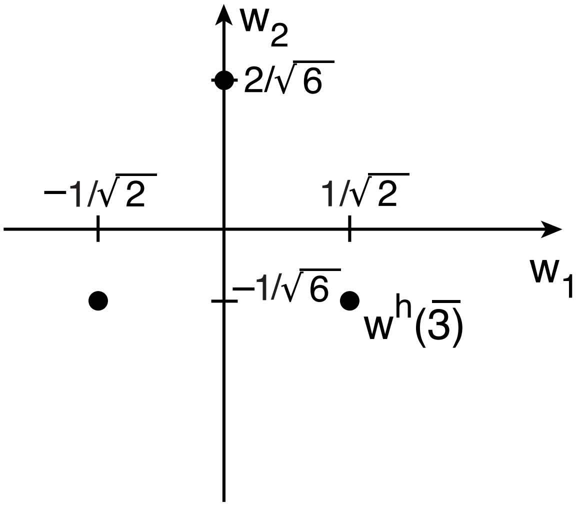


Figure 5.8.5: Weight diagram for the representation $\overline{3} = \Gamma(0,1)$.

Figures 8.6 through 8.8 show the weight diagrams for the representations $\Gamma(2,0)$, $\Gamma(0,2)$, and $\Gamma(1,1)$. The highest weights in these cases are

$$\boldsymbol{w}^h(6) = (\sqrt{2})\boldsymbol{e}^1 + (2/\sqrt{6})\boldsymbol{e}^2, \tag{5.8.25}$$

$$\boldsymbol{w}^h(\overline{6}) = (\sqrt{2})\boldsymbol{e}^1 - (2/\sqrt{6})\boldsymbol{e}^2, \tag{5.8.26}$$

$$\boldsymbol{w}^h(8) = (\sqrt{2})\boldsymbol{e}^1 = \boldsymbol{\alpha}. \tag{5.8.27}$$

According to (8.21) the dimensionality of these representations are 6, 6, and 8, respectively. Observe that Figures 8.6 and 8.7 for $\Gamma(2,0)$ and $\Gamma(0,2)$ each contain 6 weights. Correspondingly, since 6 is the dimension of each of these representations, we conclude that the corresponding eigenvector for each weight $\boldsymbol{w}$ is unique. By contrast, Figure 8.8 for $\Gamma(1,1)$

only contains 7 weights while we know that the dimension of $\Gamma(1,1)$ is 8. It can be shown that the eigenvectors corresponding to the 6 weight vectors $\boldsymbol{w}$ in the diagram at the hexagonal vertices are nondegenerate. However, there are *two* linearly independent eigenvectors corresponding to the weight at the origin. That is, an additional label, beyond the weight itself, is necessary to completely specify these vectors. Note that $6 + 2 = 8$, the dimension of $\Gamma(1,1)$.

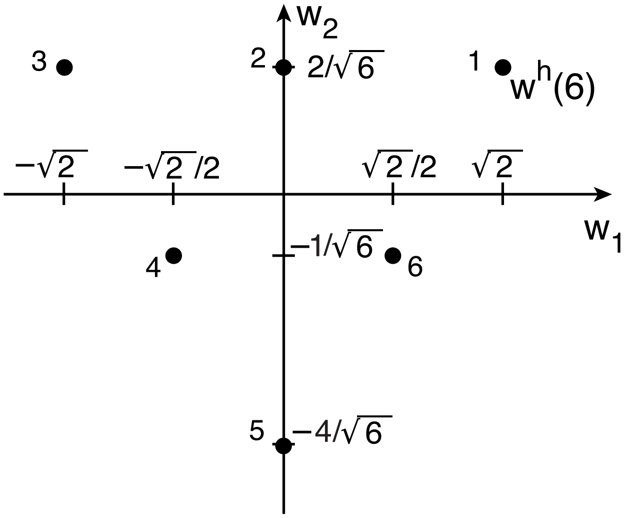


Figure 5.8.6: Weight diagram for the representation $6 = \Gamma(2,0)$.

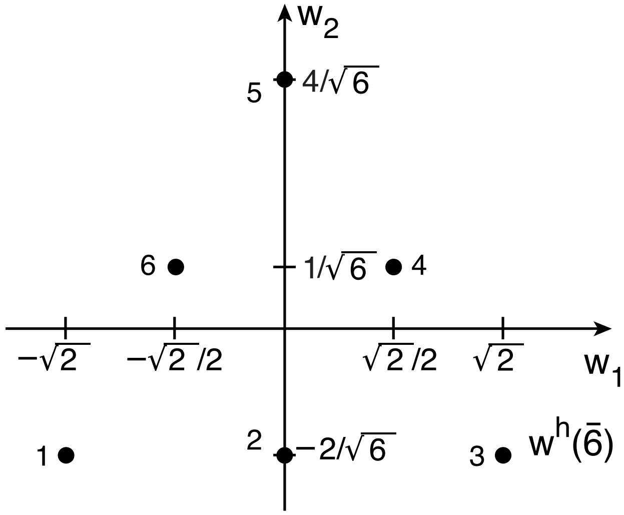


Figure 5.8.7: Weight diagram for the representation $\overline{6} = \Gamma(0,2)$.

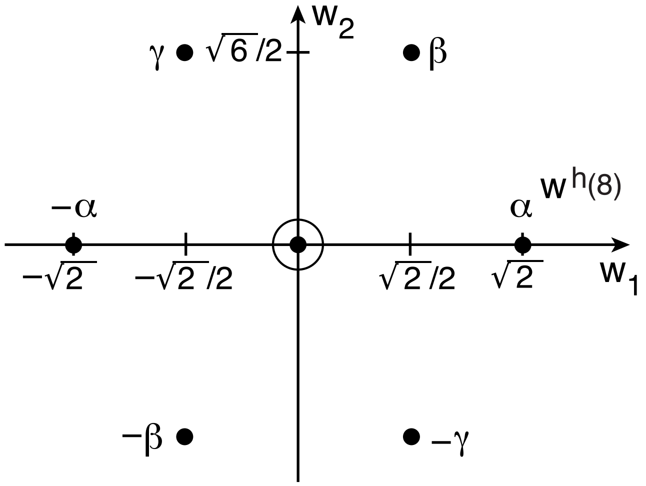


Figure 5.8.8: Weight diagram for the adjoint representation $8 = \Gamma(1,1)$. The 6 weights at the hexagonal vertices lie at the tips of the root vectors $\pm\boldsymbol{\alpha}$, $\pm\boldsymbol{\beta}$, $\pm\boldsymbol{\gamma}$ shown in Figure 8.1. The highest weight lies at the tip of the vector $\boldsymbol{\alpha}$. There are two eigenvectors corresponding to the weight at the origin.

5.8.7 Weight Diagram for the General $su(3)$ Representation

Consider the general representation $\Gamma(m,n)$. Figure 8.9 shows the general form of the weight diagram for this representation. It consists of concentric layers that may be constructed as follows:

1. Find and plot the highest weight $\boldsymbol{w}^h$ using (8.19).

2. Plot the points $\boldsymbol{w}^h{+}\boldsymbol{\gamma}, \boldsymbol{w}^h{+}2\gamma, \cdots \boldsymbol{w}^h{+}n\boldsymbol{\gamma}$ and the points $\boldsymbol{w}^h{-}\boldsymbol{\beta}$, $\boldsymbol{w}^h{-}2\boldsymbol{\beta}, \cdots \boldsymbol{w}^h{-}m\boldsymbol{\beta}$.

3. Reflect the points obtained in step 2 above across the w_2 axis and plot them.

4. Taken together, steps 2 and 3 produce the weights that lie on the left and right boundaries of the weight diagram. Now we need to fill in the top and bottom boundaries. They are the points $\boldsymbol{w}^h + n\boldsymbol{\gamma} - \boldsymbol{\alpha}, \boldsymbol{w}^h + n\boldsymbol{\gamma} - 2\boldsymbol{\alpha}, \cdots \boldsymbol{w}^h + n\boldsymbol{\gamma} - m\boldsymbol{\alpha}$ and $\boldsymbol{w}^h - m\boldsymbol{\beta} - \boldsymbol{\alpha}, \boldsymbol{w}^h - m\boldsymbol{\beta} - 2\boldsymbol{\alpha}, \cdots \boldsymbol{w}^h - m\boldsymbol{\beta} - n\boldsymbol{\alpha}$.

5. The weights on the boundary have now been found, and only the weights in the interior remain to be determined. Next, if the boundary is not triangular, find the point $\boldsymbol{w}^h - \boldsymbol{\alpha}$. Starting from this point, form the next outermost layer by repeating steps 2 through 4 with (m,n) replaced by $(m-1, n-1)$.

6. Form successive concentric layers following step 5 starting from the points $\boldsymbol{w}^h - 2\boldsymbol{\alpha}, \boldsymbol{w}^h - 3\boldsymbol{\alpha}$, etc., until a triangular layer (or the origin) is reached. If a triangular layer is reached, all successive layers will also be triangles, and the innermost layer will be either the point at the origin or a triangle that is the same as one of those shown in Figures 8.4 through 8.7.

The result of this process is a set of weights that are related under translation in the directions $\pm\boldsymbol{\alpha}, \pm\boldsymbol{\beta}, \pm\boldsymbol{\gamma}$. It can be shown that all eigenvectors $|\boldsymbol{w}\rangle$ corresponding to weights $\boldsymbol{w}$ on a given layer have the same multiplicity. Those on the boundary have multiplicity 1 (are nondegenerate). If the boundary is not triangular, the eigenvectors $|\boldsymbol{w}\rangle$ corresponding to weights on the next outermost layer will have multiplicity 2. The multiplicity will continue to increase by 1 for each consecutive layer until a triangular layer (which may be the origin) is reached. The vectors $|\boldsymbol{w}\rangle$ corresponding to this layer will also have a multiplicity one unit larger than those of the previous layer. However, all vectors $|\boldsymbol{w}\rangle$ corresponding to consecutive layers *inside* the triangle will have the same multiplicity as those of the outermost triangle. That is, the multiplicity remains constant after a triangular layer is reached. For example, referring to Figure 8.9, the multiplicity of the eigenvectors $|\boldsymbol{w}\rangle$ corresponding to the boundary layer is 1, and the multiplicities for the next two layers in are 2 and 3 respectively. The multiplicity of the eigenvectors for the first triangular layer is 4, and the multiplicity for the triangular layer inside it is also 4.

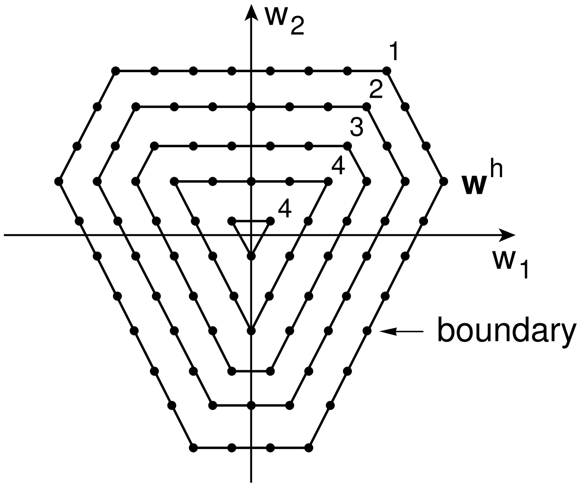


Figure 5.8.9: General form of the weight diagram for the representation $\Gamma(m,n)$. Shown here is the case $(m,n) = (7,3)$. All eigenvectors $|\boldsymbol{w}\rangle$ corresponding to weights $\boldsymbol{w}$ on a given layer have the same multiplicity. Those corresponding to sites on the boundary have multiplicity 1. Those corresponding to sites on the next two layers have multiplicities 2 and 3, respectively. Those corresponding to sites on the two triangular layers have multiplicity 4.

5.8.8 The Clebsch-Gordan Series for $su(3)$

Review of $su(2)$ Results

In the quantum-mechanical treatment of angular momentum, which is essentially an exercise in the properties of $su(2)$, there is the result that two spin 1/2 entities can be combined to form entities with spin 0 and spin 1. If we denote the spin j representation of $su(2$ by the symbols $\Gamma(j)$, then we may summarize this result by writing

$$\Gamma(1/2) \otimes \Gamma'(1/2) = \Gamma(0) \oplus \Gamma(1). \tag{5.8.28}$$

Here we have denoted the second spin 1/2 entity on the left side of (8.28) by a prime to acknowledge that the two spin 1/2 entities my be different. Indeed, the spin 0 combination [appearing on the right side of (8.28) and called the *singlet* state] is odd under the interchange/permutation of the spin 1/2 entities, and the spin 1 combination (called the *triplet* state) is even under the interchange of the spin 1/2 entities. Consequently, if the two spin 1/2 entities are the same, the $\Gamma(0)$ entry on the right side of (8.28) is empty.

Similarly, two spin 1 entities can be combined to form entities with spins 0, 1, and 2; and we may summarize this result by writing

$$\Gamma(1) \otimes \Gamma'(1) = \Gamma(0) \oplus \Gamma(1) \oplus \Gamma(2). \tag{5.8.29}$$

If we view the two spin 1 entities on the left side of (8.29) as being the three-component vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ then, for the entries on the right side of (8.29), there are the correspondences

$$\Gamma(0) \leftrightarrow \boldsymbol{u} \cdot \boldsymbol{v}, \tag{5.8.30}$$

$$\Gamma(1) \leftrightarrow \boldsymbol{u} \times \boldsymbol{v}, \tag{5.8.31}$$

$$\Gamma(2) \leftrightarrow (1/2)(u_a v_b + u_b v_a) - (1/3)\delta_{ab}(\boldsymbol{u} \cdot \boldsymbol{v}). \tag{5.8.32}$$

Note that $\Gamma(1)$ as given by (8.31) is odd under the interchange of $\boldsymbol{u}$ and $\boldsymbol{v}$. Consequently the $\Gamma(1)$ entry is empty in the case that $\boldsymbol{u} = \boldsymbol{v}$. By contrast $\Gamma(2)$ (which is a symmetric traceless tensor) and $\Gamma(0)$ are even under the interchange of $\boldsymbol{u}$ and $\boldsymbol{v}$.

We have given specific instances of the Clebsch-Gordan series for $su(2)$. It can be shown that for any two spins there is the general Clebsch-Gordan relation

$$\Gamma(j) \otimes \Gamma'(j') = \Gamma(j+j') \oplus \Gamma(j+j'-1) \oplus \Gamma(j+j'-2) \oplus \cdots \oplus \Gamma(|j-j'|). \tag{5.8.33}$$

Here all the representations on the right side occur once and only once unless some happen to be empty in the case of some possible interchange symmetry.

The Case of $su(3)$

We have briefly reviewed the Clebsch-Gordan series for representations of $su(2)$. There are similar combination rules for representations of $su(3)$ and, indeed, for all the representations of all the simple groups.[8] The remaining task of this subsection is to review these rules for

[8] See Chapter 27 for the case of the symplectic group.

the case of $su(3)$. We begin with some specific cases. For some of the first few representations of $su(3)$ there are the results

$$3 \otimes 3' = \overline{3} \oplus 6, \tag{5.8.34}$$

$$\overline{3} \otimes \overline{3}' = 3 \oplus \overline{6}, \tag{5.8.35}$$

$$3 \otimes \overline{3} = 1 \oplus 8, \tag{5.8.36}$$

$$6 \otimes 6' = \Gamma(2,0) \otimes \Gamma'(2,0) = \Gamma(0,2) \oplus \Gamma(2,1) \oplus \Gamma(4,0), \tag{5.8.37}$$

$$6 \otimes \overline{6} = 1 \oplus 8 \oplus 27, \tag{5.8.38}$$

$$8 \otimes 6 = \Gamma(1,1) \otimes \Gamma(2,0) = \Gamma(0,1) \oplus \Gamma(2,0) \oplus \Gamma(1,2) \oplus \Gamma(3,1), \tag{5.8.39}$$

$$8 \otimes 8' = 1 \oplus 8 \oplus 8 \oplus 10 \oplus \overline{10} \oplus 27. \tag{5.8.40}$$

In (8.37) and (8.39) the $\Gamma(m,n)$ notation is used to specify a representation since not all the representations appearing in (8.37) and (8.39) are uniquely specified by their dimensions. See Table 8.3. We also remark that if the two factors appearing on the left side of a Clebsch-Gordan relation are potentially the same, as for example in (8.34), (8.35), (8.37), and (8.40), then some of the terms appearing on the right side are potentially empty. See, for example, Exercises 8.21 and 8.23.

Just as is the case for $su(2)$ where (8.33) provides gives an explicit result for combining any two spins, there is also an explicit formula for the general case of $su(3)$. Use the shorthand notation (j_1, j_2) to denote the $su(3)$ representation $\Gamma(j_1, j_2)$. The Clebsch-Gordan series for $su(3)$ in the general case is given by the relation

$$(j_1, j_2) \otimes (j_1', j_2')' = \sum_{i=0}^{min(j_1, j_2')} \sum_{k=0}^{min(j_2, j_1')} (j_1 - i, j_1' - k; j_2 - k, j_2' - i), \tag{5.8.41}$$

where the quantity $(n, n'; m, m')$ is defined by the relation

$$(n, n'; m, m') = (n + n', m + m') \ \oplus \sum_{i=1}^{min(n,n')} (n + n' - 2i, m + m' + i)$$

$$\oplus \sum_{k=1}^{min(m,m')} (n + n' + k, m + m' - 2k). \tag{5.8.42}$$

All the sums in the expressions above are direct sums.

5.8.9 Representations of $su(3)$: the Approach of Schur and Weyl

Subsections 8.4 through 8.7 have illustrated how, following Cartan, the representations of $su(3)$ can be described in terms of ladder operators and weight diagrams. We will employ the same approach in Chapter 27 for the case of $sp(2n)$. However, we take here the opportunity to mention an alternate approach due to Schur (1875-1941) and Weyl.

The method of Cartan has the feature that the properties of any given representation are described without reference to the properties of any other representation. Each representation is treated in isolation. By contrast, the approach of Schur and Weyl capitalizes on the fact (as illustrated by the Clebsch-Gordan series) that suitable tensor products of low-dimensional representations contain higher-dimensional representations. In particular it can be shown for the case of $su(3)$ that by forming suitable tensor products of multiple copies of $3 = \Gamma(1,0)$ and $\bar{3} = \Gamma(0,1)$ one obtains a multi-index tensor representation of $su(3)$ that contains any desired irreducible representation. With this result in hand, the remaining task is to extract and label from such a representation the desired irreducible representation. This is done by possibly tracing over some index pairs and by forming linear combinations over various permutations of other indices with these linear combinations being described by *Young* (1873-1940) *tableaux*. Thus, in the approach of Schur and Weyl, each representation is labeled by a Young tableau.

5.8.10 Remaining Polynomials

With this brief background on representation theory for $su(3)$, we are prepared to construct 12 second-degree polynomials corresponding to the matrices JS^a in such a way that these polynomials have convenient transformation properties under $su(3)$.

su(3) Decomposition of Homogenous Polynomials

First we state a general result: Let f_ℓ be a homogeneous polynomial of degree ℓ in the six phase-space variables $z_1 \cdots z_6$. Then it is easily verified that the quantities $:b^1: f_\ell$ through $:b^8: f_\ell$ are also homogeneous polynomials of degree ℓ. Consequently, the subspace of homogeneous polynomials of degree ℓ is sent into itself under the action of $su(3)$. Next, it can be shown that each f_ℓ subspace can itself be decomposed into smaller subspaces that are each sent into themselves separately under the action of $su(3)$. Indeed, this can be done in such a way that each smaller subspace forms an irreducible representation of $su(3)$. See Section 34.2.4. When this is done, the following results are found:

1. Suppose ℓ is even. Then f_ℓ has the direct sum decomposition

$$f_\ell = \sum_{m+n=\ell} \Gamma(m,n) \oplus \sum_{m+n=\ell-2} \Gamma(m,n) \oplus \sum_{m+n=\ell-4} \Gamma(m,n) \oplus \cdots \oplus \Gamma(0,0). \tag{5.8.43}$$

   Each representation listed in (8.43) occurs once and only once. For example, f_0, f_2, and f_4 have the decompositions

$$f_0 = \Gamma(0,0), \tag{5.8.44}$$

$$f_2 = \Gamma(2,0) \oplus \Gamma(1,1) \oplus \Gamma(0,2) \oplus \Gamma(0,0), \tag{5.8.45}$$

$$f_4 = \Gamma(4,0) \oplus \Gamma(3,1) \oplus \Gamma(2,2) \oplus \Gamma(1,3) \oplus \Gamma(0,4) \oplus \Gamma(2,0) \oplus \Gamma(1,1) \oplus \Gamma(0,2) \oplus \Gamma(0,0). \tag{5.8.46}$$

2. Suppose ℓ is odd. Then f_ℓ has the direct sum decomposition

$$\begin{aligned} f_\ell = \quad & \sum_{m+n=\ell} \Gamma(m,n) \oplus \sum_{m+n=\ell-2} \Gamma(m,n) \oplus \sum_{m+n=\ell-4} \Gamma(m,n) \oplus \cdots \\ & \oplus \Gamma(1,0) \oplus \Gamma(0,1). \end{aligned} \tag{5.8.47}$$

Each representation listed in (8.47) occurs once and only once. For example, f_1 and f_3 have the decompositions

$$f_1 = \Gamma(1,0) \oplus \Gamma(0,1), \tag{5.8.48}$$

$$f_3 = \Gamma(3,0) \oplus \Gamma(2,1) \oplus \Gamma(1,2) \oplus \Gamma(0,3) \oplus \Gamma(1,0) \oplus \Gamma(0,1). \tag{5.8.49}$$

We will use these results below for the special case of quadratic polynomials. But we remark that these results are also useful for the construction of Cremona maps and determining the long-term behavior of particles in storage rings. Again see Section 34.2.4.

Explicit Results for Remaining Quadratic Polynomials

For our present discussion we are interested in the case of quadratic polynomials, the generators of $sp(6)$. According to the previous paragraph, they have the decomposition (8.45). It can be shown that the $\Gamma(0,0)$ part in (8.45) corresponds to a b^0 part as given in (8.5), and the $\Gamma(1,1)$ part corresponds to the b^1 through b^8 parts given in (8.5). See Exercise 8.19. What remains is the $\Gamma(2,0) \oplus \Gamma(0,2)$ part. It has dimension $6+6=12$, which is the dimension of the set of matrices JS^a. This circumstance suggests that the second-degree polynomials corresponding to the matrices JS^a might be arranged to transform under the action of $su(3)$ according to the representation $\Gamma(2,0) \oplus \Gamma(0,2) = 6 \oplus \overline{6}$. This is indeed the case. As a sanity check on our hypothesis, let us do a dimension count. We know that $sp(6)$ has dimension 21. Together b^0 and the b^1 through b^8 span a space of dimension $1+8=9$. Observe that $9+6+6=21$, as desired.

Let the symbols c^j and $r(\boldsymbol{\mu})$ denote the second-degree polynomials corresponding to the C^j and $R(\boldsymbol{\mu})$. They are selected and normalized in such a way that their Lie algebra (with the Poisson bracket as the Lie product) is the same as the Lie algebra of the C^j and $R(\boldsymbol{\mu})$ (with the commutator as the Lie product). Calculation shows that they are given by the relations

$$c^1 = -(i/\sqrt{2})b^3 = (-i/\sqrt{8})(q_1^2 + p_1^2 - q_2^2 - p_2^2), \tag{5.8.50}$$

$$c^2 = -(i/\sqrt{2})b^8 = (-i/\sqrt{24})(q_1^2 + p_1^2 + q_2^2 + p_2^2 - 2q_3^2 - 2p_3^2);$$

$$r(\pm\boldsymbol{\alpha}) = (i/2)(b^1 \pm ib^2) = (i/2)(q_1 \pm ip_1)(q_2 \mp ip_2), \tag{5.8.51}$$

$$r(\pm\boldsymbol{\beta}) = (i/2)(b^4 \pm ib^5) = (i/2)(q_1 \pm ip_1)(q_3 \mp ip_3),$$

$$r(\pm\boldsymbol{\gamma}) = (i/2)(b^6 \pm ib^7) = (i/2)(q_2 \pm ip_2)(q_3 \mp ip_3).$$

Define six weight vectors $\boldsymbol{w}^1 \cdots \boldsymbol{w}^6$ for $\Gamma(2,0)$ by the rules

$$\boldsymbol{w}^1 = \boldsymbol{w}^h(6) \quad , \quad \boldsymbol{w}^2 = \boldsymbol{w}^1 - \boldsymbol{\alpha},$$

$$\boldsymbol{w}^3 = \boldsymbol{w}^2 - \boldsymbol{\alpha} \quad , \quad \boldsymbol{w}^4 = \boldsymbol{w}^3 - \boldsymbol{\gamma},$$

$$\boldsymbol{w}^5 = \boldsymbol{w}^4 - \boldsymbol{\gamma} \quad , \quad \boldsymbol{w}^6 = \boldsymbol{w}^5 + \boldsymbol{\beta}. \tag{5.8.52}$$

See Figures 8.1 and 8.6, and note that the weights shown in Figure 8.6 are numbered in accord with (8.52). Define six corresponding polynomials $h^1 \cdots h^6$ by the relations

$$h^1 = (1/2)(q_1 + ip_1)^2,$$

$$\begin{aligned}
h^2 &= (q_1 + ip_1)(q_2 + ip_2),\\
h^3 &= (1/2)(q_2 + ip_2)^2,\\
h^4 &= (q_2 + ip_2)(q_3 + ip_3),\\
h^5 &= (1/2)(q_3 + ip_3)^2,\\
h^6 &= (q_3 + ip_3)(q_1 + ip_1). \qquad (5.8.53)
\end{aligned}$$

It is easy to check that the h^k are all simultaneous eigenvectors of the $: c^j :$ with eigenvalues corresponding to the weights $\boldsymbol{w}^k$,

$$: c^j : h^k = (\boldsymbol{e}^j \cdot \boldsymbol{w}^k)h^k. \qquad (5.8.54)$$

Also, there are ladder relations, corresponding to the relations (8.52), of the form

$$h^2 \propto : r(-\boldsymbol{\alpha}) : h^1,$$

$$h^3 \propto : r(-\boldsymbol{\alpha}) : h^2 \ \ , \ \ h^4 \propto : r(-\boldsymbol{\gamma}) : h^3,$$
$$h^5 \propto : r(-\boldsymbol{\gamma}) : h^4 \ \ , \ \ h^6 \propto : r(+\boldsymbol{\beta}) : h^5. \qquad (5.8.55)$$

Finally, calculation shows that the action of b^0 is given by the relation

$$: b^0 : h^k = 2ih^k. \qquad (5.8.56)$$

We conclude that the six polynomials $h^1 \cdots h^6$ transform according to the representation 6 under the action of $su(3)$, and also are transformed among each other under the action of the full $u(3)$. See Exercise 8.14.

With the h^k determined, the construction of a second set of six polynomials corresponding to the representation $\overline{6}$ is easy. Take the complex conjugate of both sides of the relations (8.54) and (8.56). Doing so gives the results

$$: \overline{c}^j : \overline{h}^k = (\boldsymbol{e}^j \cdot \boldsymbol{w}^k)\overline{h}^k, \qquad (5.8.57)$$

$$: \overline{b}^0 : \overline{h}^k = -2i\overline{h}^k. \qquad (5.8.58)$$

However, inspection of (8.5) and (8.50) gives the relations

$$\overline{b}^0 = b^0 \ \ , \ \ \overline{c}^j = -c^j. \qquad (5.8.59)$$

Consequently, we also have the results

$$: c^j : \overline{h}^k = -(\boldsymbol{e}^j \cdot \boldsymbol{w}^k)\overline{h}^k, \qquad (5.8.60)$$

$$: b^0 : \overline{h}^k = -2i\overline{h}^k. \qquad (5.8.61)$$

Upon comparing the weight diagrams in Figures 8.6 and 8.7 for the representations 6 and $\overline{6}$, we conclude that the polynomials $\overline{h}^k$ transform according to the representation $\overline{6}$.

Our task of finding a suitable set of polynomials corresponding to the matrices JS^a is almost finished. For physical applications, we will want to work with real polynomials instead of the complex polynomials h^k given by (8.53). This task is easily accomplished. Define real polynomials f^k and g^k by writing the relations

$$h^k = f^k + ig^k. \tag{5.8.62}$$

Doing so gives the results

$$\begin{aligned}
f^1 &= (1/2)(q_1^2 - p_1^2),\\
f^2 &= q_1q_2 - p_1p_2,\\
f^3 &= (1/2)(q_2^2 - p_2^2),\\
f^4 &= q_2q_3 - p_2p_3,\\
f^5 &= (1/2)(q_3^2 - p_3^2),\\
f^6 &= q_3q_1 - p_3p_1;\\
g^1 &= q_1p_1,\\
g^2 &= q_1p_2 + q_2p_1,\\
g^3 &= q_2p_2,\\
g^4 &= q_2p_3 + q_3p_2,\\
g^5 &= q_3p_3,\\
g^6 &= q_3p_1 + q_1p_3.
\end{aligned} \tag{5.8.63}$$

Since the h^k are transformed among themselves under the action of the full $u(3)$, the Poisson brackets $[b^j, h^k]$ can be written in the form

$$[b^j, h^k] = \sum_\ell \zeta_{jk\ell} h^\ell. \tag{5.8.64}$$

The results for the cases $j = 0, 3, 8$ follow from (8.56), (8.54), and (8.50):

$$\begin{aligned}
[b^0, h^k] &= 2ih^k, \\
[b^3, h^k] &= i(\sqrt{2})(\boldsymbol{e}^1 \cdot \boldsymbol{w}^k)h^k,\\
[b^8, h^k] &= i(\sqrt{2})(\boldsymbol{e}^2 \cdot \boldsymbol{w}^k)h^k.
\end{aligned} \tag{5.8.65}$$

Calculation of the other Poisson brackets requires somewhat more work, the results of which will be presented shortly in tabular form. Suppose the coefficients $\zeta_{jk\ell}$ are decomposed into real and imaginary parts by writing the relations

$$\zeta_{jk\ell} = \xi_{jk\ell} + i\eta_{jk\ell}. \tag{5.8.66}$$

Then equating real and imaginary parts of (8.64), observing that the b^j are real, and using the decomposition (8.62) give the results

$$\begin{aligned}
[b^j, f^k] &= \sum_\ell \xi_{jk\ell} f^\ell - \eta_{jk\ell} g^\ell,\\
[b^j, g^k] &= \sum_\ell \eta_{jk\ell} f^\ell + \xi_{jk\ell} g^\ell.
\end{aligned} \tag{5.8.67}$$

The nonzero values of the $\xi_{jk\ell}$ and $\eta_{jk\ell}$ are tabulated below.

Table 5.8.4: Some Structure Constants of $sp(6)$.

$jk\ell$	$\xi_{jk\ell}$	$\eta_{jk\ell}$	$jk\ell$	$\xi_{jk\ell}$	$\eta_{jk\ell}$	$jk\ell$	$\xi_{jk\ell}$	$\eta_{jk\ell}$
011	0	2	311	0	2	634	0	1
022	0	2	333	0	-2	643	0	2
033	0	2	344	0	-1	645	0	2
044	0	2	366	0	1	654	0	1
055	0	2	416	0	1	662	0	1
066	0	2	424	0	1	726	-1	0
112	0	1	442	0	1	734	-1	0
121	0	2	456	0	1	743	2	0
123	0	2	461	0	2	745	-2	0
132	0	1	465	0	2	754	1	0
146	0	1	516	-1	0	762	1	0
164	0	1	524	-1	0	811	0	$2/\sqrt{3}$
212	-1	0	542	1	0	822	0	$2/\sqrt{3}$
221	2	0	556	1	0	833	0	$2/\sqrt{3}$
223	-2	0	561	2	0	844	0	$-1/\sqrt{3}$
232	1	0	565	-2	0	855	0	$-4/\sqrt{3}$
246	1	0	626	0	1	866	0	$-1/\sqrt{3}$
264	-1	0						

It remains to compute the Poisson brackets of the f's and g's with themselves. First we observe, as can be easily verified, that the h^k are all in involution,

$$[h^j, h^k] = 0. \tag{5.8.68}$$

Next, since the commutator of two matrices of the form JS^a is a matrix of the form JS^c (see Exercise 3.9.1), we must have a relation of the form

$$[h^j, \overline{h}^k] = \sum_\ell \tau_{jk\ell} b^\ell. \tag{5.8.69}$$

Using the decomposition (8.62) and taking real and imaginary parts of (8.68) give the results

$$[f^j, f^k] = [g^j, g^k], \tag{5.8.70}$$

$$[f^j, g^k] = [f^k, g^j].$$

To complete the calculation, decompose $\tau_{jk\ell}$ into real and imaginary parts by writing the relations

$$\tau_{jk\ell} = \rho_{jk\ell} + i\sigma_{jk\ell}. \tag{5.8.71}$$

Then taking real and imaginary parts of (8.69) and using (8.70) give the results

$$[f^j, f^k] = [g^j, g^k] = +(1/2) \sum_\ell \rho_{jk\ell} b^\ell, \tag{5.8.72}$$

$$[f^j, g^k] = -(1/2)\sum_{\ell} \sigma_{jk\ell} b^{\ell}.$$

Note that by (8.70) and (8.72), $\rho_{jk\ell}$ is antisymmetric in its first two indices, and $\sigma_{jk\ell}$ is symmetric,

$$\rho_{jk\ell} = -\rho_{kj\ell}, \tag{5.8.73}$$

$$\sigma_{jk\ell} = \sigma_{kj\ell}.$$

The table below lists the needed values of $\rho_{jk\ell}$ and $\sigma_{jk\ell}$. All the rest are zero, or can be obtained from the symmetry conditions (8.73). Taken together, (8.6), (8.7), (8.67), and (8.72) specify the Lie algebra $sp(6)$ in all its beauty.

Table 5.8.5: Remaining Structure Constants of $sp(6)$.

$jk\ell$	$\rho_{jk\ell}$	$\sigma_{jk\ell}$
110	0	$-4/3$
113	0	-2
118	0	$-2/\sqrt{3}$
121	0	-2
122	2	0
164	0	-2
165	2	0
220	0	$-8/3$
228	0	$-4/\sqrt{3}$
231	0	-2
232	2	0
244	0	-2
245	2	0
266	0	-2
267	2	0
330	0	$-4/3$
333	0	2
338	0	$-2/\sqrt{3}$
346	0	-2
347	2	0
440	0	$-8/3$
443	0	2
448	0	$2/\sqrt{3}$
456	0	-2
457	2	0
461	0	-2
462	-2	0
550	0	$-4/3$
558	0	$4/\sqrt{3}$
564	0	-2
565	-2	0
660	0	$-8/3$
663	0	2
668	0	$2/\sqrt{3}$

Closing Remarks

We close this section with two remarks. First, we note that the $sp(6)$ polynomials c^1, $r(\pm\boldsymbol{\alpha})$, h^1, h^2, and h^3 are the same (up to normalizations) as the $sp(4)$ polynomials c, $r(\pm)$, $h^{\pm}$, and h^0. This correspondence indicates, as expected, that $sp(6)$ contains $sp(4)$ as a subgroup.

The second remark concerns $su(3)$. As mentioned earlier, it can be shown that the quadratic polynomials b^0 through b^8, which correspond to the Gell-Mann matrices λ^0 through λ^8, transform under the action of $su(3)$ according to the representations $1 = \Gamma(0,0)$ and $8 = \Gamma(1,1)$. See Exercise 8.19.

Now look at the relation (8.3) and compare it with (8.40). The left side of (8.3) may be viewed as the antisymmetric part of a second-order tensor consisting of ingredients which each transform according to the representation 8, and the right side of (8.3) is a sum of entities which also transform according to the representation 8. This result is consistent with the relation (8.40), which states that the the tensor product of an 8 and an 8 is expected to contain an 8. Moreover the coefficients $f_{jk\ell}$, which are the structure constants for $su(3)$, specify which $su(3)$ elements occur in each entry in the antisymmetric part of the second-order tensor product. Therefore these coefficients may also be viewed as particular instances of the Clebsch-Gordan coefficients for $su(3)$, namely those associated with the tensor product of the adjoint representation with itself! Indeed, the structure constants of any Lie algebra may be viewed as particular instances of the Clebsch-Gordan coefficients for that algebra, specifically those associated with the tensor product of the adjoint representation with itself.

Next look at (8.4). The left side of (8.4) may be viewed as the symmetric part of a second-order tensor consisting of ingredients which each transform according to the representation 8, and the right side of (8.4) is a sum of entities which transform according to the representations 1 and 8. This result is also consistent with the relation (8.40), which states that the the tensor product of an 8 and an 8 is also expected to contain a 1 and a second 8. Thus the coefficients δ_{jk} and $d_{jk\ell}$ are also particular instances of the Clebsch-Gordan coefficients for $su(3)$.

What about the entries 10, $\overline{10}$, and 27 which occur in (8.40) but do not occur in (8.3) and (8.4) and their composite (8.78)? They do not occur because both factors on the left sides of (8.3), (8.4), and (8.78) involve the same 8 rather than an 8 and an $8'$.

Note that in general the kind of Clebsch-Gordan analysis we have been making only predicts what representations can possibly occur in a product when each factor in the product has known transformation properties. For example, consider all entities of the form b^ib^j where i and j range from 1 to 8. Since each factor belongs to an 8, the product can possibly contain the representations appearing on the right side of (8.40). But each entity is also a homogeneous polynomial of degree 4, and therefore can potentially have the $su(3)$ content given by (8.46). Observe that $10 = \Gamma(3,0)$ and $\overline{10} = \Gamma(0,3)$ do not occur in (8.46), but $27 = \Gamma(2,2)$ does.

Exercises

5.8.1. Suppose, for the purposes of this exercise, that λ^0 is redefined by the relation

$$\lambda^0 = (2/3)^{1/2} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{5.8.74}$$

Show that the matrices $\lambda^0 \cdots \lambda^8$ span the vector space of all 3×3 Hermitian matrices. Show that λ^0, together with the Gell-Mann matrices, obey the relations

$$tr(\lambda^j\lambda^k) = 2\delta_{jk};\ j,k = 0,1\cdots 8. \tag{5.8.75}$$

5.8.2. Show that the commutator of two Gell-Mann matrices must be of the general form (8.3) with the $f_{jk\ell}$ real. Use (8.3) and (8.75) to derive the relation

$$f_{jk\ell} = -(i/4)tr(\{\lambda^j,\lambda^k\}\lambda^\ell);\ j,k,\ell = 1,2,\cdots 8. \tag{5.8.76}$$

Prove from this relation that $f_{jk\ell}$ is antisymmetric under the interchange of any two (adjacent) indices. Verify Table 8.1.

5.8.3. Show that the anticommutator of two Gell-Mann matrices must be of the general form (8.4) with the $d_{jk\ell}$ real. Use (8.4) and (8.75) to derive the relation

$$d_{jk\ell} = (1/4)\text{tr}(\{\lambda^j, \lambda^k\}_+ \lambda^\ell);\ j, k, \ell = 1, 2, \cdots 8. \tag{5.8.77}$$

Prove from this relation that $d_{jk\ell}$ is symmetric under the interchange of any two indices. Verify Table 8.2.

5.8.4. Show that the Gell-Mann matrices obey the multiplication rules

$$\lambda^j \lambda^k = (2/3)\delta_{jk}\lambda^0 + \sum_\ell (d_{jk\ell} + i f_{jk\ell})\lambda^\ell;\ j, k, \ell = 1, 2, \cdots 8. \tag{5.8.78}$$

5.8.5. Verify the results (8.5).

5.8.6. Verify the Poisson bracket rules (8.6) and (8.7).

5.8.7. Show that the Cartan-basis matrices given by (8.8) and (8.11) are *real* and satisfy the relations

$$(C^j)^\dagger = C^j, \tag{5.8.79}$$

$$R(\boldsymbol{\mu})^\dagger = R(-\boldsymbol{\mu}). \tag{5.8.80}$$

5.8.8. Verify the Cartan-basis commutation rules (8.9) and (8.12) through (8.14).

5.8.9. Verify that the root vectors given by (8.10) and their negatives satisfy the relation (8.22).

5.8.10. Verify that the Clebsch-Gordan relations (8.34) through (8.40) are specific cases of the general Clebsch-Gordan relation given by (8.41) and (8.42).

5.8.11. Look at the Clebsch-Gordan relations (8.34) through (8.40). As a sanity check, the dimensions of the left and right sides should agree. For example, the dimension (the number of entities) on the left side of (8.34) is $3 \times 3 = 9$. And the dimension of the right side of (8.34) is $3 + 6 = 9$. Verify that analogous results hold for (8.35) through (8.40). If you are algebraically ambitious, verify that analogous results hold in the general case described by (8.41) and (8.42) using (8.21).

5.8.12. Look at the relations (8.43) through (8.49) for the case of a six-dimensional phase space. As a sanity check, verify that the dimensions of the left and right sides agree using (7.3.40) and (8.21).

5.8.13. Review the relations (8.50) and (8.51). Verify the relation

$$r(-\boldsymbol{\mu}) = -\overline{r}(\boldsymbol{\mu}). \tag{5.8.81}$$

Next consider the Lie operators associated with the c^j and $r(\boldsymbol{\mu})$. It can be shown that a suitable scalar product can be defined so that, in analogy to (8.79) and (8.80), they satisfy the relations (7.3.22) and (7.3.23). See Section 7.3. Review Exercise 8.8. Show that the Lie operators associated with the c^j and $r(\boldsymbol{\mu})$ obey the same commutation rules as the C^j and $R(\boldsymbol{\mu})$.

5.8.14. The purpose of this exercise is to construct the polynomials (8.53) corresponding to matrices of the form JS^a. Consider some second-degree polynomial corresponding to some matrix of the form JS^a. For example, we may set $B = 0$ and $A = I_3$ in (7.10). Show that use of (7.10) and (5.1) then produces a polynomial, call it a^1, given by the relation

$$a^1 = (1/2)(q_1^2 - p_1^2 + q_2^2 - p_2^2 + q_3^2 - p_3^2). \tag{5.8.82}$$

If our suspicions about polynomials associated with the JS^a belonging to the representation $\Gamma(2,0) \oplus \Gamma(0,2)$ are correct, the polynomial a^1 should be some linear combination of polynomials corresponding to the weights of Figures 8.6 and 8.7. With luck, it may be possible to produce a polynomial corresponding to the highest weights $\boldsymbol{w}^h(6)$ and $\boldsymbol{w}^h(\overline{6})$ by repeatedly applying $: r(\boldsymbol{\alpha}) :$ to a^1. See Figure 8.1. Verify the results

$$a^2 =: r(\boldsymbol{\alpha}) : a^1 = [r(\boldsymbol{\alpha}), a^1] = (-i)(q_1p_2 + q_2p_1), \tag{5.8.83}$$

$$a^3 =: r(\boldsymbol{\alpha}) : a^2 = (1/2)[(q_1 + ip_1)^2 + (q_2 - ip_2)^2], \tag{5.8.84}$$

$$a^4 =: r(\boldsymbol{\alpha}) : a^3 = 0. \tag{5.8.85}$$

The relation (8.85) shows that a^3 cannot be raised any further in the $\boldsymbol{\alpha}$ direction, and suggests that a^3 is, as desired, a polynomial corresponding to the highest weights $\boldsymbol{w}^h(6)$ and $\boldsymbol{w}^h(\overline{6})$. Indeed, show that

$$: c^1 : a^3 = [c^1, a^3] = (\sqrt{2})a^3 = [\boldsymbol{e}^1 \cdot \boldsymbol{w}^h(6)]a^3 = [\boldsymbol{e}^1 \cdot \boldsymbol{w}^h(\overline{6})]a^3. \tag{5.8.86}$$

Finally, the components of a^3 corresponding to $\boldsymbol{w}^h(6)$ and $\boldsymbol{w}^h(\overline{6})$ separately can be removed from a^3 by using the operators $[: c^2 : -\boldsymbol{e}^2 \cdot \boldsymbol{w}^h(6)]$ and $[: c^2 : -\boldsymbol{e}^2 \cdot \boldsymbol{w}^h(\overline{6})]$, respectively. Do so by defining further polynomials a^5 and a^6 by the rules

$$a^5 = [: c^2 : -\boldsymbol{e}^2 \cdot \boldsymbol{w}^h(6)]a^3 = -(2/\sqrt{6})(q_2 - ip_2)^2, \tag{5.8.87}$$

$$a^6 = [: c^2 : -\boldsymbol{e}^2 \cdot \boldsymbol{w}^h(\overline{6})]a^3 = (2/\sqrt{6})(q_1 + ip_1)^2. \tag{5.8.88}$$

Verify (8.87) and (8.88), and show that these polynomials are simultaneous eigenvectors of both the $: c^j :$,

$$\begin{aligned} : c^j : a^5 &= [\boldsymbol{e}^j \cdot \boldsymbol{w}^h(\overline{6})]a^5, \\ : c^j : a^6 &= [\boldsymbol{e}^j \cdot \boldsymbol{w}^h(6)]a^6. \end{aligned} \tag{5.8.89}$$

Now verify that

$$h^1 \propto a^6, \tag{5.8.90}$$

and verify the results (8.55). The particular normalizations used in defining the h^k as given in (8.53) have been chosen for convenience.

5.8.15. Verify the relations (8.54) and (8.56).

5.8.16. Verify the relations (8.57) through (8.61).

5.8.17. Verify Table 8.4.

5.8.18. Verify Table 8.5.

5.8.19. Thanks to the work of Subsection 8.10, we know that the quadratic polynomials h^k transform under $su(3)$ according to the representation 6, and the $\overline{h}^k$ transform according to $\overline{6}$. The purpose of this exercise is to study how the remaining polynomials b^j transform. Recall that the c^j and $r(\boldsymbol{\mu})$ defined by (8.50) and (8.51) satisfy the same Lie algebra as the C^j and $R(\boldsymbol{\mu})$. See Exercise 8.13.

Begin by considering the polynomial b^0. The relations (8.6) can be rewritten in the form

$$:b^j:b^0 = [b^j, b^0] = 0. \tag{5.8.91}$$

The relations (8.91) may be understood to say that b^0 transforms according to the representation $\Gamma(0,0)$. Show from (8.21) that $\dim\Gamma(0,0) = 1$, and from (8.19) that $\boldsymbol{w}^h(1) = 0$.

What about the remaining 8 polynomials $b^1, \cdots b^8$? Show that $\dim\Gamma(1,1) = 8$, and that $\boldsymbol{w}^h(8) = \boldsymbol{\alpha}$. Figure 8.8 displays the weight diagram for the representation 8=Γ(1,1). There are 6 points on the vertices of a hexagon at the ends of the vectors $\pm\ \boldsymbol{\alpha}$, $\pm\ \boldsymbol{\beta}$, $\pm\ \boldsymbol{\gamma}$. In addition, there are 2 eigenvectors corresponding to the weight at the origin (indicated by a dot and a concentric circle) to make a total of 6+2=8 states. Verify the relations

$$:c^j:r(\boldsymbol{\nu}) = (\boldsymbol{e}^j\cdot\boldsymbol{\nu})r(\boldsymbol{\nu}), \tag{5.8.92}$$

$$:r(\boldsymbol{\mu}):r(\boldsymbol{\nu}) = N(\boldsymbol{\mu},\boldsymbol{\nu})r(\boldsymbol{\mu}+\boldsymbol{\nu}),\quad \text{(when } \boldsymbol{\mu}+\boldsymbol{\nu} \text{ is a root vector)}. \tag{5.8.93}$$

The relations (8.92) indicate that each $r(\boldsymbol{\nu})$ has a weight $\boldsymbol{\nu}$ corresponding to a particular vertex of the hexagon, and the relations (8.93) indicate that the $:r(\boldsymbol{\mu}):$ act on the $r(\boldsymbol{\nu})$ to produce polynomials with raised and lowered weights. Also, verify the relations

$$:c^j:c^k = 0, \tag{5.8.94}$$

$$:r(\boldsymbol{\mu}):c^k = -(\boldsymbol{e}^k\cdot\boldsymbol{\mu})r(\boldsymbol{\mu}). \tag{5.8.95}$$

The relations (8.94) indicate that c^1 and c^2 correspond to the two eigenvectors for the weight at the origin of the weight diagram, and the relations (8.95) indicate that the $:r(\boldsymbol{\mu}):$ raise and lower these eigen vectors. Finally, show that the polynomials $b^1, \cdots b^8$ are related to the c^j and $r(\boldsymbol{\mu})$ by a nonsingular matrix.

In summary, we conclude that the polynomial b^0 transforms under $su(3)$ according to the representation 1, and the 8 polynomials $b^1, \cdots b^8$ transform according to the representation 8. The representation $8 = \Gamma(1,1)$ is called the *adjoint* or *regular* representation because it arises from the action of the Lie algebra on itself. See the discussion at the end of Section 3.7.

5.8.20. Consider the first-degree polynomials t^1, t^2, t^3 defined by the relations

$$t^j = q_j + ip_j. \tag{5.8.96}$$

Consider also the representation $\Gamma(1,0)$. Show that $\dim\Gamma(1,0) = 3$ and compute $\boldsymbol{w}^h(3)$. Figure 8.4 shows the weight vectors for $\Gamma(1,0)$. Show that the t^j transform under $su(3)$ according to the representation 3. Also compute $:b^0:t^j$. Figure 8.5 shows the weight vectors for $\Gamma(0,1)$. Show that the $\overline{t}^j$ transform under $su(3)$ according to the representation $\overline{3}$. Also compute $:b^0:\overline{t}^j$. It follows that the 6 monomials $q_1, q_2, q_3, p_1, p_2, p_3$ transform according to the representation $3 \oplus \overline{3}$. This is in accord with (8.48).

5.8.21. Consider the vector space spanned by the quadratic polynomials of the form $t^j \times t^k$. See (8.96). Show that this vector space is 6 dimensional, and is spanned by the polynomials h^ℓ of (8.53). Exercise 8.20 showed that the t^j transform under $su(3)$ according to the representation 3, and the $\bar{t}^j$ transform under $su(3)$ according to the representation $\overline{3}$. It follows from group theory and the derivation property (3.7) that the products $t^j \times t^k$ must transform as some portion of the direct product representation $3 \otimes 3$. In the case of $su(3)$, the general direct product representation $3 \otimes 3'$ has the Clebsch-Gordan series decomposition (8.34). For the present application, both "3" factors on the left of (8.34) are the same, and correspondingly the $\overline{3}$ portion of the direct product representation is absent. (The $\overline{3}$ portion is antisymmetric under the interchange of the two 3 factors, and the 6 portion is symmetric.) It follows that the h^ℓ should transform according to the representation 6, which is indeed the case. Similarly, the general direct product representation $\overline{3} \otimes \overline{3}'$ has the Clebsch-Gordan series decomposition (8.35). It follows that the $\overline{h}^\ell$ should transform according to the representation $\overline{6}$, which is also the case.

5.8.22. Consider the vector space spanned by the quadratic polynomials of the form $t^j \times \overline{t}^k$. See (8.96). Show that this vector space is 9 dimensional, and is spanned by the polynomials b^ℓ of (8.5). Exercise 8.20 showed that the t^j and $\overline{t}^k$ transform under $su(3)$ according to the representations 3 and $\overline{3}$, respectively. It follows from group theory and the derivation property (3.7) that the products $t^j \times \overline{t}^k$ must transform according to the direct product representation $3 \otimes \overline{3}$. In the case of $su(3)$, the general direct product representation $3 \otimes \overline{3}$ has the Clebsch-Gordan series decomposition (8.36). It follows that the b^ℓ should transform according to the representations 1 and 8. This surmise is indeed the case since Exercise 8.19 showed that b^0 transforms according to the representation 1, and the remaining b's transform according to the representation 8.

5.8.23. Verify the relations (8.68). The polynomials h^j transform according to the representation 6. Also, the Poisson bracket operation may be viewed as a kind of multiplication. It follows from group theory and the derivation property (3.9) that the Lie products $[h^j, h^k]$ must transform as some portion of the direct product representation $6 \otimes 6'$. In the case of $su(3)$, the general direct product representation $6 \otimes 6'$ has the Clebsch-Gordan series decomposition (8.37). On the other hand, from the structure of $sp(6)$, the Poisson brackets $[h^j, h^k]$ can only yield terms of the form b^ℓ, which transform according to $1 = \Gamma(0,0)$ and $8 = \Gamma(1,1)$. See (3.9.3) and Exercise 8.19. We seem to have arrived at an apparent contradiction because $\Gamma(0,0)$ and $\Gamma(1,1)$ do not appear on the right side of (8.37). The only resolution to this apparent dilemma is for the Poisson brackets (8.68) to vanish, which they indeed do.

By contrast the general direct product representation $6 \otimes \overline{6}$ has the Clebsch-Gordan series decomposition (8.38). It follows that the Lie products $[h^j, \bar{h}^k]$ must transform as some portion of the representations $1 \oplus 8 \oplus 27$. This surmise is indeed the case since the Poisson brackets $[h^j, \bar{h}^k]$ yield terms of the form b^ℓ [see (8.53)], and these terms transform according to 1 and 8. Show that similar considerations apply to the Poisson bracket relation (8.64). The relevant Clebsch-Gordan series decomposition in this cases is (8.39).

5.8.24. Verify that the polynomials b^1, b^2, and b^3 form a Lie subalgebra under the Poisson bracket operation. This subalgebra is the Lie algebra for an $su(2)$ subalgebra of $su(3)$.

5.8.25. Verify that the polynomials b^2, b^5, and b^7 form a Lie subalgebra under the Poisson bracket operation. This subalgebra is the Lie algebra for an $so(3)$ subalgebra of $su(3)$. See Exercises 7.2.5 and 7.2.6.

5.8.26. For the case of a 6-dimensional phase space, consider the quadratic polynomials defined by the relations

$$T_{jk} = q_j q_k + p_j p_k, \tag{5.8.97}$$

$$L_j = \sum_{k\ell} \epsilon_{jk\ell} q_k p_\ell. \tag{5.8.98}$$

Show that there are 9 such elements, and that they can be written as linear combinations of the quantities b^0 through b^8, and vice versa. The quantities T and L therefore provide an alternate basis for $u(3)$. Note that T is symmetric. Relate b^0 to the trace of T. Show that the quantities L_j form a basis for $so(3)$. Show that the quantities T transform as a tensor under $so(3)$. That is, evaluate the Poisson brackets $[L_j, L_k]$ and $[L_j, T_{k\ell}]$. Finally, evaluate the Poisson brackets $[T_{k\ell}, T_{mn}]$.

5.8.27. The relation (5.3) associates a matrix JS with every quadratic polynomial. Find the matrices B^i (for $i = 0, 1, \cdots 8$) associated with the polynomials b^i. Find the matrices F^j and G^j (for $j = 1, 2, \cdots 6$) associated with the polynomials f^j and g^j. Use (5.21) if you wish. The B^i, F^j, and G^j provide a basis for the 6×6 matrix representation of $sp(6)$. Find their commutation rules.
Partial Answer:

$$B^0 = J = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \end{pmatrix}, \quad B^1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{5.8.99}$$

$$B^2 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad B^3 = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$B^4 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad B^5 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \end{pmatrix},$$

$$B^6 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad B^7 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix},$$

$$B^8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -2 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \end{pmatrix}, \quad F^1 = \begin{pmatrix} 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$F^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad F^3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$F^4 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad F^5 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \end{pmatrix},$$

$$F^6 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad G^1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$G^2 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad G^3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$G^4 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix}, \quad G^5 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix},$$

$$G^6 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \end{pmatrix}.$$

Note that these generators/matrices are given for the case that J has the form (3.1.1). In the case that J' as given by (3.2.10) is employed, one may find the related generators by means of the permutation matrix P given by (3.2.19).

5.8.28. The purpose of this exercise is to explore the *dynamic* role of the basis polynomials b^0 through b^8, f^1 through f^6, and g^1 through g^6.

a) Consider first the $u(3)$ basis polynomials b^0 through b^8. In place of b^0, b^3, and b^8 it is convenient to use the polynomials $(p_1^2 + q_1^2)/2$, $(p_2^2 + q_2^2)/2$, $(p_3^2 + q_3^2)/2$. Show that $(p_1^2 + q_1^2)/2$ generates rotations in the q_1, p_1 plane, etc. See Exercise 5.4.5. Next consider b^2, b^5, and b^7. Show that b^2 generates rotations in the p_1, p_2 and q_1, q_2 planes simultaneously, etc. See Section 7.2. Finally consider b^1, b^4, and b^6. Show that b^1 generates rotations in the q_1, p_2 and q_2, p_1 planes simultaneously, etc.

b) Next consider the remaining polynomials f^1 through f^6 and g^1 through g^6. The polynomials f^1, f^3, and f^5 are analogous. Show that f^1 generates motions on hyperbolas in the q_1, p_1 plane, etc. The polynomials g^1, g^3, and g^5 are also analogous. Show that g^1 also generates motions on hyperbolas in the q_1, p_1 plane, etc. See Exercise 5.4.4. The polynomials f^2, f^4, and f^6 are analogous. Show that f^2 generates motions on hyperbolas in the q_1, p_2 and q_2, p_1 planes simultaneously, etc. Finally, the polynomials g^2, g^4, and g^6 are analogous. Show that g^2 generates motions on hyperbolas in the p_1, p_2 and q_1, q_2 planes simultaneously, etc.

5.8.29. The purpose of this exercise is to explore conjugacy relations for the case of $su(3)$. Review Exercise 3.7.36. Suppose, for some representation, there is a basis for which the elements in the Lie algebra $su(3)$ are anti-Hermitian matrices (and the structure constants are real). Recall that for such matrices the hatting and checking operations given by (3.7.219) and (3.7.222) have the same effect.

Specifically, consider the matrices $i\lambda^j$ that obey the $su(3)$ commutation rules (8.3). Form the associated hatted representation given by (3.7.219). Next form the associated checked representation given by (3.7.222). Show that, as expected, in this case both operations produce the same result. Verify that in this case the conjugate representation is *not* equivalent to the original representation. Hint: Show that C^1 and C^2 as given by (8.8) have the eigenvalues displayed in the weight diagram of Figure 8.4. Hence the matrices iC^1 and iC^2 will have these eigenvalues multiplied by i. Consequently, the matrix pairs iC^1, iC^2, and $-iC^1$, $-iC^2$ cannot have the same eigenvalues, and therefore cannot be related by a similarity transformation. Indeed, the eigenvalues of the pair $-iC^1$, $-iC^2$ are those shown in Figure 8.5 multiplied by i.

Repeat the above analysis for the basis matrices given by (8.8) and (8.11). Hint: Review Exercise 8.7.

5.8.30. Use the quantities $f_{jk\ell}$ to form the adjoint representation of $su(3)$. See (8.3). Show that this representation has dimension 8 and is therefore $\Gamma(1,1)$. See Table 8.3 and Figure 8.8. Review Exercise 3.7.36. Show that the adjoint representation of $su(3)$ is unaffected by either the hatting operation (3.7.218) or the checking operation (3.7.219). We say that the adjoint representation is *self conjugate* in accord with the $su(3)$ representation conjugacy relation (8.20).

5.8.31. Construct the weight diagrams for the representations 10, $\overline{10}$, and 27. Indicate the multiplicity of each weight.

5.8.32. A *Chevalley* basis for a Lie algebra is one for which the structure constants are all integers. Sometimes one also requires that the entries in matrices used to represent the algebra have all integer entries. Show that the basis found for $sp(2)$ and $sp(4)$ in Sections 5.6 and 5.7 are Chevalley bases. Find Chevalley bases for $su(3)$ and $sp(6)$.

5.9 Some Topological Questions

In this section we will learn something about the topology of $Sp(2n,\mathbb{R})$ and how the stable elements of $Sp(2n,\mathbb{R})$ reside within it.

5.9.1 Nature and Connectivity of $Sp(2n,\mathbb{R})$

$Sp(2n,\mathbb{R})$ Is Connected

We begin by showing that the symplectic group $Sp(2n,\mathbb{R})$ is connected, and indeed infinitely connected. Let us start with the connected claim, which is easy to demonstrate. Suppose M and N are any two matrices in $Sp(2n,\mathbb{R})$. Define the symplectic matrix R by the rule

$$R = MN^{-1} \tag{5.9.1}$$

so that there is the relation

$$M = RN. \tag{5.9.2}$$

Since R is symplectic, from (3.8.24) we know there are symmetric matrices S^a and S^c such that R can be written in the form

$$R = \exp(JS^a)\exp(JS^c), \tag{5.9.3}$$

and the matrices $R(\lambda)$ defined by

$$R(\lambda) = \exp(\lambda J S^a)\exp(\lambda J S^c) \tag{5.9.4}$$

will form a one-parameter family of symplectic matrices with

$$R(0) = I \tag{5.9.5}$$

and

$$R(1) = R. \tag{5.9.6}$$

Now consider the one-parameter family of symplectic matrices $M(\lambda)$ defined by the rule

$$M(\lambda) = R(\lambda)N. \tag{5.9.7}$$

From this and the previous definitions we have the results

$$M(0) = N \tag{5.9.8}$$

and

$$M(1) = M. \tag{5.9.9}$$

Thus, the matrices $M(\lambda)$ provide a path, in the space of symplectic matrices, that connects N to M.

$Sp(2n, \mathbb{R})$ Is Infinitely Connected

We next turn to the harder task of examining the topology of $Sp(2n, \mathbb{R})$ in more detail and showing that the space of symplectic matrices is infinitely connected. We will begin with the case of $Sp(2, \mathbb{R})$.

Suppose the basis elements B^0, F, and G given by (6.7), (6.13), and (6.14) are used to evaluate (3.8.24). Doing so shows that the most general real 2×2 symplectic matrix can be written in the form

$$M = \exp(\phi F + \gamma G)\exp(\beta_0 B^0), \tag{5.9.10}$$

where β_0, ϕ, and γ are arbitrary real coefficients. [Note that there are indeed three coefficients as predicted by (3.7.35) evaluated for n=1.] Thus, (9.10) gives a complete parameterization of the 2×2 symplectic group. The quantities $\exp(\phi F + \gamma G)$ and $\exp(\beta_0 B^0)$ can be evaluated using (3.7.1) to give the results

$$\begin{aligned} \exp(\phi F + \gamma G) &= I\cosh[(\phi^2+\gamma^2)^{1/2}] \\ &+ [(\phi F + \gamma G)/(\phi^2+\gamma^2)^{1/2}]\sinh[(\phi^2+\gamma^2)^{1/2}], \end{aligned} \tag{5.9.11}$$

$$\exp(\beta_0 B^0) = I\cos\beta_0 + B^0\sin\beta_0 = \begin{pmatrix} \cos\beta_0 & \sin\beta_0 \\ -\sin\beta_0 & \cos\beta_0 \end{pmatrix}. \tag{5.9.12}$$

Observe that, according to (9.11), the factor $\exp(\phi F + \gamma G)$ has the topology of two-dimensional Euclidean space E^2 since ϕ and γ can each range over $\pm\infty$ without any duplication of results. By contrast, the factor $\exp(\beta_0 B^0)$, according to (9.12), has the topology of a circle T^1 since it is periodic in β_0 with period 2π. Indeed, the matrix on the far right of (9.12) represents $SO(2,\mathbb{R})$ which has the topology of T^1, and also of $U(1)$. (Here, as in Section 3.9, we use the notation T^n to denote an n-torus, the topological product of n circles. Thus, T^1 denotes a 1-torus, which is just a circle.) It follows that $Sp(2,\mathbb{R})$ has the product topology $E^2 \times T^1$.[9] Since T^1 is infinitely connected, $Sp(2,\mathbb{R})$ is infinitely connected. Finally, in view of (3.9.88, we note that there is the relation

$$W \exp(\beta_0 B^0)\, W^{-1} = \begin{pmatrix} \exp(i\beta_0) & 0 \\ 0 & \exp(-i\beta_0) \end{pmatrix}, \tag{5.9.13}$$

which shows explicitly that $\exp(\beta_0 B^0)$ is isomorphic to the representation $U(1) \oplus \overline{U(1)}$ of $U(1)$, as expected.

In an analogous way, with $m = n(n+1)$, it can be seen that $Sp(2n,\mathbb{R})$ has the product topology $E^m \times [U(n) \oplus \overline{U(n)}]$. First, again by (3.8.24), any M in $Sp(2n,\mathbb{R})$ can be written in the product form

$$M = \exp(JS^a)\exp(JS^c). \tag{5.9.14}$$

Now, according to Exercise 3.9.10, there are $m = n(n+1)$ linearly independent matrices of the form JS^a. Note that from its form, m is an *even* integer, and hence $k = m/2$ is an integer. In analogy to the cases of $sp(2,\mathbb{R}), sp(4,\mathbb{R})$, and $sp(6,\mathbb{R})$, let $F^1, \cdots F^k$ and $G^1, \cdots G^k$ be a basis for the set of matrices of the form JS^a. Then the $\exp(JS^a)$ factor can be written in the form

$$\exp(JS^a) = \exp[\sum_{j=1}^{k}(\phi_j F^j + \gamma_j G^j)]. \tag{5.9.15}$$

Since the real symmetric logarithm of a real symmetric positive definite matrix is unique, the m parameters ϕ_j and γ_j can all range from $\pm\infty$ without any duplication of results. Consequently, the factor $\exp(JS^a)$ has the topology of E^m. Finally, according to Section 3.9, matrices of the form $\exp(JS^c)$ are isomorphic to $U(n)$. Thus, as stated at the beginning of this paragraph, $Sp(2n,\mathbb{R})$ has the product topology $E^m \times [U(n) \oplus \overline{U(n)}]$.

We pause at this point to observe that some of the matrix entries on the right side of (9.11) grow in magnitude without bound as ϕ and γ range over $\pm\infty$. Thus, the group $Sp(2,\mathbb{R})$ is not compact. Moreover, since $Sp(2,\mathbb{R})$ is a subgroup of $Sp(2n,\mathbb{R})$ and $Sp(2n,\mathbb{C})$ for any n, it follows that these groups are also not compact.[10]

In this vein, what can be said about what we have called the $U(n)$ subgroup, the $[U(n) \oplus \overline{U(n)}]$ factor of $Sp(2n,\mathbb{R})$? From the discussion of Section 3.9 we know that all matrices of the form $\exp(JS^c)$ are in the orthogonal group $SO(2n,\mathbb{R})$. From the work of the first part of Section 3.6.3 we know that the rows (and columns) of an orthogonal matrix are orthonormal. In particular, the rows (and columns) are unit vectors. It follows that all entries in an orthogonal matrix are bounded in magnitude by 1. Consequently, the $U(n)$

[9] Because ϕ and γ are unrestricted, this set is sometimes referred to as a *solid* torus.

[10] *Compactness* is a topological property of sets that may be defined in a variety of ways. For our purposes, since we are generally dealing with matrices which may be viewed as being imbedded in some high dimensional Euclidean space, we will say that a set of matrices is compact if all matrix elements of these matrices are confined to lie within some closed and bounded set within this Euclidean space. (A set is *closed* if it contains all its limit points.) Conversely, if any matrix elements for some sequence of matrices in the set are unbounded (grow in magnitude without bound), we will say that the set of matrices is *noncompact.*

subgroup of $Sp(2n,\mathbb{R})$ is compact. Indeed, it can be shown to be the largest compact subgroup of $Sp(2n,\mathbb{R})$.

We have already seen that $Sp(2n,\mathbb{R})$ has the product topology $E^m \times [U(n) \oplus \overline{U(n)}]$. What remains is to study the topology of the $U(n)$ subgroup of $Sp(2n,\mathbb{R})$. In the $n = 1$ case we have already found that $Sp(2,\mathbb{R})$ has the product topology $E^2 \times U(1)$ and that $U(1)$ has the topology of T^1. We might hope to proceed in a similar fashion for the case $n > 1$. Suppose, for specificity, we consider the case $Sp(4,\mathbb{R})$ for which $n = 2$ and we are therefore interested in the topology of $U(2)$. In the 4×4 case a basis for the Lie algebra of the matrices of the form JS^c can be taken to be the matrices B^0 through B^3 given displayed in (7.45). We also note that B^0 commutes with the B^1 through B^3. See (7.5). Therefore in the 4×4 case the most general $\exp(JS^c)$ can can be written in the form

$$\exp(JS^c) = [\exp(\sum_1^3 \beta_j B^j)] \exp(\beta_0 B^0). \tag{5.9.16}$$

Matrices of the form $\exp(\sum_1^3 \beta_j B^j)$ carry the $SU(2) \oplus \overline{SU(2)}$ representation of $SU(2)$, and all the groups $SU(n)$ for $n > 1$ are known to be simply connected. For example, $SU(2)$ has the topology of the 3-sphere S^3. See Exercise 10.13. And S^3 is simply connected. What remains is to examine the factor $\exp(\beta_0 B^0)$.

A remark is in order before doing so. It is common in the physics literature to see the assertion

$$U(n) = SU(n) \otimes U(1) \tag{5.9.17}$$

where the symbol $\otimes$ denotes a direct product. [A particular case of (9.17) is the assertion that $U(2) = SU(2) \otimes U(1)$.] If this were true, since $SU(n)$ is simply connected, $U(n)$ would have the *connectivity* of $U(1)$, which is T^1. And, in particular, $U(2)$ would have the connectivity T^1. Correspondingly, $Sp(2n,\mathbb{R})$ would have the product topology $E^m \times SU(n) \times T^1$, and consequently all the $Sp(2n,\mathbb{R})$ would be *infinitely* connected. It turns out that these topological statements are correct, but the argument is wrong. The assertion (9.17) is not *globally* true. What is true is the weaker result that (9.17) holds only in some vicinity of the identity.

Let us continue. In view of the result given for B^0 in (7.45) and the relation (3.8.30) it follows that

$$\exp(\beta_0 B^0) = I \cos\beta_0 + J \sin\beta_0. \tag{5.9.18}$$

And, again in view of (3.9.88), we see that in the 4×4 case there is the result

$$W \exp(\beta_0 B^0)\, W^{-1} = \begin{pmatrix} \exp(i\beta_0) & 0 & 0 & 0 \\ 0 & \exp(i\beta_0) & 0 & 0 \\ 0 & 0 & \exp(-i\beta_0) & 0 \\ 0 & 0 & 0 & \exp(-i\beta_0) \end{pmatrix}. \tag{5.9.19}$$

We conclude, because they contain only the diagonal entries $\exp(i\beta_0)$ and $\exp(-i\beta_0)$, that for small β_0 the matrices on the right side of (9.19) behave like $U(1) \oplus \overline{U(1)}$. But observe that taking the determinants of the 2×2 matrices in the upper left and lower right blocks of (9.19) yields the results $\exp(2i\beta_0)$ and $\exp(-2i\beta_0)$, respectively. These results, when evaluated for $\beta_0 = \pi$, both have the value $+1$. Thus, when for $\beta_0 = \pi$, the 2×2 matrices are in $SU(2)$ and can be absorbed into the first factor, the $SU(2) \oplus SU(2)$ factor, on the right side of (9.16). For this value of β_0 the direct product hypothesis $U(2) = SU(2) \otimes U(1)$ has abruptly changed. Consequently the 2×2 matrices in (9.19) are *not* a global representation of $U(1) \oplus U(1)$. By an analogous analysis it is evident that the hypothesis (9.17) is not true globally for any $n \geq 2$. See, for example, Exercise 9.4.

Where does our exploration now stand? The approach we have been following has not been adequate for determining the global topology of $U(2)$, and evidently it will also fail for all $U(n)$ with $n > 1$. However, by more powerful methods beyond the scope of this book, it can be shown that all the $U(n)$ have the connectivity of T^1. Consequently $Sp(2n, \mathbb{R})$ has the product topology $E^m \times SU(n) \times T^1$. It follows, because of the presence of T^1, that the groups $Sp(2n, \mathbb{R})$ are infinitely connected for all n.

Since $Sp(2n, \mathbb{R})$ is infinitely connected, it must have a multiplicity of covering groups.[11] In particular, it has a two-fold covering group. This group is called the *metaplectic* group, and is of interest for paraxial wave optics (*Fourier* optics) and quantum mechanics.

Finally we remark that, contrary to the case of $Sp(2n, \mathbb{R})$, $Sp(2n, \mathbb{C})$ is *simply* connected.

5.9.2 Where Are the Stable Elements?

With the topology of $Sp(2n, \mathbb{R})$ in view, it would be useful to know where the stable elements (those with distinct eigenvalues on the unit circle) reside. In general this is a difficult question because $Sp(2n, \mathbb{R})$ is $n(2n + 1)$ dimensional. However, $Sp(2, \mathbb{R})$ is only 3 dimensional, and we will see that this case is tractable.

In the case of $Sp(2, \mathbb{R})$, combining (9.10) through (9.12) gives the result

$$\begin{aligned} M &= \{I \cosh[(\phi^2 + \gamma^2)^{1/2}] + [(\phi F + \gamma G)/(\phi^2 + \gamma^2)^{1/2}] \sinh[(\phi^2 + \gamma^2)^{1/2}]\} \times \\ & \quad \{I \cos\beta_0 + B^0 \sin\beta_0\}. \end{aligned} \tag{5.9.20}$$

From the work of Section 3.4.4 we know that in the 2×2 case the spectrum of M is governed by the quantity

$$A = \operatorname{tr}(M). \tag{5.9.21}$$

This quantity can be readily evaluated using (9.20) to yield the result

$$A = 2\cosh[(\phi^2 + \gamma^2)^{1/2}] \cos\beta_0. \tag{5.9.22}$$

See Exercise 9.5. Introduce a radius r in ϕ, γ space by writing

$$r^2 = \phi^2 + \gamma^2. \tag{5.9.23}$$

With this definition, (9.22) can be rewritten in the form

$$A = 2(\cosh r)(\cos\beta_0). \tag{5.9.24}$$

From (3.4.21) and Figure 3.4.3 we know that there is stability (eigenvalues of M are on the unit circle) when

$$-2 < 2(\cosh r)(\cos\beta_0) < +2. \tag{5.9.25}$$

[11] For a brief discussion of the concept of a covering group, see Exercise 8.2.11.

It follows that, when $(\cos\beta_0) > 0$, we can move away form the origin in ϕ, γ space while maintaining stability until $r = r_{\max}$ with

$$(\cosh r_{\max})(\cos\beta_0) = 1, \tag{5.9.26}$$

which is equivalent to the statement

$$r_{\max} = \cosh^{-1}[1/\cos(\beta_0)]. \tag{5.9.27}$$

On the other hand, when $(\cos\beta_0) < 0$, we can move away form the origin in ϕ, γ space while maintaining stability until

$$(\cosh r_{\max})(\cos\beta_0) = -1, \tag{5.9.28}$$

which is equivalent to the statement

$$r_{\max} = \cosh^{-1}[-1/\cos(\beta_0)]. \tag{5.9.29}$$

The two conditions (9.27) and (9.29) can be combined to give the net result

$$r_{\max} = \cosh^{-1}[1/|\cos(\beta_0)|]. \tag{5.9.30}$$

Figure 9.1 displays the relation (9.30) in the β_0, r plane. We observe that $r_{\max}(\beta_0)$ is periodic in β_0 with period π (and therefore also 2π) and that

$$r_{\max} = \infty \text{ when } \beta_0 = \pm\pi/2 \tag{5.9.31}$$

and

$$r_{\max} = 0 \text{ when } \beta_0 = 0, \pm\pi. \tag{5.9.32}$$

Note that $r = 0$ and $\beta_0 = \pm\pi/2$ correspond to tunes of $\pm 1/4$, and $r = 0$ and $\beta_0 = 0, \pm\pi$ correspond to tunes of $0, \pm 1/2$.

Suppose Γ is any closed path in $Sp(2, \mathbb{R})$ that goes once around the torus $SO(2, \mathbb{R})$. For example, it could begin at the identity I, that is $\beta_0 = \gamma = \phi = 0$, and end again at the identity with a 2π increase in β_0 so that at the end point $\beta_0 = 2\pi$ and again $\gamma = \phi = 0$. Then, somewhere along the path, the variable β_0 must take on the values $\beta_0 = \pi/2$ and $\beta_0 = 3\pi/2$. (Note that, by periodicity, the points $3\pi/2$ and $-\pi/2$ are equivalent.) At these points $r_{\max}$ is infinite. Thus, at least two stable group elements must lie on any closed path Γ that goes once around the torus. Moreover, since the eigenvalues for these elements are not ± 1 (they are $\pm i$ because $A = 0$ at these points), these elements must lie in open sets comprised of stable elements. Indeed, at these points the tunes are $\pm 1/4$.

It would be pleasant to have an analogous understanding of $Sp(2n, \mathbb{R})$ for general n, or at least for $n = 2$ and $n = 3$. Perhaps this is possible for $Sp(4, \mathbb{R})$ using the parameterization of Section 5.7 and the results associated with Figure 3.4.4. And perhaps, in a National Emergency, the case of $Sp(6, \mathbb{R})$ could also be understood. But we have not attempted to do so. However, what we already do know, thanks to the discussion of Sections 3.4 and 3.5, is that when the eigenvalues of an element lie on the unit circle and are distinct, then this element is surrounded by an open set of stable elements. We reiterate that this fact should be of comfort to accelerator designers and builders because it means that, at least in the

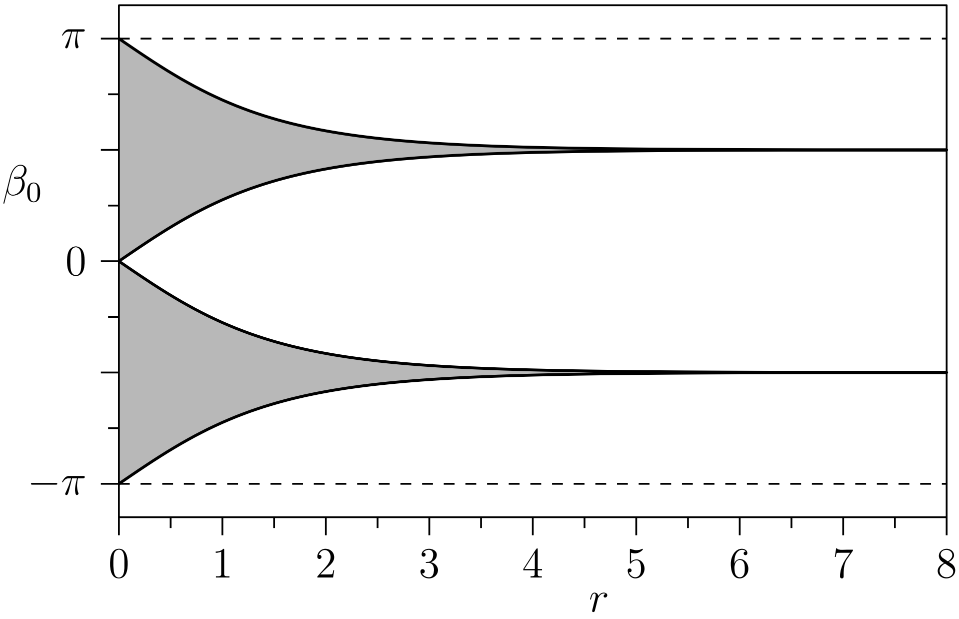


Figure 5.9.1: Stability diagram for $Sp(2,\mathbb{R})$ showing the quantity $r_{\max}$ as a function of β_0. All elements with $r < r_{\max}$ are stable, and all elements with $r > r_{\max}$ are unstable. That is, the shaded regions are stable, and the unshaded regions are unstable. In accord with toroidal topology, corresponding points on the dashed lines at the top and bottom of the figure ($\beta_0 = \pm\pi$) are to be identified.

linear approximation, the stability of orbits will not be damaged by small fabrication and control parameter errors.

We close this subsection with a remark that, perhaps, should have been made at the beginning of this subsection. We know that every symplectic matrix R has the unique factorization (9.3). Also, if S^a vanishes in this factorization, then R is diagonalizable and all its eigenvalues lie on the unit circle. Hence, all such R are stable elements. By contrast, if S^c vanishes in this factorization, then R has all its eigenvalues on the positive real axis, and some must exceed 1. Hence, all such R are unstable elements. See Exercise 3..8.12. What we have learned in this subsection is that there are cases where both S^a and S^c are non vanishing and R is stable, and other cases where both S^a and S^c are non vanishing and R is unstable.

5.9.3 Covering/Circumnavigating $U(n)$

We know that there is a $U(n)$ subgroup of $Sp(2n,\mathbb{R})$ and that any R in the $U(n)$ subgroup of $Sp(2n,\mathbb{R})$ can be written in the form

$$R(S^c) = \exp(JS^c). \tag{5.9.33}$$

Since $U(n)$ is compact, the matrices S^c cannot be arbitrarily large without some repetition occurring among the elements $R(S^c)$. Here we will find a result for how large S^c needs to be for all of the $U(n)$ subgroup to be covered.

According to the work of Section 3.9, given any R in the $U(n)$ subgroup of $Sp(2n,\mathbb{R})$, there is a $u \in U(n)$ such that

$$R = M(u). \tag{5.9.34}$$

Also, given any $u \in U(n)$ there is a $t \in U(n)$ such that

$$u = tvt^{-1} \tag{5.9.35}$$

where v is a diagonal matrix of the form (3.9.50). Since the mapping $M(u)$ is an isomorphism, we have the result

$$R = M(u) = M(tvt^{-1}) = M(t)M(v)[M(t)]^{-1} = M(t)V[M(t)]^{-1}. \tag{5.9.36}$$

Here we have used (3.9.51). But V is in the $U(n)$ subgroup and therefore there is a matrix $\hat{S}^c$ such that

$$V = \exp(J\hat{S}^c). \tag{5.9.37}$$

See (3.9.55), (3.9.63), and (3.9.64). It follows from (9.36) and (9.37) that

$$R = M(t)\exp(J\hat{S}^c)[M(t)]^{-1} = \exp\{M(t)J\hat{S}^c)[M(t)]^{-1}\}. \tag{5.9.38}$$

Upon comparing (9.33) and (9.37) we see that a suitable JS^c is given by the relation

$$JS^c = M(t)J\hat{S}^c[M(t)]^{-1}, \tag{5.9.39}$$

from which it follows that

$$S^c = J^{-1}M(t)J\hat{S}^c[M(t)]^{-1}. \tag{5.9.40}$$

From the symplectic condition $MJM^T = J$ and (3.9.30) we see that

$$J^{-1}MJ = (M^T)^{-1} = M. \tag{5.9.41}$$

It follows that

$$S^c = M\hat{S}^cM^{-1}, \tag{5.9.42}$$

and therefore

$$(S^c)^2 = M(\hat{S}^c)^2M^{-1}. \tag{5.9.43}$$

Now take the trace of both sides of (9.43) to find the result

$$\mathrm{tr}[(S^c)^2] = \mathrm{tr}[M(\hat{S}^c)^2M^{-1}] = \mathrm{tr}[(\hat{S}^c)^2]. \tag{5.9.44}$$

The right side of (9.44) can be easily evaluated using (3.9.63) and (3.9.64). We find that

$$\mathrm{tr}[(\hat{S}^c)^2] = 2\sum_{\ell=1}^{n}\phi_\ell^2. \tag{5.9.45}$$

Since each $\phi_\ell \in [-\pi,\pi]$, we see that all of the $U(n)$ subgroup is covered when

$$\mathrm{tr}[(S^c)^2] \leq 2n\pi^2. \tag{5.9.46}$$

When (9.46) holds some elements in the $U(n)$ subgroup are covered multiple times and some are covered only once. But each is covered at least once.

Exercises

5.9.1. Verify the results (9.11) and (9.12).

5.9.2. Verify that the first part of (9.12), namely

$$\exp(\beta_0 B^0) = I\cos\beta_0 + B^0\sin\beta_0, \tag{5.9.47}$$

holds in general (the $2n \times 2n$ case) for $B^0 = J$.

5.9.3. Rob, Salman, and Ivan's work on exponentiating $sp(4)$.

5.9.4. The purpose of this exercise is to make an analysis of $Sp(6,\mathbb{R})$ analogous to that provided for $Sp(4,\mathbb{R})$ in Subsection 9.1. Begin by observing that (9.14) and (9.15) hold for general n. Verify that, for $n = 3$, (9.16) takes the form

$$\exp(JS^c) = [\exp(\sum_1^8 \beta_j B^j)]\exp(\beta_0 B^0). \tag{5.9.48}$$

Matrices of the form $\exp(\sum_1^8 \beta_j B^j)$ carry the $SU(3)\oplus\overline{SU(3)}$ representation of $SU(3)$. What remains is to examine the factor $\exp(\beta_0 B^0)$ with B^0 given by the first entry in (8.99).

Verify, using (3.9.88), that for $n = 3$ there is the result

$$W\exp(\beta_0 B^0)\,W^{-1} =$$

$$\begin{pmatrix} \exp(i\beta_0) & 0 & 0 & 0 & 0 & 0 \\ 0 & \exp(i\beta_0) & 0 & 0 & 0 & 0 \\ 0 & 0 & \exp(i\beta_0) & 0 & 0 & 0 \\ 0 & 0 & 0 & \exp(-i\beta_0) & 0 & 0 \\ 0 & 0 & 0 & 0 & \exp(-i\beta_0) & 0 \\ 0 & 0 & 0 & 0 & 0 & \exp(-i\beta_0) \end{pmatrix}. \tag{5.9.49}$$

Observe that taking the determinants of the 3×3 matrices in the upper left and lower right blocks of (9.49) yields the results $\exp(3i\beta_0)$ and $\exp(-3i\beta_0)$, respectively. These results, when evaluated for $\beta_0 = 2\pi/3$, both have the value $+1$. Thus, when $\beta_0 = 2\pi/3$, the 3×3 matrices are in $SU(3)$ and can be absorbed into the first factor, the $SU(3)\oplus\overline{SU(3)}$ factor, on the right side of (9.48). For this value of β_0 the direct product hypothesis $U(3) = SU(3)\otimes U(1)$ has abruptly changed. (Verify that the same is true when $\beta_0 = 4\pi/3$.) Consequently the 3×3 matrices in (9.49) are *not* a global representation of $U(1)\oplus\overline{U(1)}$.

5.9.5. Show that carrying out the multiplication indicated in (9.20) yields a linear combination of the matrices I, F, G, B_0, FB_0, and GB_0. Show, with the exception of I, that all these matrices are traceless. Use this result to prove (9.22). Show that all matrices M of the form (9.20) satisfy

$$M^2 = -I \text{ when } \beta_0 = \pm\pi/2. \tag{5.9.50}$$

Suggestion: Use the normal form technology of Section 3.3.7.

5.9.6. Review Subsection 9.2 and Figure 9.1. Pick a value for r, say $r = 0.20$. Plot, in the complex plane, the eigenvalues of M as β_0 varies over the interval $\beta_0 \in [-\pi, \pi]$. You should find that they move about the unit circle, collide at the points ± 1, and leave and re-enter the unit circle through the collision points ± 1. Finally, when they leave the unit circle. they both lie on the positive or negative real axis. Recall Figures 3.4.1 and 3.4.3.

Consider the cases where $A = \pm 2$, but do not otherwise constrain β_0 and r, and let $M_\pm$ be the *Jordan* normal form for M in these cases. Show that, generically,

$$M_+ = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \tag{5.9.51}$$

and

$$M_- = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}. \tag{5.9.52}$$

Show that M as given by (3.5.76) with $\alpha \neq 0$ has the Jordan normal form M_+ and $-M$ [with M again given by (3.5.76) with $\alpha \neq 0$], which is also symplectic, has Jordan normal form M_-. For M given by (9.20), under what conditions can M, by a similarity transformation, be brought to the *diagonal* forms

$$M_{d\pm} = \begin{pmatrix} \pm 1 & 0 \\ 0 & \pm 1 \end{pmatrix} = \pm I? \tag{5.9.53}$$

5.10 Notational Pitfalls and Quaternions

5.10.1 The Lie Algebras $sp(2n, \mathbb{R})$ and $usp(2n)$

In the discussion of the Lie algebra $sp(6)$ we found it useful to work over the complex field even though we eventually wrote our final results in real form. The use of the complex field is both a powerful tool and a possible source of confusion. We have repeatedly made use of the particular Lie algebraic properties that the commutator of two matrices of the form JS^c is again of the same form, the commutator of a JS^c and a JS^a is a matrix of the form JS^a, and the commutator of two matrices of the form JS^a is a matrix of the form JS^c. We write these relations symbolically in the form

$$\{JS^c, JS^{c'}\} \propto JS^{c''}, \tag{5.10.1}$$

$$\{JS^c, JS^a\} \propto JS^{a'}, \tag{5.10.2}$$

$$\{JS^a, JS^{a'}\} \propto JS^c. \tag{5.10.3}$$

Here all matrices are taken to be real. That is, we are working with the Lie algebra $sp(2n, \mathbb{R})$. Now suppose that all matrices of the form JS^a are replaced by matrices of the form iJS^a, and the matrices of the form JS^c are left unchanged. Doing so converts the relations (10.1) through (10.3) into the relations

$$\{JS^c, JS^{c'}\} \propto JS^{c''}, \tag{5.10.4}$$

$$\{JS^c, (iJS^a)\} \propto (iJS^{a'}), \tag{5.10.5}$$

$$\{(iJS^a),(iJS^{a'})\} \propto JS^c. \tag{5.10.6}$$

Examination of the relations (10.4) though (10.6) shows that this replacement produces a related Lie algebra of the same dimension as before. Evidently this algebra is a subalgebra of $sp(2n,C)$. We next observe that matrices of the form JS^c and iJS^a (with S^c and S^a real) are anti-Hermitian,

$$(JS^c)^\dagger = (S^c)^\dagger J^\dagger = S^c(-J) = -JS^c, \tag{5.10.7}$$

$$\begin{aligned}(iJS^a)^\dagger &= (-i)(S^a)^\dagger(J^\dagger) = -iS^a(-J)\\ &= -iJS^a.\end{aligned} \tag{5.10.8}$$

Consequently, the Lie algebra they generate is also a subalgebra of $u(2n)$. Let us use the notation $usp(2n)$ to denote the Lie algebra generated by matrices of the form JS^c and iJS^a. Then we have the relation

$$usp(2n) = u(2n) \cap sp(2n,C). \tag{5.10.9}$$

Note that although $usp(2n)$ has a complex basis if a real basis is used for $sp(2n,\mathbb{R})$, it still has *real* structure constants in terms of this complex basis. In the language of Section 3.7, we have found that the Lie algebras $sp(2n,\mathbb{R})$ and $usp(2n)$ are equivalent over the complex field. However, they are not equivalent over the real field.

Also, let $USp(2n)$ denote the group obtained by exponentiating matrices of the form JS^c and iJS^a. These matrices belong to both $U(2n)$ and $Sp(2n,\mathbb{C})$, and we have the relation

$$USp(2n) = U(2n) \cap Sp(2n,\mathbb{C}). \tag{5.10.10}$$

This group $USp(2n)$ is called the *unitary symplectic* group.

Unfortunately for Physicists, Mathematicians often refer to this group simply as $Sp(2n)$ while we have been using the same notation as shorthand for $Sp(2n,\mathbb{R})$. This dual notation can be a source of serious confusion because $USp(2n)$ and $Sp(2n,\mathbb{R})$ have very different properties. For example, $USp(2n)$ is compact [all the entries in $USp(2n)$ matrices are bounded in absolute value by 1 since these matrices are unitary] while, as can be seen from the results of Section 5.9, the matrix elements of $Sp(2n,\mathbb{R})$ matrices can be arbitrarily large. Moreover, it can be shown that $USp(2n)$ is simply connected, and we have seen that $Sp(2n,\mathbb{R})$ is infinitely connected. Finally, for completeness, we remark that $Sp(2n,\mathbb{C})$ is noncompact and simply connected.

5.10.2 $USp(2n)$ and the Quaternion Field

The group $USp(2n)$ is of mathematical interest for at least two reasons. First, because it is compact, it is much easier to analyze than is $Sp(2n,\mathbb{R})$. And, because $sp(2n,\mathbb{R})$ and $usp(2n)$ are complex equivalent, many results obtained for $usp(2n)$ are readily transferable to $sp(2n,\mathbb{R})$. Second, $USp(2n)$ is closely related to quaternions and can be viewed as the quaternion field analog of the groups $O(n,\mathbb{R})$ and $U(n)$ for the real and complex fields.[12] We will now describe briefly how this comes about.

[12] Quaternions as an algebra were discovered by *Hamilton* in 1843, and often the quaternion field is referred to as $\mathbb{H}$. Some aspects of them were also known in some form earlier and independently to *Euler* in 1748, *Gauss* in 1819, and *Rodrigues* in 1840.

Consider an n-dimensional real vector space with the usual real inner product. To emphasize the use of the *real* field, we denote this inner product by the symbols $(,)_{\mathbb{R}}$. Then the set of real linear transformations that preserves this inner product forms the orthogonal group, $O(n, \mathbb{R})$. Specifically, if x and y are any two vectors, we require the relation

$$(Ox, Oy)_{\mathbb{R}} = (x, y)_{\mathbb{R}} \tag{5.10.11}$$

But we also have the relation

$$(Ox, Oy)_{\mathbb{R}} = (x, O^T Oy)_{\mathbb{R}} \tag{5.10.12}$$

from which it follows that O must satisfy the condition

$$O^T O = I. \tag{5.10.13}$$

Next consider an n-dimensional *complex* vector space with the usual complex inner product. We denote this inner product by the symbols $(,)_{\mathbb{C}}$ to emphasize the use of the complex field. Then the set of complex linear transformations that preserves this inner product forms the unitary group, $U(n)$. Specifically, if x and y are any two vectors, we require the relation

$$(Ux, Uy)_{\mathbb{C}} = (x, y)_{\mathbb{C}}. \tag{5.10.14}$$

But we also have the relation

$$(Ux, Uy)_{\mathbb{C}} = (x, U^\dagger Uy)_{\mathbb{C}} \tag{5.10.15}$$

from which it follows that U must satisfy the condition

$$U^\dagger U = I. \tag{5.10.16}$$

Finally, suppose we consider an n-dimensional vector space over the *quaternion* field $\mathbb{H}$ with a suitable inner product yet to be defined. Then the set of linear transformations with quaternion entries that preserves this inner product can be shown to be isomorphic to $USp(2n)$. Thus, the groups $O(n), U(n)$, and $USp(2n)$ all arise from analogous constructions over the real field, the complex field, and the quaternion field, respectively.

We will work up to this result in stages. First we will study the structure of $usp(2n)$ and $USp(2n)$. Next we will represent quaternions using Pauli matrices, and define a suitable inner product. Finally, we will show that $USp(2n)$ preserves this inner product.

5.10.3 Quaternion Matrices

Let S^c be any real $2n \times 2n$ symmetric matrix that *commutes* with J. For J we shall take the form (3.2.10). That is, J is an $n \times n$ collection of 2×2 blocks. Suppose that S^c is also written as an $n \times n$ collection of 2×2 blocks,

$$S^c = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & & \vdots \\ c_{n1} & \cdots & c_{nn} \end{pmatrix}, \tag{5.10.17}$$

where each entry c_{jk} is a 2×2 block. Then it is easily verified that requiring S^c to commute with J is equivalent to requiring that each entry c_{jk} commute with the 2×2 matrix J_2 of (3.2.11),

$$c_{jk}J_2 - J_2 c_{jk} = 0. \tag{5.10.18}$$

The condition (10.18) in turn requires that each c_{jk} be a linear combination (with arbitrary real coefficients) of σ^0 and J_2.

Similarly, let S^a be any real $2n \times 2n$ matrix that *anticommutes* with J. Suppose that S^a is written as an $n \times n$ collection of 2×2 blocks in the form

$$S^a = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}. \tag{5.10.19}$$

Then requiring that S^a anticommute with J is equivalent to requiring that each entry a_{jk} anticommute with J_2,

$$a_{jk}J_2 + J_2 a_{jk} = 0. \tag{5.10.20}$$

It is easily checked that the condition (10.20) implies in turn that each a_{jk} must be a linear combination (with real coefficients) of σ^3 and σ^1.

Now consider matrices of the form iS^a where the entries in S^a itself are real. Then, every 2×2 block in S^a must be a linear combination with real coefficients of the matrices $i\sigma^3$ and $i\sigma^1$. Also, we note that J_2 is given by the relation

$$J_2 = i\sigma^2. \tag{5.10.21}$$

We conclude that all matrices of the form S^c and iS^a, and their linear combinations with real coefficients, must have in their 2×2 blocks only matrices of the form

$$Q = w_0\sigma^0 + iw_1\sigma^1 + iw_2\sigma^2 + iw_3\sigma^3, \tag{5.10.22}$$

where the coefficients w_0 through w_3 are *real.* It is convenient to regard the quantities w_1, w_2, and w_3 as the three components of a vector $\boldsymbol{w}$, and to then write (10.22) in the more compact form

$$Q = w_0\sigma^0 + i\boldsymbol{w} \cdot \boldsymbol{\sigma}. \tag{5.10.23}$$

The reader will eventually have the pleasure of showing, in Exercise 10.15, that the set of all 2×2 matrices of the form (10.23) is isomorphic to the quaternion field $\mathbb{H}$.[13] For this reason, these matrices will be called *quaternion matrices.*

5.10.4 Properties of Quaternion Matrices

Suppose two quaternion matrices Q and Q' are multiplied together. From the relation (7.40) we find the result

$$QQ' = (w_0w_0' - \boldsymbol{w} \cdot \boldsymbol{w}')\sigma^0 + i(w_0\boldsymbol{w}' + w_0'\boldsymbol{w} - \boldsymbol{w} \times \boldsymbol{w}') \cdot \boldsymbol{\sigma} = Q'' \tag{5.10.24}$$

[13] This discovery was first made by Cayley.

with

$$Q'' = w_0''\sigma^0 + i\boldsymbol{w}'' \cdot \boldsymbol{\sigma}. \tag{5.10.25}$$

and

$$w_0'' = w_0 w_0' - \boldsymbol{w} \cdot \boldsymbol{w}', \tag{5.10.26}$$

$$\boldsymbol{w}'' = w_0 \boldsymbol{w}' + w_0' \boldsymbol{w} - \boldsymbol{w} \times \boldsymbol{w}'. \tag{5.10.27}$$

We conclude that the product of two quaternion matrices is again a quaternion matrix. Also, any linear combination with real coefficients of quaternion matrices is again a quaternion matrix,

$$\alpha Q + \beta Q' = (\alpha w_0 + \beta w_0')\sigma^0 + i(\alpha \boldsymbol{w} + \beta \boldsymbol{w}') \cdot \boldsymbol{\sigma} \tag{5.10.28}$$

We summarize the results of (10.24) and (10.28) by saying that the set of quaternion matrices is *closed* under the operators of multiplication and addition with real coefficients.

From the specific form of the Pauli matrices we find that Q can be written in the explicit form

$$Q = \begin{pmatrix} w_0 + iw_3 & iw_1 + w_2 \\ iw_1 - w_2 & w_0 - iw_3 \end{pmatrix}. \tag{5.10.29}$$

From this form we easily compute that the determinant of Q is given by the relation

$$\det(Q) = w_0^2 + w_1^2 + w_2^2 + w_3^2. \tag{5.10.30}$$

We conclude that all quaternion matrices are invertible save for the zero quaternion matrix. That is, quaternion matrices form a *division algebra*.

Given any quaternion matrix Q specified by (10.23), we define a *conjugate* quaternion matrix Q^* by the relation

$$Q^* = w_0\sigma^0 - i\boldsymbol{w} \cdot \boldsymbol{\sigma}. \tag{5.10.31}$$

It is easily verified from (10.24) that the product Q^*Q is given by the relation

$$Q^*Q = QQ^* = (w_0^2 + w_1^2 + w_2^2 + w_3^2)\sigma^0 = [\det(Q)]\sigma^0. \tag{5.10.32}$$

Consequently, the inverse of a quaternion matrix is given by the relation

$$Q^{-1} = Q^*/[\det(Q)]. \tag{5.10.33}$$

Since the Pauli matrices are Hermitian, the definition (10.31) is equivalent to the relation

$$Q^* = Q^\dagger. \tag{5.10.34}$$

We also find by explicit calculation the relation

$$Q^* = -J_2 Q^T J_2 = (J_2)^{-1} Q^T J_2. \tag{5.10.35}$$

From either (10.33), (10.34), or (10.35) we find the relation

$$(Q'Q)^* = Q^*(Q')^*. \tag{5.10.36}$$

We note that if Q is a quaternion matrix, so are the matrices Q^*, Q^{-1}, and Q^T. Finally, we observe from (10.21) that the matrix J_2 is also a quaternion matrix. It follows that the set of all nonzero quaternion matrices forms a group.

5.10.5 Quaternion Matrices and $USp(2n)$

We have learned that all matrices of the form S^c and iS^a, and their linear combinations with real coefficients, must have quaternion matrices in their 2×2 blocks when the form (3.2.10) is used for J. Moreover, since in this form J also has quaternion matrices in its 2×2 blocks, all matrices of the form JS^c and iJS^a, and their linear combinations with real coefficients, must also have quaternion matrices in their 2×2 blocks. This result follows from the fact that the set of quaternion matrices is closed under multiplication and addition with real coefficients. Finally, suppose matrices of the form JS^c and iJS^a, and their linear combinations with real coefficients, are exponentiated and multiplied together. Since these operations all reduce to the multiplication and addition (again with real coefficients) of quaternion matrices, all $n\times n$ arrays resulting from these operations must also have quaternion matrices in their 2×2 blocks. We conclude that when the form (3.2.10) is used for J, all matrices in the group $USp(2n)$ must have quaternion matrices in their 2×2 blocks.

Let M be a matrix in $USp(2n)$. Suppose M is written in an $n\times n$ arrray,

$$M = \begin{pmatrix} m_{11} & \cdots & m_{1n} \\ \vdots & & \vdots \\ m_{n1} & \cdots & m_{nn} \end{pmatrix}, \tag{5.10.37}$$

where, according to the preceding discussion, each block m_{jk} is a quaternion matrix. Then the matrices $M^\dagger$ and M^T are given by the relations

$$M^\dagger = \begin{pmatrix} m_{11}^\dagger & \cdots & m_{n1}^\dagger \\ \vdots & & \vdots \\ m_{1n}^\dagger & \cdots & m_{nn}^\dagger \end{pmatrix}, \tag{5.10.38}$$

$$M^T = \begin{pmatrix} m_{11}^T & \cdots & m_{n1}^T \\ \vdots & & \vdots \\ m_{1n}^T & \cdots & m_{nn}^T \end{pmatrix}. \tag{5.10.39}$$

Because M is unitary, it must satisfy the relation

$$M^\dagger M = I. \tag{5.10.40}$$

However, because the entries in M are quaternion matrices, use of the relations (3.2.10), (10.34), and (10.35) gives the result

$$M^\dagger = J^{-1}M^TJ. \tag{5.10.41}$$

Consequently (10.40) can be rewritten in the form

$$J^{-1}M^TJM = I \;\text{ or }\; M^TJM = J. \tag{5.10.42}$$

We conclude that a unitary matrix whose 2×2 blocks are quaternion matrices must also be a symplectic matrix. Conversely, if M is symplectic and is made of 2×2 quaternion matrix blocks, then M must also be unitary.

5.10.6 Quaternion Inner Product and Its Preservation

Let x and y be any two n-component vectors whose entries are quaternion matrices. We will call such vectors *quaternion vectors*. For quaternion vectors we define a quaternion inner product, which we denote by the symbols $(,)_{\mathbb{H}}$, by the relation

$$(x, y)_{\mathbb{H}} = \sum_{j=1}^{n} x_j^* y_j. \tag{5.10.43}$$

(Note that the result of forming a quaternion inner product is a quaternion matrix.)

Next, suppose M is any matrix in $USp(2n)$. Using the representation (10.37), we define transformed vectors x', y' by the relations

$$x'_j = \sum_k m_{jk} x_k, \tag{5.10.44}$$

$$y'_j = \sum_\ell m_{j\ell} y_\ell. \tag{5.10.45}$$

Note that the operations on the right sides of (10.44) and (10.45) involve only quaternion matrix multiplication and addition. We write (10.44) and (10.45) more compactly in the form

$$x' = Mx \ , \ y' = My. \tag{5.10.46}$$

From (10.34), (10.36), and (10.44) we find the relations

$$(x'_j)^* = \sum_k (m_{jk} x_k)^* = \sum_k x_k^* m_{jk}^* = \sum_k x_k^* (m_{jk})^\dagger. \tag{5.10.47}$$

Let us compute $(Mx, My)_{\mathbb{H}}$. We find from (10.45) and (10.47) the result

$$\begin{aligned}(Mx, My)_{\mathbb{H}} &= (x', y')_Q = \sum_{j=1}^{n} (x'_j)^* y'_j \\ &= \sum_{j,k,\ell} x_k^* (m_{jk})^\dagger m_{j\ell} y_\ell = \sum_{k,\ell} x_k^* \delta_{k\ell} \sigma^0 y_\ell \\ &= \sum_k x_k^* y_k = (x, y)_{\mathbb{H}}.\end{aligned} \tag{5.10.48}$$

Here we have used the fact that the relation (10.40) can be written in the 2 × 2 block form

$$\sum_j (m_{jk})^\dagger m_{j\ell} = \delta_{k\ell} \sigma^0. \tag{5.10.49}$$

See (10.37) and (10.38). Note that (10.49) can also be written in the form

$$\sum_j (m_{jk})^* m_{j\ell} = \delta_{k\ell} \sigma^0, \tag{5.10.50}$$

and in this form only quaternion operations are involved. We conclude from (10.48) that M preserves the quaternion inner product,

$$(Mx, My)_{\mathbb{H}} = (x, y)_{\mathbb{H}}. \tag{5.10.51}$$

Further, it can be checked that if M preserves (10.51) for arbitrary quaternion vectors x and y, then M must belong to $USp(2n)$.

At this point we might wonder if there is a connection between the the quaternion inner product (10.43) and the fundamental symplectic 2-form (3.2.3). By working Exercise 10.18 you will have the pleasure of seeing that they are closely related.

5.10.7 Discussion

Comparison of (10.11), (10.14), and (10.51) shows that $O(n), U(n)$, and $USp(2n)$ all arise from analogous constructions over the real field, the complex field, and the quaternion field, respectively. We remark that the only finite-dimensional associative normed division algebras over the real number field are the real number field itself, the complex field, and the quaternion field. (This proposition is known as Frobenius' theorem.) Thus, $O(n), U(n)$, and $USp(2n)$ are not only analogous, they are also exhaustive.

Reference to Table 3.7.1 shows that we have accounted for all the classical Lie algebras. What can be said about the exceptional algebras? After the reals, the complex numbers, and the quaternions come the *octonions* (also called *Cayley numbers*). As their name suggests, they form an eight-dimensional vector space for which multiplication can also be defined. Like the reals, complexes, and quaternions, octonions form a normed division algebra. (In fact, these four are the *only* normed division algebras.) However, octonion multliplication is *not* associative. It can be shown that, in one way or another, all the exceptional Lie algebras are related to various properties of the octonions. Moreover, the failure of the exceptional Lie algebras to form regular infinite families (like the classical Lie algebras do) is related to the nonassociativity of octonion multiplication.

Exercises

5.10.1. Verify the commutation rules (10.4) through (10.6).

5.10.2. Look at the relations (10.9) and (10.10). Strictly speaking, our discussion has only shown that $usp(2n)$ is contained in $u(2n) \cap sp(2n, \mathbb{C})$, etc. Prove that they are in fact equal. That is, prove that (10.9) and (10.10) are correct.

5.10.3. Show that W as given by (3.9.8) belongs to $USp(2n)$.

5.10.4. Show that the matrix elements of (real) orthogonal matrices and unitary matrices are less than or equal to 1 in absolute value. Show that the matrix elements of matrices in $GL(n, \mathbb{R})$ [which, as we have seen in Section (3.10), is a subgroup of $Sp(2n, \mathbb{R})$] are unbounded. That is, there are matrices in $GL(n, \mathbb{R})$ whose matrix elements are arbitrarily large.

5.10.5. Verify that requiring S^c to commute with J is equivalent to (10.18). Verify the claim that each c_{jk} must be a linear combination of σ^0 and J_2 with real coefficients. Verify that requiring S^a to anticommute with J is equivalent to (10.20). Verify the claim that each a_{jk} must be a linear combination with real coefficients of σ^3 and σ^1.

5.10.6. Verify the multiplication rule (10.24) through (10.27). Show that the multiplication of quaternion matrices is generally not commutative.

5.10.7. Verify the relations (10.29) through (10.34).

5.10.8. Verify (10.35) by explicit calculation. Find the same result using (3.1.7) and (10.33).

5.10.9. Verify (10.36) directly from (10.24) and the definition (10.31).

5.10.10. Verify (10.41). Is it true that any unitary matrix that is also symplectic with respect to (3.2.10) must have quaternion matrices in its 2×2 blocks? Prove your answer.

5.10.11. Verify (10.49).

5.10.12. Show that the quaternion inner product (10.43) has the property

$$(y, x)_{\mathbb{H}} = [(x, y)_{\mathbb{H}}]^*. \tag{5.10.52}$$

Suppose that the vector x has quaternion entries x_j. Let λ be any quaternion. Define $x\lambda$ to be the vector with quaternion entries $x_j\lambda$. Show that the quaternion inner product has the properties

$$(x, y\lambda)_{\mathbb{H}} = [(x, y)_{\mathbb{H}}]\lambda, \tag{5.10.53}$$

$$(x\lambda, y)_{\mathbb{H}} = \lambda^*(x, y)_{\mathbb{H}}. \tag{5.10.54}$$

We see that in the case of quaternion vectors with the quaternion inner product (10.43), what is the analog of scalar multiplication must take place by multiplication on the right.

5.10.13. Show that the set of nonzero quaternion matrices forms a group. Show that any nonzero quaternion matrix Q can be written in the form

$$Q = \exp(v_0\sigma^0 + i\boldsymbol{v} \cdot \boldsymbol{\sigma}), \tag{5.10.55}$$

where v_0 and $\boldsymbol{v}$ are real. Thus, these quaternion matrices form a Lie group. Find the associated Lie algebra. Consider quaternions with determinant +1. In view of (10.30) and Exercise 10.14 below, we will refer to such quaternions as *unit* quaternions. Show that the set of all unit quaternion matrices forms a subgroup that is identical to $SU(2)$. Show, in view of (10.30), that $SU(2)$ may be viewed as the manifold S^3, the 3-dimensional surface of a sphere in 4-dimensional Euclidean space, also known as the *3-sphere*. (It can be shown that among all the n-spheres, only S^1 and S^3 also have the structure of a group.) Suppose that Q is a unit quaternion matrix. Using the parameterization (10.29) and the result (3.9.20), find the matrix $M(Q)$.

5.10.14. Define a quaternion matrix norm by the relation

$$\parallel Q \parallel = \sqrt{\det(Q)}. \tag{5.10.56}$$

Show that this norm satisfies (3.7.10) through (3.7.13).

5.10.15. The purpose of this exercise is to define quaternions and to show that, as discovered by Cayley, they are faithfully represented by quaternion matrices.[14] The quaternion field $\mathbb{H}$, often called Hamilton's quaternion algebra, is a four-dimensional linear vector space over the *real* number field. Let the basis for this vector space be denoted by the symbols e, j, k, ℓ. Impose the following laws of multiplication among the basis vectors:

$$e^2 = e, \quad ej = je = j, \quad ek = ke = k, \quad e\ell = \ell e = \ell;$$

$$j^2 = k^2 = \ell^2 = -e;$$

$$jk = -kj = \ell, \quad k\ell = -\ell k = j, \quad \ell j = -j\ell = k. \tag{5.10.57}$$

Note that the quantities e, j, k, ℓ all anticommute. Since the vectors e, j, k, ℓ form a basis, the most general quaternion is a vector, which we will denote by the symbol q, of the form

$$q = ae + bj + ck + d\ell, \tag{5.10.58}$$

where the quantities a, b, c, d are real numbers.[15] Suppose q' is a second quaternion,

$$q' = a'e + b'j + c'k + d'\ell. \tag{5.10.59}$$

We then have the addition rule

$$q + q' = q'' = a''e + b''j + c''k + d''\ell \tag{5.10.60}$$

with

$$a'' = a + a',\ b'' = b + b',\ c'' = c + c',\ d'' = d + d'. \tag{5.10.61}$$

Show, using the multiplication rules (10.57), that

$$qq' = q'' = a''e + b''j + c''k + d''\ell \tag{5.10.62}$$

with

$$\begin{aligned} a'' &= aa' - bb' - cc' - dd', \\ b'' &= ab' + ba' + cd' - dc', \\ c'' &= ac' + ca' + db' - bd', \\ d'' &= ad' + da' + bc' - cb'. \end{aligned} \tag{5.10.63}$$

Now make the following correspondence ↔ between the quaternion matrices σ^0, $-i\sigma^1$, $-i\sigma^2$, $-i\sigma^3$ and the quaternion basis vectors e, j, k, ℓ:

$$\sigma^0 \leftrightarrow e,$$

[14] For faithful representations of quaternions by real 4×4 matrices, see Exercise 11.1.7.

[15] Other authors, including Hamilton, commonly use the symbols $1, i, j, k$ for our e, j, k, ℓ. Our notation is designed to avoid confusion between quaternion basis vectors and the quantities 1 and $\sqrt{-1}$. Finally we remark that if the coefficients a, b, c, d are permitted to be *complex*, the resulting object is called a *biquaternion*.

$$-i\sigma^1 \leftrightarrow j,$$
$$-i\sigma^2 \leftrightarrow k,$$
$$-\, i\sigma^3 \leftrightarrow \ell. \tag{5.10.64}$$

Make the correspondence $\leftrightarrow$ into a linear mapping by extending it from basis elements to arbitrary elements in a linear fashion. Suppose q is the quaternion (10.58). Define a corresponding quaternion matrix Q by the rule

$$Q = a\sigma^0 + b(-i\sigma^1) + c(-i\sigma^2) + d(-i\sigma^3) = \begin{pmatrix} a-id & -c-ib \\ c-ib & a+id \end{pmatrix}, \tag{5.10.65}$$

and make the correspondence

$$Q \leftrightarrow q. \tag{5.10.66}$$

Using (7.3), verify the arithmetic in (10.65). By linearity the correspondence (10.66) and the correspondence

$$Q' \leftrightarrow q' \tag{5.10.67}$$

imply the correspondence

$$Q' + Q \leftrightarrow q' + q. \tag{5.10.68}$$

Verify this assertion. Using (10.57) and the rules for matrix multiplication, show that the correspondences (10.66) and (10.67) imply the correspondence.

$$Q'Q \leftrightarrow q'q. \tag{5.10.69}$$

See (7.40). Prove that quaternion multiplication is associative.

Given any quaternion q of the form (10.58), the conjugate quaternion q^* is defined by the relation

$$q^* = ae - bj - ck - d\ell. \tag{5.10.70}$$

Show that the correspondence given by (10.65) and (10.66) implies the correspondence

$$Q^\dagger \leftrightarrow q^*. \tag{5.10.71}$$

Review Exercise 10.14. Compare q^*q and $qq^*$with $||Q||^2$.

5.10.16. Suppose M is a (possibly complex) symplectic matrix. Then according to (3.1.8) and (3.1.9), M is nonsingular. Consequently, M must have a unique polar decomposition of the form

$$M = PU, \tag{5.10.72}$$

where P is positive definite Hermitian and U is unitary. Show, in analogy to (3.8.6) and (3.8.7), that P and U are also symplectic. Next show, in analogy to the derivation of (3.9.33), that P must have determinant +1. Now consider the matrix U. Since U is both unitary and symplectic, it must belong to $USp(2n)$. Show that if U is sufficiently near the identity, then it must have determinant +1. But since $USp(2n)$ is connected (indeed, simply connected), every matrix in $USp(2n)$ can be continuously deformed to the identity while remaining within $USp(2n)$. Show, by continuity arguments, that these circumstances require that all U in $USp(2n)$ must have determinant +1. Finally, use (10.72) to show that M must have determinant +1.

5.10.17. Review Exercises 3.1.2 and 3.1.3. Show that the groups $USp(2)$ and $SU(2)$, and correspondingly the Lie algebras $usp(2)$ and $su(2)$, are the same.

5.10.18. The quantities x_j and y_j appearing in (10.46) are quaternion matrices, and therefore can be written in the form

$$y_j = \begin{pmatrix} q_j & s_j \\ p_j & r_j \end{pmatrix}, \tag{5.10.73}$$

$$x_j = \begin{pmatrix} \tilde{q}_j & \tilde{s}_j \\ \tilde{p}_j & \tilde{r}_j \end{pmatrix}, \tag{5.10.74}$$

where the various entries are (possibly complex) numbers and $\tilde{}$ is simply a mark that distinguishes quaternion matrix entries associated with an x_j from those associated with a y_j. Let z, w, $\tilde{z}$, and $\tilde{w}$ be *column* vectors with $2n$ entries of the form

$$z = (q_1, p_1, q_2, p_2, \cdots q_n, p_n)^T, \tag{5.10.75}$$

$$w = (s_1, r_1, s_2, r_2, \cdots s_n, r_n)^T, \tag{5.10.76}$$

$$\tilde{z} = (\tilde{q}_1, \tilde{p}_1, \tilde{q}_2, \tilde{p}_2, \cdots \tilde{q}_n, \tilde{p}_n)^T, \tag{5.10.77}$$

$$\tilde{w} = (\tilde{s}_1, \tilde{r}_1, \tilde{s}_2, \tilde{r}_2, \cdots \tilde{s}_n, \tilde{r}_n)^T. \tag{5.10.78}$$

The vector z is made from the entries in the first columns of the y_j and the vector w is made from the entries in the second columns of the y_j, etc. We know that the quantity $(x, y)_{\mathbb{H}}$ is a quaternion matrix. Show, using (10.38) and (10.46), that it is the quaternion matrix given by the relation

$$(x, y)_{\mathbb{H}} = \begin{pmatrix} -(\tilde{w}, J'z) & -(\tilde{w}, J'w) \\ (\tilde{z}, J'z) & (\tilde{z}, J'w) \end{pmatrix} \tag{5.10.79}$$

where J' is the matrix (3.2.10), the matrix we have been calling J in this section. Evidently the entries of $(x, y)_{\mathbb{H}}$ consist of fundamental symplectic 2-forms involving the vectors z, w, $\tilde{z}$, and $\tilde{w}$. Show that (10.79) can also be written in the more symmetric form

$$(x, y)_{\mathbb{H}} = -J_2 \begin{pmatrix} (\tilde{z}, J'z) & (\tilde{z}, J'w) \\ (\tilde{w}, J'z) & (\tilde{w}, J'w) \end{pmatrix}. \tag{5.10.80}$$

Verify that the matrix appearing in the second factor in (10.80) is a quaternion matrix. Hint: Use the fact that $(x, y)_{\mathbb{H}}$ is a quaternion matrix and that J_2 is an invertible quaternion matrix. Finally, verify that (10.45) is equivalent to the two ordinary vector and matrix relations

$$z' = Mz, \tag{5.10.81}$$

$$w' = Mw, \tag{5.10.82}$$

and that there are analogous results for (10.44). Note that, in writing relations of the form (10.73), we have not forced the y_j, etc., to be quaternion matrices. Show that to do so one should require the relations

$$s_j = -\bar{p}_j, \tag{5.10.83}$$

$$r_j = \bar{q}_j, \tag{5.10.84}$$

where the overbar denotes complex conjugation. Thus, y_j takes the form

$$y_j = \begin{pmatrix} q_j & -\bar{p}_j \\ p_j & \bar{q}_j \end{pmatrix}, \tag{5.10.85}$$

and similarly for x_j. Hint: Use (10.29), but realize the that w's appearing in it are different from those in (10.76). Finally, show that

$$(y, y)_{\mathbb{H}} = \sigma^0 \sum_j (|q_j|^2 + |p_j|^2) = \sigma^0 \sum_j \det y_j. \tag{5.10.86}$$

5.10.19. The work of Section 3.8.2 showed that any element of $Sp(2n, \mathbb{R})$ can be written uniquely in the form

$$M = \exp(JS^a)\exp(JS^c). \tag{5.10.87}$$

Also, according to Section 3.8.1, any unitary matrix U can be written in the form

$$U = \exp(A) \tag{5.10.88}$$

where A is anti-Hermitean. What can be said about matrices in $USp(2n)$?

5.11 Möbius Transformations

Möbius transformations occur in many branches of pure mathematics, and also in some areas of applied mathematics. This section defines and lays out some general properties of Möbius transformations. Two subsequent sections use Möbius transformations to provide a relation between $Sp(2n, \mathbb{R})$ and the theory of several complex variables, and to provide a relation between symplectic and symmetric matrices. This second relation generalizes the Cayley representation of Section 3.12. Later, in Section 6.7 of Chapter 6, Möbius transformations will be used to provide a fundamental connection between symplectic maps and gradient maps, thereby producing a plethora of generating functions.

5.11.1 Definition in the Context of Complex Variables

In the theory of a single complex variable z, one set of transformations of particular interest is the set of *Möbius* or *homographic* or *fractional linear* transformations given by relations of the kind[16]

$$z' = (az + b)/(cz + d). \tag{5.11.1}$$

(Some authors refer to these transformations as *linear* even though they manifestly are not.) Let M be a 2×2 matrix of the form

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{5.11.2}$$

[16] In 1778, years before Möbius and Cayley were born, Euler employed this relation and, apart from interchanging the letters a and b in the numerator and the letters c and d in the denominator, wrote the right side of (11.1) in exactly the same form as it appears here. He also considered the possibility that z was the tangent of some other quantity as in (3.11.2).

where the coefficients a through d are those appearing in (11.1). We view (11.1) as a transformation T_M associated with the matrix M, and write (11.1) in the compact form

$$z' = T_M(z). \tag{5.11.3}$$

Suppose T_N is a second Möbius transformation sending z' to z'',

$$z'' = T_N(z'). \tag{5.11.4}$$

Upon combining (11.3) and (11.4), we get the composite transformation

$$z'' = T_N(z') = T_N(T_M(z)). \tag{5.11.5}$$

Direct evaluation of (11.5) shows that the composite transformation is given by the relation

$$T_N T_M = T_{NM}. \tag{5.11.6}$$

Note also that T_I, where I is the identity matrix, is the identity transformation,

$$T_I(z) = z. \tag{5.11.7}$$

Suppose we agree to work with Möbius transformations for which the matrix (11.2) has nonzero determinant. Then (11.6) and (11.7) show that such Möbius transformations form a group. Moreover, suppose we scale all the entries in (11.2) by a common factor. In particular, suppose we select the scaling factor in such a way that the matrix (11.2) has determinant +1. [This will require scaling by a complex number if $\det(M)$ is not positive.] Examination of (11.1) shows that such scaling leaves the Möbius transformation unchanged. That is, there is the relation

$$T_{\lambda M}(z) = T_M(z) \tag{5.11.8}$$

where λ is any non vanishing scalar. We may therefore restrict our attention to matrices (11.2) that are symplectic. Thus, the Möbius transformations associated with these matrices provide a realization of $Sp(2, \mathbb{R})$ or $Sp(2, \mathbb{C})$ as a set of *nonlinear* transformations of the complex plane into itself. We remark that this realization is of mathematical interest for the construction of unitary representations of $Sp(2, \mathbb{R})$. It is of physical interest for the construction of unitary representations of the Lorentz group (needed for elementary particle physics) and for laser optics. In the case of laser optics, the Möbius transformation is essentially the so-called *ABCD law* for the propagation of axially symmetric Gaussian beams.

5.11.2 Matrix Extension

The Möbius transformation can be extended/generalized to higher dimensions in the following way. Let U be a $n \times n$ matrix and let M be a $2n \times 2n$ matrix written in terms of $n \times n$ matrices A^M through D^M in the block form

$$M = \begin{pmatrix} A^M & B^M \\ C^M & D^M \end{pmatrix}. \tag{5.11.9}$$

(Here we use the superscript M in connection with A^M through D^M to indicate that these matrices depend on M.) Define a transformation T_M associated with M that sends U to U' by the rule

$$U' = T_M(U) = (A^M U + B^M)(C^M U + D^M)^{-1}. \tag{5.11.10}$$

Then it can be verified by direct but slightly tedious matrix algebra that successive transformations again obey the composition law (11.6). Indeed, the composition law holds for any set of $n \times n$ matrices U and $2n \times 2n$ matrices M and N. Thus, if we require that the matrices M be invertible, the transformations T_M provide a representation of the group $GL(2n, \mathbb{C})$. Note that according to (11.10) these transformations are again nonlinear. Moreover, in analogy to (11.8), there is the scaling relation

$$T_{\lambda M}(U) = T_M(U) \tag{5.11.11}$$

where λ is any non vanishing scalar. We may therefore restrict our attention to matrices (11.9) that have unit determinant. Thus, if we require that the matrices M have unit determinant, the associated Möbius transformations provide a realization of $SL(2n, \mathbb{C})$ as a set of *nonlinear* transformations of the set of $n \times n$ matrices into itself.

5.11.3 Invertibility Conditions

Of course, for (11.10) to make sense, we must require that the matrix $(C^M U + D^M)$ be invertible,

$$\det(C^M U + D^M) \neq 0. \tag{5.11.12}$$

We will learn that this invertibility condition entails, and is entailed by, three others. That is, we will learn that there are *four equivalent* invertibility conditions. First we must see what these equivalent invertibility conditions might be.

The relation (11.10) can be solved for U by matrix manipulation. First multiply both sides of (11.10) on the right by $(C^M U + D^M)$. So doing gives the result

$$U'(C^M U + D^M) = A^M U + B^M \tag{5.11.13}$$

which, when multiplied out, becomes

$$U' C^M U + U' D^M = A^M U + B^M. \tag{5.11.14}$$

Now rearrange terms in (11.14) so that it becomes

$$(U' C^M - A^M) U = -U' D^M + B^M. \tag{5.11.15}$$

This relation can be rewritten in the form

$$U = (U' C^M - A^M)^{-1}(-U' D^M + B^M). \tag{5.11.16}$$

This assertion makes sense if $(U' C^M - A^M)$ is invertible,

$$\det(U' C^M - A^M) \neq 0. \tag{5.11.17}$$

On the other hand, from (11.10) and the group property (11.6), we deduce that

$$T_{M^{-1}}(U') = T_{M^{-1}}(T_M(U)) = T_{M^{-1}M}(U) = T_I(U) = U. \tag{5.11.18}$$

But, from the general definition (11.10), there is also the relation

$$T_{M^{-1}}(U') = (A^{M^{-1}}U' + B^{M^{-1}})(C^{M^{-1}}U' + D^{M^{-1}})^{-1}. \tag{5.11.19}$$

Comparison of (11.18) and (11.19) gives the result

$$U = (A^{M^{-1}}U' + B^{M^{-1}})(C^{M^{-1}}U' + D^{M^{-1}})^{-1}. \tag{5.11.20}$$

For this result to make sense, the matrix $(C^{M^{-1}}U' + D^{M^{-1}})$ must be invertible,

$$\det(C^{M^{-1}}U' + D^{M^{-1}}) \neq 0. \tag{5.11.21}$$

Also, comparison of (11.20) and (11.16) gives the identity

$$(A^{M^{-1}}U' + B^{M^{-1}})(C^{M^{-1}}U' + D^{M^{-1}})^{-1} = (U'C^M - A^M)^{-1}(-U'D^M + B^M) \tag{5.11.22}$$

which must hold for any matrices U' and M as long as (11.17) and (11.21) are satisfied. Observe that the identity (11.22) can also be written in the form

$$(A^{M^{-1}}U' + B^{M^{-1}})(C^{M^{-1}}U' + D^{M^{-1}})^{-1} = (-U'C^M + A^M)^{-1}(U'D^M - B^M). \tag{5.11.23}$$

Finally, (11.20) can be solved for U' by using the same kind of matrix manipulation that was employed to solve (11.10) for U. So doing gives the result

$$U' = (UC^{M^{-1}} - A^{M^{-1}})^{-1}(-UD^{M^{-1}} + B^{M^{-1}}). \tag{5.11.24}$$

This assertion makes sense if $(UC^{M^{-1}} - A^{M^{-1}})$ is invertible,

$$\det(UC^{M^{-1}} - A^{M^{-1}}) \neq 0. \tag{5.11.25}$$

Also, comparison of (11.10) and (11.24) gives the identity

$$(A^MU + B^M)(C^MU + D^M)^{-1} = (UC^{M^{-1}} - A^{M^{-1}})^{-1}(-UD^{M^{-1}} + B^{M^{-1}}), \tag{5.11.26}$$

which must hold for any matrices U and M as long as (11.12) and (11.25) are satisfied. Note that the identities (11.22) and (11.26) are equivalent: M is simply replaced by M^{-1}.

We will now prove that the four invertibility conditions (11.12), (11.17), (11.21), and (11.25) are equivalent. Let I^n and I^{2n} be the $n\times n$ and $2n\times 2n$ identity matrices, respectively. Then, using the representation (11.9), the equations

$$MM^{-1} = M^{-1}M = I^{2n} \tag{5.11.27}$$

are equivalent to the set of equations

$$A^MA^{M^{-1}} + B^MC^{M^{-1}} = A^{M^{-1}}A^M + B^{M^{-1}}C^M = I^n, \tag{5.11.28}$$

$$C^M A^{M^{-1}} + D^M C^{M^{-1}} = C^{M^{-1}} A^M + D^{M^{-1}} C^M = 0, \tag{5.11.29}$$

$$A^M B^{M^{-1}} + B^M D^{M^{-1}} = A^{M^{-1}} B^M + B^{M^{-1}} D^M = 0, \tag{5.11.30}$$

$$C^M B^{M^{-1}} + D^M D^{M^{-1}} = C^{M^{-1}} B^M + D^{M^{-1}} D^M = I^n. \tag{5.11.31}$$

Now examine the pair of relations (11.20) and (11.10). We claim that there is the equality

$$(C^{M^{-1}} U' + D^{M^{-1}})(C^M U + D^M) = I^n. \tag{5.11.32}$$

The proof is by direct calculation. Use of (11.10) gives the result

$$(C^{M^{-1}} U' + D^{M^{-1}}) = (C^{M^{-1}})(A^M U + B^M)(C^M U + D^M)^{-1} + D^{M^{-1}}. \tag{5.11.33}$$

Here we have assumed that (11.12) holds. From (11.33) we conclude that

$$(C^{M^{-1}} U' + D^{M^{-1}})(C^M U + D^M) = C^{M^{-1}}(A^M U + B^M) + D^{M^{-1}}(C^M U + D^M). \tag{5.11.34}$$

The terms on the right side of (11.34) can be regrouped to give the result

$$C^{M^{-1}}(A^M U + B^M) + D^{M^{-1}}(C^M U + D^M) = (C^{M^{-1}} A^M + D^{M^{-1}} C^M)U + (C^{M^{-1}} B^M + D^{M^{-1}} D^M). \tag{5.11.35}$$

By (11.29), the first factor on the right side of (11.35) vanishes, and by (11.31) the second factor equals I^n. Therefore (11.32) is correct. Now take determinants of both sides of (11.32) to find the result

$$[\det(C^{M^{-1}} U' + D^{M^{-1}})][\det(C^M U + D^M)] = 1. \tag{5.11.36}$$

We conclude that (11.12) and (11.21) are logically equivalent,

$$\det(C^{M^{-1}} U' + D^{M^{-1}}) \neq 0 \;\Leftrightarrow\; \det(C^M U + D^M) \neq 0. \tag{5.11.37}$$

In a similar way, using (11.24), it can be verified that

$$(U C^{M^{-1}} - A^{M^{-1}})(U' C^M - A^M) = I^n, \tag{5.11.38}$$

providing (11.25) holds. Hence (11.17) and (11.25) are logically equivalent,

$$\det(U' C^M - A^M) \neq 0 \;\Leftrightarrow\; \det(U C^{M^{-1}} - A^{M^{-1}}) \neq 0. \tag{5.11.39}$$

It remains to be shown that the invertibility conditions (11.12) and (11.25) are logically equivalent,

$$\det(C^M U + D^M) \neq 0 \;\Leftrightarrow\; \det(U C^{M^{-1}} - A^{M^{-1}}) \neq 0. \tag{5.11.40}$$

Once this is done we will have the complete chain of inferences

$$(11.21) \;\Leftrightarrow\; (11.12) \;\Leftrightarrow\; (11.25) \;\Leftrightarrow\; (11.17). \tag{5.11.41}$$

Specifically, in terms of matrices, (11.41) states that there is the complete chain of inferences

$$\begin{aligned} &\det(C^{M^{-1}} U' + D^{M^{-1}}) \neq 0 \;\Leftrightarrow\; \det(C^M U + D^M) \neq 0 \;\Leftrightarrow \\ &\det(U C^{M^{-1}} - A^{M^{-1}}) \neq 0 \;\Leftrightarrow\; \det(U' C^M - A^M) \neq 0. \end{aligned} \tag{5.11.42}$$

We now check the logical equivalence (11.40). It can be verified from (11.28) through (11.31) by matrix multiplication that there is the identity

$$\begin{pmatrix} I & -U \\ C^M & D^M \end{pmatrix}\begin{pmatrix} A^{M^{-1}} & B^{M^{-1}} \\ C^{M^{-1}} & D^{M^{-1}} \end{pmatrix} = \begin{pmatrix} A^{M^{-1}} - UC^{M^{-1}} & B^{M^{-1}} - UD^{M^{-1}} \\ 0 & I^n \end{pmatrix}. \tag{5.11.43}$$

Also, there is the identity

$$\begin{pmatrix} I^n & -U \\ C^M & D^M \end{pmatrix} = \begin{pmatrix} I^n & 0 \\ C^M & I^n \end{pmatrix}\begin{pmatrix} I^n & -U \\ 0 & C^M U + D^M \end{pmatrix}. \tag{5.11.44}$$

Taking the determinant of both sides of (11.44) gives the result

$$\det\begin{pmatrix} I^n & -U \\ C^M & D^M \end{pmatrix} = \det(C^M U + D^M). \tag{5.11.45}$$

Finally, take the determinant of both sides of (11.43) and use (11.45) to get the relation

$$[\det(C^M U + D^M)][\det(M^{-1})] = \det(A^{M^{-1}} - UC^{M^{-1}}) = (-1)^n \det(UC^{M^{-1}} - A^{M^{-1}}). \tag{5.11.46}$$

Since we have assumed that M is invertible, we have $[\det(M^{-1})] \neq 0$ and therefore the relation (11.46) implies the relation (11.40).

5.11.4 Transitivity

We close this section with a simple, but useful observation. Suppose U and V are any two nonsingular matrices. Then there is a nonsingular matrix M such that

$$V = T_M(U). \tag{5.11.47}$$

That is, any nonsingular matrix can be sent into any other nonsingular matrix by a suitable Möbius transformation. To verify this assertion, simply define M by the equation

$$M = \begin{pmatrix} VU^{-1} & 0 \\ 0 & I \end{pmatrix}, \tag{5.11.48}$$

and see that (11.47) is satisfied.

Exercises

5.11.1. Verify that

$$T_I(U) = U \tag{5.11.49}$$

and that (11.6) holds for the generalized Möbius transformation (11.10).

5.11.2. Verify (11.24).

5.11.3. Verify (11.28) through (11.31).

5.11.4. The critical reader might object that the proof of the logical equivalence (11.37) is incomplete. Why? Using (11.20) under the assumption (11.21), show that

$$(C^M U + D^M)(C^{M^{-1}} U' + D^{M^{-1}}) = I^n. \tag{5.11.50}$$

Similarly, complete the proof of the logical equivalence(11.39). Verify (11.38) and show that

$$(U' C^M - A^M)(U C^{M^{-1}} - A^{M^{-1}}) = I^n \tag{5.11.51}$$

using (11.16) under the assumption (11.17).

5.11.5. Verify (11.43) through (11.46).

5.11.6. Suppose two functions $f(z)$ and $g(z)$ are connected by the relation

$$g(z) = [af(z) + b]/[cf(z) + d] \tag{5.11.52}$$

which, employing the notation of (11.2) and (11.3), we also write in the form

$$g = T_M(f). \tag{5.11.53}$$

Assume that $\det M \neq 0$. Write $g \sim f$ if (11.53) holds for some M. Show that $\sim$ is an equivalence relation among functions. (For the definition of an equivalence relation, see Exercise 5.12.7.) Show that

$$f \sim g \tag{5.11.54}$$

if, and only if,

$$\mathcal{S}f = \mathcal{S}g \tag{5.11.55}$$

where $\mathcal{S}$ denotes the Schwarzian derivative (1.2.16). Show that the differential equation

$$\mathcal{S}f = 0 \tag{5.11.56}$$

has, as its most general solution, the relation

$$f(z) = (az + b)/(cz + d). \tag{5.11.57}$$

5.11.7. Exercise 3.12.5 introduced the Cayley function cay defined by

$$\text{cay}(X) = (I - X)/(I + X). \tag{5.11.58}$$

Show that

$$\text{cay}(X) = T_M(X) \tag{5.11.59}$$

with

$$M = (1/\sqrt{2}) \begin{pmatrix} -I & I \\ I & I \end{pmatrix}. \tag{5.11.60}$$

Verify that

$$M^2 = I, \tag{5.11.61}$$

in agreement with (3.12.47).

5.12 Symplectic Transformations and Siegel Space

5.12.1 Action of $Sp(2n,\mathbb{C})$ on the Space of Complex Symmetric Matrices

Suppose X and Y are $n \times n$ real matrices. Define the most general *complex* $n \times n$ matrix Z by writing the relation

$$Z = X + iY. \tag{5.12.1}$$

Now, with U replaced by Z, define a generalized Möbius transformation T_M associated with any $2n \times 2n$ matrix M and sending Z to Z' by the rule

$$Z' = T_M(Z) = (AZ + B)(CZ + D)^{-1}. \tag{5.12.2}$$

Here, for the moment, we have omitted the M superscript on the matrices A through D.

Suppose M is symplectic. Then it is a remarkable fact that Z' is symmetric if Z is symmetric. (We say that Z is symmetric if X and Y are symmetric.) Thus, if we regard complex symmetric matrices as generalizations of a single complex variable, then the transformations T_M with M symplectic can be viewed as transformations of a generalized complex variable.

Brute force verification of the assertion that Z' is symmetric if Z is symmetric (provided that M is symplectic) is difficult. However, the proof is easy if we make use of the fact that $Sp(2n,\mathbb{C})$ is generated by matrices of the form (3.3.9) through (3.3.11). The assertion can easily be verified for the transformations associated with these matrices, and use of the group property (11.6) then assures that the assertion is true for all symplectic matrices.

We now check each of the cases (3.3.9) through (3.3.11). Suppose M is of the form (3.3.9). Then we have the transformation

$$Z' = Z + B. \tag{5.12.3}$$

Because of (3.3.12), Z' is symmetric if Z is symmetric. Next suppose M is of the form (3.3.11). Then we have the transformation

$$Z' = AZA^T \tag{5.12.4}$$

where use has been made of (3.3.13). Again Z' is symmetric if Z is. Finally, suppose that M is of the form (3.3.10). Recall the conjugacy relation (3.10.8). Evidently, in view of this relation, of what has already been checked, and by the group property, verification of the case (3.3.10) is equivalent to verification of the case $M = J$. When $M = J$, we have the transformation

$$Z' = -Z^{-1}. \tag{5.12.5}$$

Again it is evident that Z' is symmetric if Z is.

5.12.2 Siegel Space and $Sp(2n,\mathbb{R})$

One of the properties of the Möbius transformation (11.1), when the coefficients a through d are real and $\det M = ad - bc > 0$, is that it maps the upper half plane $y > 0$ into itself, and is in fact the most general analytic mapping of the upper half plane into itself.

Consider all symmetric matrices of the form (12.1) with Y positive definite. Such matrices are sometimes called a *Siegel space*, and may be viewed as a *generalized upper half plane* (guhp). Remarkably, this guhp is mapped into itself by the generalized Möbius transformation (12.2) providing M is real symplectic. (See Exercise 12.3.) Indeed, it can be shown that the most general analytic mapping of the guhp into itself must be of the form (12.2) with M an element of $Sp(2n, \mathbb{R})$.

With regard to physical applications, we remark that the generalized Möbius realization of $Sp(4, \mathbb{R})$ is of interest for the propagation of generalized Gaussian laser beams when axial symmetry is not assumed.

5.12.3 Group Actions on Homogeneous Spaces

5.12.3.1 Definition of a Homogeneous Space

What is going on here? Speaking abstractly, we have a group G [$Sp(2n, \mathbb{R})$ in our case] acting on some space $\mathcal{Z}$ (the guhp in our case) by mapping it into itself,

$$G : \mathcal{Z} \rightarrow \mathcal{Z}. \tag{5.12.6}$$

Suppose the action of G on $\mathcal{Z}$ is such that given any two "points" Z and Z' in $\mathcal{Z}$, there is some group element g in G whose action sends Z to Z'. Then the action of G on $\mathcal{Z}$ is said to be *transitive*, and the space $\mathcal{Z}$ is said to be *homogeneous* with respect to the group G. The remarkable fact about a homogeneous space, as we will eventually show, is that there is a natural identification between it and the coset space of the group G with respect to some subgroup H. Moreover, the action of G on the homogeneous space is equivalent, under this identification, to the action of G (under group multiplication) on its own coset space. Thus, in a sense, homogeneous spaces are really aspects of various groups masquerading as if they were independent spaces in their own right.

5.12.3.2 Siegel Space Is a Homogeneous Space

Before continuing our general abstract discussion, we will show that the guhp is a homogeneous space with respect to $Sp(2n, \mathbb{R})$. First consider the point Z^0 given by (12.1) with $X = 0$ and $Y = I$,

$$Z^0 = iI. \tag{5.12.7}$$

Note that Z^0 is symmetric and I is positive definite. Thus, Z^0 is in the guhp, and may be viewed as the analog of the point $+i$ in the ordinary upper half plane. Next consider all matrices L in $Sp(2n, \mathbb{R})$ that leave Z^0 fixed under the action (12.2). The relation

$$T_L(Z^0) = Z^0, \tag{5.12.8}$$

with Z^0 given by (12.7), is equivalent to the relation

$$(AiI + B)(CiI + D)^{-1} = iI, \tag{5.12.9}$$

which gives the relation

$$AiI + B = iI(CiI + D). \tag{5.12.10}$$

Upon equating real and imaginary parts in (12.10), we find the results

$$B = -C \quad , \quad D = A. \tag{5.12.11}$$

Now look at (3.9.28). We see that L must be orthogonal as well as (real) symplectic. We already know from Section 3.9 that such matrices form a $U(n)$ subgroup in $Sp(2n, \mathbb{R})$. Evidently, a necessary and sufficient condition for L to satisfy (12.8) is that L belong to the $U(n)$ subgroup.

Next suppose that X is any real symmetric $n \times n$ matrix and Y is any real symmetric positive definite $n \times n$ matrix. Since Y is real symmetric positive definite, it has a square root $Y^{1/2}$ that is also real symmetric positive definite, and this matrix has an inverse $Y^{-1/2}$ that is also real symmetric positive definite. (See Exercise 12.4.) Consider the matrix $M(Z)$ defined by (12.1) and the relation

$$M(Z) = \begin{pmatrix} Y^{1/2} & 0 \\ 0 & Y^{-1/2} \end{pmatrix} \begin{pmatrix} I & Y^{-1/2}XY^{-1/2} \\ 0 & I \end{pmatrix} = \begin{pmatrix} Y^{1/2} & XY^{-1/2} \\ 0 & Y^{-1/2} \end{pmatrix}. \tag{5.12.12}$$

Look at the two factors in (12.12). The first factor is of the form (3.3.11) and satisfies (3.3.13). Therefore it is symplectic. The second factor is of the form (3.3.9) and satisfies the first of the relations (3.3.12). Therefore it is also symplectic. It follows that M is symplectic. Now let M act on Z^0. We find the result

$$\begin{aligned} T_M(Z^0) &= (Y^{1/2}iI + XY^{-1/2})(0iI + Y^{-1/2})^{-1} \\ &= (iY^{1/2} + XY^{-1/2})Y^{1/2} = X + iY = Z, \end{aligned} \tag{5.12.13}$$

where, according to (12.1), Z is an arbitrary point in the guhp. Moreover we have the inverse relation

$$T_{M^{-1}}(Z) = T_{M^{-1}}(T_M(Z^0)) = T_{M^{-1}M}(Z^0) = T_I(Z^0) = Z^0. \tag{5.12.14}$$

This result follows from (11.37), which also appears in the logical equivalence chain (11.42), and from (11.6). We note that (11.37) insures the mutual invertibility of any relation of the form (11.10) and its inverse. Finally, let Z' be any other point in the guhp. Define a symplectic matrix M' associated with Z' by the analog of (12.2). Then we have the result

$$Z' = T_{M'}(Z^0) = T_{M'}(T_{M^{-1}}(Z)) = T_{M'M^{-1}}(Z). \tag{5.12.15}$$

We conclude that the symplectic matrix $M'M^{-1}$ sends the arbitrary point Z in the guhp to the arbitrary point Z' in the guhp. Therefore the guhp is indeed a homogeneous space with respect to $Sp(2n, \mathbb{R})$.

5.12.4 Homogeneous Spaces and Cosets

We now resume our general abstract discussion of homogeneous spaces. As before, $\mathcal{Z}$ will denote some space on which some group G acts according to the relation

$$Z' = T_g(Z). \tag{5.12.16}$$

Here g is any element in G and T_g is some transformation rule that depends on g. In analogy with (11.6), we require that the transformation rule satisfy the group representation property

$$T_{g_1}(T_{g_2}(Z)) = T_{g_1 g_2}(Z) \tag{5.12.17}$$

for all "points" Z in $\mathcal{Z}$ and all elements g_1 and g_2 in G. We also require the relation

$$T_e(Z) = Z \text{ for all } Z \text{ in } \mathcal{Z}, \tag{5.12.18}$$

where e is the identity element in G.

5.12.4.1 Definition of Stability Group

Now pick some point in $\mathcal{Z}$ and call it Z^0.[17] Consider all elements h in G that keep Z^0 *fixed*. That is, consider all elements h such that

$$T_h(Z^0) = Z^0. \tag{5.12.19}$$

If h_1 is such an element, it follows from (12.17) through (12.19) that h_1^{-1} is also such an element,

$$T_{h_1^{-1}}(Z^0) = Z^0. \tag{5.12.20}$$

Also, if h_1 and h_2 are two such elements, it follows from (12.17) that the product $h_1 h_2$ is also such an element. We conclude that the elements h from a subgroup of G, which we will call H. This subgroup is often referred to as the *stability* (stationary, isotropy, little) *group* of Z^0.

Suppose we had selected some other point Z^1 instead of Z^0. Then, since we assume that the action of G is transitive, there is some g_1 in G such that

$$Z^1 = T_{g_1}(Z^0). \tag{5.12.21}$$

Consider the subgroup of elements in G that keep Z^1 fixed. From (12.17), (12.19), and (12.21) we have the relations,

$$\begin{aligned} T_{g_1 h g_1^{-1}}(Z^1) &= T_{g_1 h}(T_{g_1^{-1}}(Z^1)) = T_{g_1 h}(Z^0) \\ &= T_{g_1}(T_h(Z^0)) = T_{g_1}(Z^0) = Z^1. \end{aligned} \tag{5.12.22}$$

We conclude that the subgroup that keeps Z^1 fixed is conjugate (under g_1) to the subgroup that keeps Z^0 fixed. See Exercise 12.6. Therefore, it does not really matter what point in $\mathcal{Z}$ we choose to be a fixed point.

[17] We reiterate that in this subsection and the next we are again working in the abstract. That is, we are dealing with some general point Z^0 in some abstract space $\mathcal{Z}$ and not necessarily the specific point Z^0 given by (12.7) in the concrete case $\mathcal{Z} = \text{guhp}$.

5.12.4.2 Use of Stability Group to Define Cosets

We will now use the subgroup H to define an equivalence relation among the elements of G. Suppose g_1 and g_2 are any two elements in G. We say that g_2 is *equivalent* to g_1 (and write $g_2 \sim g_1$) if there exists an h in H such that

$$g_1^{-1} g_2 = h \text{ or, put another way, } g_2 = g_1 h. \tag{5.12.23}$$

This equivalence relation can be used to partition the elements of G into disjoint equivalence classes. These equivalence classes are called the *left cosets* of G with respect to H.[18] The collection of all of these cosets is called the *left coset space*, and is customarily denoted by the symbols G/H. See Exercises 12.7 and 12.15.

5.12.4.3 Identification of a Homogeneous Space with Cosets

Suppose two group elements g_1 and g_2 both send Z^0 to the same point Z',

$$T_{g_1}(Z^0) = Z',\ T_{g_2}(Z^0) = Z'. \tag{5.12.24}$$

Then from (12.24) we have the result

$$T_{g_1^{-1} g_2}(Z^0) = T_{g_1^{-1}}(T_{g_2}(Z^0)) = T_{g_1^{-1}}(Z') = Z^0. \tag{5.12.25}$$

It follows from (12.25) and the definition of H that there is an h in H such that (12.23) is satisfied, and we include that g_1 and g_2 are in the same equivalence class (coset). Conversely, if g_1 and g_2 are equivalent (in the same coset), then it follows from (12.23) and (12.19) that

$$T_{g_2}(Z^0) = T_{g_1 h}(Z^0) = T_{g_1}(T_h(Z^0)) = T_{g_1}(Z^0). \tag{5.12.26}$$

Thus, they then both send Z^0 to the same point Z'. We conclude that the points Z' may be used to label the cosets, and that the correspondence between cosets in G/H and points Z' in $\mathcal{Z}$ is one-to-one. Put another way, to see what coset a particular group element g belongs to, simply compute $T_g(Z^0)$. We conclude that there is a natural identification between points in $\mathcal{Z}$ and cosets in G/H:

$$\mathcal{Z} \leftrightarrow G/H. \tag{5.12.27}$$

5.12.5 Group Action on Cosets Equals Group Action on a Homogeneous Space

Next, suppose that g_1 is some element in G. All elements in the same coset as g_1 are of the form $g_1 h$ with h being an arbitrary element in H. For any element in this coset we have the result

$$T_{g_1 h}(Z^0) = T_{g_1}(T_h(Z^0)) = T_{g_1}(Z^0) = Z^1. \tag{5.12.28}$$

[18] Note that while, for the second relation in (12.23), the elements of H act by multiplication on the right, the elements of G that are being acted on appear on the *left*.

Let g be any element in G. Consider the element gg_1. It must belong to some coset. Suppose g_2 also belongs to this coset. Then we must have the relation

$$gg_1 = g_2h', \tag{5.12.29}$$

where h' is some element in H. We note that (12.29) can also be written in the form

$$g(g_1h) = g_2h'', \tag{5.12.30}$$

where h and h'' are elements in H. In this form we see that the effect of g in (12.30) is to send the coset containing g_1 to the coset containing g_2. That is, elements g in G act on "points" (cosets) in G/H by left multiplication.

Finally, suppose the coset g_1 is labelled by Z^1 as in (12.28), and that the coset containing g_2 is labelled by Z^2,

$$T_{g_2}(Z^0) = Z^2. \tag{5.12.31}$$

Let us compute the action of g on Z^1. We find the result

$$\begin{aligned} T_g(Z^1) &= T_g(T_{g_1h}(Z^0)) = T_{gg_1h}(Z^0) \\ &= T_{g_2h''}(Z^0) = T_{g_2}(T_{h''}(Z^0)) = T_{g_2}(Z^0) = Z^2. \end{aligned} \tag{5.12.32}$$

Upon comparing (12.32) and (12.30), we see that the action of G on $\mathcal{Z}$ is equivalent to the left multiplicative action of G on G/H.

5.12.6 Application of Results to Action of $Sp(2n, \mathbb{R})$ on Siegel Space

How do these results work out in the case of $Sp(2n, \mathbb{R})$ and its action on the guhp? From the discussion surrounding (12.8) we learned that the subgroup H of $Sp(2n, \mathbb{R})$ that keeps Z^0 fixed [with Z^0 given by (12.7)] is $U(n)$. It follows that points Z in the guhp $\mathcal{Z}$ are in one-to-one correspondence with the cosets in $Sp(2n, \mathbb{R})/U(n)$. Indeed, suppose we are given a real symplectic matrix N. In accord with the previous discussion, to find out what coset it belongs to we simply compute $T_N(Z^0)$. From (12.2) and (12.7) we find the result

$$Z = T_N(Z^0) = (iA + B)(iC + D)^{-1}. \tag{5.12.33}$$

Here N is assumed to be written in the block form (3.3.1). Thanks to Exercise 12.3, we know that Z is in the guhp. Finally, we note that the generalized Möbius transformation (12.2) is equivalent to the left multiplicative action of $Sp(2n, \mathbb{R})$ on $Sp(2n, \mathbb{R})/U(n)$.

Let Z be an arbitrary point in the guhp. We know that it labels a coset of $Sp(2n, \mathbb{R})/U(n)$ and that, according to (12.13), the matrix $M(Z)$ given by (12.12) belongs to this coset. The most general matrix belonging to this coset, call it N, is of the form $M(Z)L$ where L is in $U(n)$. We also know that the general element in $U(n)$ can be written in the form $\exp(JS^{c'})$. It follows that the general element N in $Sp(2n, \mathbb{R})$ can be written uniquely in the form

$$N = M(Z)\exp(JS^{c'}) \tag{5.12.34}$$

for some (unique) point Z in the guhp and some $S^{c'}$. Since the general element in $U(n)$ can also be written in the form (3.9.19) with m unitary, it follows that N can just as well be written uniquely in the form

$$N = M(Z)M(m) \tag{5.12.35}$$

for some (unique) point Z in the guhp and some (unique) $n \times n$ unitary matrix m. The factorization (12.34) or (12.35) provides what we will call a *partial Iwasawa decomposition* or *factorization* for $Sp(2n, \mathbb{R})$. For a discussion of the associated partial Iwasawa decomposition of the Lie algebra $sp(2n, \mathbb{R})$, see Exercise 7.2.12. For a discussion of what is usually called the (full) Iwasawa decomposition, see Section *. For a variant of the partial Iwasawa factorization, see Exercise 12.11.

Let us try to write $M(Z)$ in factorized Lie form. Look again at the two factors in (12.12). The second factor can be written in the form

$$\begin{pmatrix} I & Y^{-1/2}XY^{-1/2} \\ 0 & I \end{pmatrix} = \exp\begin{pmatrix} 0 & Y^{-1/2}XY^{-1/2} \\ 0 & 0 \end{pmatrix} = \exp(JS), \tag{5.12.36}$$

where S is given by the relation

$$S = \begin{pmatrix} 0 & 0 \\ 0 & Y^{-1/2}XY^{-1/2} \end{pmatrix}. \tag{5.12.37}$$

The first factor in (12.12) can be written in the form

$$\begin{pmatrix} Y^{1/2} & 0 \\ 0 & Y^{-1/2} \end{pmatrix} = \exp\begin{pmatrix} (1/2)\log Y & 0 \\ 0 & (-1/2)\log Y \end{pmatrix} = \exp(JS) \tag{5.12.38}$$

where S is given by the relation

$$S = \begin{pmatrix} 0 & (1/2)\log Y \\ (1/2)\log Y & 0 \end{pmatrix}. \tag{5.12.39}$$

For an explanation of the meaning of $\log Y$, see Exercise 12.9. Of course, we also know that $M(Z)$ has a factorization of the form

$$M(Z) = \exp(JS^a)\exp(JS^{c''}), \tag{5.12.40}$$

where each of the factors on the right side of (12.40) is unique, and hence uniquely determined by Z. Upon combining (12.34) and (12.40) we find the result

$$\begin{aligned} N &= \exp(JS^a)\exp(JS^{c''})\exp(JS^{c'}) \\ &= \exp(JS^a)\exp(JS^c), \end{aligned} \tag{5.12.41}$$

which should be compared with (3.8.24).

5.12.7 Action of $Sp(2n, \mathbb{R})$ on the Generalized Real Axis

We have seen that points Z in the guhp $\mathcal{Z}$ are in one-to-one correspondence with the cosets in $Sp(2n, \mathbb{R})/U(n)$. There is second coset/homogeneous space construction that will be of future use. Consider the space of all real $n \times n$ symmetric matrices X. That is, consider all matrices of the form (12.1) with $Y = 0$. This space may be viewed as a *generalized real axis* (gra). Moreover, if we let any $Sp(2n, \mathbb{R})$ element M act on X by the rule

$$X' = T_M(X) = (AX + B)(CX + D)^{-1}, \tag{5.12.42}$$

then we know from the previous discussion that X' will also be symmetric. Thus, Möbius transformations T_M, with M symplectic and real, send the gra into itself. Moreover, if M is of the form (3.3.9), then we have the transformation

$$X' = X + B. \tag{5.12.43}$$

Consequently, since B can be any symmetric matrix, we see that the action of T_M on the gra is transitive. Therefore the gra is a homogeneous space.

In analogy with our previous discussion, take as a representative element in the gra the matrix X^0 defined by the equation

$$X^0 = 0, \tag{5.12.44}$$

and consider all matrices L in $Sp(2n, \mathbb{R})$ that leave X^0 fixed under the action (12.42). The relation

$$T_L(X^0) = X^0, \tag{5.12.45}$$

with X^0 given by (12.44), is equivalent to the relation

$$(A0 + B)(C0 + D)^{-1} = 0, \tag{5.12.46}$$

which gives the relation

$$B = 0. \tag{5.12.47}$$

We have already learned at the end of Section 3.10 that symplectic matrices with $B = 0$ form a subgroup. See (3.10.19) and (3.10.20). This subgroup does not seem to have an established name, but let us call it $H(2n, \mathbb{R})$ or $H(2n, \mathbb{C})$ depending on the field that is being employed. Then we know from the standard construction discussed earlier that elements in the gra are in one-to-one correspondence with cosets in $Sp(2n, \mathbb{R})/H(2n, \mathbb{R})$.

As a sanity check, let us compare dimensions. First compute the dimension of $H(2n, \mathbb{R})$. Since the block A in (3.10.20) is an arbitrary $n \times n$ matrix, its dimension is n^2. Since the block C is also $n \times n$, and symmetric, its dimension is $n(n + 1)/2$. Finally, the block D is completely specified by (3.3.8), and therefore does not contribute to the dimension count. We conclude that the dimension of $H(2n, \mathbb{R})$ is given by the relation

$$\dim H(2n, \mathbb{R}) = n^2 + n(n + 1)/2 = n(3n + 1)/2. \tag{5.12.48}$$

We already know that the dimension of $Sp(2n, \mathbb{R})$ is $n(2n+1)$. Therefore we have the count

$$\begin{aligned} \dim[Sp(2n, \mathbb{R})/H(2n, \mathbb{R})] &= \dim[Sp(2n, \mathbb{R})] - \dim[H(2n, \mathbb{R})] \\ &= n(2n + 1) - n(3n + 1)/2 = n(n + 1)/2. \end{aligned} \tag{5.12.49}$$

However, since the gra consists of $n \times n$ symmetric matrices, its dimension must also be $n(n+1)/2$,

$$\dim \text{gra} = n(n+1)/2. \tag{5.12.50}$$

Comparison of (12.49) and (12.50) gives the result

$$\dim[Sp(2n,\mathbb{R})/H(2n,\mathbb{R})] = \dim \text{gra}, \tag{5.12.51}$$

as expected.

5.12.8 Symplectic Modular Groups

We close this section with a final remark. Generally, the entries in a matrix belonging to $Sp(2n,\mathbb{R})$ can be any real numbers subject only to the symplectic condition. We might wonder whether there are subgroups of $Sp(2n,\mathbb{R})$ for which all the entries in the various matrices in a given subgroup are *integers* (positive, negative, or zero). Such subgroups do indeed exist, and are called symplectic *modular* groups. The symplectic modular groups and their associated generalized Möbius transformations are important for the theory of automorphic, theta, and elliptic functions.[19] Automorphic and theta functions are among the most important tools of analytic number theory. Moreover, theta functions and the elliptic functions they generate are key to many soluble problems in nonlinear dynamics.

Exercises

5.12.1. For the transformations (12.2), show that $T_{-M} = T_M$. See (11.11). Consequently, the group of Möbius transformations described by M is only homomorphic to $Sp(2n,\mathbb{R})$, and does not provide a faithful representation. [It does provide a faithful representation of the quotient group G/H where $G = Sp(2n,\mathbb{R})$ and H is the invariant subgroup consisting of $\pm I$. This quotient group is called the *projective* symplectic group, and is denoted by the symbols $PSp(2n,\mathbb{R})$.]

5.12.2. Verify the relations (12.3) through (12.5). With regard to (12.5), also show that Z' is symmetric if Z is, and vice versa.

5.12.3. Suppose Z is given by (12.1) with X real symmetric, and Y real symmetric and positive definite. That is, suppose Z is in the guhp. Also, assume that M is real symplectic.

a) Show that Z' given by (12.3) is also in the guhp.

b) Show that Z' given by (12.4) is also in the guhp.

c) Show that Z is invertible, and that Z' given by (12.5) is also in the guhp.

d) Show that Z' given by (12.2) is also in the guhp.

[19] Poincaré's thesis (he was a student of Hermite) was devoted to what he called Fuchsian functions, but are now called automorphic functions.

Hint for part c: Since Y is real symmetric positive definite, there is a real orthogonal matrix O such that

$$OYO^T = D, \tag{5.12.52}$$

where D is diagonal and has positive entries. Define $D^{1/2}$ to be a diagonal matrix with entries equal to the positive square root of the corresponding entries in D. Then we have the relation

$$(D^{1/2})^{-1}OZO^T(D^{1/2})^{-1} = X' + iI, \tag{5.12.53}$$

where X' is given by the relation

$$X' = (D^{1/2})^{-1}OXO^T(D^{1/2})^{-1}. \tag{5.12.54}$$

Verify that X' is real symmetric. Since X' is real symmetric, there is a real orthogonal matrix R such that

$$RX'R^T = D', \tag{5.12.55}$$

where D' is diagonal. Thus, we have the result

$$R(D^{1/2})^{-1}OZO^T(D^{1/2})^{-1}R^T = D' + iI. \tag{5.12.56}$$

Show that $(D' + iI)$ is invertible and that $-(D' + iI)^{-1}$ is in the guph. Finally, show that

$$-Z^{-1} = -O^T(D^{1/2})^{-1}R^T(D' + iI)^{-1}R(D^{1/2})^{-1}O \tag{5.12.57}$$

is in the guhp.

5.12.4. Suppose Y is a real symmetric positive definite matrix. Study the hint to part c of Exercise 12.3. Show that $Y^{1/2}$ defined by the relation

$$Y^{1/2} = O^T D^{1/2} O \tag{5.12.58}$$

satisfies

$$(Y^{1/2})^2 = Y, \tag{5.12.59}$$

and is real symmetric positive definite and invertible, and that its inverse $Y^{-1/2}$ defined by

$$Y^{-1/2} = O^T(D^{1/2})^{-1}O \tag{5.12.60}$$

is also real symmetric positive definite.

5.12.5. Verify (12.12) through (12.15).

5.12.6. Let H^0 be the subgroup that keeps Z^0 fixed, and H^1 be the subgroup that keeps Z^1 fixed. What (12.22) really shows is that all elements of the form $g_1 h g_1^{-1}$, with h in H^0, are in H^1. We write this inclusion relation, using set theoretic notation, in the form

$$g_1 H^0 g_1^{-1} \subset H^1. \tag{5.12.61}$$

Show that there is also the relation

$$g_1 H^0 g_1^{-1} \supset H^1, \tag{5.12.62}$$

and therefore

$$g_1 H^0 g_1^{-1} = H^1. \tag{5.12.63}$$

5.12.7. Let X be some (possibly abstract) set, and let $\sim$ be some relation (something that can be true or false) among pairs of elements in X. The relation $\sim$ is said to be an *equivalence* relation if it satisfies three properties:

a) $x \sim x$ for all x in X (reflexive property).

b) $x_1 \sim x_2$ implies $x_2 \sim x_1$ for all x_1, x_2 in X (symmetric property) .

c) $x_1 \sim x_2$ and $x_2 \sim x_3$ implies $x_1 \sim x_3$ for all x_1, x_2, x_3 in X (transitive property).

The set of all elements in X that are equivalent (under some given equivalence relation $\sim$) to a given x in X is called the *equivalence class* of x. Given an equivalence relation $\sim$ on some set X, show that each x in X belongs to one and only one equivalence class. Thus, under an equivalence relation, a set divides up in a natural way into disjoint subsets. Show that both conjugacy and symplectic conjugacy are equivalence relations. See Exercise 3.5.7. Let G be a group having a subgroup H. Show that (12.23) defines (satisfies the properties of) an equivalence relation among the elements of G.

5.12.8. Verify (12.36) and (12.37).

5.12.9. Let D be the diagonal matrix of Exercise 12.3. Define $\log(D)$ to be a diagonal matrix whose entries are the logarithms of the diagonal entries of D. Since the entries of D are positive, these logarithms can all be taken to be real. Define $\log(Y)$ by the rule

$$\log(Y) = O^T \log(D) O, \tag{5.12.64}$$

where O is the real orthogonal matrix of Exercise 12.3. Show that this matrix satisfies the relation

$$\exp[\log(Y)] = Y. \tag{5.12.65}$$

Show also that $\log(Y)$ is real and symmetric.

5.12.10. In Exercise 3.9.10 you should have found that the dimension of the vector space spanned by all $2n \times 2n$ real matrices of the form JS^a is $n(n+1)$. Use (7.17) and (7.18) to obtain this result. If $Z = X + iY$ is $n \times n$ and symmetric (with X and Y real), show that this space also has real dimension $n(n+1)$. Use (12.34) or (12.35) to derive the relation

$$NN^T = M(Z)M^T(Z). \tag{5.12.66}$$

Use (12.41) to derive the relation

$$NN^T = \exp(2JS^a). \tag{5.12.67}$$

Show from (12.66) and (12.67) that a knowledge of Z completely determines JS^a. Use (12.8), (12.13), and (12.34) to derive the relation

$$T_N(Z^0) = Z. \tag{5.12.68}$$

[Here we are working in the guhp with Z^0 given by (12.7).] Show that a knowledge of JS^a also completely determines Z. That is, show that the $\exp(JS^c)$ part of N in (12.41) makes no contribution to (12.68).

5.12.11. The representation (12.35) might be called a *lower left* partial Iwasawa decomposition or factorization of N because the lower left block of $M(Z)$ is empty. See (12.12). The purpose of this exercise is to show that the general symplectic matrix N also has what we will call an *upper right* partial Iwasawa decomposition or factorization of the form

$$N = \bar{M}(\bar{Z})M(\bar{m}), \tag{5.12.69}$$

where $\bar{M}(\bar{Z})$ is a matrix of the form

$$\bar{M}(\bar{Z}) = \begin{pmatrix} \bar{Y}^{-1/2} & 0 \\ 0 & \bar{Y}^{1/2} \end{pmatrix} \begin{pmatrix} I & 0 \\ -\bar{Y}^{-1/2}\bar{X}\bar{Y}^{1/2} & I \end{pmatrix} = \begin{pmatrix} \bar{Y}^{-1/2} & 0 \\ -\bar{X}\bar{Y}^{-1/2} & \bar{Y}^{1/2} \end{pmatrix}. \tag{5.12.70}$$

(Here, as a test of the reader's mental agility, the overbar does not denote complex conjugation, but rather is used only as a distinguishing mark.) To prove (12.69) and (12.70), consider the matrix $\bar{N}$ defined by the relation

$$\bar{N} = JNJ^{-1}. \tag{5.12.71}$$

Since $\bar{N}$ is symplectic, it must have the factorization (12.35),

$$\bar{N} = M(\bar{Z})M(\bar{m}). \tag{5.12.72}$$

Suppose that N and $\bar{N}$ are written in $n \times n$ block form,

$$N = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \tag{5.12.73}$$

$$\bar{N} = \begin{pmatrix} \bar{A} & \bar{B} \\ \bar{C} & \bar{D} \end{pmatrix}. \tag{5.12.74}$$

Use (12.71) to find the relation between A, B, C, D and $\bar{A}, \bar{B}, \bar{C}, \bar{D}$. Show that

$$\begin{aligned} \bar{Z} &= \bar{X} + i\bar{Y} = (i\bar{A} + \bar{B})(i\bar{C} + \bar{D})^{-1} \\ &= -(C - iD)(A - iB)^{-1} = -Z^{-1}. \end{aligned} \tag{5.12.75}$$

Now solve (12.71) for N and use (12.72) to find the relation

$$N = J^{-1}\bar{N}J = J\bar{N}J^{-1} = JM(\bar{Z})J^{-1}JM(\bar{m})J^{-1}. \tag{5.12.76}$$

Use (12.12) and (3.9.19) to find the results

$$JM(\bar{Z})J^{-1} = \bar{M}(\bar{Z}), \tag{5.12.77}$$

$$JM(\bar{m})J^{-1} = M(\bar{m}). \tag{5.12.78}$$

5.12.12. In the theory of a single complex variable z, the domain $z\bar{z} < 1$ [or, equivalently, $(1 - z\bar{z}) > 0$] is the (open) unit disk. Let Z be a matrix of the form (12.1) with both X and Y real and symmetric. In the space of such matrices we may define a *generalized unit disk* (gud) by the relation

$$(I - ZZ^{\dagger}) > 0, \tag{5.12.79}$$

where here > 0 means positive definite. (Note that since Z is symmetric, $Z^\dagger = \overline{Z}$.) Suppose that Z is in the guhp. In analogy with the case of a single complex variable, it can be shown that W given by

$$W = (Z - iI)(Z + iI)^{-1} \tag{5.12.80}$$

is then in the gud. Conversely, it can be shown that if W is in the gud, then Z given by the inverse of (12.80),

$$Z = i(I + W)(I - W)^{-1}, \tag{5.12.81}$$

is in the guhp. Show that (12.80) sends the point Z^0 given by (12.7) to the origin of the gud. Note that (12.80) and (12.81) are transformations of the form (12.2) with complex entries in M. Indeed, they can be written in the form

$$W = [(2i)^{-1/2}Z - i(2i)^{-1/2}I][(2i)^{-1/2}Z + i(2i)^{-1/2}I]^{-1}, \tag{5.12.82}$$

$$Z = [i(2i)^{-1/2}W + i(2i)^{-1/2}I][-(2i)^{-1/2}W + (2i)^{-1/2}I]^{-1}. \tag{5.12.83}$$

Show that the matrices M associated with the transformations (12.82) and (12.83), see (12.2), are inverses of each other, and are both in $Sp(2n, \mathbb{C})$.

5.12.13. Perform for the guhp a dimension sanity check analogous to that given by (12.51) for the gra. That is, verify the relation

$$\dim[Sp(2n, \mathbb{R})/U(n)] = \dim \text{guhp}. \tag{5.12.84}$$

5.12.14. Suppose M is a symplectic matrix with integer entries. Then the same is true of its powers. Moreover, if N is any other such matrix, products made from M and N have integer entries. Also, the matrices I and J are symplectic matrices with integer entries. Finally, according to (3.1.9), inverse powers of M then also have integer entries. Thus, any set of symplectic matrices with integer entries must form or be part of some group. As described in Subsection 12.8, such groups are called symplectic modular groups. Show that the matrices J_2 [see (3.2.11)] and M and M^T with M given by

$$M = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \tag{5.12.85}$$

generate (by multiplication) a symplectic modular subgroup. Show that this group has an *infinite* number of elements, and exhibit some of them. Consider $n \times n$ orthogonal matrices with integer entries. Do they form a group? Show that there are only a *finite* number of such matrices. Hint: Use Exercise 10.4.

5.12.15. Subsection 12.4 used the subgroup H to define an equivalence relation among the elements of G. It involved right multiplication of elements g of G by elements h of H. Recall (12.23). One can also set up an equivalence relation among elements of G using left multiplication by elements of H. Suppose g_1 and g_2 are any two elements in G. We may say that g_2 is *equivalent* to g_1 (and write $g_2 \sim g_1$) if there exists an h in H such that

$$g_2 g_1^{-1} = h \text{ or, put another way, } g_2 = h g_1. \tag{5.12.86}$$

Verify that (12.86) is indeed an equivalence relation. See Exercise 12.7. This equivalence relation can also be used to partition the elements of G into disjoint equivalence classes. These equivalence classes are called the *right cosets* of G with respect to H. The collection of all of these right cosets is customarily denoted by the symbols $H\backslash G$.

5.13 Möbius Transformations Relating Symplectic and Symmetric Matrices

5.13.1 Overview

Möbius transformations can also be used to show that there is an intimate connection between symplectic matrices and symmetric matrices. Subsequently, as mentioned before, in Section 6.7 the results of this section will be generalized to show that there is a fundamental connection between symplectic maps and gradient maps.

To proceed, we will have to change notation. In the previous sections M was a $2n \times 2n$ matrix that *characterized* a Möbius transformation. In this section M will be a $2n \times 2n$ symplectic matrix that describes the *outcome* of a Möbius transformation acting on some other $2n \times 2n$ matrix W. Conversely, W will also be the outcome of the inverse Möbius transformation *acting* on M. The Möbius transformation itself will be described by a $4n \times 4n$ matrix.

Inspection of (3.11.5) shows that the Cayley representation for M in terms of the $2n \times 2n$ symmetric matrix W is actually a Möbius transformation, and the inverse relation (3.11.12) is also a Möbius transformation. Moreover, the relation between R and W displayed in (4.8.26), and arising from a F_2 generating function, is a Möbius transformation. In both cases a Möbius transformation provides a relation between symplectic and symmetric matrices.

There is a deep reason why, as evinced by these two examples, there is a connection between symplectic and symmetric matrices. And an understanding of this reason will reveal that there are a great many ways of relating symplectic and symmetric matrices by Möbius transformations. This understanding arises as follows: First, introduce a $4n$ dimensional space and define two different symplectic forms on this space. Next show that these forms are congruent under a Darboux transformation. Then show that one of these symplectic forms is related to symplectic matrices in $2n$ dimensional space, and the other is related to symmetric matrices in $2n$ dimensional space. The congruency of the two forms then leads to a connection between symplectic and symmetric matrices. Finally, we will show that this connection is given by Möbius transformations. Along the way we will make use of Lagrangian planes.

5.13.2 The Cayley Möbius Transformation

Before beginning this trek, we pause to extract a useful matrix from the Cayley transformation. Inspection shows that the Cayley transformation (3.12.5) can be written in the form

$$M = T_\tau(W). \tag{5.13.1}$$

where τ is the $4n \times 4n$ matrix

$$\tau = \begin{pmatrix} A^\tau & B^\tau \\ C^\tau & D^\tau \end{pmatrix} \tag{5.13.2}$$

and the matrices A^τ through D^τ are given by the relations

$$A^\tau = J, \tag{5.13.3}$$

$$B^\tau = I, \tag{5.13.4}$$

$$C^\tau = -J, \tag{5.13.5}$$

$$D^\tau = I. \tag{5.13.6}$$

More compactly, we may write

$$\tau = \begin{pmatrix} J & I \\ -J & I \end{pmatrix}. \tag{5.13.7}$$

Here I is the $2n \times 2n$ identity matrix I^{2n}; and J, which we will sometimes write as J^{2n}, is the standard $2n \times 2n$ fundamental matrix given by (3.1.1). We note, as is easily checked, that τ has the pleasing property

$$\tau^T \tau = 2I^{4n} \text{ or } \tau^{-1} = \tau^T/2. \tag{5.13.8}$$

Conversely, W can be written as a Möbius transformation of M in the form

$$W = T_{\tau^{-1}}(M). \tag{5.13.9}$$

See (3.12.19), and make use of (13.8) to write

$$\tau^{-1} = \begin{pmatrix} -J/2 & J/2 \\ I/2 & I/2 \end{pmatrix}. \tag{5.13.10}$$

We now define, for future use, a matrix σ given in terms of τ by the equation

$$\sigma = \sqrt{2}\tau^{-1} = \tau^T/\sqrt{2} = (1/\sqrt{2}) \begin{pmatrix} -J^{2n} & J^{2n} \\ I^{2n} & I^{2n} \end{pmatrix}. \tag{5.13.11}$$

By this definition and (13.8), σ is orthogonal,

$$\sigma^{-1} = \sigma^T = (1/\sqrt{2}) \begin{pmatrix} J^{2n} & I^{2n} \\ -J^{2n} & I^{2n} \end{pmatrix}. \tag{5.13.12}$$

It can also be verified that

$$\det \sigma = 1. \tag{5.13.13}$$

See Exercise 13.1. Finally we note that, in terms of σ, the Cayley relations (3.12.19) and (3.12.5) take the form

$$W = T_\sigma(M) \tag{5.13.14}$$

and

$$M = T_{\sigma^{-1}}(W). \tag{5.13.15}$$

[Note that, in view of (13.11) and with use of a scaling relation of the form (11.11) with the substitutions $M \to \sigma$ or σ^{-1} or τ or τ^{-1} and $U \to M$ or W, the pair (13.1) and (13.9) and the pair (13.14) and (13.15) are equivalent.] We will call T_σ the *Cayley Möbius* transformation.

5.13.3 Two Symplectic Forms and Their Relation by a Darboux Transformation

Now we are ready to continue. As outlined above, we begin by introducing two different symplectic forms in $4n$ dimensional Euclidean space. The first, which we will denote by J^{4n}, is the standard $4n \times 4n$ antisymmetric matrix given by (3.1.1). The second is the $4n \times 4n$ antisymmetric matrix $\tilde{J}^{4n}$ defined by the equation

$$\tilde{J}^{4n} = \begin{pmatrix} J^{2n} & 0^{2n} \\ 0^{2n} & -J^{2n} \end{pmatrix}. \tag{5.13.16}$$

Here 0^{2n} denotes the $2n \times 2n$ null matrix. Evidently $\tilde{J}^{4n}$ has similar properties to those of J^{4n}. It is nonsingular, and in fact satisfies the relation

$$(\tilde{J}^{4n})^2 = -I^{4n}. \tag{5.13.17}$$

It also satisfies, as direct calculation shows, the relation

$$\det \tilde{J}^{4n} = 1. \tag{5.13.18}$$

Our future discussion will capitalize on the properties of J^{4n} and $\tilde{J}^{4n}$ and their block structure.

Is there a relation between J^{4n} and $\tilde{J}^{4n}$? Since both are $4n \times 4n$, antisymmetric, and nonsingular, they must be congruent. That is, they must be related by a Darboux transformation. See Section 3.12. Indeed, it is easily verified that there is the congruency relation

$$\sigma^T(J^{4n})\sigma = \tilde{J}^{4n}. \tag{5.13.19}$$

Because σ is orthogonal, J^{4n} and $\tilde{J}^{4n}$ are also *conjugate* (*similar*). We also remark that (13.17) and (13.18) now follow directly from (13.12) and (13.19) and the already established properties of J^{4n}. Finally, we will call σ the *Cayley Darboux* matrix.

5.13.4 The Infinite Family of Darboux Transformations

Moreover, there is a $2n(4n + 1)$ parameter family of Darboux transformations that connect J^{4n} and $\tilde{J}^{4n}$. Suppose that J^{4n} and $\tilde{J}^{4n}$ are congruent under the action of two Darboux matrices α and β,

$$\alpha^T(J^{4n})\alpha = \tilde{J}^{4n}, \tag{5.13.20}$$

$$\beta^T(J^{4n})\beta = \tilde{J}^{4n}. \tag{5.13.21}$$

By taking determinants of both sides of (13.20) and (13.21) it is easy to see that both α and β are *invertible*. Indeed, they both have determinant +1. All Darboux matrices have determinant +1. See Exercise 13.1. We can say even more. For example, if (13.20) holds, it follows from (13.20) and (13.17) that there is the relation

$$\alpha^{-1} = -\tilde{J}^{4n}\alpha^T J^{4n}. \tag{5.13.22}$$

To continue, from (13.20) and (13.21) we conclude that

$$\alpha^T(J^{4n})\alpha = \beta^T(J^{4n})\beta, \tag{5.13.23}$$

from which it follows that

$$(\beta\alpha^{-1})^T(J^{4n})\beta\alpha^{-1} = (\alpha^{-1})^T\beta^T(J^{4n})\beta\alpha^{-1} = J^{4n}. \tag{5.13.24}$$

Define a matrix γ by the rule

$$\gamma = \beta\alpha^{-1} \text{ or, equivalently, } \beta = \gamma\alpha. \tag{5.13.25}$$

We then see from the far left and far right sides of (13.24) that γ is an element of the group $Sp(4n)$, $\gamma \in Sp(4n)$. Conversely, if β is any element of the form

$$\beta = \gamma\alpha \tag{5.13.26}$$

where γ is an element of $Sp(4n)$ and α satisfies (13.20), then this β satisfies (13.21). We may, for example, take for α the Cayley Darboux matrix σ and write

$$\beta = \gamma\sigma. \tag{5.13.27}$$

Then we get all possible matrices β satisfying (13.21) by using the representation (13.27) and letting γ range over $Sp(4n)$. Therefore the parameter count cited above, which is the dimension of $sp(4n)$, is correct:

$$\text{dimension of set of } 4n \times 4n \text{ Darboux matrices} = \dim sp(4n) = 2n(4n+1). \tag{5.13.28}$$

See (3.7.35) and Table 3.7.1. Thus, for example, in the simplest case of a two-dimensional phase space, there is a 10 parameter family of Darboux matrices/transformations; and in the case of a six-dimensional phase space there is a 78 parameter family of Darboux matrices/transformations.

There is a variant of the argument just made that is also useful. Rewrite (13.27) in the form

$$\beta = \sigma\mu \tag{5.13.29}$$

where μ is yet to be determined. Now require that β be a Darboux transformation so that (13.21) is satisfied. Then, from (13.19) and (13.29) we see that μ must obey the relation

$$\mu^T(\tilde{J}^{4n})\mu = \tilde{J}^{4n}. \tag{5.13.30}$$

We will describe such μ matrices as being $\tilde{J}^{4n}$ symplectic. They form a group which we will refer to as $\tilde{Sp}(4n)$. According to Section 3.12, this group is related to $Sp(4n)$ by a similarity transformation, and therefore has the same dimension as $sp(4n)$. See also Exercise 13.2. Thus, we also get all possible matrices β satisfying (13.21) by using the representation (13.29) and letting μ range over $\tilde{Sp}(4n)$. Either of the representations (13.27) and (13.29) may be used, but sometimes one is more convenient than the other.

Finally, suppose $\hat{\alpha}$ is any matrix of the form

$$\hat{\alpha} = \hat{\gamma}\sigma\hat{\mu} \tag{5.13.31}$$

where

$$\hat{\gamma} \in Sp(4n) \tag{5.13.32}$$

and

$$\hat{\mu} \in \tilde{S}p(4n). \tag{5.13.33}$$

Then, we find that

$$\begin{aligned} \hat{\alpha}^T(J^{4n})\hat{\alpha} &= (\hat{\gamma}\sigma\hat{\mu})^T(J^{4n})\hat{\gamma}\sigma\hat{\mu} = \hat{\mu}^T\sigma^T\hat{\gamma}^T(J^{4n})\hat{\gamma}\sigma\hat{\mu} \\ &= \hat{\mu}^T\sigma^T(J^{4n})\sigma\hat{\mu} = \hat{\mu}^T(\tilde{J}^{4n})\hat{\mu} \\ &= \tilde{J}^{4n}. \end{aligned} \tag{5.13.34}$$

Here we have used relations of the forms (3.1.2), (13.19), and (13.30). We conclude that $\hat{\alpha}$ is a Darboux matrix.

5.13.5 Isotropic Vectors and Lagrangian Planes

5.13.5.1 Construction and Definitions

Next we will introduce and employ the concept of a Lagrangian plane. Suppose M is a $2n \times 2n$ symplectic matrix. View M as a collection of $2n$ column vectors by writing it in the form

$$M = (m^1, m^2, m^3, \cdots m^{2n}) \tag{5.13.35}$$

where each vector m^j is the jth column of M,

$$m_i^j = M_{ij}. \tag{5.13.36}$$

(We will say that each vector m^j is of *length*/dimension $2n$ because each has $2n$ entries.) The vectors m^j form a symplectic basis and are therefore linearly independent. See Section 3.6.3. We will also need the $2n$ column vectors e^j, also of length $2n$, that form the columns of I^{2n}. They have the components

$$e_i^j = \delta_{ij}. \tag{5.13.37}$$

See (3.6.4). Now construct $2n$ column vectors u^j, each of length $4n$, by adjoining the entries of each e^j to the bottom of the entries of each m^j. Thus we have

$$u^1 = (m_1^1, m_2^1, m_3^1, \cdots m_{2n}^1; 1, 0, 0, \cdots)^T = (M_{1,1}, M_{2,1}, M_{3,1}, \cdots M_{2n,1}; 1, 0, 0, \cdots)^T, \tag{5.13.38}$$

$$u^2 = (m_1^2, m_2^2, m_3^2, \cdots m_{2n}^1; 0, 1, 0, \cdots)^T = (M_{1,2}, M_{2,2}, M_{3,2}, \cdots M_{2n,2}; 0, 1, 0, \cdots)^T \text{ etc.} \tag{5.13.39}$$

Put another way, the vectors u^j (for $j = 1$ to $2n$) have the components

$$u_i^j = m_i^j \text{ for } i = 1 \text{ to } 2n, \tag{5.13.40}$$

$$u_i^j = \delta_{i-2n,j} \text{ for } i = 2n+1 \text{ to } 4n. \tag{5.13.41}$$

Even more compactly, we may write

$$u^j = (m^j; e^j)^T. \tag{5.13.42}$$

Evidently the u^j are linearly independent. Indeed, the m^j are linearly independent and so are the e^j. Now compute the quantities $(u^i, \tilde{J}^{4n}u^j)$. Because of the form of $\tilde{J}^{4n}$ and the form of the u^k, there is the result

$$(u^i, \tilde{J}^{4n}u^j) = (m^i, J^{2n}m^j) - (e^i, J^{2n}e^j) = J^{2n}_{ij} - J^{2n}_{ij} = 0. \tag{5.13.43}$$

[Here we have used the fact that the m^j form a symplectic basis. See (3.6.34).] Because all the $(u^i, \tilde{J}^{4n}u^j)$ vanish, the vectors u^k are said to be *isotropic* with respect to the symplectic form $\tilde{J}^{4n}$. More succinctly, we will say that the u^k are $\tilde{J}^{4n}$ isotropic. Finally, a set of $2n$ linearly independent isotropic vectors in a $4n$ dimensional space is said to span a *Lagrangian* plane. In this case we will say that the u^k span a $\tilde{J}^{4n}$ Lagrangian plane. (For the reason why such a plane is called Lagrangian, see Section 6.7.2.)

5.13.5.2 Forming Linear Combinations

Suppose we create a new set of linearly independent vectors $\acute{u}^i$ by forming linear combinations of the u^j. We write

$$\acute{u}^i = \sum\nolimits_j a_{ji}u^j \tag{5.13.44}$$

where the a_{ji} are various coefficients, not all zero, which we may view as the entries in a $2n \times 2n$ matrix a. It is easily verified that the $\acute{u}^k$ are also $\tilde{J}^{4n}$ isotropic,

$$(\acute{u}^i, \tilde{J}^{4n}\acute{u}^j) = 0, \tag{5.13.45}$$

because they are linear combinations of the u^ℓ. They therefore span the same $\tilde{J}^{4n}$ Lagrangian plane as the u^i.

Suppose we also require that any u^k can be expressed as a linear combination of the $\acute{u}^\ell$,

$$u^k = \sum\nolimits_\ell b_{\ell k}\acute{u}^\ell. \tag{5.13.46}$$

Inserting (13.44) into (13.46) gives the relation

$$u^k = \sum\nolimits_\ell b_{\ell k}\sum\nolimits_j a_{j\ell}u^j = \sum\nolimits_j\sum\nolimits_\ell a_{j\ell}b_{\ell k}u^j = \sum\nolimits_j (ab)_{jk}u^j. \tag{5.13.47}$$

Upon comparing both sides of (13.47), and recalling that the u^i are linearly independent, we conclude that there must be the relation

$$(ab)_{jk} = \delta_{jk}. \tag{5.13.48}$$

That is, we must require that a be invertible so that we may write

$$b = a^{-1}. \tag{5.13.49}$$

5.13.6 Connection between Symplectic Matrices and Lagrangian Planes for the Symplectic Form $\tilde{J}^{4n}$

We will now see that there is a close connection between symplectic matrices and $\tilde{J}^{4n}$ Lagrangian planes. Suppose we view the first $2n$ entries of the u^i as column vectors of a $2n \times 2n$ matrix E and the last $2n$ entries as column vectors of a $2n \times 2n$ matrix F. Then we have the relations

$$E = M \tag{5.13.50}$$

and

$$F = I^{2n}. \tag{5.13.51}$$

We will call the collection $\{E, F\} = \{M, I^{2n}\}$ a *standard symplectic pair*. The first matrix in the pair is symplectic, and the second is the identity, which is also symplectic.

Recall the vectors $\acute{u}^i$ given by (13.44). If we use the $\acute{u}^i$ to construct $2n \times 2n$ matrices $\acute{E}$ and $\acute{F}$ in the same way E and F were constructed from the u^i, we find from (13.44) the relations

$$\acute{E} = Ea = Ma \tag{5.13.52}$$

and

$$\acute{F} = Fa = I^{2n}a. \tag{5.13.53}$$

We will call the collection $\{\acute{E}, \acute{F}\} = \{Ea, Fa\} = \{Ma, I^{2n}a\}$ an *equivalent* symplectic pair and write

$$\{Ea, Fa\} \sim \{E, F\} \tag{5.13.54}$$

because multiplication on the right of a pair of matrices by a nonsingular matrix can be shown to set up an equivalence relations among pairs of matrices. See Exercise 13.3. And, because such multiplication does set up an equivalence relation and we have assumed a is nonsingular, we may also write

$$\{Ma, I^{2n}a\} \sim \{M, I^{2n}\}. \tag{5.13.55}$$

Moreover, suppose we are given any set of $2n$ linearly independent $\tilde{J}^{4n}$ isotropic vectors $\acute{u}^i$ in a $4n$ dimensional space and from them we form the associated matrices $\acute{E}$ and $\acute{F}$. Then it is easy to verify from the definition (13.16) and the block structure of $\tilde{J}^{4n}$ that the $\tilde{J}^{4n}$ isotropy condition (13.45) is equivalent to the matrix relation

$$\acute{E}^T J^{2n} \acute{E} = \acute{F}^T J^{2n} \acute{F}. \tag{5.13.56}$$

If $\acute{F}$ is invertible, (13.56) can be rewritten in the equivalent form

$$(\acute{E}\acute{F}^{-1})^T J^{2n} (\acute{E}\acute{F}^{-1}) = J^{2n}, \tag{5.13.57}$$

and we conclude that the matrix M defined by

$$M = \acute{E}\acute{F}^{-1} \tag{5.13.58}$$

is symplectic. In terms of our equivalence relation, we may write

$$\{\acute{E}, \acute{F}\} \sim \{\acute{E}\acute{F}^{-1}, \acute{F}\acute{F}^{-1}\} = \{M, I^{2n}\} \tag{5.13.59}$$

with M given by (13.58). Thus, any set of basis vectors spanning a $\tilde{J}^{4n}$ Lagrangian plane whose associated "F" matrix is invertible produces an equivalent standard symplectic pair. Conversely, we have already seen that any symplectic matrix M produces a set of basis vectors spanning a $\tilde{J}^{4n}$ Lagrangian plane and a standard symplectic pair. In this latter case, their associated F matrix is trivially invertible because it is the identity.

5.13.7 Connection between Symmetric Matrices and Lagrangian Planes for the Symplectic Form J^{4n}

There is an analogous construction that can be carried out for the symplectic form J^{4n}, but now using symmetric matrices W. Suppose W is a $2n \times 2n$ symmetric matrix. View W as a collection of $2n$ column vectors, each of length $2n$, by writing it in the form

$$W = (w^1, w^2, w^3, \cdots w^{2n}) \tag{5.13.60}$$

where each vector w^j is the jth column of W,

$$w_i^j = W_{ij}. \tag{5.13.61}$$

Again we will also employ the $2n$ column vectors e^j, also of length $2n$, that form the columns of I^{2n}. Now construct $2n$ column vectors v^j, each of length $4n$, by adjoining the entries of each e^j to the bottom of the entries of each w^j. This procedure will again yield $2n$ linearly independent vectors because the e^j are linearly independent. Using the compact notation introduced earlier, we may write the v^j in the form

$$v^j = (w^j; e^j)^T. \tag{5.13.62}$$

Let us now compute the quantities $(v^i, J^{4n}v^j)$. From the block form of J^{4n} and (13.62) it is easily checked that the result is given by the relation

$$(v^i, J^{4n}v^j) = (w^i, e^j) - (e^i, w^j) = (e^j, w^i) - (e^i, w^j). \tag{5.13.63}$$

But from (13.61) we have the result

$$(e^i, w^j) = w_i^j = W_{ij}. \tag{5.13.64}$$

Combining these results gives the relation

$$(v^i, J^{4n}v^j) = W_{ji} - W_{ji} = 0. \tag{5.13.65}$$

Here we have used the fact that W is assumed to be symmetric. We conclude that the v^j are J^{4n} isotropic, and span a J^{4n} Lagrangian plane.

As before, from the v^j construct two $2n \times 2n$ matrices, call them G and H. Construct G using the first $2n$ entries in the v^j, and construct H using the last $2n$ entries. In this case we evidently get the results

$$G = W \tag{5.13.66}$$

and

$$H = I^{2n}. \tag{5.13.67}$$

We will call the collection $\{G, H\} = \{W, I^{2n}\}$ a *standard symmetric pair.* The first matrix in the pair is symmetric, and the second is the identity, which is also symmetric.

We can also form vectors $\acute{v}^i$ by taking linear combinations of the v^j. These vectors will also be J^{4n} isotropic,

$$(\acute{v}^i, J^{4n}\acute{v}^j) = 0, \tag{5.13.68}$$

and span the same J^{4n} Lagrangian plane. Moreover, their associated $2n \times 2n$ matrices are given by the relations

$$\acute{G} = Ga = Wa \tag{5.13.69}$$

and

$$\acute{H} = Ha = I^{2n}a. \tag{5.13.70}$$

We will call the collection $\{\acute{G}, \acute{H}\} = \{Ga, Ha\} = \{Wa, I^{2n}a\}$ an equivalent symmetric pair and write

$$\{Ga, Ha\} \sim \{G, H\} = \{W, I^{2n}\}. \tag{5.13.71}$$

Finally, suppose we are given any set of $2n$ linearly independent J^{4n} isotropic vectors $\acute{v}^i$ in a $4n$ dimensional space and from them we form the associated matrices $\acute{G}$ and $\acute{H}$. Then it is easy to verify from the definition (3.1.1) and the block structure of J^{4n} that the J^{4n} isotropy condition (13.68) is equivalent to the matrix relation

$$\acute{G}^T\acute{H} - \acute{H}^T\acute{G} = 0. \tag{5.13.72}$$

If $\acute{H}$ is invertible, (13.72) can be rewritten in the equivalent form

$$\acute{G}\acute{H}^{-1} = (\acute{H}^{-1})^T\acute{G}^T. \tag{5.13.73}$$

Therefore the matrix W defined by the equation

$$W = \acute{G}\acute{H}^{-1} \tag{5.13.74}$$

is symmetric,

$$W^T = W. \tag{5.13.75}$$

In terms of our equivalence relation, we may write

$$\{\acute{G}, \acute{H}\} \sim \{\acute{G}\acute{H}^{-1}, \acute{H}\acute{H}^{-1}\} = \{W, I^{2n}\} \tag{5.13.76}$$

with W given by (13.74). Thus, any set of basis vectors spanning a J^{4n} Lagrangian plane whose associated "H" matrix is invertible produces an equivalent standard symmetric pair. Conversely, we have already seen that any symmetric matrix W produces a set of basis vectors spanning a J^{4n} Lagrangian plane and a standard symmetric pair. In this latter case, their associated H matrix is trivially invertible because it is the identity.

5.13.8 Relation between Symplectic and Symmetric Matrices and the Role of Darboux Möbius Transformations

The stage is set to discover the relation between symplectic and symmetric matrices. Suppose we are given some set of $2n$ vectors u^i that span a $\tilde{J}^{4n}$ Lagrangian plane. Form associated vectors v^i by the rule

$$v^i = \alpha u^i \tag{5.13.77}$$

where α is any $4n \times 4n$ matrix. If we now require that α be a Darboux matrix that satisfies the relation (13.20), then we find the result

$$(v^i, J^{4n} v^j) = (\alpha u^i, J^{4n} \alpha u^i) = (u^i, \alpha^T J^{4n} \alpha u^i) = (u^i, \tilde{J}^{4n} u^j) = 0. \tag{5.13.78}$$

That is, the vectors v^i are J^{4n} isotropic , and span a J^{4n} Lagrangian plane. Next, construct $2n \times 2n$ matrices G and H from the first $2n$ and the last $2n$ entries in the v^i, respectively. Suppose that the matrix H turns out to be invertible. Then we can write

$$\{G, H\} \sim \{GH^{-1}, I^{2n}\}. \tag{5.13.79}$$

From the previous discussion we know that the W given by

$$W = GH^{-1}, \tag{5.13.80}$$

will be symmetric.

What do Möbius transformations have to do with this discussion? Watch. Suppose we are given a symplectic matrix M and from it construct the vectors u^i by (13.42). That is, the u^i are the vectors associated with the standard symplectic pair $\{M, I^{2n}\}$. Define associated vectors v^i using (13.77). Let E and F be the matrices associated with the u^i, and let G and H be the matrices associated with the v^i. Suppose also that we write α in the block form

$$\alpha = \begin{pmatrix} A^\alpha & B^\alpha \\ C^\alpha & D^\alpha \end{pmatrix}. \tag{5.13.81}$$

Then, in terms of the matrices A^α through D^α and the matrices E through G, the relation (13.77) is equivalent to the relations

$$G = A^\alpha E + B^\alpha F, \tag{5.13.82}$$

$$H = C^\alpha E + D^\alpha F. \tag{5.13.83}$$

Therefore we have the result

$$W = GH^{-1} = (A^\alpha E + B^\alpha F)(C^\alpha E + D^\alpha F)^{-1}. \tag{5.13.84}$$

Now use the explicit forms of E and F given by (13.50) and (13.51) to rewrite (13.84). So doing gives the result

$$W = GH^{-1} = (A^\alpha M + B^\alpha)(C^\alpha M + D^\alpha)^{-1}. \tag{5.13.85}$$

5.13.8.1 Mapping of Symplectic Matrices into Symmetric Matrices

We see that W is related to M by the Möbius transformation associated with α,

$$W = T_\alpha(M). \tag{5.13.86}$$

Thus, W has been expressed as the Möbius transformation of a symplectic matrix. Of course, for (13.86) to be well defined, the matrix $(C^\alpha M + D^\alpha)$ must be invertible,

$$\det\,(C^\alpha M + D^\alpha) \neq 0. \tag{5.13.87}$$

[Note that (13.87) is precisely the condition for H to be invertible. See (13.83), (13.50), and (13.51).] That is, given M, we must find some Darboux matrix α satisfying (13.20) such that its associated C^α and D^α also satisfy (13.87). Once this is achieved (and we will verify subsequently that it can be achieved), we know from out previous work that the W given by (13.80) and hence by (13.86) will be symmetric.

Suppose we now hold α fixed (thereby holding its associated C^α and D^α fixed), and vary M. It can be verified by continuity that, for small enough variations in M, (13.87) will continue to hold. Correspondingly, again based on our previous work, we know that the varied W associated with the varied M will continue to be symmetric. Thus we get a local mapping of symplectic matrices into symmetric matrices. Since the α appearing in T_α is a Darboux transformation, we might call T_α a Darboux Möbius transformation.

5.13.8.2 Mapping of Symmetric Matrices into Symplectic Matrices

Conversely, suppose we are given a symmetric matrix W, which may be the W of (13.86). From it construct the vectors v^i using(13.61) and (13.62). That is, the v^i are the vectors associated with the standard symmetric pair $\{W, I^{2n}\}$. Define associated vectors u^i in terms of the v^i by the rule

$$u^i = \alpha^{-1} v^i \tag{5.13.88}$$

where, as before, α is any $4n \times 4n$ Darboux matrix that satisfies (13.20). [Note that (13.88) is equivalent to (13.77).] For the u^i we find the relation

$$(u^i, \tilde{J}^{4n} u^j) = (\alpha^{-1} v^i, \tilde{J}^{4n} \alpha^{-1} v^j) = (v^i, (\alpha^T)^{-1} \tilde{J}^{4n} \alpha^{-1} v^j) = (v^i, J^{4n} v^j) = 0. \tag{5.13.89}$$

We see that the vectors u^i are $\tilde{J}^{4n}$ isotropic. Now construct $2n \times 2n$ matrices E and F from the first $2n$ and the last $2n$ entries in the u^i, respectively. Suppose that the matrix F turns out to be invertible. Then we can write

$$\{E, F\} \sim \{EF^{-1}, I^{2n}\}. \tag{5.13.90}$$

From the previous discussion we know that the M given by

$$M = EF^{-1}, \tag{5.13.91}$$

will be symplectic. Also, let G and H be the matrices associated with the v^i and write the matrix α^{-1} in the block form form

$$\alpha^{-1} = \begin{pmatrix} A^{\alpha^{-1}} & B^{\alpha^{-1}} \\ C^{\alpha^{-1}} & D^{\alpha^{-1}} \end{pmatrix}. \tag{5.13.92}$$

In terms of the matrices $A^{\alpha^{-1}}$ through $D^{\alpha^{-1}}$ and the matrices E through G, the relation (13.88) is equivalent to the relations

$$E = A^{\alpha^{-1}}G + B^{\alpha^{-1}}H, \tag{5.13.93}$$

$$F = C^{\alpha^{-1}}G + D^{\alpha^{-1}}H. \tag{5.13.94}$$

Therefore we have the result

$$M = EF^{-1} = (A^{\alpha^{-1}}G + B^{\alpha^{-1}}H)(C^{\alpha^{-1}}G + D^{\alpha^{-1}}H)^{-1}. \tag{5.13.95}$$

Now use the explicit forms of G and H given by (13.66) and (13.67) to rewrite (13.95) in the form

$$M = (A^{\alpha^{-1}}W + B^{\alpha^{-1}})(C^{\alpha^{-1}}W + D^{\alpha^{-1}})^{-1}. \tag{5.13.96}$$

We see, consistent with (13.86), that M is related to W by the Möbius transformation associated with α^{-1},

$$M = T_{\alpha^{-1}}(W). \tag{5.13.97}$$

Thus, M has been expressed as the Möbius transformation of a symmetric matrix. Of course, for this relation to make sense, the matrix $(C^{\alpha^{-1}}W + D^{\alpha^{-1}})$ must be invertible,

$$\det\,(C^{\alpha^{-1}}W + D^{\alpha^{-1}}) \neq 0. \tag{5.13.98}$$

Observe that (13.98) is exactly the condition for F to be invertible.

For fixed α, and hence fixed $C^{\alpha^{-1}}$ and fixed $D^{\alpha^{-1}}$, the relation (13.98) describes an open set in W space. Therefore $T_{\alpha^{-1}}$ provides a local mapping of symmetric matrices into symplectic matrices. Finally we know from the work of Subsection 11.3 that if (13.87) holds (thereby making it possible to find a symmetric W given a symplectic M), then (13.98) also holds (thereby making it possible to find a symplectic M given a symmetric W), and vice versa. That is, there is the logical equivalence

$$\det\,(C^{\alpha^{-1}}W + D^{\alpha^{-1}}) \neq 0 \Leftrightarrow \det\,(C^{\alpha}M + D^{\alpha}) \neq 0. \tag{5.13.99}$$

To very this claim, make in the first line of (11.42) the substitutions $M \to \alpha$, $U' \to W$, and $U \to M$.

We close this subsection with the observation that to find α^{-1} it is not actually necessary to carry out the inversion of a $4n \times 4n$ matrix. Instead one can use the inversion relation (13.22), which only involves matrix multiplication. Indeed, it is easily verified that its use gives the results

$$A^{\alpha^{-1}} = J^{2n}(C^{\alpha})^T, \tag{5.13.100}$$

$$B^{\alpha^{-1}} = -J^{2n}(A^{\alpha})^T, \tag{5.13.101}$$

$$C^{\alpha^{-1}} = -J^{2n}(D^{\alpha})^T, \tag{5.13.102}$$

$$D^{\alpha^{-1}} = J^{2n}(B^{\alpha})^T. \tag{5.13.103}$$

5.13.9 Completion of Tasks

5.13.9.1 Verification of Möbius Transformation Invertibility Conditions

Several uncompleted tasks remain. The first is to verify that, given a symplectic matrix M, a Darboux matrix α satisfying (13.20) can be found such that the conditions (13.87) and (13.98) are also satisfied. We have already seen, as stated in (13.99), that these conditions are logically equivalent. Now we will learn more. Actually, for notational convenience, we will find a Darboux matrix β with these desired properties.

Suppose L is a symplectic matrix near M so that we may write

$$M = LN \tag{5.13.104}$$

where N is a symplectic matrix near the identity. Inspection of the Cayley Möbius transformation T_σ given by (13.14), see also (3.11.12), shows that it is ideally suited to matrices M near the identity I. What we would like to find is a choice of β such that the Darboux Möbius transformation T_β is ideally suited to matrices near L. This is easily done using group properties. We first find a Möbius transformation that sends L to I and then follow it by a Cayley Möbius transformation. Of course, in so doing, we must ensure that the resulting β is also a Darboux transformation. Let μ be the $4n \times 4n$ matrix defined by the rule

$$\mu = \begin{pmatrix} L^{-1} & 0 \\ 0 & I^{2n} \end{pmatrix}. \tag{5.13.105}$$

Then, analogous to the relations (11.47) and (11.48), we have the result

$$T_\mu(L) = I. \tag{5.13.106}$$

Furthermore, we have the relation

$$T_\mu(M) = N. \tag{5.13.107}$$

Also we observe that μ is an element of $\tilde{Sp}(4n)$ since I is symplectic and L^{-1} is symplectic (because L is assumed to be symplectic). Therefore the β given by (13.29) will be a Darboux matrix. Its associated Möbius transformation will have the property

$$W = T_\beta(M) = T_{\sigma\mu}(M) = T_\sigma(T_\mu(M)) = T_\sigma(N) = (-JN + J)(N + I)^{-1}. \tag{5.13.108}$$

Here we have used the group property of Möbius transformations. Evidently the matrix $(N + I)$ will be invertible for N sufficiently near the identity. Indeed, all that is required is that -1 not be an eigenvalue of N. Correspondingly, $T_\beta(M)$ is well defined. Finally we see from (13.11), (13.29), and (13.105) that β has the explicit form

$$\beta = (1/\sqrt{2}) \begin{pmatrix} -J^{2n}L^{-1} & J^{2n} \\ L^{-1} & I^{2n} \end{pmatrix}. \tag{5.13.109}$$

Therefore, if we write out $T_\beta(M)$ explicitly, we find the result

$$W = T_\beta(M) = (A^\beta M + B^\beta)(C^\beta M + D^\beta)^{-1} \tag{5.13.110}$$

with

$$(A^\beta M + B^\beta) = (1/\sqrt{2})(-J^{2n}L^{-1}M + J^{2n} = (1/\sqrt{2})(-JN + J), \tag{5.13.111}$$

and

$$(C^\beta M + D^\beta) = (1/\sqrt{2})(L^{-1}M + I^{2n}) = (1/\sqrt{2})(N + I). \tag{5.13.112}$$

We see that

$$\det(C^\beta M + D^\beta) \neq 0 \tag{5.13.113}$$

provided N is sufficiently near I.

The result inverse to (13.108) is given by the relation

$$\begin{aligned} M &= T_{\beta^{-1}}(W) = T_{(\sigma\mu)^{-1}}(W) = T_{\mu^{-1}\sigma^{-1}}(W) \\ &= T_{\mu^{-1}}(T_{\sigma^{-1}}(W)) = T_{\mu^{-1}}(N) = LN. \end{aligned} \tag{5.13.114}$$

Here we have again used the group property of Möbius transformations and the fact that, consistent with (13.108), there is the relation

$$N = T_{\sigma^{-1}}(W) = (JW + I)(-JW + I)^{-1}. \tag{5.13.115}$$

Note that (13.115) is well defined provided

$$\det(-JW + I) \neq 0. \tag{5.13.116}$$

This condition is met for W sufficiently near 0. The matrix W will, in turn, be near 0 if N is sufficiently near I. See (13.108). We can also evaluate $T_{\beta^{-1}}(W)$ directly. For β^{-1} we find the result

$$\beta^{-1} = (1/\sqrt{2}) \begin{pmatrix} LJ^{2n} & L \\ -J^{2n} & I^{2n} \end{pmatrix}. \tag{5.13.117}$$

See Exercise 13.4. Consequently, we obtain the result

$$M = T_{\beta^{-1}}(W) = (LJW + L)(-JW + I)^{-1}. \tag{5.13.118}$$

Again we see that the condition (13.116) arises and is satisfied.

In summary, for an arbitrary symplectic matrix L, and for $M = L$ or M in the neighborhood of L, the Möbius transformations $W = T_\beta(M)$ and $M = T_{\beta^{-1}}(W)$ are well defined when β is the Darboux transformation defined by (13.11), (13.29), and (13.105).

5.13.9.2 Verification of Full Family of Darboux Transformations

The second task is to verify that, given a symplectic matrix M, we have the full advertised $2n(4n + 1)$ degrees of freedom in the choice of α. Suppose we hold M fixed and replace α by the β given by (13.26). Write (13.86) more explicitly as

$$W(\alpha) = T_\alpha(M) \tag{5.13.119}$$

to indicate that W depends on α as well as on M. Then we have the relation

$$W(\beta) = T_\beta(M) = T_{\gamma\alpha}(M) = T_\gamma(T_\alpha(M)) = T_\gamma(W(\alpha)). \tag{5.13.120}$$

Note that T_γ is a Möbius transformation associated with a symplectic transformation since γ is assumed to be in $Sp(4n)$. From the work of Subsection 12.7 and earlier we know that such Möbius transformations send symmetric matrices into symmetric matrices. Therefore the matrix $W(\beta)$ will also be symmetric. Of course we must again worry about the inversion of the matrix occurring in the second factor of the Möbius transformation. In the case where T_γ is applied to $W(\alpha)$ this matrix will be $[C^\gamma W(\alpha) + D^\gamma]$. It is easy to check that, for γ sufficiently near the identity, the matrix C^γ is small and the matrix D^γ is near the identity. Thus the required inverse exists, and we get a full $2n(4n+1)$ parameter family of Möbius transformations T_β that send the symplectic matrix M to the symmetric matrix $W(\beta)$.

To finish this aspect of our discussion, we should also explore what happens in the map (13.97) when W is held fixed, and α is varied. Again we will replace α by β with β given by (13.26). Then we may view M as depending on β and write

$$M(\beta) = T_{\beta^{-1}}(W) = T_{(\gamma\alpha)^{-1}}(W) = T_{\alpha^{-1}\gamma^{-1}}(W) = T_{\alpha^{-1}}(T_{\gamma^{-1}}(W)) = T_{\alpha^{-1}}(W') \quad (5.13.121)$$

where

$$W' = T_{\gamma^{-1}}(W). \quad (5.13.122)$$

Moreover we know that W' will symmetric because W is symmetric, γ^{-1} is in $Sp(4n)$, and Möbius transformations corresponding to symplectic matrices send symmetric matrices into symmetric matrices. Again see Subsection 12.7. For γ near the identity W' will be near W and consequently, by the argument of the previous paragraph, $M(\beta)$ will be well defined and symplectic. To verify this claim, examine W'. Writing out (13.122) in detail gives the result

$$W' = (A^{\gamma^{-1}}W + B^{\gamma^{-1}})(C^{\gamma^{-1}}W + D^{\gamma^{-1}})^{-1}. \quad (5.13.123)$$

For γ near the identity, The matrices $A^{\gamma^{-1}}$ and $D^{\gamma^{-1}}$ will be near the identity, and the matrices $B^{\gamma^{-1}}$ and $C^{\gamma^{-1}}$ will be small. Therefore W' will be well defined, and indeed will be near W. Thus we get a full $2n(4n+1)$ parameter family of Möbius transformations $T_{\beta^{-1}}$ that send the symmetric matrix W to the symplectic matrix $M(\beta)$.

5.13.9.3 Freedom in the Choice of Darboux Transformation

Finally, for some semblance of completeness, we should address the question of what freedom exists in the choice of α for the relations (13.86) and (13.97) when both M and W are held fixed. Suppose we require that

$$M(\beta) = M(\alpha) \quad (5.13.124)$$

or, equivalently,

$$T_{\beta^{-1}}(W) = T_{\alpha^{-1}}(W). \quad (5.13.125)$$

Then we conclude that

$$W = T_I(W) = T_\beta(T_{\beta^{-1}}(W)) = T_\beta(T_{\alpha^{-1}}(W)) = T_{\beta\alpha^{-1}}(W) = T_\gamma(W). \quad (5.13.126)$$

Here we have used the definition (13.25). Upon comparing the far left and far right sides of (13.126) we see that W must be a fixed point of T_γ. From the work of Subsection 12.4.1 we know that such γ form a subgroup, the stability group of W.

Even more can be said. Suppose δ is the $Sp(4n)$ element

$$\delta = \begin{pmatrix} I^{2n} & W \\ 0^{2n} & I^{2n} \end{pmatrix}. \tag{5.13.127}$$

From the discussion of Section 3.3 we know that δ is indeed in $Sp(4n)$ because W is symmetric. Moreover, by direct calculation or from the discussion surrounding (12.43), we know that this δ has the property

$$T_\delta\left(0^{2n}\right) = W. \tag{5.13.128}$$

Consequently, (13.126) can be rewritten in the form

$$T_\delta\left(0^{2n}\right) = T_\gamma(T_\delta\left(0^{2n}\right)) \text{ or, equivalently, } T_\gamma(T_\delta\left(0^{2n}\right)) = T_\delta\left(0^{2n}\right) \tag{5.13.129}$$

from which it follows that

$$T_{\delta^{-1}}(T_\gamma(T_\delta\left(0^{2n}\right))) = 0^{2n} \text{ or, equivalently, } T_{(\delta^{-1}\gamma\delta)}\left(0^{2n}\right) = 0^{2n}. \tag{5.13.130}$$

Let ϵ be the $Sp(4n)$ element specified by the definition

$$\epsilon = \delta^{-1}\gamma\delta \text{ or, equivalently, } \gamma = \delta\epsilon\delta^{-1}. \tag{5.13.131}$$

From (13.130) we see that ϵ must be in the stability group of the zero matrix,

$$T_\epsilon\left(0^{2n}\right) = 0^{2n}. \tag{5.13.132}$$

We have encountered this group before in Subsection 12.7. Using the notation introduced there, it is the group $H(4n, \mathbb{R})$ with dimension $(6n^2 + n)$. Combining (13.26) and (13.131) gives the relation

$$\beta = \delta\epsilon\delta^{-1}\alpha. \tag{5.13.133}$$

We conclude there is a $(6n^2 + n)$ parameter set of matrices β that satisfy, for fixed M and fixed W, the relation

$$M = T_{\beta^{-1}}(W) \tag{5.13.134}$$

and its inverse

$$W = T_\beta(M). \tag{5.13.135}$$

5.13.9.4 Explicit Construction of the Most General Darboux Transformation

To explore the implications of the relations (13.133) through (13.135) in a concrete case, let us begin by constructing a particular Darboux transformation ϕ such that its associated Möbius transformation T_ϕ sends any specified symplectic matrix L into any specified symmetric matrix V:

$$V = T_\phi(L) \tag{5.13.136}$$

and

$$L = T_{\phi^{-1}}(V). \tag{5.13.137}$$

This is easily done. Let θ be the $4n \times 4n$ matrix defined by the rule

$$\theta = \begin{pmatrix} I^{2n} & V \\ 0 & I^{2n} \end{pmatrix}. \tag{5.13.138}$$

Its associated Möbius transformation has the property

$$T_\theta(0^{2n}) = V. \tag{5.13.139}$$

See (12.3) or the V analog of (13.127) and (13.128). Moreover, θ is J^{4n} symplectic. See (3.3.9). Now define a $4n \times 4n$ matrix ϕ by the rule

$$\phi = \theta\sigma\mu. \tag{5.13.140}$$

Here σ and μ are defined by (13.11) and (13.105), respectively. Since θ is J^{4n} symplectic and μ is $\tilde{J}^{4n}$ symplectic, ϕ will be a Darboux transformation. See the discussion associated with (13.31) through (13.34). Also, by construction and the group property, we have the relation

$$T_\phi(L) = T_{\theta\sigma\mu}(L) = T_\theta(T_\sigma(T_\mu(L))) = T_\theta(T_\sigma(I^{2n})) = T_\theta(0^{2n}) = V. \tag{5.13.141}$$

Here we have used (13.106), (13.139), and the Cayley transformation property

$$T_\sigma(I^{2n}) = 0^{2n}. \tag{5.13.142}$$

Evaluation of (13.140) gives the explicit result

$$\phi = (1/\sqrt{2}) \begin{pmatrix} [-JL^{-1} + VL^{-1}] & [J+V] \\ L^{-1} & I \end{pmatrix}. \tag{5.13.143}$$

Let us continue by setting $\alpha = \phi$ in (13.133) and $W = V$ in (13.127) so that $\delta = \theta$. Then we find for β the result

$$\beta = \delta\epsilon\delta^{-1}\theta\sigma\mu = \theta\epsilon\sigma\mu. \tag{5.13.144}$$

This β will be the most general Darboux transformation such that

$$V = T_\beta(L) \tag{5.13.145}$$

and

$$L = T_{\beta^{-1}}(V). \tag{5.13.146}$$

In view of the discussion at the beginning of Section 3.3 and (3.10.20), the general ϵ in the group $H(4n,\mathbb{R})$ can be written in the form

$$\epsilon = \begin{pmatrix} A & 0 \\ 0 & (A^T)^{-1} \end{pmatrix} \begin{pmatrix} I^{2n} & 0 \\ C & I^{2n} \end{pmatrix}. \tag{5.13.147}$$

Carrying out the indicated multiplications (13.144) gives for β the explicit form

$$\beta = (1/\sqrt{2}) \begin{pmatrix} -AJL^{-1} + V(A^T)^{-1}(-CJ+I)L^{-1} & AJ + V(A^T)^{-1}(CJ+I) \\ (A^T)^{-1}(-CJ+I)L^{-1} & (A^T)^{-1}(CJ+I) \end{pmatrix}, \tag{5.13.148}$$

and for its inverse the explicit form

$$\beta^{-1} = (1/\sqrt{2}) \begin{pmatrix} LJA^{-1} - LCA^{-1} & -LJA^{-1}V + L(CA^{-1}V + A^T) \\ -JA^{-1} - CA^{-1} & JA^{-1}V + (CA^{-1}V + A^T) \end{pmatrix}. \tag{5.13.149}$$

(See Exercise 13.5.) Here J stands for J^{2n}. It is readily verified by direct calculation that (13.145) and (13.146) are satisfied. We observe, since $H(4n, \mathbb{R})$ is a $(6n^2 + n)$ dimensional group, that there is a $(6n^2+n)$ dimensional family of Darboux matrices that relate a specified L to a specified V. Note also that we have written (parameterized) the general $4n \times 4n$ Darboux matrix β in terms of a general $2n \times 2n$ symplectic matrix L, two general $2n \times 2n$ symmetric matrices C and V, and a general $GL(2n)$ matrix A. Exercise 13.8 shows that this parameterization must in fact have some redundancy because the parameter count for this parameterization exceeds the dimensionality of $sp(4n)$.

5.13.9.5 Two Convenient Simpler Choices

There are two convenient simpler choices for Darboux matrices whose associated Möbius transformations relate any specified symplectic matrix L to any specified symmetric matrix V. The first, call it $\tilde{\beta}$, is the Darboux matrix obtained by setting $A = I$ in (13.148) to yield the result

$$\tilde{\beta} = (1/\sqrt{2}) \begin{pmatrix} [-JL^{-1} + V(-CJ + I)L^{-1}] & [J + V(CJ + I)] \\ (-CJ + I)L^{-1} & (CJ + I) \end{pmatrix} \tag{5.13.150}$$

with the inverse

$$\tilde{\beta}^{-1} = (1/\sqrt{2}) \begin{pmatrix} LJ - LC & -LJV + L(CV + I) \\ -J - C & JV + (CV + I) \end{pmatrix}. \tag{5.13.151}$$

[That $\tilde{\beta}$ is a Darboux matrix follows from the fact β as given by (13.148) is a Darboux matrix for all choices of A.] It is easily verified by direct calculation that

$$V = T_{\tilde{\beta}}(L) \tag{5.13.152}$$

and

$$L = T_{\tilde{\beta}^{-1}}(V). \tag{5.13.153}$$

A still simpler choice, call it $\tilde{\tilde{\beta}}$, is the Darboux matrix obtained by setting $A = I$ and $C = 0$ in (13.148) to yield the result

$$\tilde{\tilde{\beta}} = (1/\sqrt{2}) \begin{pmatrix} [-JL^{-1} + VL^{-1}] & [J + V] \\ L^{-1} & I \end{pmatrix} = \phi \tag{5.13.154}$$

with the inverse

$$\tilde{\tilde{\beta}}^{-1} = (1/\sqrt{2}) \begin{pmatrix} LJ & -LJV + L \\ -J & JV + I \end{pmatrix}. \tag{5.13.155}$$

[Note that this choice amounts to setting $\epsilon = I$. That $\tilde{\tilde{\beta}}$ is also a Darboux matrix follows from the fact β as given by (13.148) is a Darboux matrix for all choices of A and C.] It is easily verified by direct calculation that

$$V = T_{\tilde{\tilde{\beta}}}(L) \tag{5.13.156}$$

and

$$L = T_{\tilde{\tilde{\beta}}^{-1}}(V). \tag{5.13.157}$$

Exercises

5.13.1. Verify the matrix identity

$$\begin{pmatrix} -J & J \\ I & I \end{pmatrix} \begin{pmatrix} I & I \\ 0 & I \end{pmatrix} = \begin{pmatrix} -J & 0 \\ I & 2I \end{pmatrix}. \tag{5.13.158}$$

Use this identity to prove (13.13). Verify (13.19). Use the representation (13.27) to show that all Darboux matrices, i.e. all matrices that satisfy (13.20) or (13.21), must have determinant +1.

5.13.2. Show, using the representations (13.27) and (13.29), that there are the relations

$$\mu = \sigma^{-1}\gamma\sigma \;\text{ or }\; \gamma = \sigma\mu\sigma^{-1} \tag{5.13.159}$$

which demonstrate that $\tilde{Sp}(4n)$ and $Sp(4n)$ are related by a similarity transformation.

5.13.3. Review Exercise 12.7. Verify that (13.54) is an equivalence relation.

5.13.4. Verify by direct calculation that β as given by (13.109) satisfies (13.21). Verify (13.117) both by direct calculation and by use of (13.11), (13.29), and (13.105).

5.13.5. Verify (13.143) by working out the product (13.140). Verify by direct calculation that ϕ satisfies (13.136) and (13.137). Verify that ϕ is a Darboux matrix/transformation.

5.13.6. Verify (13.148) and (13.149) using (13.144) and (13.147).

5.13.7. Verify by direct calculation that β and β^{-1} as given by (13.148) and (13.149) satisfy (13.145) and (13.146).

5.13.8. The Darboux matrix β given by (13.148) is parameterized in terms of a general $2n \times 2n$ symplectic matrix L, two general $2n \times 2n$ symmetric matrices C and V, and a general $GL(2n)$ matrix A. Verify that the dimensionality of the space of all $2n\times 2n$ symmetric matrices is $n(2n+1)$, which is also the dimensionality of $sp(2n)$. Also, the dimensionality of $GL(2n)$ is evidently $(2n)^2$. Verify that the dimension count for the parameterization (13.148) of Darboux matrices in terms of a $2n\times 2n$ symplectic matrix, two $2n\times 2n$ symmetric matrices, and a general $GL(2n)$ matrix is given by the sum

$$n(2n+1) + 2[n(2n+1)] + 4n^2 = 10n^2 + 3n. \tag{5.13.160}$$

By comparison, the dimensionality of the set of $4n \times 4n$ Darboux matrices is the same as the dimensionality of $sp(4n)$, which is given by the relation

$$\dim sp(4n) = 2n(4n+1) = 8n^2 + 2n. \tag{5.13.161}$$

Thus, the parameterization (13.148) must have some redundancy. Determine what this redundancy is in the simplest case of 4×4 Darboux matrices. Verify that the number of parameters in $\tilde{\beta}$ as given by (13.150) is $6n^2 + 3n$. Verify that the number of parameters in $\tilde{\tilde{\beta}}$ as given by (13.154) is $4n^2 + 2n$..

5.13.9. Suppose N is a J^{2n} symplectic matrix and μ is a $\tilde{J}^{4n}$ symplectic matrix. Show, using (13.159), that M given by

$$M = T_\mu(N) \tag{5.13.162}$$

is also a J^{2n} symplectic matrix.

5.13.10. Verify that the relation (12.80) between the guhp and the gud can be rewritten in the form

$$-W = (iZ+I)(-iZ+I)^{-1} = T_\phi(Z) \tag{5.13.163}$$

where

$$\phi = \begin{pmatrix} iI & I \\ -iI & I \end{pmatrix}. \tag{5.13.164}$$

Verify that the Cayley relation (3.11.5) between symmetric and symplectic matrices can be written in the form

$$M = [(-J)(-W)+I][(J)(-W)+I]^{-1} = T_\psi(-W) \tag{5.13.165}$$

where

$$\psi = \begin{pmatrix} -J & I \\ J & I \end{pmatrix}. \tag{5.13.166}$$

Relate ϕ and ψ. Hint: Review Exercise 3.2.6.

5.13.11. Let ν be the matrix defined by the relation

$$\nu = (1/\sqrt{2})\begin{pmatrix} -I & I \\ I & I \end{pmatrix}. \tag{5.13.167}$$

Verify that the Cayley relation (3.11.5) between a symplectic matrix M and a Hamiltonian matrix JW, a matrix in the symplectic Lie algebra, can be written in the form

$$M = T_\nu(-JW). \tag{5.13.168}$$

Verify that ν has the property

$$\nu^2 = I, \tag{5.13.169}$$

from which it follows that T_ν is an *involution.*[20] That is, by the composition law (11.6), there is the relation

$$T_\nu T_\nu = T_{\nu^2} = T_I = I. \tag{5.13.170}$$

[20] In this context, an involution is a map whose square is the identity map.

Consequently, show that the relation (13.168) has the inverse relation

$$-JW = T_\nu(M), \tag{5.13.171}$$

in agreement with (3.11.12). We have learned that T_ν provides a map between the *group* $Sp(2n, \mathbb{R})$ and its *Lie algebra* $sp(2n, \mathbb{R})$. Show that it does the same for $Sp(2n, \mathbb{C})$ and $sp(2n, \mathbb{C})$.

Show that T_ν has analogous properties for the orthogonal and unitary (but not special unitary) groups. For example, if A is antisymmetric, show that $M = T_\nu(-A)$ is orthogonal, etc.

5.14 Uniqueness of Cayley Möbius Transformation

The Cayley Möbius transformation has three properties that make it essentially unique. The first is that

$$T_\sigma(M^{-1}) = -T_\sigma(M). \tag{5.14.1}$$

The second, consistent with the first, is that

$$T_\sigma(I) = 0. \tag{5.14.2}$$

The third is the relation

$$JT_\sigma(N^{-1}MN) = N^{-1}JT_\sigma(M)N, \tag{5.14.3}$$

from which it follows that

$$T_\sigma(N^{-1}MN) = -JN^{-1}JT_\sigma(M)N. \tag{5.14.4}$$

Here $J = J^{2n}$ and N is any invertible matrix. Now suppose that N is symplectic. From the symplectic condition written in the form

$$NJN^T = J \tag{5.14.5}$$

we infer the relation

$$-JN^{-1}J = N^T, \tag{5.14.6}$$

so that we also have for symplectic N the result

$$T_\sigma(N^{-1}MN) = N^T T_\sigma(M)N. \tag{5.14.7}$$

These properties are easily shown to follow from the form of σ as given in (13.11), and lead to the inversion and symplectic similarity invariance properties of Cayley matrix symplectification described by (4.7.14) and (4.7.15). They will also be important for the work of Chapter 34 on Optimal Evaluation of Symplectic Maps.

We now verify that essentially only the Cayley Möbius transformation, among all Darboux Mobius transformations, has these properties. To begin suppose, in analogy with (14.1), we seek Darboux Mobius transformations β with the property

$$T_\beta(M^{-1}) = -T_\beta(M). \tag{5.14.8}$$

Also assume that $T_\beta(I)$ is well defined. Then it follows from (14.8) that there is the condition

$$T_\beta(I) = 0, \tag{5.14.9}$$

and hence $L = I$ and $V = 0$ in (13.145) so that β as given by (13.148) takes the form

$$\beta = (1/\sqrt{2}) \begin{pmatrix} -AJ & AJ \\ (A^T)^{-1}(-CJ + I) & (A^T)^{-1}(CJ + I) \end{pmatrix}. \tag{5.14.10}$$

Correspondingly $\tilde{\beta}$, which is obtained from (14.10) by setting $A = I$, is given by the relation

$$\tilde{\beta} = (1/\sqrt{2}) \begin{pmatrix} -J & J \\ (-CJ + I) & (CJ + I) \end{pmatrix}. \tag{5.14.11}$$

So far, we have actually only used the fact that (14.9) is a consequence of (14.8) to put restrictions on β. Now let us make further direct use of (14.8). To do so it is useful to compute $T_{\tilde{\beta}}(M)$ and $T_\beta(M)$. Begin by computing $T_{\tilde{\beta}}(M)$. From (14.11) we find the result

$$T_{\tilde{\beta}}(M) = [-JM + J][(-CJ + I)M + (CJ + I)]^{-1}. \tag{5.14.12}$$

We will also need the result

$$T_{\tilde{\beta}}(M^{-1}) = [-JM^{-1} + J][(-CJ + I)M^{-1} + (CJ + I)]^{-1} \tag{5.14.13}$$

which follows from (14.12) upon replacing M by M^{-1}. Next compute $T_\beta(M)$ using (14.10) and manipulate the result to find the relation

$$\begin{aligned} T_\beta(M) &= \{-AJM + AJ\}\{(A^T)^{-1}(-CJ + I)M + (A^T)^{-1}(CJ + I)\}^{-1} \\ &= A\{-JM + J\}\{(A^T)^{-1}[(-CJ + I)M + (CJ + I)]\}^{-1} \\ &= A\{-JM + J\}\{(-CJ + I)M + (CJ + I)\}^{-1}A^T \\ &= A[T_{\tilde{\beta}}(M)]A^T. \end{aligned} \tag{5.14.14}$$

It follows from (14.14) that

$$T_{\tilde{\beta}}(M) = A^{-1}[T_\beta(M)](A^T)^{-1}. \tag{5.14.15}$$

Similarly, there is the result

$$T_{\tilde{\beta}}(M^{-1}) = A^{-1}[T_\beta(M^{-1})](A^T)^{-1}. \tag{5.14.16}$$

Now add (14.15) and (14.16) to obtain the relation

$$T_{\tilde{\beta}}(M) + T_{\tilde{\beta}}(M^{-1}) = A^{-1}[T_\beta(M) + T_\beta(M^{-1})](A^T)^{-1}. \tag{5.14.17}$$

Upon making use of (14.8) in (14.17) we find the result

$$T_{\tilde{\beta}}(M) + T_{\tilde{\beta}}(M^{-1}) = 0 \text{ or, equivalently, } T_{\tilde{\beta}}(M^{-1}) = -T_{\tilde{\beta}}(M). \tag{5.14.18}$$

We are almost done with this part of the argument. In (14.13) multiply both the numerator and denominator on the right by M to obtain the result

$$T_{\tilde{\beta}}(M^{-1}) = [JM - J][(CJ + I)M + (-CJ + I)]^{-1}. \tag{5.14.19}$$

Upon employing (14.12) and (14.19) in the second version of (14.18) we now find the result

$$[JM - J][(CJ + I)M + (-CJ + I)]^{-1} = [JM - J][(-CJ + I)M + (CJ + I)]^{-1}, \tag{5.14.20}$$

from which it follows that

$$(CJ + I)M + (-CJ + I) = (-CJ + I)M + (CJ + I). \tag{5.14.21}$$

(Here we have assumed $M \neq I$.) Finally, upon canceling like terms on the left and right sides of (14.21), we find the relation

$$CJM - CJ = -CJM + CJ \text{ or, equivalently } 2CJ(M - I) = 0 \tag{5.14.22}$$

from which it follows that

$$C = 0. \tag{5.14.23}$$

Employing (14.23) in (14.11) gives the result

$$\tilde{\beta} = (1/\sqrt{2})\begin{pmatrix} -J & J \\ I & I \end{pmatrix} = \sigma. \tag{5.14.24}$$

Correspondingly (14.14) can be rewritten as

$$T_\beta(M) = A[T_\sigma(M)]A^T. \tag{5.14.25}$$

In analogy to (14.7) let us now invoke the further requirement that

$$T_\beta(N^{-1}MN) = N^T T_\beta(M)N. \tag{5.14.26}$$

With the aid of (14.25) and (14.7) we find for the left side of (14.26) the result

$$T_\beta(N^{-1}MN) = A[T_\sigma(N^{-1}MN)]A^T = AN^T[T_\sigma(M)]NA^T. \tag{5.14.27}$$

For the right side of (14.26), again using (14.25)), we find the result

$$N^T T_\beta(M)N = N^T A T_\sigma(M) A^T N. \tag{5.14.28}$$

Therefore (14.26) is equivalent to the condition

$$AN^T[T_\sigma(M)]NA^T = N^T A[T_\sigma(M)]A^T N. \tag{5.14.29}$$

In order for (14.29) to hold for all matrices M, there must be the relation

$$AN^T = N^T A. \tag{5.14.30}$$

That is, N^T and A must commute. See Exercise 15.3.

Moreover, since N is an arbitrary symplectic matrix, N^T is also an arbitrary symplectic matrix. It can be shown that a matrix A that commutes with all symplectic matrices must be a multiple of the identity. See Exercise 21.14.1. Therefore, the requirement (14.26) yields the conclusion that A must be of the form

$$A = \lambda I. \tag{5.14.31}$$

Correspondingly, β takes the form

$$\beta = (1/\sqrt{2}) \begin{pmatrix} -\lambda J & \lambda J \\ (1/\lambda)I & (1/\lambda)I \end{pmatrix}. \tag{5.14.32}$$

We see that, apart from a scaling factor λ, the matrix β is essentially σ. If the scaling factor is set to one, β becomes σ. Another possible choice is to set $\lambda = \sqrt{2}$. When this is done, β takes the rational form

$$\beta = \begin{pmatrix} -J & J \\ I/2 & I/2 \end{pmatrix}. \tag{5.14.33}$$

However, for this λ choice β is not orthogonal, and the simplicity of the orthogonality feature possessed by σ is lost.

Exercises

5.14.1. Verify the relations (14.1) through (14.7) for the Cayley Möbius transformation.

5.14.2. Verify that (14.21) follows from (14.20).

5.14.3. The purpose of this exercise is to verify that (14.29) implies (14.30). Define matrices X and Y by the relations

$$X = AN^T, \tag{5.14.34}$$

$$Y = N^T A. \tag{5.14.35}$$

Verify that with these definitions (14.29) can be rewritten in the form

$$X[T_\sigma(M)]X^T = Y[T_\sigma(M)]Y^T. \tag{5.14.36}$$

Next show that, for some finite (but perhaps small) number δ and *any* symmetric matrix S, there is a symplectic M such that

$$T_\sigma(M) = \delta S. \tag{5.14.37}$$

Thus, show that (14.36) is equivalent to the relation

$$XSX^T = YSY^T. \tag{5.14.38}$$

Define a matrix Λ by the equation

$$Y = X\Lambda. \tag{5.14.39}$$

Verify, because A and N are invertible, that this equation actually defines Λ. Using this definition, show that (14.38) is equivalent to the relation

$$\Lambda S \Lambda^T = S. \tag{5.14.40}$$

If we put $S = I$, which is a possibility, we conclude that Λ must satisfy the relation

$$\Lambda \Lambda^T = I, \tag{5.14.41}$$

and therefore Λ is orthogonal. Verify that (14.40) and (14.41) entail the relation

$$\Lambda S = S \Lambda. \tag{5.14.42}$$

Show that (14.41), together with (14.42) holding for all symmetric matrices S, requires that

$$\Lambda = \pm I. \tag{5.14.43}$$

Hint: First show that Λ must be diagonal by considering the cases for which S is diagonal and has only one nonzero entry. Next show that all diagonal entries in Λ must be equal by considering the cases for which S has only two nonzero entries located at symmetric places above and below the diagonal. Finally, use (14.41).

Show that taking the minus sign in (14.43) leads to the condition $AN^T = -N^T A$, and that if this condition holds for all N as it must, it also holds for $N = I$ leading to the conclusion $A = 0$, which is not possible since A is assumed to be nonsingular. Show that taking the plus sign in (14.43) yields the advertised relation (14.30).

5.14.4. Review Exercise 3.11.4. Suppose (13.86) is rewritten in the form

$$W = J(-J)T_\alpha(M) \tag{5.14.44}$$

and we define $g(M)$ by the rule

$$g(M) = -JT_\alpha(M). \tag{5.14.45}$$

Then we have the relation

$$W = Jg(M). \tag{5.14.46}$$

Define a quadratic form $Q_\alpha(z)$ by the rule

$$Q_\alpha(z) = (z, Wz). \tag{5.14.47}$$

Exercise 3.11.4 showed that $Q_\alpha(z)$ is invariant when $\alpha = \sigma$. Show that many other choices of α do not yield a Q_α that has this property. Are there any other choices that do?

5.15 Matrix Symplectification Revisited

Section 4.7 described the use of the Cayley representation to carry out matrix symplectification, and Section 4.8 described the use of generating functions for the same purpose. As pointed out earlier, both procedures are examples of the use of Möbius transformations. Moreover, it was also remarked that there were cases for which the Cayley representation could not be used, and cases for which none of the generating functions F_1 through F_4 could be used. The purpose of this section is to show that, given any nearly symplectic matrix M, there is a symplectification procedure employing Möbius transformations that will succeed. (Subsequently, Exercise 6.7.1 shows that there is an associated quadratic generating function that produces any such Möbius transformation.)

Suppose M is a matrix that is nearly symplectic. Let β be some appropriate Darboux matrix. Use it to define a matrix U in terms of M by the rule

$$U = T_\beta(M). \tag{5.15.1}$$

Since M is nearly symplectic, and by the properties of Darboux Möbius transformations, U will be nearly symmetric. Define a matrix W in terms of U by the rule

$$W = (U + U^T)/2. \tag{5.15.2}$$

Since U is nearly symmetric, W will be near U. Finally, define a matrix R in terms of W by the rule

$$R = T_{\beta^{-1}}(W). \tag{5.15.3}$$

Since W is symmetric by construction, and by the properties of Darboux Möbius transformations, R will be symplectic. Moreover, because W is near U, R will be near M. Note also that $R = M$ if M is symplectic. Therefore R may be viewed as a symplectification of M.

We still have to demonstrate that there is a choice of β such that (15.1) and (15.3) are well defined. Suppose we write M in the form

$$M = LN \tag{5.15.4}$$

where L is symplectic and N is a matrix in the vicinity (in a sense to be made more precise shortly) of the identity. Use for β the Darboux matrix given by (13.148) with the same L that appears in (15.4). From (15.1) we then find for U a result of the form

$$U = T_\beta(M) = \{\text{Numerator}\} \times \{\text{Denominator}\}^{-1} \tag{5.15.5}$$

where the denominator is given by the equation

$$\text{Denominator} = (A^T)^{-1}\{[(-CJ + I)L^{-1}]M + (CJ + I)\}. \tag{5.15.6}$$

Inserting the representation (15.4) for M into (15.6) gives the result

$$\text{Denominator} = (A^T)^{-1}\{[(-CJ + I)N + (CJ + I)\}, \tag{5.15.7}$$

which can be rewritten in the form

$$\begin{aligned}\text{Denominator} &= (A^T)^{-1}\{[(-CJ+I)(N-I)+2I)\}\\ &= 2(A^T)^{-1}\{[I+(1/2)(-CJ+I)(N-I)]\}.\end{aligned} \tag{5.15.8}$$

Compute the determinant of the denominator to find the result

$$\det\,(\text{Denominator}) = \det\,[2(A^T)^{-1}] \times \det\,\{[I+(1/2)(-CJ+I)(N-I)]\}. \tag{5.15.9}$$

We see from the second factor in (15.9) that the determinant of the denominator cannot vanish as long as N is reasonably near I. Correspondingly we conclude that there is a choice of β such that, for any given M that is nearly symplectic, (15.1) is well defined for this M and for all matrices near this M. Conversely, from the work of Section 11, we know that $T_{\beta^{-1}}(U)$ will then be well defined for the U given by (15.1) and for all matrices near this U. We have already seen that W is near U. Therefore R as given by (15.3) is well defined. The symplectification procedure has succeeded.

The last item to be considered in this section is the extent to which the symplectification procedure given by (15.1) through (15.3) depends on the choice of the Darboux transformation β. We might suspect some redundancy because the procedure (15.1) through (15.3) involves the use of both β and β^{-1}, and therefore there is some possibility for compensation or cancellation.

To explore this question, we will need to study the properties of β as given by (13.148) in some more detail. As the result of some preliminary monkeying around, we observe that the terms in the two upper blocks of β can be rewritten in the form

$$-AJL^{-1} + V(A^T)^{-1}(-CJ+I)L^{-1} = A[-JL^{-1} + A^{-1}V(A^T)^{-1}(-CJ+I)L^{-1}], \tag{5.15.10}$$

$$AJ + V(A^T)^{-1}(CJ+I) = A[J + A^{-1}V(A^T)^{-1}(CJ+I)]. \tag{5.15.11}$$

Define a new matrix $\acute{V}$ by the rule

$$\acute{V} = A^{-1}V(A^{-1})^T. \tag{5.15.12}$$

The matrix $\acute{V}$ will also be symmetric because V is symmetric. Moreover, since A is assumed invertible and V is an arbitrary symmetric matrix, the matrix $\acute{V}$ may be taken to be an arbitrary symmetric matrix. With this definition, (15.10) and (15.11) can be written in the more compact forms

$$-AJL^{-1} + V(A^T)^{-1}(-CJ+I)L^{-1} = A[-JL^{-1} + \acute{V}(-CJ+I)L^{-1}], \tag{5.15.13}$$

$$AJ + V(A^T)^{-1}(CJ+I) = A[J + \acute{V}(CJ+I)]. \tag{5.15.14}$$

Now β can be expressed in the form

$$\beta = (1/\sqrt{2})\begin{pmatrix} A[-JL^{-1} + \acute{V}(-CJ+I)L^{-1}] & A[J + \acute{V}(CJ+I)] \\ (A^T)^{-1}(-CJ+I)L^{-1} & (A^T)^{-1}(CJ+I) \end{pmatrix}. \tag{5.15.15}$$

When the form for β given by (15.15) is used to compute $T_\beta(M)$, we find the result (15.5) with the denominator given by (15.6) and the numerator given by

$$\text{Numerator} = A\{[-JL^{-1} + \acute{V}(-CJ+I)L^{-1}]M + [J + \acute{V}(CJ+I)]\}. \tag{5.15.16}$$

Note that the numerator has a common factor of A and, as (15.6) shows, the denominator has a common factor of $(A^T)^{-1}$. Therefore we have the *identity* that $T_\beta(M)$ for any M can be written in the factored form

$$T_\beta(M) = A[T_{\tilde\beta}(M)]A^T. \tag{5.15.17}$$

Here $\tilde\beta$ is a Darboux matrix defined in terms of β by writing

$$\tilde\beta = (1/\sqrt{2})\begin{pmatrix} [-JL^{-1} + \acute{V}(-CJ + I)L^{-1}] & [J + \acute{V}(CJ + I)] \\ (-CJ + I)L^{-1} & (CJ + I) \end{pmatrix}. \tag{5.15.18}$$

[That $\tilde\beta$ is a Darboux matrix follows from the fact β as given by (15.15) is a Darboux matrix for all choices of A, and $\tilde\beta$ is simply the result of putting $A = I$ in (15.15).]

We note in passing that $\tilde\beta$ has the property

$$T_{\tilde\beta}(L) = \acute{V}. \tag{5.15.19}$$

That is, $T_{\tilde\beta}$ sends the arbitrary symplectic matrix L to the arbitrary symmetric matrix $\acute{V}$.

We will now learn that for fixed L, $\acute{V}$, and C in (15.15), the symplectification procedure given by (15.1) through (15.3) employing the β of (15.15) yields a result that is *independent* of the choice of the matrix A. To see this, suppose that the matrix $\tilde\beta$ is used to symplectify M using a procedure analogous to (15.1) through (15.3). As just pointed out, this amounts to setting $A = I$ in (15.15). Then we find the results

$$\tilde U = T_{\tilde\beta}(M), \tag{5.15.20}$$

$$\tilde W = (\tilde U + \tilde U^T)/2, \tag{5.15.21}$$

$$\tilde R = T_{\tilde\beta^{-1}}(\tilde W). \tag{5.15.22}$$

Now carry out some manipulations using previous results. From (15.1), (15.17), and (15.20) it follows that

$$U = T_\beta(M) = A[T_{\tilde\beta}(M)]A^T = A\tilde U A^T. \tag{5.15.23}$$

Consequently we have the relations

$$U^T = A\tilde U^T A^T, \tag{5.15.24}$$

$$W = (U + U^T)/2 = A[(\tilde U + \tilde U^T)/2]A^T = A\tilde W A^T. \tag{5.15.25}$$

But we also have, from (15.3) and application of the identity (15.17), the result

$$W = T_\beta(R) = A[T_{\tilde\beta}(R)]A^T. \tag{5.15.26}$$

Upon comparing (15.25) and (15.26) we conclude there is the relation

$$T_{\tilde\beta}(R) = \tilde W, \tag{5.15.27}$$

and therefore there is also the inverse relation

$$R = T_{\tilde\beta^{-1}}(\tilde W). \tag{5.15.28}$$

Finally (15.22) and (15.28) taken together show that

$$R = \tilde{R}. \tag{5.15.29}$$

Thus R is indeed independent of the choice of A. Since the first factor in (13.147), the factor that involves A, produces an *arbitrary* linear transformation on the coordinate-space variables, we may say that the Darboux Möbius symplectification procedure is *invariant* under linear transformations of the coordinate-space variables.

Exercises

5.15.1. Verify (15.1) through (15.9).

5.15.2. Verify (15.10) through (15.18).

5.15.3. Verify (15.20) through (15.29).

5.15.4. By studying various examples, explore how the choice of L, $\acute{V}$, and C in (15.15) affects the outcome of the symplectification procedure. Study, for example, the use of $\tilde{\beta}$ to symplectify matrices of the form λI where λ is a parameter near 1.

Bibliography

Group Theory, $U(3)$, and $SU(3)$

[1] H. Weyl, *The Classical Groups: Their Invariants and Representations*, Princeton University Press (1946).

[2] H. Georgi, *Lie Algebras in Particle Physics*, Perseus Books (1999).

[3] A. Zee, *Group Theory in a Nutshell for Physicists*, Princeton University Press (2016).

[4] M. Gell-Mann and Y. Ne'eman, *The Eightfold Way*, pp. 49-50, Benjamin (1964).

[5] R.E. Behrends et al., *Rev. of Mod. Phys.* **34**, p. 1 (1962).

[6] S. Gasiorowicz, *Elementary Particle Physics*, p. 257, John Wiley (1966).

[7] S. Gasiorowicz, "A Simple Graphical Method in the Analysis of SU(3)", Argonne National Laboratory Report, ANL-6729 (1963).

[8] S. Coleman, *Aspects of Symmetry*, Cambridge University Press (1985).

[9] W. Greiner and B. Muller, *Quantum Mechanics—Symmetries*, Springer-Verlag (1994).

[10] A. Dragt, "Classification of Three-Particle States According to $SU(3)$", *Journal of Mathematical Physics* **6**, 533 (1965).

[11] P. J. Olver, *Equivalence, Invariants, and Symmetry*, Cambridge University Press (1995).

Topology of $Sp(2, \mathbb{R})$ and $Sp(2n, \mathbb{R})$

[12] M. Levi, "Stability of the Inverted Pendulum - a Topological Explanation", *SIAM Review* **30** 639 (1988).

[13] A. Abbondandolo, *Morse theory for Hamiltonian systems*, Chapman & Hall/CRC (2001).

Metaplectic Group

[14] G.B. Folland, *Harmonic Analysis in Phase space*, Annals of Mathematics Studies Number 122, Princeton University Press (1989).

Quaternions, Octonions, and Lie Groups

[15] J.B. Kuipers, *Quaternions and Rotation Sequences: A Primer with Application to Orbits, Aerospace and Virtual Reality*, Princeton University Press (2002).

[16] S. Altmann, *Rotations, Quaternions, and Double Groups*, Dover (2005).

[17] B.L. van der Waerden, *Modern Algebra, Volume I*, Frederick Ungar (1953).

[18] I.L. Kantor and A.S. Solodovnik, *Hypercomplex Numbers*, Springer-Verlag (1989).

[19] C. Chevalley, *Theory of Lie Groups*, Princeton University Press (1946). There is also a Dover reprint/edition (2018).

[20] C. Chevalley, *The Algebraic Theory of Spinors and Clifford Algebras: Collected Works of Claude Chevalley (v. 2)*, Springer (1996).

[21] H.-D. Ebbinghaus et al., *Numbers*, Springer Verlag (1991).

[22] J.C. Baez, "The octonions", *Bull. American Mathematical Society* **39**, p. 145 (2002) and **42**, p. 213 (2005).

[23] J.H. Conway and D.A. Smith, *On Quaternions and Octonions: Their Geometry, Arithmetic, and Symmetry*, A.K. Peters (2003).

[24] J.C. Baez, *My favorite Numbers: 8*, The Rankin Lectures (2008), http://theoryoforder.com/img/8.pdf.

[25] J.C. Baez, *Bull. Amer. Math. Soc.* **42**, p. 229, (2005). Also available at http://math.ucr.edu/home/baez/octonions/conway_smith/.

[26] I. Porteous, *Clifford Algebras and the Classical Groups*, Cambridge (1995).

[27] P. Lounesto, *Clifford Algebras and Spinors*, 2nd ed., Cambridge (2001).

[28] F. Reese Harvey, *Spinors and Calibrations*, Academic Press (1990).

[29] P. Cvitanović, *Group Theory: Birdtracks, Lie's, and Exceptional Groups*, Princeton University Press (2008).

[30] D. Hestenes, *Space-Time Algebra*, second edition, Birkhäuser (2015).

Group Theory, Möbius Transformations, Theta Functions, Etc.

(See also the Lie Group Theory sections of the Bibliographies for Chapters 3 and 27.)

[31] H. Bateman, A. Erdelyi, W. Magnus, F. Oberhettinger, F.G. Tricomi, *Higher Transcendental Functions*, Vol. III, Chapter 14: Automorphic Functions, McGraw-Hill (1955).

[32] C.L. Siegel, *Symplectic Geometry*, Academic Press (New York, 1964).

[33] C.L. Siegel, *Topics in Complex Function Theory, Vols. I-III*, Wiley-Interscience (New York, 1971).

[34] L.K. Hua, “On the theory of automorphic functions of a matrix variable I,II”, *Amer. J. Math.* **66**, 470-488, 531-563 (1944).

[35] Feng Kang, Wu Hua-mo, Qin Meng-shao, and Wang Dao-liu, “Construction of Canonical Difference Schemes for Hamiltonian Formalism via Generating Functions”, *Journal of Computational Mathematics* **11**, p. 71 (1989).

[36] Feng Kang, “The Calculus of Generating Functions and the Formal Energy for Hamiltonian Algorithms”, *Journal of Computational Mathematics* **16**, p. 481 (1998).

[37] M. Eichler, *Introduction to the Theory of Algebraic Numbers and Functions*, Academic Press (1966).

[38] J. Lehner, *Discontinuous Groups and Automorphic Functions, Mathematical Surveys and Monographs # 8*, American Mathematical Society (1964).

[39] D. Mumford, *Tata Lectures on Theta I-III*, Birkhäuser (1984).

[40] D. Mumford, C. Series, and D. Wright, *Indra's Pearls: The Vision of Felix Klein*, Cambridge University Press (2002).

[41] R.E. Bellman, *A Brief Introduction to Theta Functions*, Holt, Rinehart, and Winston (1961).

[42] I.M. Gel'fand, M.I. Graev, and I.I. Pyatetskii-Shapiro, *Representation Theory and Automorphic Functions*, W.B. Saunders Co. (1969).

[43] N. Koblitz, *Introduction to Elliptic Curves and Modular Forms*, (Springer-Verlag 1984).

[44] S. Lang, *SL(2,R)*, Springer-Verlag (1985).

[45] R. Howe and E.C. Tan, *Non-Abelian Harmonic Analysis*, Springer-Verlag (1992).

[46] A. Terras, *Harmonic Analysis on Symmetric Spaces and Applications I and II*, Springer-Verlag (1985 and 1988).

[47] A. Perelomov, *Generalized Coherent States and Their Applications*, Springer-Verlag (1986).

[48] S. Helgason, *Differential Geometry, Lie Groups, and Symmetric Spaces*, Academic Press (1978).

[49] S. Helgason, *Groups and Geometric Analysis: Integral Geometry, Invariant Differential Operators, and Spherical Functions*, Second Edition, American Mathematical Society (2002).

[50] S. Helgason, *Geometric Analysis on Symmetric Spaces*, American Mathematical Society (2008).

[51] L. Ahlfors, "Clifford Numbers and Möbius Transformations in R^n", published in *Clifford Algebras and their Applications in Mathematical Physics*, J. Chisholm and A. Common, Edit., Proceedings of NATO and SERC Workshop, Canterbury, Kent, 1985, NATO ASI Series (Reidel 1986).

[52] J. Gray, *Linear Differential Equations and Group Theory from Riemann to Poincaré*, Birkhäuser (1986).

Lorentz Group and Laser Optics

[53] I.M. Gel'fand, R.A. Minlos, and Z.Y. Shapiro, *Representations of the rotation and Lorentz groups and their applications*, Pergamon Press and Macmillan Co. (New York, 1963).

[54] V. Bargmann, "Irreducible Unitary Representations of the Lorentz Group", *Annals of Math.* **48**, no. 3, p. 568 (1947).

[55] A. Yariv, *Quantum Electronics*, John Wiley (New York, 1989); *Optical Electronics*, Holt, Rinehart, and Winston (1985).

[56] A.J. Dragt, "Lie Algebraic Methods for Ray and Wave Optics" (University of Maryland, 1995).

Lie Series and Lie Transformations

[57] W. Gröbner, *Die Lie-Reihen and Ihre Anwendungen*, Deutscher Verlag der Wissenschaften (Berlin 1960).

[58] W. Gröbner and H. Knapp, *Contributions to the Methods of Lie Series*, Bibliographisches Institut, (Manheim 1967).

[59] G. Hori, "Theory of general perturbations with unspecified canonical variables", *Publications of the Astronomical Society of Japan* **18**, p. 287 (1966).

[60] A. Deprit, *Celest. Mech.* **1**, 12 (1969).

[61] A. A. Kamel, "Perturbation Method in the Theory of Nonlinear Oscillations", *Celest. Mech.* **3**, 90 (1970).

[62] A. A. Kamel, "Lie Transforms and the Hamiltonization of Non-Hamiltonian Systems", *Celest. Mech.* **4**, 397 (1971).

[63] J. Henrard, "On Perturbation Theory Using Lie Transforms", *Celest. Mech.* **3**, 107 (1970).

[64] A.H. Nayfeh, *Perturbation Methods*, Wiley (New York, 2000).

[65] E. Leimanis, *The General Problem of the Motion of Coupled Rigid Bodies About a Fixed Point*, p. 121, Springer (1965).

[66] G.E.O. Giacaglia, *Perturbation Methods in Non-Linear Systems*, Springer-Verlag (1972).

[67] J.R. Cary, "Lie Transform Perturbation Theory for Hamiltonian Systems", *Physics Reports* **79**, p. 129 (North Holland 1981).

[68] K. Kowalski and W. Steeb, *Nonlinear Dynamical Systems and Carleman Linearization*, (World Scientific 1991).

[69] K. Meyer, G. Hall, and D. Offin, *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, Second Edition, Springer (2009).

[70] D. Boccaletti and G. Pucacco, *Theory of Orbits*, 2 vols., Springer-Verlag (1996).

[71] G.J. Sussman, J. Wisdom, and M.E. Meyer, *Structure and Interpretation of Classical Mechanics*, MIT Press (2001).

[72] A.J. Lichtenberg and M.A. Lieberman, *Regular and Stochastic Motion*, Springer-Verlag (1983).

[73] K. Meyer, "Lie transform tutorial — II", *Computer Aided Proofs in Analysis*, K. Meyer and D. Schmidt, Eds., Springer-Verlag (1991). Or see the Web site http://math.uc.edu/~meyer/capa91.pdf.

[74] S. Coffey, A. Deprit, E. Deprit, L. Healy, and B. Miller, "A toolbox for nonlinear dynamics", *Computer Aided Proofs in Analysis*, K. Meyer and D. Schmidt, Eds., Springer-Verlag (1991).

[75] A. Dragt, "A Lie Algebraic Theory of Geometrical Optics and Optical Aberrations", *J. Opt. Sci. Am.* **72**, p. 372 (1982).

[76] A. Dragt, E. Forest, and K. Wolf, "Foundations of a Lie Algebraic Theory of Geometrical Optics", *Lie Methods in Optics*, J.S. Mondragon and K.B. Wolf, Edit., Springer-Verlag (1986).

[77] A. Dragt and E. Forest, "Lie Algebraic Theory of Charged Particle Optics and Electron Microscopes", *Advances in Electronics and Electron Physics* **67**, P. Hawkes, edit., Academic Press (1986). NB: The journals *Advances in Electronics and Electron Physics* and *Advances in Optical and Electron Microscopy* have been merged to form the journal *Advances in Imaging and Electron Physics*.

[78] A. Dragt and J. Finn, "Normal Form for Mirror Machine Hamiltonians", *J. Math. Physics* **20**, 2649-2660 (1979).

Chapter 6

Symplectic Maps

This chapter defines symplectic maps and explores some of their properties. They form an infinite dimensional Lie group whose Lie algebra (as will become clear in Chapter 7) is the Poisson bracket Lie algebra of all phase-space functions. It is shown that Hamiltonian flows produce symplectic maps, and essentially any family of symplectic maps arises from an associated Hamiltonian. Thus, Hamiltonian Dynamics *is* the study of symplectic maps, and vice versa. It is also shown that, just as symplectic and symmetric matrices are closely related, symplectic and gradient maps are closely related, and this relation provides a general theory of generating functions. Finally, an introductory discussion is given of symplectic invariants.

6.1 Preliminaries and Definitions

Let $z_1 \cdots z_{2n}$ be a set of canonical coordinates for a $2n$-dimensional space. By *canonical* we mean that we wish to view the $2n$-dimensional space as a *phase* space and, as in (1.7.9), have identified the first n of the z's as being q's and the remaining n as being p's. Suppose a transformation is made that sends the point z with coordinates $z_1 \cdots z_{2n}$ to some other point $\overline{z}$ with coordinates $\overline{z}_1(z,t) \cdots \overline{z}_{2n}(z,t)$. Such a transformation will be called a mapping, and will be denoted by the symbol $\mathcal{M}$,

$$\mathcal{M} : z \to \overline{z}(z,t). \tag{6.1.1}$$

See Figure 1.1. In this discussion, the time t simply plays the role of a parameter. It is included in the notation to indicate that the transformation may depend on the time. That is, the map $\mathcal{M}$ may be different at different times.

Let $M(z,t)$ be the *Jacobian matrix* of the map $\mathcal{M}$. It is defined by the equation

$$M_{ab}(z,t) = \partial \overline{z}_a / \partial z_b. \tag{6.1.2}$$

The Jacobian matrix describes the small changes produced in the *final* quantities $\overline{z}_a$ when small changes are made in the *initial* quantities z_b. See Exercise 1.4.6.

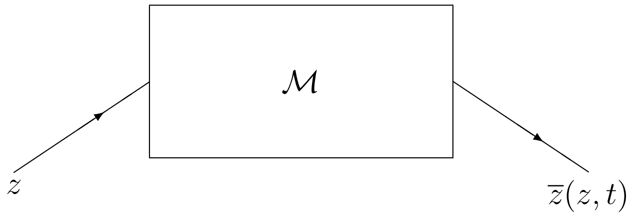


Figure 6.1.1: The map $\mathcal{M}$ sends z to $\overline{z}(z,t)$.

6.1.1 Gradient Maps

As the title to this chapter indicates, it is mostly about symplectic maps. However, we shall subsequently need *gradient* maps as well. Indeed, in Section 6.7 we will learn that there is an intimate connection between gradient and symplectic maps. Therefore, this is a convenient place to make a detour to define gradient maps.

Suppose $g(u,t)$ is some function of the $2n$ variables $u_1 \cdots u_{2n}$ and possibly some parameter t. Use g to define a map $\mathcal{G}$ by the rule

$$\mathcal{G} : u \rightarrow \overline{u}(u,t), \tag{6.1.3}$$

with

$$\overline{u}(u,t)_a = \partial g / \partial u_a. \tag{6.1.4}$$

Note that a gradient is involved in the definition of $\mathcal{G}$, hence the name *gradient* map. We also note that a *single* function, namely $g(u,t)$, has been used to produce the $2n$ functions $\overline{u}(u,t)$. We will refer to g as a *source* function.[1]

Let $G(u,t)$ be the Jacobian matrix of the map $\mathcal{G}$. In accord with the spirit of (1.2), it is given by the equation

$$G_{ab}(u,t) = \partial \overline{u}_a / \partial u_b. \tag{6.1.5}$$

If we now make use of (1.4), we find the relation

$$G_{ab}(u,t) = \partial^2 g / \partial u_b \partial u_a = \partial^2 g / \partial u_a \partial u_b. \tag{6.1.6}$$

That is, G is the Hessian of g. We observe that G is symmetric because the order of partial differentiation is immaterial for functions with continuous derivatives,

$$[G(z,t)]^T = G(z,t). \tag{6.1.7}$$

Conversely, for any map sending u to $\overline{u}$, consider the differential form

$$\sum_a \overline{u}(u,t)_a \, du_a. \tag{6.1.8}$$

It will be *closed* if

$$\partial \overline{u}_a / \partial u_b = \partial \overline{u}_b / \partial u_a. \tag{6.1.9}$$

[1] Since (1.4) involves a gradient, some authors refer to g as a *potential* function.

See Exercise 1.1. But (1.9) is simply the condition that G be symmetric. Thus if the Jacobian matrix of a map is symmetric, there is a function g such that

$$dg = \sum_a \overline{u}(u,t)_a \, du_a. \tag{6.1.10}$$

Indeed, g is given by the path integral

$$g(u,t) = \int^u \sum_a \overline{u}(u',t)_a \, du'_a, \tag{6.1.11}$$

where the integral is to be taken over any path with some fixed initial point and variable end point u. Moreover, it is evident from (1.10) that (1.4) holds. We conclude that a necessary and sufficient condition for a map $\mathcal{G}$ to be a gradient map is that its Jacobian matrix G be symmetric.

We also note that, although we have been working with an even number of variables, namely $2n$, gradient maps are also defined for an odd number of variables. Finally we note that a necessary and sufficient condition for a gradient map to be (locally) invertible is that $\det G \neq 0$, in which case it can be shown that the inverse map is also a gradient map. See Exercise 2.9.[2]

6.1.2 Symplectic Maps

With our detour complete, let us return to the main subject of symplectic maps. The map $\mathcal{M}(t)$ is said to be *symplectic* if its Jacobian matrix M is a symplectic matrix for all values of z and t,

$$M^T J M = J \;\; \text{or} \;\; M J M^T = J, \;\; \forall z, t. \tag{6.1.12}$$

Note that in general M depends on z and t. However, the particular combinations $M^T JM$ or MJM^T must be z and t *independent*. Therefore, a symplectic map must have very special properties.

To appreciate the significance of a symplectic mapping, consider the Poisson brackets of the various $\overline{z}$'s with each other. Using (5.1.3), we find the result

$$[\overline{z}_a, \overline{z}_b] = \sum_{c,d} (\partial \overline{z}_a / \partial z_c) J_{cd} (\partial \overline{z}_b / \partial z_d). \tag{6.1.13}$$

By using the definition (1.2) of the Jacobian matrix M, (1.13) can also be written in the form

$$\begin{aligned} [\overline{z}_a, \overline{z}_b] &= \sum_{c,d} M_{ac} J_{cd} M_{bd} \\ &= \sum_{c,d} M_{ac} J_{cd} (M^T)_{db} = (MJM^T)_{ab}. \end{aligned} \tag{6.1.14}$$

[2] For example in the context of Lagrangian/Hamiltonian dynamics, (1.5.7) is a gradient map from velocity space to momentum space with the Lagrangian L serving as source function, and the first relation in (1.5.11) is the inverse gradient map from momentum space to velocity space with the Hamiltonian H serving as source function. Finally, H and L are Legendre transforms of each other.

Finally, upon using the symplectic condition (1.12), we find the result

$$[\bar{z}_a, \bar{z}_b] = (MJM^T)_{ab} = J_{ab} = [z_a, z_b]. \tag{6.1.15}$$

Consequently, a necessary and sufficient condition for a map $\mathcal{M}$ to be symplectic is that it preserve the fundamental Poisson brackets (1.7.10). As will be shown in Subsection 3, this statement is equivalent, in turn, to the condition that the map $\mathcal{M}$ must preserve the Poisson bracket Lie algebra of all dynamical variables.

Symplectic mappings also have a geometrical aspect. Let z^0 be some point in phase space, and suppose it is sent to the point $\bar{z}^0$ under the action of a symplectic map $\mathcal{M}$. Also, let dz and δz be two small vectors originating at the point z^0. Under the action of $\mathcal{M}$, they are sent to two vectors $d\bar{z}$ and $\delta\bar{z}$. See Figure 1.2.

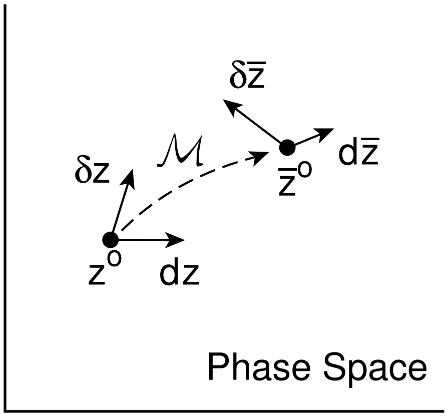


Figure 6.1.2: The action of a symplectic map $\mathcal{M}$ on phase space. The general point z^0 is mapped to the point $\bar{z}^0$, and the small vectors dz and δz are mapped to the small vectors $d\bar{z}$ and $\delta\bar{z}$. The figure is only schematic since in general phase space has a large number of dimensions.

From calculus, we have the relation

$$d\bar{z}_a = \sum_b (\partial \bar{z}_a / \partial z_b) dz_b, \tag{6.1.16}$$

or more compactly, using (1.2),

$$d\bar{z} = M dz. \tag{6.1.17}$$

Similarly, the vectors δz and $\delta\bar{z}$ are related by the equation

$$\delta\bar{z} = M\delta z. \tag{6.1.18}$$

Now use the two vectors $\delta\bar{z}, d\bar{z}$ and the matrix J to form the quantity $(\delta\bar{z}, Jd\bar{z})$. As described in Section 3.2, this quantity is called the *fundamental symplectic 2-form.* Suppose the relations (1.17) and (1.18) are inserted into the 2-form $(\delta\bar{z}, Jd\bar{z})$. Then, using matrix manipulation and the symplectic condition (1.12), we find the relation

$$(\delta\bar{z}, Jd\bar{z}) = (M\delta z, JMdz) = (\delta z, M^T JMdz) = (\delta z, Jdz). \tag{6.1.19}$$

That is, the value of the fundamental symplectic 2-form is *unchanged* by a symplectic map. Evidently, a necessary and sufficient condition for a map to be symplectic is that it preserve the fundamental symplectic 2-form at all points of phase space and for all time.[3]

There is a third aspect of symplectic mappings that should already be familiar. In the usual treatments of Classical Mechanics, an important topic is that of canonical transformations. Canonical transformations are usually defined as those transformations that either

a. preserve the Hamiltonian form of the equations of motion for all Hamiltonian dynamical systems, or

b. preserve the fundamental Poisson brackets.

In case b, according to the previous discussion, canonical transformations and symplectic maps are the same thing. In case a, it can be shown that the most general canonical transformation is a map $\mathcal{M}$ whose Jacobian matrix satisfies the condition

$$M^T JM = \lambda J, \tag{6.1.20}$$

where λ is some real nonzero constant *independent* of z and t. Furthermore, it can be shown that $\mathcal{M}$ in this case consists of a symplectic map followed or preceded by a simple scaling of phase-space variables. See Appendix D. Therefore, in either case, the central object of interest is a symplectic map.

From our perspective, and as will be shown in Subsections 4.1 through 4.3, the most important property of symplectic maps is that Hamiltonian flows produce symplectic maps, and vice versa. Thus, the study of Hamiltonian Dynamics is equivalent to the study of Symplectic Maps.

Exercises

6.1.1. This is an exercise on key properties of *differential forms.* Consider the differential form

$$\sum_{b=1}^{m} C_b(z)dz_b \tag{6.1.21}$$

where the $C_b(z)$ are specified functions of the m variables $z_1, z_2, \cdots z_m$. Before going any further, it is convenient to give a differential form a name so that it is not necessary to always write it out in full. We, as is common, will use the symbol ω to denote the differential form (1.21) and write

$$\omega = \sum_{b=1}^{m} C_b(z)dz_b. \tag{6.1.22}$$

We are now prepared to make some definitions and demonstrate some results about differential forms:

[3] The fundamental symplectic 2-form is also sometimes called the *Lagrange invariant.*

a) A differential form is called *exact* or *perfect* if there exists a function $f(z)$ such that

$$\omega = df. \tag{6.1.23}$$

We know that for any differentiable function f there is the relation

$$df = \sum_{b=1}^{m} (\partial f/\partial z_b) dz_b. \tag{6.1.24}$$

Now suppose the differential form ω is exact. Then, from (1.23), and comparing (1.22) and (1.24), we find the result

$$C_b = \partial f/\partial z_b. \tag{6.1.25}$$

Show that (1.25) implies the result

$$\partial C_b/\partial z_a - \partial C_a/\partial z_b = 0. \tag{6.1.26}$$

A differential form ω that satisfies this relation is called *closed.* Thus, being exact implies being closed.

b) Conversely, suppose (1.26) holds in some *simply-connected* region $\mathcal{R}$. That is, assume the form ω is closed in $\mathcal{R}$. Let z^i and z^f be two arbitrary points in $\mathcal{R}$, and let P be some path in $\mathcal{R}$ joining them. Consider the integral

$$I[P] = \int_{z^i}^{z^f} \sum_b C_b(z) dz_b \tag{6.1.27}$$

evaluated over the path P. In view of (1.22), we may also employ the notation

$$I[P] = \int_{z^i}^{z^f} \omega. \tag{6.1.28}$$

We may regard (1.27) as a *functional* on paths, and write

$$I[z(\tau)] = \int_{\tau^i}^{\tau^f} \{\sum_b C_b(z)\dot{z}_b\} d\tau \tag{6.1.29}$$

where $z(\tau)$ is some parameterization of the path and

$$\dot{z}_b = dz_b/d\tau. \tag{6.1.30}$$

Define a "Lagrangian" L by writing

$$L(z, \dot{z}) = \sum_b C_b(z)\dot{z}_b. \tag{6.1.31}$$

With this definition, (1.29) takes the form

$$I = \int_{\tau^i}^{\tau^f} L(z, \dot{z}) d\tau. \tag{6.1.32}$$

Show, using standard variational calculus, that

$$\delta I = \int_{\tau^i}^{\tau^f} d\tau \{ \sum_a \left(-\frac{d}{d\tau}\frac{\partial L}{\partial \dot{z}_a} + \frac{\partial L}{\partial z_a} \right) \delta z_a \} \tag{6.1.33}$$

for a varied path with the same end points. Show from its definition (1.31) that L satisfies Lagrange's equation,

$$\frac{d}{d\tau}\frac{\partial L}{\partial \dot{z}_a} - \frac{\partial L}{\partial z_a} = 0, \tag{6.1.34}$$

if (1.26) holds, and therefore

$$\delta I = 0 \text{ for all } \delta z_a. \tag{6.1.35}$$

The relation (1.35) shows that I is unchanged in first order when infinitesimal variations (with end points fixed) are made in the path.

From this result, show that I is in fact path independent. In particular, suppose $z(\tau)$ and $\tilde{z}(\tau)$ are two paths in $\mathcal{R}$ with the same end points. Consider the family of paths $z(\tau, \lambda)$ defined by

$$z(\tau, \lambda) = (1 - \lambda)z(\tau) + \lambda\tilde{z}(\tau). \tag{6.1.36}$$

Evidently there are the relations

$$z(\tau, 0) = z(\tau)\ ,\ z(\tau, 1) = \tilde{z}(\tau). \tag{6.1.37}$$

Verify that all the paths in the family have the same end points. Assuming that $z(\tau, \lambda)$ remains in $\mathcal{R}$ for $\tau \in [\tau^i, \tau^f]$ and $\lambda \in [0, 1]$, show from (1.35) that

$$(\partial/\partial\lambda)I[z(\tau, \lambda)] = 0, \tag{6.1.38}$$

and therefore $I[z(\tau, \lambda)]$ is independent of λ so that

$$I[z(\tau)] = I[\tilde{z}(\tau)]. \tag{6.1.39}$$

Now that it has been established that the integral (1.27) is path independent, and therefore depends only on the end points, show that one can define a function $f(z)$ by the rule

$$f(z) = \int_{z^i}^{z} \sum_b C_b(z')dz'_b. \tag{6.1.40}$$

Show, by selecting and sketching a suitable path, that

$$\partial f/\partial z_a = C_a(z). \tag{6.1.41}$$

Hint: To verify (1.41), select a path such that only z'_a varies near the upper integration limit. That is, near and at the final end of this path, the z'_b for $b \neq a$ have already taken on the values $z'_b = z_b$.

Show that

$$df = \sum_b (\partial f/\partial z_b)dz_b = \sum_b C_b(z)dz_b. \tag{6.1.42}$$

Therefore, (1.26) is both necessary and sufficient for a differential to be exact: An exact form is closed, and a form that is closed in a simply-connected region is exact. This result is sometimes called the *Poincaré lemma.* Note that it is an m-dimensional generalization of the familiar 3-dimensional theorem that a vector field can be written as the gradient of a scalar field if and only if the vector field has vanishing curl.

c) Finally, show that if ω is exact, then

$$\int_\Gamma \omega = \int_\Gamma \sum_b C_b(z) dz_b = 0 \tag{6.1.43}$$

where Γ is any closed path in $\mathcal{R}$, and vice versa.

6.1.2. Consider a two-dimensional phase space consisting of the variables q, p. Evaluate the quantity $(\delta z, Jdz)$ and show that it is related to the area formed by the small parallelogram with sides δz and dz. Note that $(\delta z, Jdz)$ can be either positive or negative. Thus, the area is "signed". Consider a $2n$ dimensional phase space. Show that the points $z(\sigma, \tau)$ given by the relation

$$z(\sigma, \tau) = z^0 + \sigma dz + \tau \delta z \text{ with } \sigma, \tau \in [0, 1] \tag{6.1.44}$$

form a two dimensional surface in phase space that can be viewed as a generalized parallelogram with sides δz and dz. Show that this generalized parallelogram has projections into the z_a, z_b planes that are "ordinary" parallelograms (each z_a, z_b plane is two dimensional). In particular, the projections of the generalized parallelogram into the q_i, p_i planes are parallelograms. Finally, show that $(\delta z, Jdz)$ is related to the sum of the *signed* areas of the parallelograms in the q_i, p_i planes. Hint: Use (3.2.3).

6.2 Group Properties

6.2.1 The General Case

Let $\mathcal{M}$ be a symplectic mapping of z to $\overline{z}$, and suppose it has an inverse $\mathcal{M}^{-1}$,

$$\mathcal{M} : z \to \overline{z}, \tag{6.2.1}$$

$$\mathcal{M}^{-1} : \overline{z} \to z. \tag{6.2.2}$$

According to (1.17), the relation between a small change dz in z, and the associated small change $d\overline{z}$ in $\overline{z}$, is given by the Jacobian matrix M of $\mathcal{M}$. Since M is symplectic, it has an inverse M^{-1}. Therefore, (1.17) can be inverted to give the relation

$$dz = M^{-1} d\overline{z}. \tag{6.2.3}$$

But now, comparison of (2.2) and (2.3) shows that the Jacobian matrix of $\mathcal{M}^{-1}$ is M^{-1}. Note also that the local existence of $\mathcal{M}^{-1}$ did not really have to be assumed, but follows instead from the inverse function theorem since M^{-1} is known to exist from the symplectic condition. Finally, the matrix M^{-1} is symplectic since the inverse of a symplectic matrix is also a symplectic matrix. It follows that $\mathcal{M}^{-1}$ is a symplectic map. What has been shown is that if $\mathcal{M}$ is a symplectic map, then $\mathcal{M}^{-1}$ exists (at least locally) and is also a symplectic map.

Next suppose that $\mathcal{M}^{(1)}$ is a symplectic mapping of z to $\overline{z}$ and $\mathcal{M}^{(2)}$ is a symplectic mapping of $\overline{z}$ to another set of variables $\overline{\overline{z}}$. Now consider the composite mapping $\mathcal{M} = \mathcal{M}^{(2)}\mathcal{M}^{(1)}$, which sends z to $\overline{\overline{z}}$.

$$\mathcal{M} = \mathcal{M}^{(2)}\mathcal{M}^{(1)}, \tag{6.2.4}$$

$$\mathcal{M}^{(1)} : z \rightarrow \overline{z}, \tag{6.2.5}$$

$$\mathcal{M}^{(2)} : \overline{z} \rightarrow \overline{\overline{z}}, \tag{6.2.6}$$

$$\mathcal{M}^{(2)}\mathcal{M}^{(1)} : z \rightarrow \overline{\overline{z}} \, . \tag{6.2.7}$$

According to the chain rule, the Jacobian matrix M of the composite mapping $\mathcal{M}$ is the product of the Jacobian matrices of $\mathcal{M}^{(2)}$ and $\mathcal{M}^{(1)}$,

$$M = M^{(2)}M^{(1)}. \tag{6.2.8}$$

However, the matrices $M^{(2)}$ and $M^{(1)}$ are symplectic since they are the Jacobian matrices of symplectic maps. It follows from (2.8) and the group property for symplectic matrices that M is also a symplectic matrix. Consequently, the composite mapping $\mathcal{M}$ is also a symplectic map. What has been shown is that if $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ are symplectic maps, so is their product $\mathcal{M}^{(2)}\mathcal{M}^{(1)}$.

It is also obvious that the identity mapping, which sends each z into itself, is a symplectic map because the Jacobian matrix of this map is evidently the identity matrix, and the identity matrix is symplectic.

The previous discussion has shown that the set of symplectic maps has properties very analogous to the group properties of the group of symplectic matrices. As defined earlier, the concept of a group applied only to matrices. However, it is clear that the concept of a group can be enlarged to include the possibility of general mappings. When this is done, the set of all symplectic maps is entitled to be called a group. The set of all differentiable maps forms a group called the group of all diffeomorphisms. Because of the symplectic restriction, the set of all symplectic maps is a subgroup of the group of all diffeomorphisms.

6.2.2 Various Subgroups and Their Names

We found in Section 3.6.1 that the set of all real $2n \times 2n$ symplectic matrices forms a group. We have denoted this group and its Lie algebra by the symbols $Sp(2n, \mathbb{R})$ and $sp(2n, \mathbb{R})$. Equivalently, in the present context, the subset of all symplectic maps that send the origin into itself (preserve the origin) and are *linear* is a subgroup of the group of all symplectic maps, and this subgroup is $Sp(2n, \mathbb{R})$. See Exercise 2.1. Evidently, the subset of all symplectic maps that send the origin into itself, but are not necessarily linear, is also a subgroup of the group of all symplectic maps. For lack of any standard terminology, we will refer to this group as $SpM(2n, \mathbb{R})$. Here the M in the name stands either for the word *map* or, to please the French, the word *morphism* because those of Gallic bent often refer to maps as *morphisms.* The underlying Lie algebra of $SpM(2n, \mathbb{R})$, see Section 7.7, will be referred to as $spm(2n, \mathbb{R})$.

Next consider mappings of the form

$$\bar{z}_a = z_a + c_a, \tag{6.2.9}$$

where the quantities c_a are constants. It is easily verified that such maps are symplectic, and form a group. This group is called the phase-space *translation* group. Now consider phase-space mappings of the form

$$\bar{z}_a = c_a + \sum_b M_{ab} z_b, \tag{6.2.10}$$

where the matrices M are symplectic. Such maps are also symplectic and form a group. This group is called the *inhomogeneous* symplectic group, and will be referred to by the symbols $ISp(2n, \mathbb{R})$. The underlying Lie algebra of $ISp(2n, \mathbb{R})$, see Sections 7.7 and 9.2, will be referred to as $isp(2n, \mathbb{R})$.

Finally, consider the group of all symplectic maps (also called *symplectomorphisms*) that do not necessarily preserve the origin and are not necessarily linear. This group will be referred to as $ISpM(2n, \mathbb{R})$ and its Lie algebra, see Exercise 7.7.2, will be referred to as $ispm(2n, \mathbb{R})$.[4]

We close this section by noting that gradient maps do *not* form a group. In the case of gradient maps there is again a relation like (2.8) for the Jacobian of the product of two maps. However, the product of two symmetric matrices is generally not a symmetric matrix. Therefore, the product of two gradient maps is generally not a gradient map. Gradient maps belong to the group of all diffeomorphisms, but do not form a subgroup. However, it can be shown that the identity map is a gradient map; and the inverse of a gradient map, if the inverse exists, is also a gradient map. See Exercise 2.9.

Exercises

6.2.1. Consider phase-space mappings of the form

$$\bar{z} = Mz \tag{6.2.11}$$

where M is a symplectic matrix. Show that such maps are symplectic, and form a group. Show that the symplectic map for $M = J$ interchanges (with a minus sign) coordinates and momenta.

6.2.2. Consider phase-space mappings of the form (2.9). Show that such maps are symplectic, and form a group. Consider phase-space mappings of the form (2.10). Show that such maps are symplectic, and also form a group.

6.2.3. Use (1.2) and the chain rule to verify (2.8).

6.2.4. Consider the nonrelativistic motion of a particle of mass m described by Cartesian coordinates $\boldsymbol{q}(t)$. In the usual way, define the momentum $\boldsymbol{p}(t)$ by the relation

$$\boldsymbol{p} = m\dot{\boldsymbol{q}}. \tag{6.2.12}$$

[4] Some authors refer to $ISpM(2n, \mathbb{R})$ simply as $Symp(n)$ and to $Sp(2n, \mathbb{R})$ as $Sp(n)$.

The *Euclidean* group consists of spatial transformations of the form

$$\bar{\boldsymbol{q}} = R\boldsymbol{q} + \boldsymbol{d} \tag{6.2.13}$$

where R is a 3×3 rotation matrix and $\boldsymbol{d}$ is a fixed vector. It describes rotations and translations (displacements) in 3-dimensional space. These transformations are extended to phase space by the rule

$$\bar{\boldsymbol{p}} = R\boldsymbol{p}. \tag{6.2.14}$$

Setting $z = (\boldsymbol{q};\boldsymbol{p})$, verify that (2.13) and (2.14) specify a symplectic map. That is, verify that

$$[\bar{z}_a, \bar{z}_a] = J_{ab}. \tag{6.2.15}$$

Thus, the Euclidean group is a subgroup of the group of all symplectic maps.

To the Euclidean group add the further spatial transformations

$$\bar{\boldsymbol{q}}(t) = \boldsymbol{q}(t) + \boldsymbol{u}t. \tag{6.2.16}$$

These transformations describe the (nonrelativistic) coordinate relation between two inertial frames moving with (fixed) relative velocity $\boldsymbol{u}$. In accord with (2.12), these transformations may be extended to phase space by the rule

$$\bar{\boldsymbol{p}} = \boldsymbol{p} + m\boldsymbol{u}. \tag{6.2.17}$$

Show that the transformations described by (2.16) and (2.17) are also symplectic maps. Together the transformations described by (2.13), (2.14) and (2.16),(2.17) form the group of all *Galilean* transformations. You have shown that the Galilean group is a subgroup of the group of all symplectic maps.[5]

Suppose we extend phase space to include t as a coordinate and p_t as its conjugate momentum, in which case some parameter τ becomes the independent variable. See Exercises 1.6.4 and 1.6.5. Can the Galilean group be extended to act on this extended phase space? Implicit in the nonrelativistic approach is the assumption that time is the same in all inertial frames,

$$\bar{t} = t. \tag{6.2.18}$$

How should we define $\bar{p}_t$? As motivation, consider the free particle case for which we have the relation

$$p_t = -\boldsymbol{p} \cdot \boldsymbol{p}/(2m). \tag{6.2.19}$$

If we write

$$\bar{p}_t = -\bar{\boldsymbol{p}} \cdot \bar{\boldsymbol{p}}/(2m) \tag{6.2.20}$$

and use (2.17), we find the result

$$\bar{p}_t = p_t - \boldsymbol{u} \cdot \boldsymbol{p} - (m/2)\boldsymbol{u} \cdot \boldsymbol{u}, \tag{6.2.21}$$

which we take to be the rule for how p_t transforms.[6]

[5] Note that all these transformations are in fact a subset of the inhomogeneous symplectic group, and are therefore automatically symplectic. See (2.9) and (2.10).

[6] See also Exercises 2.5 and 2.7 below.

Show that the relations (2.13),(2.14) and (2.16),(2.17) and (2.18),(2.21) also yield a group, which we might call the extended Galilean group. Are these transformations symplectic maps on the extended phase space? Show that they are, and therefore the extended Galilean group is a subgroup of the group of all symplectic maps on extended phase space. To make this demonstration, write $z = (\boldsymbol{q}, t; \boldsymbol{p}, p_t)$ and again set up the usual rules

$$[z_a, z_b] = J_{ab}. \tag{6.2.22}$$

Show it then follows that (2.15) also holds on the extended phase space. In particular, you will need to verify that

$$[\bar{p}_t, \bar{\boldsymbol{q}}] = 0. \tag{6.2.23}$$

Finally, note that the fact that a particular group can be realized as a set of phase-space transformations does not necessarily say anything about the invariance properties of the dynamics of any particular system. What is needed for invariance is for trajectories to be sent into trajectories under the action of the group. For example, see Exercise 1.6.9.

6.2.5. Read Exercise 2.4. The reader may be dubious about the use of the free particle case to motivate the transformation rule (2.21). Here is another approach. For simplicity, consider the case in which phase space is two dimensional. Suppose that the transformation rule for p_t is of the form

$$\bar{p}_t = p_t - up + \alpha(u) \tag{6.2.24}$$

where $\alpha(u)$ is a function yet to be determined. Verify that the $-up$ term in (2.24) is necessary to satisfy (2.23), but that the symplectic condition is satisfied for any choice of α. Now make the requirement that the extended Galilean transformations form a group. Make two successive transformations with relative velocities u_1 and u_2 to obtain the relations

$$\bar{q} = q + u_1 t, \tag{6.2.25}$$

$$\bar{t} = t, \tag{6.2.26}$$

$$\bar{p} = p + mu_1, \tag{6.2.27}$$

$$\bar{p}_t = p_t - u_1 p + \alpha(u_1); \tag{6.2.28}$$

$$\bar{\bar{q}} = \bar{q} + u_2 \bar{t}, \tag{6.2.29}$$

$$\bar{\bar{t}} = \bar{t}, \tag{6.2.30}$$

$$\bar{\bar{p}} = \bar{p} + mu_2, \tag{6.2.31}$$

$$\bar{\bar{p}}_t = \bar{p}_t - u_2 \bar{p} + \alpha(u_2). \tag{6.2.32}$$

Show that combining the relations (2.25) through (2.32) yields the net relations

$$\bar{\bar{q}} = q + (u_1 + u_2)t, \tag{6.2.33}$$

$$\bar{\bar{t}} = t, \tag{6.2.34}$$

$$\bar{\bar{p}} = p + m(u_1 + u_2), \tag{6.2.35}$$

$$\bar{\bar{p}}_t = p_t - (u_1 + u_2)p - mu_1 u_2 + \alpha(u_1) + \alpha(u_2). \tag{6.2.36}$$

We see that (2.33) through (2.35) are of the standard Galilean transformation form corresponding to a relative velocity u_1+u_2. Therefore, if we wish (2.36) to also be of the standard form (2.24), we must require the relation

$$-mu_1u_2 + \alpha(u_1) + \alpha(u_2) = \alpha(u_1 + u_2). \tag{6.2.37}$$

Show that (2.37) implies the relation

$$\alpha(0) = 0. \tag{6.2.38}$$

To make further progress, assume that α is differentiable.[7] Set

$$u_1 = u \tag{6.2.39}$$

and

$$u_2 = \epsilon. \tag{6.2.40}$$

Then, assuming differentiability, show that there are the relations

$$\alpha(u_1 + u_2) = \alpha(u + \epsilon) = \alpha(u) + \alpha'(u)\epsilon + O(\epsilon)^2, \tag{6.2.41}$$

$$\alpha(u_2) = \alpha(\epsilon) = \alpha(0) + \alpha'(0)\epsilon + O(\epsilon)^2. \tag{6.2.42}$$

Next, show that inserting (2.38) through (2.42) into (2.37) and equating like powers of ϵ yields the differential equation

$$\alpha'(u) = \alpha'(0) - mu \tag{6.2.43}$$

with the solution

$$\alpha(u) = u\alpha'(0) - (1/2)mu^2. \tag{6.2.44}$$

Here, in solving (2.43), we have taken into account the boundary condition (2.38). Finally, let us apply the transformation (2.25) through (2.28) to the phase-space origin $q = p = 0$. Doing so gives the result

$$\bar{p}_t = p_t + \alpha(u) = p_t + u\alpha'(0) - (1/2)u^2. \tag{6.2.45}$$

If we now require that $\bar{p}_t$ be independent of the sign of u, which seems reasonable since there is no preferred direction when $q = p = 0$, we conclude that we should demand the further condition

$$\alpha'(0) = 0. \tag{6.2.46}$$

Thus, under reasonable assumptions, we again arrive at (2.21).

6.2.6. Study Exercises 1.6.7, 1.6.8, and 1.7.5. Suppose x and y are any two space-time points. The *interval* $D^2(x, y)$ between them is given by the relation

$$D^2(x, y) = g_{\mu\nu}(x - y)^\mu(x - y)^\nu = ([x - y], g[x - y]) = (x - y) \cdot (x - y). \tag{6.2.47}$$

[7] Actually, it is sufficient to assume continuity. It is a remarkable property of Lie groups that the assumption of continuity implies differentiability, and indeed, also the far stronger condition of analyticity.

Here we use the metric g given by (1.6.45) and $(*,*)$ denotes the usual/ordinary scalar product. Consider the set of all transformations that send space-time into itself. Special Relativity asserts that if two events can occur at the points x, y (i.e. the events are consonant with physical law), then they can also occur at the points $\tilde{x}, \tilde{y}$ provided

$$D^2(\tilde{x}, \tilde{y}) = D^2(x, y). \tag{6.2.48}$$

Transformations that satisfy (2.48) are called *Poincaré* transformations.

In studying Poincaré transformations it is often assumed from the outset that they are of the form

$$\tilde{x}^\alpha = \sum_{\beta=1}^{4} \Lambda^{\alpha\beta} x^\beta + d^\alpha \Leftrightarrow \tilde{x} = \Lambda x + d. \tag{6.2.49}$$

That is, they are assumed to consist of a *linear* transformation described by the matrix Λ followed by a space-time translation described by the 4-vector d. Such an assumption is not necessary. It can be proved that the most general transformation satisfying (2.48) for all pairs of points must be of the form (2.49). See Exercise 7.3.26. Assuming the form (2.49), show from (2.48) that the matrix Λ must satisfy the relation

$$\Lambda^T g \Lambda = g. \tag{6.2.50}$$

Show that the matrices Λ form a group (called the *Lorentz* group).[8] Show that Poincaré transformations also form a group (called the Poincaré group).[9] Show that there are the logical implications

$$\Lambda^T g \Lambda = g \Leftrightarrow \Lambda g \Lambda^T = g \Leftrightarrow (\Lambda^T)^T g \Lambda^T = g. \tag{6.2.51}$$

Suppose that Λ is an element of the Lorentz group. Show that then Λ^{-1} and Λ^T and $(\Lambda^T)^{-1} = (\Lambda^{-1})^T$ are elements of the Lorentz group, and vice versa.

Suppose space-time coordinates are transformed according to (2.49) and the action of the Lorentz (and Poincaré) group is *extended* to act on momenta by the rule

$$\tilde{p} = \Lambda p \Leftrightarrow \tilde{p}^\alpha = \sum_{\beta=1}^{4} \Lambda^{\alpha\beta} p^\beta. \tag{6.2.52}$$

Define canonical coordinates in an eight-dimensional phase space to consist of the pairs $(x^\mu,\ p_\nu)$. It can be shown that (2.49) and (2.52) produce a symplectic map in this phase

[8] The finite dimensional representations of the Lorentz group formed by the matrices Λ are described in Exercise 7.3.27. Remarkably, as shown in Exercise 7.3.29, the identity component of the Lorentz group is homomorphic to the group $SL(2, \mathbb{C})$. Indeed, $SL(2, \mathbb{C})$ is the covering group of the Lorentz group.

We also take this occasion to make a comment about nomenclature and notation. Under a Lorentz transformation space-time transforms according to the rule $\tilde{x} = \Lambda x$ from which it follows that $d\tilde{x} = \Lambda dx$. There is also the chain-rule relation $d\tilde{x}^\alpha = \sum_\beta (\partial \tilde{x}^\alpha / \partial x^\beta) dx^\beta$ and therefore $\Lambda^{\alpha\beta} = \partial \tilde{x}^\alpha / \partial x^\beta$. Any collection of four elements V^α is defined to be a *four-vector* if there is the transformation rule $\tilde{V} = \Lambda V$. Similarly, any set of sixteen elements $T^{\alpha\beta}$ is defined to be a *second-rank tensor* if there is the transformation rule $\tilde{T}^{\alpha\beta} = \sum_{\mu\nu} \Lambda^{\alpha\mu} \Lambda^{\beta\nu} T^{\mu\nu}$. Note that these transformation rules can also be written in the less compact forms $\tilde{V}^\alpha = \sum_\mu (\partial \tilde{x}^\alpha / \partial x^\mu) V^\mu$ and $\tilde{T}^{\alpha\beta} = \sum_{\mu\nu} (\partial \tilde{x}^\alpha / \partial x^\mu)(\partial \tilde{x}^\beta / \partial x^\nu) T^{\mu\nu}$, etc., as is frequently done.

[9] The Poincaré group could also be called the inhomogeneous Lorentz group.

space. See Exercise 2.13. That is, (extended) Poincaré transformations are symplectic maps, and therefore form a subgroup of the group of all symplectic maps. Indeed, since Lorentz transformations are linear, the (extended) Lorentz group is a subgroup of $Sp(8,\mathbb{R})$. And, according to (2.49), the (extended) Poincaré group is a subgroup of $ISp(8,\mathbb{R})$.

6.2.7. Read Exercises 2.4 and 2.6 above if you have not already done so. The Poincaré group involves the parameter c. The aim of this exercise is to show that in the limit $c \to \infty$ the Poincaré group becomes the extended Galilean group plus translations in time. Consider, for simplicity, a velocity transformation along the z axis with velocity u. In this case Λ is the matrix

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \gamma(u) & \beta(u)\gamma(u) \\ 0 & 0 & \beta(u)\gamma(u) & \gamma(u) \end{pmatrix} \tag{6.2.53}$$

where

$$\beta(u) = u/c \tag{6.2.54}$$

and

$$\gamma(u) = 1/\sqrt{1-[\beta(u)]^2}. \tag{6.2.55}$$

Then, from (2.49), we find for the space-time coordinate variables x^μ the relations

$$\tilde{x}^1 = x^1, \tag{6.2.56}$$

$$\tilde{x}^2 = x^2, \tag{6.2.57}$$

$$\tilde{x}^3 = \gamma(u)x^3 + \gamma(u)\beta(u)x^4, \tag{6.2.58}$$

$$\tilde{x}^4 = \gamma(u)\beta(u)x^3 + \gamma(u)x^4. \tag{6.2.59}$$

Show that, with the aid of (1.6.41), these last two relations can be rewritten in the form

$$\tilde{z} = \gamma(u)z + \gamma(u)\beta(u)ct = \gamma(u)z + \gamma(u)ut, \tag{6.2.60}$$

$$\tilde{t} = \gamma(u)\beta(u)z/c + \gamma(u)t. \tag{6.2.61}$$

Show that in the limit $c \to \infty$ the relations (2.60) and (2.61) become

$$\tilde{z} = z + ut, \tag{6.2.62}$$

$$\tilde{t} = t, \tag{6.2.63}$$

which, along with (2.56) and (2.57), are a special case of the Galilean transformation given by (2.16) and (2.18).

The limiting case of the momentum relations (2.52) is a bit more delicate because the quantities p^μ are c dependent. Verify that for Λ given by (2.53) the relations (2.52) have the component form

$$\tilde{p}^1 = p^1, \tag{6.2.64}$$

$$\tilde{p}^2 = p^2, \tag{6.2.65}$$

$$\tilde{p}^3 = \gamma(u)p^3 + \gamma(u)\beta(u)p^4, \tag{6.2.66}$$

$$\tilde{p}^4 = \gamma(u)\beta(u)p^3 + \gamma(u)p^4. \tag{6.2.67}$$

Show that use of (1.6.82), (1.6.96), and (1.6.97) in (2.66) gives the result

$$\begin{aligned} \gamma(\tilde{v})m\tilde{v}_z + q\tilde{A}_z &= \gamma(u)[\gamma(v)mv_z + A_z] + \gamma(u)\beta(u)[qA^4 + \gamma(v)mc] \\ &= \gamma(u)[\gamma(v)mv_z + A_z] + \gamma(u)\gamma(v)mu + \gamma(u)\beta(u)qA^4. \end{aligned} \tag{6.2.68}$$

Show that in the limit $c \to \infty$ (2.68) becomes

$$m\tilde{v}_z + q\tilde{A}_z = mv_z + A_z + mu, \tag{6.2.69}$$

which can be written in the form

$$\tilde{p}_z^{\text{nr}} = p_z^{\text{nr}} + mu \tag{6.2.70}$$

where $\boldsymbol{p}^{\text{nr}}$ is the nonrelativistic canonical momentum. Observe that (2.64), (2.65), and (2.70) are a special case of (2.17). Thus we have obtained in the $c \to \infty$ limit the Galilean transformations for the spatial components of p^μ.

What remains is the temporal component of p^μ. Show that use of (1.6.82), (1.6.96), and (1.7.20) in (2.67) gives the result

$$\tilde{p}_t = -\gamma(u)\beta(u)cp^3 + \gamma(u)p_t, \tag{6.2.71}$$

which can be rewritten in the form

$$\tilde{p}_t = -\gamma(u)up^3 + \gamma(u)p_t. \tag{6.2.72}$$

Next we need to expand p_t as given by (1.7.22). Show that

$$\begin{aligned} p_t &= -q\psi - \gamma(v)mc^2 = -q\psi - mc^2 - (1/2)mv^2 + c^2O(v/c)^4 \\ &= p_t^{\text{nr}} - mc^2 + c^2O(v/c)^4 \end{aligned} \tag{6.2.73}$$

where we have defined a nonrelativistic p_t by the rule

$$p_t^{\text{nr}} = -q\psi - (1/2)mv^2 = p_t + mc^2 + c^2O(v/c)^4. \tag{6.2.74}$$

Now insert (2.73) into (2.72) to obtain the result

$$\begin{aligned} \tilde{p}_t^{\text{nr}} &= -\gamma(u)up^3 + \gamma(u)[p_t^{\text{nr}} - mc^2] + mc^2 + c^2O(\tilde{v}/c)^4 + c^2O(v/c)^4 \\ &= -\gamma(u)up^3 + \gamma(u)p_t^{\text{nr}} + mc^2[1-\gamma(u)] + c^2O(\tilde{v}/c)^4 + c^2O(v/c)^4. \end{aligned} \tag{6.2.75}$$

Next verify that

$$p^3 = p_z^{\text{nr}} + O(v/c)^2 \tag{6.2.76}$$

and

$$mc^2[1-\gamma(u)] = -(1/2)mu^2 + c^2O(u/c)^4. \tag{6.2.77}$$

Now we are ready to take the $c \to \infty$ limit of (2.75). Verify that this limit is

$$\tilde{p}_t^{\text{nr}} = p_t^{\text{nr}} - up_z^{\text{nr}} - (m/2)u^2. \tag{6.2.78}$$

Observe that (2.78) is a special case of (2.21).

We have achieved our goal. We have seen that under a suitable limiting process the Lorentz group reduces to the extended Galilean group.[10] Correspondingly, the Poincaré group reduces to the extended Galilean group plus translations in time, $\tilde{t} = t + a^4$.

6.2.8. Study Exercises 1.6.7 through 1.6.10 and Exercise 1.7.5, and adopt the phase-space coordinates of Exercise 2.6. Show that gauge transformations produce symplectic maps. Note that if it is necessary to append a gauge transformation to the Lorentz transformation (2.52), as described in a footnote to Exercise 1.6.7, the net result is still a symplectic map. Let $\phi(x)$ be any scalar field, and let $\mathcal{A}$ be the symplectic (see Section 7.1) map

$$\mathcal{A} = \exp(-q : \phi :). \tag{6.2.79}$$

Let $\mathcal{A}$ act on the H_R given by (1.6.92), and demonstrate that $\mathcal{A}$ produces gauge transformations.

6.2.9. Let $\mathcal{S}_0$ be the set of all diffeomorphisms (in $2n$ dimensions), $\mathcal{S}_1$ be the set of all orientation preserving diffeomorphisms (see Exercise 1.4.6), and $\mathcal{S}_2$ be the set of all symplectic maps. Show that each of these sets forms a group, and that there is the inclusion relation

$$\mathcal{S}_0 \supset \mathcal{S}_1 \supset \mathcal{S}_2. \tag{6.2.80}$$

Review Section 6.1.1. Show, by an example, that the product of two symmetric matrices need not itself be symmetric, and thereby demonstrate, in view of (2.4) through (2.8), that gradient maps do not form a subgroup of the group of all diffeomorphisms. Review Exercise 5.3.7. Show that the maps produced by integrating gradient vector fields over some time interval are diffeomorphisms, but generally do not form a subgroup of the group of all diffeomorphisms. Show that, despite the use of the adjective *gradient* to describe the underlying vector fields, such maps are also generally not gradient maps.

Show that the identity map $\mathcal{I}$ defined by

$$\overline{u} = \mathcal{I}u = u \tag{6.2.81}$$

has the identity matrix I for its Jacobian matrix, and therefore $\mathcal{G} = \mathcal{I}$ is a gradient map. Using (1.11), show that the associated source function $g(u)$ that produces this $\mathcal{G}$ is given by

$$g(u) = (u, u)/2, \tag{6.2.82}$$

and verify that use of this g in (1.4) yields $\mathcal{G} = \mathcal{I}$.

[10] This reduction of the Lorentz group to the Galilean group is an example of a process that can be applied to many groups and is called *Inönü-Wigner contraction*. The inverse process to contraction is called *deformation*. For example, it can be shown that the Quantum Mechanical commutator Lie algebra of functions of the quantum variables Q and P is a deformation of the Poisson bracket Lie algebra of functions of the associated classical variables q and p. We may say that Quantum Mechanics is a deformation of Classical Mechanics, and Classical Mechanics is a contraction of Quantum Mechanics in the limit $\hbar \to 0$. See Appendix Y. Similarly, ray optics is a contraction of wave optics in the limit that the wavelength $\lambda \to 0$, and wave optics is a deformation of ray optics.

Suppose $\mathcal{G}$ is a gradient map. Write

$$\bar{u} = \mathcal{G}u, \tag{6.2.83}$$

and require that its Jacobian matrix G be symmetric as in (1.7) so that $\mathcal{G}$ is indeed a gradient map. Suppose further that $\mathcal{G}$ is invertible, and let $\mathcal{H}$ be its inverse so that we may write

$$u = \mathcal{H}\bar{u}, \tag{6.2.84}$$

or

$$\mathcal{H} = \mathcal{G}^{-1}. \tag{6.2.85}$$

Your first task is to show that $\mathcal{H}$ is also a gradient map.

Show from (1.5) that there is the relation

$$d\overline{u} = G(u)du. \tag{6.2.86}$$

By the inverse function theorem, the condition for $\mathcal{G}$ to be invertible is

$$\det G \neq 0. \tag{6.2.87}$$

Show under the assumption (2.87) that (2.86) can be rewritten in the form

$$du = H(\bar{u})d\bar{u} \tag{6.2.88}$$

with

$$H(\bar{u}) = [G(u)]^{-1}. \tag{6.2.89}$$

Thus, H is the Jacobian matrix for $\mathcal{H}$. Verify it follows that there are the series of relations

$$HG = I, \tag{6.2.90}$$

$$G^T H^T = I, \tag{6.2.91}$$

$$GH^T = I, \tag{6.2.92}$$

$$H^T = G^{-1} = H. \tag{6.2.93}$$

You have shown that $\mathcal{H}$ is also a gradient map.

What is the source function h for $\mathcal{H}$? Let g be the source function for $\mathcal{G}$. Show that the Ansatz

$$h(\overline{u}) = (u, \overline{u}) - g(u) \tag{6.2.94}$$

produces a well-defined function $h(\overline{u})$ when u is viewed as a function of $\overline{u}$,

$$u = \mathcal{G}^{-1}\overline{u}. \tag{6.2.95}$$

Show that the differential of h is given by the relation

$$dh = \sum_a [\overline{u}_a(du_a) + u_a(d\overline{u}_a) - (\partial g/\partial u_a)(du_a)]. \tag{6.2.96}$$

Next use (1.4) to show that dh as given by (2.96) can also be written in the form

$$dh = \sum_a u_a(d\overline{u}_a), \tag{6.2.97}$$

and thereby conclude that

$$u_a = \partial h/\partial \overline{u}_a. \tag{6.2.98}$$

Comparison of (2.84) and (2.98) shows that h is the source function for $\mathcal{H}$.

Show from (2.88) and (2.98) that

$$H_{ab} = \partial^2 h/\partial \overline{u}_a \partial \overline{u}_b. \tag{6.2.99}$$

The map $\mathcal{H}$ will be invertible iff

$$\det H \neq 0. \tag{6.2.100}$$

Show from (2.90) that

$$(\det H)(\det G) = 1. \tag{6.2.101}$$

Therefore $\mathcal{H}$ is invertible if $\mathcal{G}$ is invertible, and vice versa. Moreover, the relation (2.94) can also be written in the form

$$g(u) = (u, \overline{u}) - h(\overline{u}) \tag{6.2.102}$$

where

$$\overline{u} = \mathcal{H}^{-1}u. \tag{6.2.103}$$

Equation (2.94) shows that h is the Legendre transform of g and, conversely, (2.102) shows that g is the Legendre transform of h. In this context, a Legendre transformation is the relation between the source function of a gradient map and the source function of its inverse.

The production of one function from another by performing a Legendre transformation may be viewed as the result of some operator $\mathcal{O}$ acting on *function* space. Observe that g is associated with $\mathcal{G}$ and h is associated with $\mathcal{G}^{-1}$. Since $(\mathcal{G}^{-1})^{-1} = \mathcal{G}$, it follows that $\mathcal{O}^2$ is the *identity* operator on function space. An operator whose square is the identity is called an *involution*.[11] We have learned that the act of performing a Legendre transformation is an involution.

As a sanity check on this claim, work out an example. Suppose $f(x)$ is a function of a single variable x given by the rule

$$f(x) = \lambda x^n \tag{6.2.104}$$

where λ is some positive constant, and let $g(\overline{x})$ be its Legendre transform. Show that

$$g(\overline{x}) = \overline{\lambda}(\overline{x})^{\overline{n}} \tag{6.2.105}$$

where

$$\overline{n} = n/(n-1) \tag{6.2.106}$$

and

$$\overline{\lambda} = (n-1)\lambda(n\lambda)^{-\overline{n}}. \tag{6.2.107}$$

[11] Note: this use of the word *involution* is not to be confused with that in Section 5.2.

Let $h(\bar{\bar{x}})$ be the Legendre transform of $g(\overline{x})$. Find $h(\bar{\bar{x}})$ and verify that

$$h(\bar{\bar{x}}) = f(\bar{\bar{x}}). \tag{6.2.108}$$

Here is an alternate, but equivalent, definition of the Legendre transform. Given the function $g(u)$, show that $h(\bar{u})$ can be defined by the rule

$$h(\bar{u}) = \max_u [(u, \overline{u}) - g(u)]. \tag{6.2.109}$$

This relation holds when g is *convex*, i.e. the Hessian of g is a positive definite matrix at each point u. More generally, one should look for an extremum rather than a maximum. Extrema will exist and be locally isolated as long as the Hessian of g is nonsingular.

In the case of a function $f(x)$ of a single variable x with Legendre transform $g(\overline{x})$, how are the graphs $y = f(x)$ and $\overline{y} = g(\overline{x})$ related geometrically? Suggestion: See the Legendre transformation references listed in the bibliography at the end of this chapter.

Review the passage from a Lagrangian to a Hamiltonian employed in Section 1.5. Observe that the relation (1.5.7) between p and $\dot{q}$ is a gradient map produced by using L as a source function, with the remaining variables q and t simply going along for the ride. Also, (1.5.8) is a Legendre transformation that produces H from L, again with the variables q and t going along for the ride. Finally, the first of the equations (1.5.11), the one yielding the $\dot{q}_i$, is simply the inverse of the gradient map (1.5.7). Evidently, the only additional “physics” in the relations (1.5.11) is that given by the second equation which yields the $\dot{p}_i$.

6.2.10. Let $\mathcal{M}$ be a map and let M be its Jacobian matrix. By the inverse function theorem, $\mathcal{M}$ is invertible in the vicinity of a point if M is invertible at that point. Suppose that M is invertible everywhere. Does it then follow that $\mathcal{M}$ is globally invertible? Consider the following two-dimensional counter example: Suppose the map $\mathcal{M}$ sends the points x, y to the points u, v by the rule

$$u = e^x \cos y, \tag{6.2.110}$$

$$v = e^x \sin y. \tag{6.2.111}$$

Show that in this case M is the matrix

$$M = \begin{pmatrix} e^x \cos y & -e^x \sin y \\ e^x \sin y & e^x \cos y \end{pmatrix}. \tag{6.2.112}$$

Verify that

$$\det M = e^{2x} \neq 0, \tag{6.2.113}$$

and therefore M is globally invertible. Show that nevertheless $\mathcal{M}$ is not globally invertible. Hint: Introduce complex variables z, w by writing the relations

$$z = x + iy, \tag{6.2.114}$$

$$w = u + iv. \tag{6.2.115}$$

Show that $\mathcal{M}$ as given by (2.110) and (2.111) is equivalent to the complex relation

$$w = e^z, \tag{6.2.116}$$

and therefore $\mathcal{M}^{-1}$ is given by the *multivalued* relation

$$z = \log w. \tag{6.2.117}$$

6.2.11. Consider a two-dimensional phase space with Cartesian coordinates q, p. Define new coordinates Q, P implicitly by the rules

$$q = (2P)^{1/2}\cos(Q), \tag{6.2.118}$$

$$p = -(2P)^{1/2}\sin(Q). \tag{6.2.119}$$

Verify that (2.118) and (2.119) have the inverse relations

$$Q = \tan^{-1}(-p/q) = -\tan^{-1}(p/q), \tag{6.2.120}$$

$$P = (1/2)(p^2 + q^2). \tag{6.2.121}$$

Evidently the quantities $(2P)^{1/2}$ and Q assign what are essentially cylindrical polar coordinates to the phase-space point having Cartesian coordinates q, p. The only difference is that increasing Q produces a clockwise rotation whereas, in the usual convention, increasing the polar angle θ produces a counterclockwise rotation. We also remark that commonly the symbols J, ϕ, called *action-angle variables*, are used instead of P, Q.

Let C be the closed circular path of radius R about the origin of q, p phase space obtained by using (2.118) and (2.119) with $(2P)^{1/2} = R$ and $Q \in [0, 2\pi]$. Verify that the *action* A associated with this circular path, defined by the relation

$$A = \oint_C p\, dq = \int_{p^2+q^2 \le R^2} dp\, dq, \tag{6.2.122}$$

has the value

$$A = \pi R^2 = 2\pi P. \tag{6.2.123}$$

Show that there is the relation

$$[Q, P] = [q, p] = 1 \tag{6.2.124}$$

so that the quantities Q and P may be thought of as position-like and momentum-like coordinates, respectively. Correspondingly, the relations (2.120) and (2.121) describe a symplectic map $\mathcal{M}$, and the relations (2.118) and (2.119) describe its inverse. Verify that $\mathcal{M}$ and $\mathcal{M}^{-1}$ are *not* analytic at the origin. This is to be expected because polar coordinates are ill defined at the origin.

Let L be the harmonic oscillator Lagrangian

$$L = (1/2)(\dot{q}^2 - q^2). \tag{6.2.125}$$

Show that the associated Hamiltonian is

$$H = (1/2)(p^2 + q^2). \tag{6.2.126}$$

Show that under the symplectic map $\mathcal{M}$ the Hamiltonian H becomes the transformed Hamiltonian K given by the relation

$$K = P. \tag{6.2.127}$$

Show that there is no Lagrangian whose associated Hamiltonian is K. See Exercise 2.9 and Section 1.5. In particular, see Exercises 1.5.13 and 1.5.14. Thus, under the action of symplectic maps, it is possible that one may move beyond the realm of Lagrangian mechanics.

6.2.12. Exercise to find Λ from F and $\bar{F}$ given that $I_1(\bar{F}) = I_1(F)$, etc. See Exercise 1.6.17.

Suppose $\boldsymbol{E}$ and $\boldsymbol{B}$ are electric and magnetic fields having (at some point x^μ in space-time) arbitrary magnitude and direction.

6.2.13. Review the last paragraph of Exercise 2.6. The purpose of this exercise is to show that extended Lorentz/Poincaré transformations are symplectic maps and to study various features of the relation (1.6.287) that connects contravariant and covariant transformation properties. Hints: Show that the Poisson bracket relations (1.7.17) are preserved by (extended) Lorentz/Poincaré transformations. Show that the linear part M of the phase-space map is of the form (3.3.11) and that the condition (3.3.13) is satisfied. Given a Lorentz transformation, what is the f_2 for the corresponding symplectic map associated with the (extended) Lorentz transformation? Also see Exercise 3.7.36.

6.3 Preservation of General Poisson Brackets

Let $\mathcal{M}$ be a symplectic mapping of z to $\overline{z}$, and let $\mathcal{M}^{-1}$ be its inverse,

$$\mathcal{M}: \quad z \to \overline{z} = \overline{z}(z,t), \tag{6.3.1}$$

$$\mathcal{M}^{-1}: \quad \overline{z} \to z = z(\overline{z},t). \tag{6.3.2}$$

Suppose we view these relations as a transformation of variables. Let $f(z,t)$ be any dynamical variable. Then the map $\mathcal{M}^{-1}$ given by (3.2) produces a *transformed* dynamical variable $f^*(\overline{z},t)$ (a function of the transformed phase-space variables $\overline{z}$ and perhaps the time t) by the rule

$$\mathcal{M}^{-1}: \quad f(z,t) \to f^*(\overline{z},t) = f(z(\overline{z},t),t). \tag{6.3.3}$$

Conversely, if $f^*(\overline{z},t)$ is any function of the transformed phase-space variables $\overline{z}$ and perhaps the time t, then the map $\mathcal{M}$ given by (3.1) produces the dynamical variable $f(z,t)$, involving the original phase-space variables, by the rule

$$\mathcal{M}: \quad f^*(\overline{z},t) \to f(z,t) = f^*(\overline{z}(z,t),t). \tag{6.3.4}$$

In either case, we have the common relation

$$f^*(\overline{z},t) = f(z,t). \tag{6.3.5}$$

The matter of transforming dynamical variables can also be viewed from a somewhat different perspective. Suppose $f^{\text{old}}(z,t)$ is some dynamical variable involving the original phase-space variables and perhaps the time t. Then the map $\mathcal{M}$ given by (3.1) produces a *new* dynamical variable $f^{\text{new}}(z,t)$ of the *original* phase-space variables by the rule

$$\mathcal{M}: \quad f^{\text{old}}(z,t) \to f^{\text{new}}(z,t) = f^{\text{old}}(\overline{z}(z,t),t). \tag{6.3.6}$$

In this case, $\overline{z}$ is to be regarded as a transformed point in the same phase space as the original point z. By contrast, in the relations (3.3) through (3.5), the points $\overline{z}$ and z may be regarded as members of two different phase spaces.

We now examine the relation between original and transformed dynamical variables and their Poisson brackets. Suppose f and g are any two dynamical variables. For clarity, it is convenient to introduce the notation

$$[f, g]_z = (\partial_z f, J\partial_z g), \tag{6.3.7}$$

$$[f, g]_{\overline{z}} = (\partial_{\overline{z}} f, J\partial_{\overline{z}} g). \tag{6.3.8}$$

See (5.1.4). Then the relation (1.15), and the fact that $\mathcal{M}^{-1}$ is a symplectic map if (and only if) $\mathcal{M}$ is also a symplectic map, can be written in the more precise form

$$J_{ab} = [\overline{z}_a, \overline{z}_b]_z = [z_a, z_b]_z = [\overline{z}_a, \overline{z}_b]_{\overline{z}} = [z_a, z_b]_{\overline{z}}. \tag{6.3.9}$$

Now consider $f^*(\overline{z}, t)$, and its counterpart $g^*(\overline{z}, t)$, as defined by (3.3). We claim that corresponding to the relations (3.3) and (3.4) there are the relations

$$[f^*, g^*]_{\overline{z}} = [f, g]_z|_{z=z(\overline{z},t)}, \tag{6.3.10}$$

$$[f, g]_z = [f^*, g^*]_{\overline{z}}|_{\overline{z}=\overline{z}(z,t)}, \tag{6.3.11}$$

respectively. We will prove (3.10) in a moment, and the proof of (3.11) is similar. Note that the relations (3.10) and (3.11) indicate that the operations of transforming variables and Poisson bracketing are interchangeable. That is, we may first Poisson bracket two functions and then change variables, or we may first change variables, and then Poisson bracket with respect to the transformed variables. In this sense, Poisson brackets in general are preserved under symplectic maps.

The proof of (3.10) makes used of the chain rule and the symplectic condition. From (3.3) and (3.8) we have the relations

$$[f^*, g^*]_{\overline{z}} = (\partial_{\overline{z}} f^*, J\partial_{\overline{z}} g^*) = (\partial_{\overline{z}} f, J\partial_{\overline{z}} g). \tag{6.3.12}$$

By the chain rule there is the relation

$$\partial f/\partial \overline{z}_a = \sum_b (\partial f/\partial z_b)(\partial z_b/\partial \overline{z}_a). \tag{6.3.13}$$

However, (2.3) can be rewritten in the form

$$dz_b = \sum_c (M^{-1})_{bc} d\overline{z}_c, \tag{6.3.14}$$

and it follows that there is the relation

$$\partial z_b/\partial \overline{z}_a = (M^{-1})_{ba}. \tag{6.3.15}$$

Consequently, (3.13) can be written also in the form

$$\partial f/\partial \overline{z}_a = \sum_b [(M^T)^{-1}]_{ab} (\partial f/\partial z_b). \tag{6.3.16}$$

This relation has the compact form

$$\partial_{\overline{z}} f = (M^T)^{-1} \partial_z f. \tag{6.3.17}$$

Upon inserting (3.17) and its counterpart for g into (3.12), we find the advertised result,

$$\begin{aligned} [f^*, g^*]_{\overline{z}} &= (\partial_{\overline{z}} f, J \partial_{\overline{z}} g) = ([M^T]^{-1} \partial_z f, J [M^T]^{-1} \partial_z g) \\ &= (\partial_z f, [M^{-1} J (M^T)^{-1}] \partial_z g) = (\partial_z f, J \partial_z g) \\ &= [f, g]_z. \end{aligned} \tag{6.3.18}$$

Here we have used the symplectic condition for M^{-1} in the form

$$M^{-1} J (M^T)^{-1} = J. \tag{6.3.19}$$

The reader is urged to prove (3.11) in an analogous fashion. Finally, she or he should also prove the related result for *old* and *new* functions as given by (3.6),

$$[f^{\text{new}}(z,t), g^{\text{new}}(z,t)]_z = [f^{\text{old}}(\overline{z},t), g^{\text{old}}(\overline{z},t)]_{\overline{z}}|_{\overline{z}=\overline{z}(z,t)}. \tag{6.3.20}$$

Exercises

6.3.1. Prove the relations (3.11) and (3.20).

6.3.2. Suppose that $h(z,t)$ is any function of the phase-space variables z and the time t. Let z be related to $\overline{z}$ by the symplectic map $\mathcal{M}^{-1}$ as in (3.2). Prove the relation

$$[\overline{z}_a(z,t), h(z,t)]_z = [\overline{z}_a, h(z(\overline{z},t),t)]_{\overline{z}}. \tag{6.3.21}$$

Prove also the relation

$$[z_a(\overline{z},t), h(\overline{z},t)]_{\overline{z}} = [z_a, h(\overline{z}(z,t),t)]_z. \tag{6.3.22}$$

Hint: Use (5.1.4), (1.2), and (3.15) to show that the left side of (3.21) can be written in the form

$$[\overline{z}_a(z,t), h(z,t)]_z = (MJ\partial_z h)_a, \tag{6.3.23}$$

and the right side can be written in the form

$$[\overline{z}_a, h(z(\overline{z},t),t)]_{\overline{z}} = (J(M^T)^{-1}\partial_z h)_a. \tag{6.3.24}$$

Then demonstrate and use the relation

$$MJ = J(M^T)^{-1}, \tag{6.3.25}$$

which is a consequence of the symplectic condition. Alternatively, use the identity

$$\overline{z}_a^*(\overline{z},t) = \overline{z}_a(z(\overline{z},t),t) = \overline{z}_a \tag{6.3.26}$$

and the relation (3.11), etc.

6.4 Relation to Hamiltonian Flows

Let $H(z,t)$ be the Hamiltonian for some dynamical system. Consider a large Euclidean space with $2n+1$ axes labeled by the phase-space variables $z_1 \cdots z_{2n}$ and the time t. We will call this construction *augmented phase space*.[12] See Figure 4.1. Suppose the $2n$ quantities $z_1(t^i) \cdots z_{2n}(t^i)$ are specified at some *initial* time t^i. Then the quantities $z_1(t) \cdots z_{2n}(t)$ at some other time t are uniquely determined by the initial conditions $z_1(t^i) \cdots z_{2n}(t^i)$ and Hamilton's equations of motion (5.2.2). Recall Theorem 1.3.1 and review Section 1.4. The set of all trajectories in augmented phase space for all possible initial conditions will be called a *Hamiltonian flow*. Indeed we know that no two trajectories can intersect. (See Exercise 1.3.6.) Therefore the behavior of the trajectories in augmented phase space is analogous to fluid flow in a high dimensional space.

6.4.1 Hamiltonian Flows Generate Symplectic Maps

Let t^i be some *initial* time, and let t^f be some other *final* time. Also, let z^i denote the set of quantities $z_1(t^i) \cdots z_{2n}(t^i)$, and let z^f denote the corresponding set $z_1(t^f) \cdots z_{2n}(t^f)$. We have already seen in Section 1.4 that the relation between z^i and z^f can be viewed as a transfer map $\mathcal{M}(t^i, t^f)$ depending on the parameters t^i and t^f. [Indeed, since the set of trajectories in augmented phase space is equivalent to a knowledge of $\mathcal{M}(t^i, t^f)$ for variable t^f, a flow for a differential equation may be equally well, and often is, *defined* to be the family of such maps.] What we will now see is that $\mathcal{M}$ is a *symplectic* map.

Theorem 4.1 Let $H(z,t)$ be the Hamiltonian for some dynamical system, and let z^i denote a set of initial conditions at some initial time t^i. Also, let z^f denote the coordinates at some final time t^f of the trajectory with initial conditions z^i. Finally, let $\mathcal{M}$ denote the mapping from z^i to z^f obtained by following the Hamiltonian trajectory specified by H,

$$\mathcal{M} : z^i \rightarrow z^f. \tag{6.4.1}$$

Then the mapping $\mathcal{M}$ is symplectic.

Proof Suppose the flow takes place for a time interval of duration T so that t^i and t^f are related by the equation

$$t^f = t^i + T. \tag{6.4.2}$$

Divide the interval T into N small steps each of duration h. Evidently, T, N, and h are related by the equation

$$T = Nh. \tag{6.4.3}$$

Also, define intermediate times t^m at each step by the rule

$$t^0 = t^i,$$

[12] Some authors, following Cartan, call it *state space*. But other authors use *state space* and *phase space* interchangeably. Still other authors call this construction *extended phase space*. However, we have already used that term to describe ordinary phase space augmented by the *two* additional variables t and p_t. Review Exercise 1.6.5.

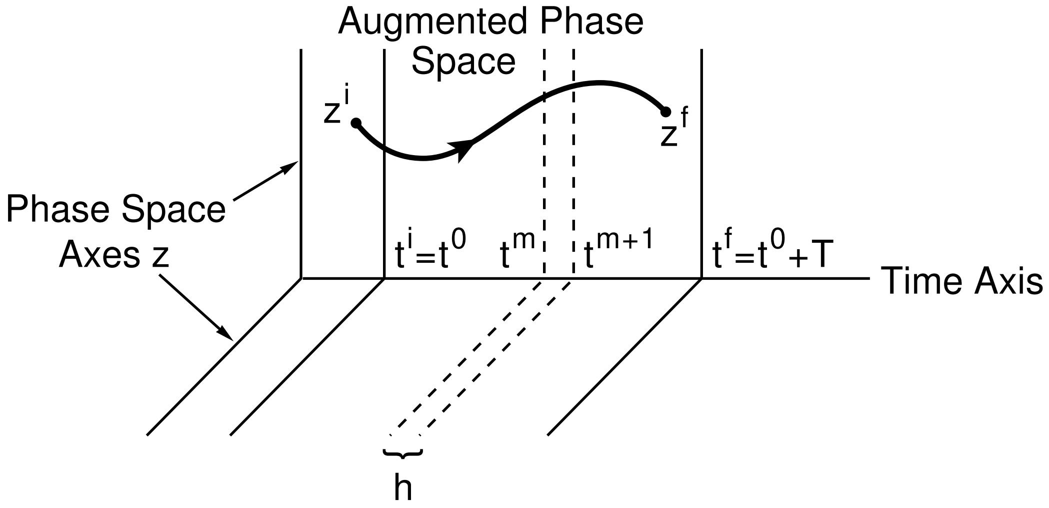


Figure 6.4.1: A trajectory in augmented phase space. Under the Hamiltonian flow specified by a Hamiltonian H, the general phase-space point z^i is mapped into the phase-space point z^f. The mapping $\mathcal{M}$ is symplectic for any Hamiltonian.

$$t^m = t^0 + mh \ \ , \ \ m = 0, 1, \cdots N, \qquad (6.4.4)$$
$$t^N = t^f.$$

Suppose that the mapping $\mathcal{M}$ is viewed as a composite of mappings between adjacent times t^m and t^{m+1}. That is, $\mathcal{M}$ is written in the form

$$\mathcal{M} = \mathcal{M}^{t^f \leftarrow t^{N-1}} \cdots \mathcal{M}^{t^{m+1} \leftarrow t^m} \cdots \mathcal{M}^{t^1 \leftarrow t^i} \qquad (6.4.5)$$

with the notation that $\mathcal{M}^{t^{m+1} \leftarrow t^m}$ denotes the mapping between the quantities

$$z^m = \{z_1(t^m), \cdots z_{2n}(t^m)\}$$

and

$$z^{m+1} = \{z_1(t^{m+1}), \cdots z_{2n}(t^{m+1})\}. \qquad (6.4.6)$$

Corresponding to the relation (4.5), the Jacobian matrix M of the mapping $\mathcal{M}$ can be written using the chain rule in the product form

$$M = M^{t^f \leftarrow t^{N-1}} \cdots M^{t^{m+1} \leftarrow t^m} \cdots M^{t^1 \leftarrow t^i}, \qquad (6.4.7)$$

where, as the notation is meant to indicate, $M^{t^{m+1} \leftarrow t^m}$ is the Jacobian matrix for the map $\mathcal{M}^{t^{m+1} \leftarrow t^m}$.

Next it will be shown that each matrix in the product (4.7) is symplectic at least through terms of order h. According to Taylor's series, the relation between z^{m+1} and z^m can be written in the form

$$\begin{aligned} z_a^{m+1} &= z_a(t^{m+1}) = z_a(t^m + h) \\ &= z_a(t^m) + h\dot{z}_a(t^m) + O(h^2) \\ &= z_a^m + h(J\partial_z H)_a + O(h^2). \end{aligned} \qquad (6.4.8)$$

Here use has also been made of the equations of motion (5.2.2). Suppose (4.8) is used to compute the associated Jacobian matrix. The result of this computation is the relation

$$\begin{aligned} M_{ab}^{t^{m+1}\leftarrow t^m} &= \partial z_a^{m+1}/\partial z_b^m \\ &= \delta_{ab} + h\sum_c J_{ac}\partial^2 H/\partial z_c \partial z_b + O(h^2). \end{aligned} \tag{6.4.9}$$

Using matrix notation, (4.9) can be written more compactly in the form

$$M^{t^{m+1}\leftarrow t^m} = I + hJS + O(h^2), \tag{6.4.10}$$

where S is the *symmetric* matrix

$$S_{cb} = \partial^2 H/\partial z_c \partial z_b. \tag{6.4.11}$$

Now compare (4.10) with (3.7.28) and (3.7.34). Evidently, the Jacobian matrix (4.10) is a *symplectic* matrix at least through terms of order h.

The desired proof is almost complete. Since symplectic matrices form a group, the product matrix M given by (4.7) differs from a symplectic matrix by terms at most of order Nh^2 because each of the N terms in the product differs from a symplectic matrix by terms at most of order h^2. Now take the limit $h \to 0$ and $N \to \infty$. In this limit terms proportional to Nh^2 vanish since, using (4.3),

$$Nh^2 = (T/h)h^2 = Th, \tag{6.4.12}$$

and the quantity Th vanishes as h goes to zero. It follows that M is a symplectic matrix, and $\mathcal{M}$ is a symplectic map.

What has been shown is that the problem of describing and following Hamiltonian trajectories, which is one of the fundamental aspects of classical mechanics, is equivalent to the problem of representing and calculating symplectic maps. We remark that there is another proof of the result just obtained based on the use of variational equations. It is shorter, but perhaps less instructive. See Exercise 4.3.

In the proof just given, suppose we regard the final time t^f as a general time t. What we have found is that following trajectories specified by H produces a symplectic map $\mathcal{M}(t^i, t)$ for each value of t. Thus, we have produced a *one-parameter family* of symplectic maps $\mathcal{M}(t^i, t)$. Moreover, we have the initial condition

$$\mathcal{M}(t^i, t^i) = \mathcal{I} \tag{6.4.13}$$

where $\mathcal{I}$ denotes the identity map. We describe this state of affairs by saying that the family $\mathcal{M}(t^i, t)$ is *generated* by the Hamiltonian $H(z, t)$ starting from the identity map $\mathcal{I}$ when $t = t^i$.

It can be verified that the set of symplectic maps generated by Hamiltonians forms a group which may be regarded as a subgroup of the set of all symplectic maps. This group is sometimes referred to as $Ham(n)$ where $2n$ is the dimensionality of the phase space under consideration.

6.4.2 Any Family of Symplectic Maps Is Hamiltonian Generated

Consider the space of all symplectic maps. We may regard the $\mathcal{M}(t^i, t)$, for variable t, as a path in this space. And, according to (4.13), the starting point of this path, namely $\mathcal{M}(t^i, t^i)$, is the identity map. Therefore the $\mathcal{M}(t^i, t)$ form a one-parameter family continuously connected to the identity. Is there a converse result? There is.

Theorem 4.2 Suppose we are given a one-parameter family of symplectic maps $\mathcal{N}(t)$ for $t \in [t^i, t^f]$. Let $\mathcal{N}_i$ denote the map

$$\mathcal{N}_i = \mathcal{N}(t^i). \tag{6.4.14}$$

Then there is a *generating* Hamiltonian G that generates this family starting from the map $\mathcal{N}_i$.[13] See Figure 4.2. It depicts *augmented symplectic map space*, which consists of a time axis and multiple additional axes that provide coordinates for points in the space of all symplectic maps. Let $\bar{z}(z, t)$ be the result of $\mathcal{N}(t)$ acting on the general phase-space point z,

$$\mathcal{N}(t): \ z \rightarrow \bar{z}(z, t). \tag{6.4.15}$$

What we want to show is that there is a function $G(\bar{z}; t)$ such that

$$(\partial \bar{z}_a / \partial t)|_z = [\bar{z}_a, G(\bar{z}; t)]_{\bar{z}}. \tag{6.4.16}$$

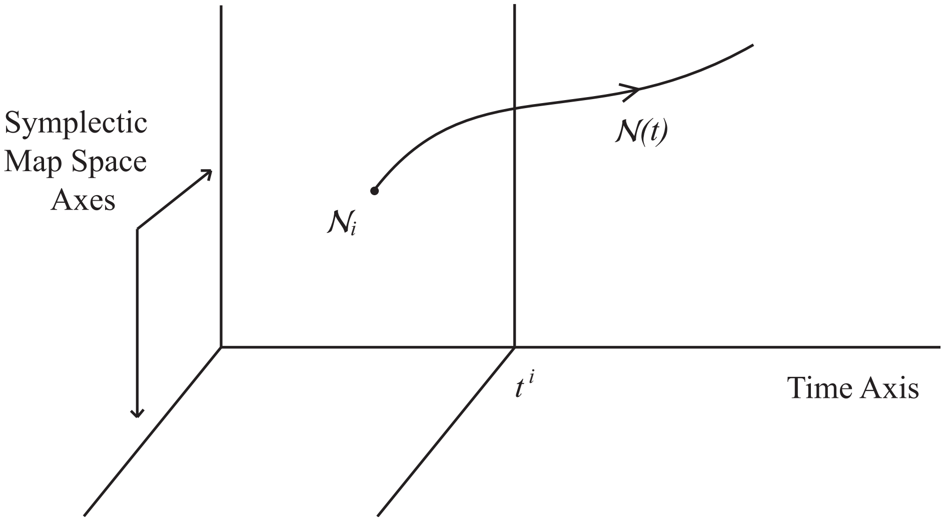


Figure 6.4.2: The symplectic map family $\mathcal{N}(t)$ in augmented symplectic map space.

Proof The proof proceeds by construction. Express the mapping (4.15) in the explicit component form

$$\bar{z}_a(t) = u_a(z, t) \tag{6.4.17}$$

[13] Here we apologize that the symbol G has also been used in Subsection 1.1, and again will be used subsequently, to denote the Jacobian of the gradient map $\mathcal{G}$. There are not always enough letters to go around. The reader should be able to determine from the context what is meant in any particular case.

where the u_a with $a = 1$ to $2n$ are assumed to be known functions of z and t. Next form the functions $w_a(z,t)$ defined by the relations

$$w_a(z,t) = \partial u_a(z,t)/\partial t. \tag{6.4.18}$$

By these definitions we have the equivalent statements

$$(\partial \bar{z}_a/\partial t)|_z = w_a(z,t). \tag{6.4.19}$$

Suppose that $\mathcal{N}$ has an inverse, as will be the case if $\mathcal{N}$ is symplectic. Then, the relations (4.15) or (4.17) can be inverted to give relations of the form

$$z_b = v_b(\bar{z},t). \tag{6.4.20}$$

Now form the functions $g_a(\bar{z},t)$ by using (4.20) in the arguments of the w_a and writing

$$g_a(\bar{z},t) = w_a(z(\bar{z},t),t). \tag{6.4.21}$$

The net result of these steps is the set of relations

$$\partial \bar{z}_a/\partial t = g_a(\bar{z},t). \tag{6.4.22}$$

That is, we have produced a *vector field* $\mathcal{L}_{\boldsymbol{g}}$ defined by the relation

$$\mathcal{L}_{\boldsymbol{g}} = \sum_a g_a(\partial/\partial \bar{z}_a) \tag{6.4.23}$$

and having the property

$$\dot{\bar{z}}_a = \mathcal{L}_{\boldsymbol{g}} \bar{z}_a. \tag{6.4.24}$$

We will now show that this vector field is Hamiltonian. (See Section 5.3.). Consider the quantities $\bar{z}_a(t+\epsilon)$ where ϵ is small. According to Taylor there is the expansion

$$\bar{z}_a(t+\epsilon) = \bar{z}_a(t) + \epsilon(\partial \bar{z}_a/\partial t) + O(\epsilon^2) \tag{6.4.25}$$

or, in view of (4.19),

$$\bar{z}_a(t+\epsilon) = \bar{z}_a(t) + \epsilon w_a(z,t) + O(\epsilon^2). \tag{6.4.26}$$

Let us compute $[\bar{z}_a(t+\epsilon), \bar{z}_b(t+\epsilon)]$. Using (4.26) we find the result

$$[\bar{z}_a(t+\epsilon), \bar{z}_b(t+\epsilon)]_z = [\bar{z}_a(t), \bar{z}_b(t)]_z + \epsilon[\bar{z}_a, w_b]_z + \epsilon[w_a, \bar{z}_b]_z + O(\epsilon^2). \tag{6.4.27}$$

Since $\mathcal{N}$ is symplectic for all $t \in [t^i, t^f]$, there must be the relations

$$[\bar{z}_a(t+\epsilon), \bar{z}_b(t+\epsilon)]_z = [\bar{z}_a(t), \bar{z}_b(t)]_z = J_{ab}. \tag{6.4.28}$$

See (1.15). Therefore, upon equating powers of ϵ, (4.27) provides the relation

$$[\bar{z}_a, w_b]_z + [w_a, \bar{z}_b]_z = 0. \tag{6.4.29}$$

Since symplectic maps preserve Poisson brackets, see Section 3, we may also write (4.29) in the form

$$[\bar{z}_a, g_b(\bar{z}, t)]_{\bar{z}} = [\bar{z}_b, g_a(\bar{z}, t)]_{\bar{z}}. \tag{6.4.30}$$

Here we have used (4.21) and the antisymmetry of the Poisson bracket. Finally, expand out the Poisson brackets using (5.1.3). So doing, for example, gives the result

$$\begin{aligned} [\bar{z}_a, g_b]_{\bar{z}} &= \sum_{cd} (\partial \bar{z}_a/\partial \bar{z}_c) J_{cd} (\partial g_b/\partial \bar{z}_d) \\ &= \sum_{cd} \delta_{ac} J_{cd} (\partial g_b/\partial \bar{z}_d) = \sum_d J_{ad}(\partial g_b/\partial \bar{z}_d). \end{aligned} \tag{6.4.31}$$

The net result is that (4.30) is equivalent to the relation

$$\sum_d J_{ad}(\partial g_b/\partial \bar{z}_d) = \sum_d J_{bd}(\partial g_a/\partial \bar{z}_d). \tag{6.4.32}$$

To make sense of (4.32), multiply both sides by $J_{ac}J_{be}$, sum over a and b, and manipulate to produce the relations

$$\sum_{abd} J_{ac}J_{be}J_{ad}(\partial g_b/\partial \bar{z}_d) = \sum_{abd} J_{ac}J_{be}J_{bd}(\partial g_a/\partial \bar{z}_d), \tag{6.4.33}$$

$$\sum_{abd} (J^T)_{ca}J_{ad}J_{be}(\partial g_b/\partial \bar{z}_d) = \sum_{abd} (J^T)_{eb}J_{bd}J_{ac}(\partial g_a/\partial \bar{z}_d), \tag{6.4.34}$$

$$\sum_{bd} (J^TJ)_{cd}J_{be}(\partial g_b/\partial \bar{z}_d) = \sum_{ad} (J^TJ)_{ed}J_{ac}(\partial g_a/\partial \bar{z}_d), \tag{6.4.35}$$

$$\sum_{bd} \delta_{cd}J_{be}(\partial g_b/\partial \bar{z}_d) = \sum_{ad} \delta_{ed}J_{ac}(\partial g_a/\partial \bar{z}_d), \tag{6.4.36}$$

$$\sum_{b} J_{be}(\partial g_b/\partial \bar{z}_c) = \sum_{a} J_{ac}(\partial g_a/\partial \bar{z}_e), \tag{6.4.37}$$

$$(\partial/\partial \bar{z}_c)\sum_{b} J_{be}g_b = (\partial/\partial \bar{z}_e)\sum_{a} J_{ac}g_a. \tag{6.4.38}$$

Here use has been made of (3.1.6). Now introduce quantities η_c by the rule

$$\eta_c = \sum_a J_{ac}g_a. \tag{6.4.39}$$

In terms of these quantities (4.38) yields the relations

$$\partial \eta_e/\partial \bar{z}_c = \partial \eta_c/\partial \bar{z}_e. \tag{6.4.40}$$

[Note that this condition is a restatement of (5.3.26).] It follows that the quantity $\sum_a \eta_a d\bar{z}_a$ is an *exact* differential. See Exercise 1.1.

Now define G by the phase-space path integral

$$G(\bar{z};t) = \int^{\bar{z}} \sum_a \eta_a dz'_a. \tag{6.4.41}$$

Since the integrand is an exact differential, the integral is path independent and satisfies the relation

$$\partial G/\partial \bar{z}_a = \eta_a. \tag{6.4.42}$$

Let us put everything together. The relations (4.39) can be solved for the g_a to give the result

$$g_a = \sum_b J_{ab}\eta_b. \tag{6.4.43}$$

Upon combining (4.22), (4.42), and (4.43), we find the net result

$$\partial \bar{z}_a/\partial t = \sum_b J_{ab}\partial G/\partial \bar{z}_b, \tag{6.4.44}$$

or, more compactly,

$$\partial \bar{z}/\partial t = J\partial_{\bar{z}}G = [\bar{z}, G(\bar{z};t)], \tag{6.4.45}$$

which is the desired result (4.16).

Even a bit more can be said. Consider the straight-line path $\bar{z}'_a(\tau)$ in phase space that connects the origin to $\bar{z}$. It has the parametric form

$$\bar{z}'_a(\tau) = \tau\bar{z}_a \ , \ \tau \in [0,1]. \tag{6.4.46}$$

Suppose the integrability condition (4.40) holds in a simply-connected region that surrounds this path. Then we may employ this path in (4.41) to obtain the result

$$G(\bar{z};t) = \int_0^1 d\tau \sum_a \eta_a(\tau\bar{z},t)\bar{z}_a = -\int_0^1 d\tau \sum_{ab} J_{ab}\bar{z}_a g_b(\tau\bar{z},t). \tag{6.4.47}$$

Let us recapitulate what has been done. By differentiating the map $\mathcal{N}(t)$ with respect to t, the steps involved in (4.17) through (4.21) produced the vector field $\boldsymbol{\mathcal{L}_g}$ with components g_a. These steps can be carried out for any invertible one-parameter family of maps $\mathcal{N}(t)$. Then we used the symplectic condition in the steps associated with (4.25) through (4.41) to show that the vector field was Hamiltonian and to explicitly construct the Hamiltonian.

There are a few more steps that can be made. In analogy to the work of Subsection 4.1, let $\mathcal{M}(t^i,t)$ be the map generated by the $G(z;t)$ constructed in this subsection and with the initial condition (4.13). Then, using the methods of Section 10.1, it can be verified that there is the relation

$$\mathcal{N}(t) = \mathcal{N}_i\mathcal{M}(t^i,t). \tag{6.4.48}$$

We have found an explicit expression for $\mathcal{N}$ in terms of the map generated by its associated Hamiltonian. Moreover, differentiating (4.48), and again using the methods of Section 10.1, gives the result

$$\dot{\mathcal{N}}(t) = \mathcal{N}_i\dot{\mathcal{M}}(t^i,t) = \mathcal{N}_i\mathcal{M}(t^i,t) : -G := \mathcal{N}(t) : -G :, \tag{6.4.49}$$

from which we conclude that

$$: -G := \mathcal{N}^{-1}\dot{\mathcal{N}}(t). \tag{6.4.50}$$

In summary, given any family of symplectic maps, we have shown that this family is generated by a Hamiltonian G starting from an initial map $\mathcal{N}_i$, have explicitly constructed G, and have found a representation for $\mathcal{N}$.

6.4.3 Almost All Symplectic Maps Are Hamiltonian Generated

In Section 7.9 we will learn that under rather general low-order differentiability conditions any symplectic map can be connected to the identity map by a one-parameter family of symplectic maps. This one-parameter family can then be used to construct a Hamiltonian, and we can also set

$$\mathcal{N}_i = \mathcal{I}. \tag{6.4.51}$$

We conclude that, under mild assumptions, any symplectic map can be generated by a Hamiltonian starting from the identity map. Thus, for our purposes, we will generally not make a distinction between $ISpM(2n, \mathbb{R})$ [$= Symp(n)$] and $Ham(n)$.

6.4.4 Transformation of a Hamiltonian Under the Action of a Symplectic Map

Suppose $H(z;t)$ is the Hamiltonian governing the motion of some system described by the canonical coordinates z. Suppose we wish to introduce new canonical coordinates $\bar{z}(z,t)$ that are related to the old coordinates z by a (possibly time dependent) symplectic map $\mathcal{N}(t)$ as in (4.15). The purpose of this subsection is to show that the motion of the system, when described by the new coordinates $\bar{z}$, is also governed by a new Hamiltonian that we will call $K(\bar{z};t)$, and to find the relation between K and the old Hamiltonian H.

We proceed as follows: View the quantities $\bar{z}_a(z,t)$ as dynamical variables. Then, by the work of Section 1.7 that defined the Poisson bracket, there is the result

$$d\bar{z}_a(z,t)/dt = \partial\bar{z}_a(z,t)/\partial t + [\bar{z}_a(z,t), H(z;t)]_z. \tag{6.4.52}$$

Here we have placed a subscript z on the Poisson bracket to make it clear that the Poisson bracket is taken with respect to the variable z. But, since $\bar{z}$ and z are related by the *symplectic* map $\mathcal{N}$, it follows from the invariance property of the Poisson bracket that

$$[\bar{z}_a(z,t), H(z;t)]_z = [\bar{z}_a, H(z(\bar{z},t);t)]_{\bar{z}}. \tag{6.4.53}$$

See Section 3. Moreover, from the work of Subsection 4.2, we know that there is a generating Hamiltonian $G(\bar{z};t)$ for $\mathcal{N}$ such that

$$\partial\bar{z}_a(z,t)/\partial t = [\bar{z}_a, G(\bar{z};t)]_{\bar{z}}. \tag{6.4.54}$$

Upon combining (4.52) through (4.54) we see that there is the relation

$$d\bar{z}_a/dt = [\bar{z}_a, K(\bar{z};t)]_{\bar{z}} \tag{6.4.55}$$

where

$$K(\bar{z};t) = H(z(\bar{z},t);t) + G(\bar{z};t) \tag{6.4.56}$$

so that $K(\bar{z};t)$ is the desired new Hamiltonian. We note that if $\mathcal{N}$ is time independent, then $G = 0$. See (4.50). In this case the new Hamiltonian is simply the old Hamiltonian expressed in terms of the new variables,

$$K(\bar{z};t) = H(z(\bar{z});t). \tag{6.4.57}$$

Exercises

6.4.1. Use (1.2) and the chain rule to verify (4.7).

6.4.2. Show that the matrix S defined by (4.11) is indeed symmetric.

6.4.3. The purpose of this exercise is to provide another proof of Theorem 4.1: Hamiltonian flows generate symplectic maps. Refer to Exercise 1.4.6. From Hamilton's equations of motion written in the form (5.2.3), show, for the associated variational equations, that the A matrix of (1.4.51) is given by

$$A = JS \tag{6.4.58}$$

with S given by (4.11). Next show that in terms of a general final time t the Jacobian matrix M satisfies the differential equation

$$\dot{M}(t) = JS(t)M(t) \tag{6.4.59}$$

with the initial condition

$$M(t^i) = I. \tag{6.4.60}$$

Now consider the matrix product M^TJM. Because of (4.59), it satisfies the differential equation

$$\begin{aligned}(d/dt)[M^T(t)JM(t)] &= \dot{M}^TJM + M^TJ\dot{M}\\ &= [JSM]^TJM + M^TJJSM = -M^TSJJM + M^TJJSM\\ &= M^TSM - M^TSM = 0.\end{aligned} \tag{6.4.61}$$

Thus, in view of (4.60), this equation has the unique solution

$$M^T(t)JM(t) = J, \tag{6.4.62}$$

and we conclude that the Jacobian matrix must be symplectic.

6.4.4. Show that the maps between q, p and Q, P given by (1.4.9) is symplectic. Show that the map given by (1.4.13) between Q^i, P^i and Q^f, P^f is symplectic. In both cases, find the associated Jacobian matrix M and verify that it is symplectic.

6.4.5. Show that the maps given by (1.4.22), (1.4.23) and (1.4.24), (1.4.25) are symplectic. In both cases, find the associated Jacobian matrix M and verify that it is symplectic.

6.4.6. Suppose $H(z,t)$ is a possibly time-dependent quadratic Hamiltonian written, without loss of generality, in the form

$$H(z,t) = (1/2)(z, Sz) \tag{6.4.63}$$

where S is a symmetric and possibly time-dependent matrix. Verify that the equations of motion generated by this H are linear, and therefore that the associated transfer map is linear and can be described by a matrix M. Use the machinery of Exercise 4.3 above to show that $M(t,t_0)$ is symplectic. Here t is a general time and t_0 is some initial time such that

$$M(t_0,t_0) = I. \tag{6.4.64}$$

Let u_0 and v_0 be two initial conditions. Then, for these initial conditions, verify that the associated solutions to the equations of motion are given by the relations

$$u(t) = M(t,t_0)u_0, \tag{6.4.65}$$

$$v(t) = M(t,t_0)v_0. \tag{6.4.66}$$

Form the quantity $C(u,v)$ by the rule

$$C(u,v) = (u, Jv). \tag{6.4.67}$$

Show that

$$C(u,v) = C(u_0,v_0), \tag{6.4.68}$$

and therefore C is *constant* (time independent) and depends only on the initial conditions.

Consider the set of differential equations arising from any Hamiltonian and, for any particular trajectory, form the associated variational equations. Show that any two solutions u and v to the variational equations also satisfy (4.68).

6.4.7. Consider the one-parameter family of maps

$$\bar{z}_1(z,t) = z_1 \cos t - z_2 \sin t, \tag{6.4.69}$$

$$\bar{z}_2(z,t) = z_1 \sin t + z_2 \cos t. \tag{6.4.70}$$

Verify that these maps are symplectic. Find the Hamiltonian that generates this family of maps.

6.4.8. Consider the two-parameter family of maps (called the general Hénon map, see Section 19.7) given by the relations

$$\bar{q} = 1 + p - aq^2, \tag{6.4.71}$$

$$\bar{p} = bq. \tag{6.4.72}$$

Show that the inverse of this map is given by the relations

$$q = \bar{p}/b, \tag{6.4.73}$$

$$p = \bar{q} - 1 + a(\bar{p}/b)^2. \tag{6.4.74}$$

Show that if b is held fixed and a is treated as a variable parameter, then the resulting one-parameter family of maps is generated by the vector field

$$\mathcal{L} = -(\bar{p}/b)^2(\partial/\partial\bar{q}) =: [1/(3b^2)]\bar{p}^3 : . \tag{6.4.75}$$

Note that this vector field is Hamiltonian even though the general Hénon map is symplectic only when $b = -1$. Show that if a is held fixed and b is treated as a variable parameter, then the resulting one-parameter family of maps is generated by the vector field

$$\mathcal{L} = (1/b)\bar{p}(\partial/\partial\bar{p}). \tag{6.4.76}$$

This vector field is not Hamiltonian. See Section 18.3.

6.4.9. Newton's equation of motion for an harmonic oscillator consisting of a mass m and a spring with spring constant k is given by the relation

$$d^2x/dt^2 + (k/m)x = 0 \tag{6.4.77}$$

where x is the difference between the actual and natural lengths of the spring. Introduce the notation

$$K = k/m, \tag{6.4.78}$$

and consider the possibility that K is time dependent so that (4.77) becomes

$$d^2x/dt^2 + K(t)x = 0. \tag{6.4.79}$$

If K is in fact time dependent, the harmonic oscillator is said to be *parametrically driven.*

The purpose of this exercise is to explore some aspects of the behavior of a parametrically driven harmonic oscillator. The behavior of a parametrically driven harmonic oscillator can be very complicated, and there is a vast literature on the subject. If K is *periodic*, (4.79) is a form of *Hill's* equation. If K is periodic and consists of only a constant term and a square wave, (4.79) becomes *Meissner's* equation. If K is periodic and consists of only a constant term and a rectangular wave, (4.79) becomes the *Kronig-Penney* model. If K is periodic and consists of only a constant term and a string of equally spaced delta function spikes, (4.79) becomes the *Dirac comb* or *periodic* delta function model. If K is periodic and consists of only a constant term and a sinusoidal term, (4.79) becomes a form of *Mathieu's* equation. Mathieu functions will play an important role in Section 17.4. The general periodic case, in essence Hill's equation, is important for the subject of *strong focussing* in Accelerator Physics. It is also important for many other areas of physics including band theory in Condensed Matter Physics, the motion of the Moon (the context in which Hill formulated and studied his equation), and wave-guide theory. [14]

[14] It is interesting to note that George William Hill (1838-1914) did not hold any permanent academic appointment. For ten years of his life he was a clerk at the U. S. National Bureau of Standards (NBS, now NIST, the National Institute of Standards and Technology) working long hours and doing his own research at home at night. Much of the rest of his life was spent working only at home. He went unappreciated by his colleagues for many years. When Poincaré (along with Darboux, Picard, and Boltzmann) visited the United States in 1904 to lecture at the St. Louis Mathematics Congress held in connection with St. Louis hosting

Your first task is to show that the equation of motion (4.79) arises from a Hamiltonian. Define p by the rule

$$p = dx/dt. \tag{6.4.80}$$

Show that (4.79) and (4.80) are generated by the Hamiltonian

$$H = (1/2)[p^2 + K(t)x^2] \tag{6.4.81}$$

where p and x are taken to be canonically conjugate.

Consider, as a specific example, the Mathieu case for which the Fourier series for $K(t)$ only has two terms,

$$K(t) = K_0 + K_1 \cos(\Omega t + \phi). \tag{6.4.82}$$

In this case (4.79) takes the form

$$d^2x/dt^2 + [K_0 + K_1 \cos(\Omega t + \phi)]x = 0. \tag{6.4.83}$$

The quantity K_0 describes the natural frequency of the oscillator,

$$\omega = \sqrt{K_0}, \tag{6.4.84}$$

where it is assumed that $K_0 > 0$, and K_1 describes the parametric driving strength. Compare (4.83) with the standard form of the Mathieu equation given by (17.4.22). Make the change of variable

$$\tau = \Omega t + \phi, \tag{6.4.85}$$

and verify that this change of variable brings (4.83) to the form

$$d^2x/d\tau^2 + [\bar{K}_0 + \bar{K}_1 \cos(\tau)]x = 0 \tag{6.4.86}$$

where

$$\bar{K}_0 = K_0/\Omega^2 = \omega^2/\Omega^2, \tag{6.4.87}$$

$$\bar{K}_1 = K_1/\Omega^2. \tag{6.4.88}$$

In the case that $K(t)$ is periodic, the solution to (4.79), and hence also to (4.86), can be described in terms of a stroboscopic map. See Section 1.4.3. Moreover, since the equations of motion are linear and are generated by a Hamiltonian, namely (4.81), the stroboscopic map will be linear and symplectic. See Exercise 4.6 above. Introduce the notation

$$z = (x, p). \tag{6.4.89}$$

a World's Fair, the one American mathematician he sought out was Hill. [After the congress these four foreign speakers boarded a train to Washington D.C. (where NBS was located) to attend a reception hosted by President Theodore Roosevelt, followed by subsequent stops at Harvard and Columbia Universities, before sailing back to Europe.] Ernest Brown, in his 1915 National Academy of Sciences Biographical Memoir of Hill, wrote "Hill's 1877 publication 'Researches in the Lunar Theory' of but fifty quarto pages has become fundamental for the development of celestial mechanics in three different directions. It would be difficult to say as much for any other publication of its length in the whole range of modern mathematics, pure or applied. Poincaré's remark that in it we may perceive the germ of all the progress which has been made in celestial mechanics since its publication is doubtless fully justified".

Let z^i be the initial condition at the beginning of a drive period ($\tau = 0$) and let z^f be the final condition at the end of a drive period ($\tau = 2\pi$). Then we may write, in the case of the equation of motion (4.86), the relation

$$z^f = Mz^i \tag{6.4.90}$$

where M is a 2×2 symplectic matrix to be determined. In writing (4.90), because of the variable change (4.85), we take the associated Hamiltonian to be that given by (4.81) with K replaced by $\bar{K}$ where $\bar{K} = K/\Omega^2$. Also, (4.80) is replaced by $p = dx/d\tau$.

In the case that $\bar{K}_1 = 0$, the equation of motion (4.86) can be solved in terms of trigonometric functions. Show that in this case the matrix M, which describes the stroboscopic map, is given by the relation

$$M = \begin{pmatrix} \cos 2\pi\bar{\omega} & (1/\bar{\omega}) \sin 2\pi\bar{\omega} \\ -\bar{\omega} \sin 2\pi\bar{\omega} & \cos 2\pi\bar{\omega} \end{pmatrix} \tag{6.4.91}$$

where

$$\bar{\omega} = \sqrt{\bar{K}_0} = \omega/\Omega. \tag{6.4.92}$$

Verify that the eigenvalues of M lie on the unit circle and have the values

$$\lambda_{\pm} = \exp(\pm 2\pi i \bar{\omega}). \tag{6.4.93}$$

Now supposed that $\bar{K}_1$ takes on small nonzero values. Then the eigenvalues of M will remain on the unit circle provided they were originally not too close to the values ± 1. On the other hand, they could leave the unit circle if originally they were close to or had the values ± 1. Recall Figures 3.4.1 and 3.4.3. The eigenvalues have the value $+1$ when

$$2\pi\bar{\omega} = 2n\pi \iff \bar{\omega} = n, \tag{6.4.94}$$

and have the value -1 when

$$2\pi\bar{\omega} = \pi + 2n\pi \iff \bar{\omega} = n + 1/2. \tag{6.4.95}$$

Here $n = 0, 1, 2, \cdots$. Finally, verify that combining (4.92), (4.94), and (4.95) yields the conditions

$$\omega = n\Omega \text{ or } \Omega = \omega/n \text{ with } n = 1, 2, \cdots , \tag{6.4.96}$$

$$\omega = (n + 1/2)\Omega \text{ or } \Omega = \omega/(n + 1/2) \text{ with } n = 0, 1, 2, \cdots . \tag{6.4.97}$$

Note that in (4.96) we have excluded the case $n = 0$ since the case $\omega = 0$ requires more refined analysis.

When its eigenvalues are off the unit circle, repeated application of the matrix M leads to exponential growth. See Subsections 3.4.5 and 3.5.8. Verify that the conditions (4.96) and (4.97) for possible instability can be combined to yield the *parametric resonance* conditions

$$\Omega = 2\omega/m \Leftrightarrow \omega = (m/2)\Omega \Leftrightarrow 1/\Omega = m(1/2)(1/\omega) \text{ with } m = 1, 2, \cdots . \tag{6.4.98}$$

Verify in this latter formulation that odd values of m correspond to the possibility of the eigenvalues leaving the unit circle through the value -1, and even values of m correspond to the possibility of the eigenvalues leaving the unit circle through the value $+1$.

A pendulum of length ℓ in a gravitational field g has a small-amplitude natural frequency $\omega = (g/\ell)^{1/2}$. Show that, according to (4.98), the trapeze artist Jules Léotard (1838-1870) could increase the amplitude of his swing by alternatively crouching down and then standing up with frequency $\Omega = 2\omega$.[15] Also, like a child on a swing, he could do so with frequency $\Omega = \omega$. Remarkably, he could also do so with the subharmonic frequencies $\Omega = (2/3)\omega$, $\Omega = (2/4)\omega$, $\cdots$. The first choice $\Omega = 2\omega$ is used by professionals and is the most effective. We know from childhood experience with pumping swings that the second choice also works pretty well, and is easier for mortals. The other choices produce successively slower amplitude growths.

6.4.10. The purpose of this exercise is to explore the difference between *forcefully* and parametrically driven harmonic oscillators. Review Exercise 4.9 above. By forcefully driven we mean an oscillator described by an equation of motion of the form

$$d^2x/dt^2 + \beta dx/dt + \omega^2 x = d\cos(\Omega t + \psi). \tag{6.4.99}$$

When $\beta > 0$ the motion of this oscillator is bounded for all values of Ω. It is also bounded when $\beta = 0$ provided $\Omega \neq \omega$. See Section 28.2. Verify, when $\beta = 0$ and $\Omega = \omega$, that (4.99) has the solution

$$x(t) = [d/(2\omega)]t\sin(\omega t + \psi). \tag{6.4.100}$$

Thus, *exactly* at resonance and in the absence of damping, the amplitude of a forcefully driven harmonic oscillator grows *linearly* in time. By contrast it can be shown, in accord with the results of Exercise 4.9, that the amplitude of a parametrically driven oscillator described by the Mathieu equation grows *exponentially* in time when K_1 is small and any of the parametric resonance conditions (4.98) is approximately satisfied.

Moreover, even if the parametrically driven oscillator is damped by adding a term of the form $\beta dx/dt$ (with $\beta > 0$) to the left side of (4.83), it can be shown that there is still a range of K_1 and Ω values for which the amplitude grows exponentially in time. Thus, parametric driving can overcome damping, and can do so even for a range of K_1 and Ω values. That is, there is no resonance condition that needs to be met exactly. Rather, there is a whole band of parameter values for which there is exponential growth. Alternatively, suppose the parametrically driven oscillator is *anti-damped* by adding a term of the form $\beta dx/dt$ with $\beta < 0$ to the left side of (4.83), or suppose $K_0 < 0$. Then the solution would grow exponentially when $K_1 = 0$. However, there is a now a range of K_1 and Ω values for which parametric driving can *stabilize* the oscillator. That is, when $\beta < 0$ or $K_0 < 0$, it can be shown that there is range of K_1 and Ω values for which $x(t)$ is nevertheless bounded.

Finally we remark, as seems plausible from the arguments made in Exercise 4.9, that the behavior we have found/claimed for the Mathieu case will occur quite generally for other cases of Hill's equation.

6.4.11. Consider the motion of a charged particle in an electromagnetic field. With time as the independent variable, suppose one integrates the first-order set of differential equations

[15] "He'd fly through the air with the greatest of ease, a daring young man on the flying trapeze. His movements were graceful, all girls he could please. And my love he purloined away". The *leotard* garment is named after Léotard who invented and first wore it in his performances.

(1.6.69) and (1.6.70) for the quantities $\boldsymbol{r}$ and $\boldsymbol{p}$ from $t = t^{in}$ to $t = t^{fin}$. Recall that here the quantity $\boldsymbol{p}$ is the *mechanical* momentum. Therefore, to be more precise, we will use the notation $\boldsymbol{r}^{\text{mech}}$, $\boldsymbol{p}^{\text{mech}}$ and $\boldsymbol{r}^{\text{can}}$, $\boldsymbol{p}^{\text{can}}$ to refer to mechanical and canonical quantities, respectively. With this notation in mind, is the relation between the initial conditions $(\boldsymbol{r}^{\text{mech}})^{in}$, $(\boldsymbol{p}^{\text{mech}})^{in}$ and the final conditions $(\boldsymbol{r}^{\text{mech}})^{fin}$, $(\boldsymbol{p}^{\text{mech}})^{fin}$ a symplectic map? You are to show that in general the answer is *no*.

More precisely, let $d(\boldsymbol{r}^{\text{mech}})^{in}$, $d(\boldsymbol{p}^{\text{mech}})^{in}$ denote small changes in the initial conditions, and let $d(\boldsymbol{r}^{\text{mech}})^{fin}$, $d(\boldsymbol{p}^{\text{mech}})^{fin}$ be the corresponding changes in the final conditions. By definition they are connected by the Jacobian matrix relation

$$\begin{pmatrix} d(\boldsymbol{r}^{\text{mech}})^{fin} \\ d(\boldsymbol{p}^{\text{mech}})^{fin} \end{pmatrix} = N \begin{pmatrix} d(\boldsymbol{r}^{\text{mech}})^{in} \\ d(\boldsymbol{p}^{\text{mech}})^{in} \end{pmatrix}. \tag{6.4.101}$$

Your task is to show that in general N is *not* a symplectic matrix.

The mechanical and canonical quantities are connected by the relations

$$\boldsymbol{r}^{\text{mech}} = \boldsymbol{r}^{\text{can}}, \tag{6.4.102}$$

$$\boldsymbol{p}^{\text{mech}} = \boldsymbol{p}^{\text{can}} - q\boldsymbol{A}. \tag{6.4.103}$$

Recall (1.5.30). Use (4.102) and (4.103) to obtain the relations

$$d(\boldsymbol{r}^{\text{mech}})^{in} = d(\boldsymbol{r}^{\text{can}})^{in}, \tag{6.4.104}$$

$$d(\boldsymbol{r}^{\text{mech}})^{fin} = d(\boldsymbol{r}^{\text{can}})^{fin}, \tag{6.4.105}$$

$$d(p_j^{\text{mech}})^{in} = d(p_j^{\text{can}})^{in} - q\sum_k A_{j,k}[(\boldsymbol{r}^{\text{can}})^{in}, t^{in}]d(x_k^{\text{can}})^{in}, \tag{6.4.106}$$

$$d(p_j^{\text{mech}})^{fin} = d(p_j^{\text{can}})^{fin} - q\sum_k A_{j,k}[(\boldsymbol{r}^{\text{can}})^{fin}, t^{fin}]d(x_k^{\text{can}})^{fin}, \tag{6.4.107}$$

where

$$A_{j,k}(\boldsymbol{r}^{\text{can}}, t) = \partial A_j(\boldsymbol{r}^{\text{can}}, t)/\partial x_k^{\text{can}}. \tag{6.4.108}$$

Next verify that (4.104), (4.106) and (4.105), (4.107) can be written in the more compact matrix form

$$\begin{pmatrix} d(\boldsymbol{r}^{\text{mech}})^{in} \\ d(\boldsymbol{p}^{\text{mech}})^{in} \end{pmatrix} = V^{\text{in}} \begin{pmatrix} d(\boldsymbol{r}^{\text{can}})^{in} \\ d(\boldsymbol{p}^{\text{can}})^{in} \end{pmatrix}, \tag{6.4.109}$$

$$\begin{pmatrix} d(\boldsymbol{r}^{\text{mech}})^{fin} \\ d(\boldsymbol{p}^{\text{mech}})^{fin} \end{pmatrix} = V^{\text{fin}} \begin{pmatrix} d(\boldsymbol{r}^{\text{can}})^{fin} \\ d(\boldsymbol{p}^{\text{can}})^{fin} \end{pmatrix}, \tag{6.4.110}$$

where V is the matrix

$$V(\boldsymbol{r}^{\text{can}}, t) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ -qA_{1,1} & -qA_{1,2} & -qA_{1,3} & 1 & 0 & 0 \\ -qA_{2,1} & -qA_{2,2} & -qA_{2,3} & 0 & 1 & 0 \\ -qA_{3,1} & -qA_{3,2} & -qA_{3,3} & 0 & 0 & 1 \end{pmatrix}, \tag{6.4.111}$$

and V^{in} and V^{fin} are the matrices

$$V^{in} = V[(\boldsymbol{r}^{\rm can})^{in}, t^{in}], \tag{6.4.112}$$

$$V^{fin} = V[(\boldsymbol{r}^{\rm can})^{fin}, t^{fin}]. \tag{6.4.113}$$

Finally, explain why there is a relation of the form

$$\begin{pmatrix} d(\boldsymbol{r}^{\rm can})^{fin} \\ d(\boldsymbol{p}^{\rm can})^{fin} \end{pmatrix} = M \begin{pmatrix} d(\boldsymbol{r}^{\rm can})^{in} \\ d(\boldsymbol{p}^{\rm can})^{in} \end{pmatrix} \tag{6.4.114}$$

where M is a symplectic matrix. Recall the Hamiltonian (1.5.31).

We are now ready for some matrix manipulation. For calculational convenience write (4.111) in the block form

$$V = \begin{pmatrix} I & 0 \\ C & I \end{pmatrix} \tag{6.4.115}$$

where C is the matrix with entries

$$C_{jk} = -qA_{j,k}. \tag{6.4.116}$$

Verify that V is invertible, and its inverse is given by the formula

$$V^{-1} = \begin{pmatrix} I & 0 \\ -C & I \end{pmatrix}. \tag{6.4.117}$$

Verify that (4.101), (4.109), and (4.110) can be combined to yield the relations

$$V^{fin} \begin{pmatrix} d(\boldsymbol{r}^{\rm can})^{fin} \\ d(\boldsymbol{p}^{\rm can})^{fin} \end{pmatrix} = NV^{in} \begin{pmatrix} d(\boldsymbol{r}^{\rm can})^{in} \\ d(\boldsymbol{p}^{\rm can})^{in} \end{pmatrix}, \tag{6.4.118}$$

or, equivalently,

$$\begin{pmatrix} d(\boldsymbol{r}^{\rm can})^{fin} \\ d(\boldsymbol{p}^{\rm can})^{fin} \end{pmatrix} = (V^{fin})^{-1}NV^{in} \begin{pmatrix} d(\boldsymbol{r}^{\rm can})^{in} \\ d(\boldsymbol{p}^{\rm can})^{in} \end{pmatrix}. \tag{6.4.119}$$

Verify that comparison of (4.114) and (4.119) yields the matrix relations

$$(V^{fin})^{-1}NV^{in} = M, \tag{6.4.120}$$

or, equivalently,

$$N = V^{fin}M(V^{in})^{-1}. \tag{6.4.121}$$

You are ready for the final steps. Begin by showing that in general V is not symplectic. In particular verify, using the results of Section 3.3.2 , that the condition for V to be symplectic is that

$$C - C^T = 0. \tag{6.4.122}$$

Show that in fact for the present case there is the result

$$C - C^T = q\boldsymbol{B} \cdot \boldsymbol{L} \tag{6.4.123}$$

where $\boldsymbol{B}(\boldsymbol{r}^{\mathrm{can}},t)$ is the magnetic field and $\boldsymbol{L}$ denotes the collection of matrices given by (3.7.177) through (3.7.179). We see that V is not symplectic unless the magnetic field vanishes, which should not be too surprising in view of (1.7.18). Show that in fact V satisfies the relation

$$V^T JV = J \left(\begin{array}{cc} I & 0 \\ q\boldsymbol{B}\cdot\boldsymbol{L} & I \end{array} \right). \tag{6.4.124}$$

Verify that the product of a symplectic and a nonsymplectic matrix is nonsymplectic. Since V^{in} and V^{fin} are in general different, it follows from (4.121) that in general N is not symplectic. To strengthen the argument further, suppose that M is of the form

$$M = \left(\begin{array}{cc} \lambda I & 0 \\ 0 & \lambda^{-1} I \end{array} \right) \tag{6.4.125}$$

where λ is any scalar. According to Section 3.3.2 such an M is symplectic. Verify in this case that

$$N = M \left(\begin{array}{cc} I & 0 \\ C' & I \end{array} \right) \tag{6.4.126}$$

where

$$C' = \lambda^2 C^{fin} - C^{in}. \tag{6.4.127}$$

Show that

$$C' - (C')^T = q(\lambda^2 \boldsymbol{B}^{fin} - \boldsymbol{B}^{in})\cdot\boldsymbol{L}, \tag{6.4.128}$$

and therefore in general the second matrix on the right side of (4.126) is not symplectic. Correspondingly, in this case N is generally not symplectic even if V^{in} and V^{fin} are the same. Verify that in this case

$$N^T JN = J \left(\begin{array}{cc} I & 0 \\ q\boldsymbol{B}'\cdot\boldsymbol{L} & I \end{array} \right) \tag{6.4.129}$$

where

$$\boldsymbol{B}' = \lambda^2 \boldsymbol{B}^{fin} - \boldsymbol{B}^{in}. \tag{6.4.130}$$

6.4.12. Recall Section 4.3 and review Exercise 4.3.24. Find the symplectic polar decomposition for the matrix V given by (4.115).

6.5 Mixed-Variable Generating Functions

It is well known that canonical transformations/symplectic maps can be produced by the use of mixed-variable *generating* functions, the most familiar of which are traditionally referred to as F_1 through F_4. The generating functions are called *mixed* because they involve both "old" and "new" variables.[16]

In this section we will verify that the generating functions F_1 through F_4 produce symplectic maps. Conversely, given a symplectic map, we will find possible associated generating

[16] In the field of light ray optics, where the use of generating functions was first introduced in the seminal work of Hamilton, generating functions are sometimes referred to as *characteristic* functions.

functions F_1 through F_4. A time-dependent generating function F_j produces a one-parameter family of symplectic maps. In that case we will find the associated generating Hamiltonian.

In Section 7 we will find that the functions F_1 through F_4 are but four examples of an *infinite* set of generating functions. Until then, the term *generating functions* will refer simply to the functions F_1 through F_4.

6.5.1 Generating Functions Produce Symplectic Maps

6.5.1.1 Background

Since the use of generating functions does not treat coordinate and momentum variables on a common footing, it is convenient (as in Section 4.8) to introduce the notation

$$z = (q_1 \cdots q_n, p_1 \cdots p_n), \tag{6.5.1}$$

$$Z = (Q_1 \cdots Q_n, P_1 \cdots P_n). \tag{6.5.2}$$

In this notation the symplectic map $\mathcal{M}$ sends z to Z,

$$\mathcal{M}:\; z \to Z. \tag{6.5.3}$$

We begin with the mixed-variable generating functions $F_1(q, Q, t)$, $F_2(q, P, t)$, $F_3(p, Q, t)$, and $F_4(p, P, t)$. These four functions produce maps by the (implicit) relations

$$\begin{gathered} p_k = \partial F_1/\partial q_k,\;\; P_k = -\partial F_1/\partial Q_k; \\ \text{assumes }\; \det(\partial^2 F_1/\partial q_k \partial Q_\ell) \neq 0,\;\; \text{yields }\; \det(B) \neq 0, \end{gathered} \tag{6.5.4}$$

$$\begin{gathered} p_k = \partial F_2/\partial q_k,\;\; Q_k = \partial F_2/\partial P_k; \\ \text{assumes }\; \det(\partial^2 F_2/\partial q_k \partial P_\ell) \neq 0,\;\; \text{yields }\; \det(D) \neq 0, \end{gathered} \tag{6.5.5}$$

$$\begin{gathered} q_k = -\partial F_3/\partial p_k,\;\; P_k = -\partial F_3/\partial Q_k; \\ \text{assumes }\; \det(\partial^2 F_3/\partial p_k \partial Q_\ell) \neq 0,\;\; \text{yields }\; \det(A) \neq 0, \end{gathered} \tag{6.5.6}$$

$$\begin{gathered} q_k = -\partial F_4/\partial p_k,\;\; Q_k = \partial F_4/\partial P_k; \\ \text{assumes }\; \det(\partial^2 F_4/\partial p_k \partial P_\ell) \neq 0,\;\; \text{yields }\; \det(C) \neq 0. \end{gathered} \tag{6.5.7}$$

The matrices A through D will be specified shortly. It can be verified that each relation pair produces a symplectic map subject to only mild restrictions on the functional behavior of the associated mixed-variable generating function. However, as mentioned at the end of Section 4.8 and will be proved subsequently, there are symplectic maps that cannot be produced by any of the generating functions F_j.

6.5.1.2 Use of F_2

Consider, for example, the use of F_2. The equations (5.5) are implicit,

$$p_k = p_k(q, P, t), \tag{6.5.8}$$

$$Q_k = Q_k(q, P, t), \tag{6.5.9}$$

and have to be brought to the explicit form

$$Q_k = Q_k(q, p, t), \tag{6.5.10}$$

$$P_k = P_k(q, p, t). \tag{6.5.11}$$

(The fact that the map equations are initially implicit and subsequently, often with considerable effort, have to be made explicit is one of the drawbacks of using generating functions to produce symplectic maps. By contrast, as we will see in Chapter 7, Lie transformations can be used to produce symplectic maps that are immediately in explicit form.)

Take differentials of both sides of (5.8) and (5.9), and use (5.5), to get the relations

$$\begin{aligned} dp_k &= \sum_\ell (\partial p_k/\partial q_\ell)dq_\ell + (\partial p_k/\partial P_\ell)dP_\ell \\ &= \sum_\ell (\partial^2 F_2/\partial q_k \partial q_\ell)dq_\ell + (\partial^2 F_2/\partial q_k \partial P_\ell)dP_\ell, \end{aligned} \tag{6.5.12}$$

$$\begin{aligned} dQ_k &= \sum_\ell (\partial Q_k/\partial q_\ell)dq_\ell + (\partial Q_k/\partial P_\ell)dP_\ell \\ &= \sum_\ell (\partial^2 F_2/\partial P_k \partial q_\ell)dq_\ell + (\partial^2 F_2/\partial P_k \partial P_\ell)dP_\ell. \end{aligned} \tag{6.5.13}$$

These relations can be written in the matrix form

$$dp = \alpha dq + \beta dP, \tag{6.5.14}$$

$$dQ = \gamma dq + \delta dP = \beta^T dq + \delta dP, \tag{6.5.15}$$

where α through δ are the matrices

$$\alpha_{k\ell} = \partial p_k/\partial q_\ell = \partial^2 F_2/\partial q_k \partial q_\ell, \tag{6.5.16}$$

$$\beta_{k\ell} = \partial p_k/\partial P_\ell = \partial^2 F_2/\partial q_k \partial P_\ell, \tag{6.5.17}$$

$$\gamma_{k\ell} = \partial Q_k/\partial q_\ell = \partial^2 F_2/\partial P_k \partial q_\ell, \tag{6.5.18}$$

$$\delta_{k\ell} = \partial Q_k/\partial P_\ell = \partial^2 F_2/\partial P_k \partial P_\ell. \tag{6.5.19}$$

(Note here that the matrix δ is not to be confused with the Kronecker delta.) By inspection, these matrices have the properties

$$\alpha^T = \alpha \ , \ \beta^T = \gamma \ , \ \delta^T = \delta; \tag{6.5.20}$$

and we have used the second property in (5.20) to write the terms on the far right side of (5.15).

Solve (5.14) for dP to find the result

$$dP = -\beta^{-1}\alpha dq + \beta^{-1}dp, \tag{6.5.21}$$

and insert this result in (5.15) to get the complementary relation

$$dQ = (\gamma - \delta\beta^{-1}\alpha)dq + \delta\beta^{-1}dp. \tag{6.5.22}$$

Note that these manipulations require that β be invertible. [See (5.17) and the assumption made in (5.5).] By the inverse function theorem, this invertibility is equivalent to requiring that the first set of equations in (5.5), see (5.8), can be solved for the P_ℓ to find $P_\ell(q,p,t)$.

Next write (5.21) and (5.22) in the compact matrix form

$$dZ = Mdz. \tag{6.5.23}$$

Comparison of (5.23) with (5.21) and (5.22) shows that M has the block form

$$M = \begin{pmatrix} \gamma - \delta\beta^{-1}\alpha & \delta\beta^{-1} \\ -\beta^{-1}\alpha & \beta^{-1} \end{pmatrix}. \tag{6.5.24}$$

As in Section 3.3, it is convenient to employ the notation

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}. \tag{6.5.25}$$

Thus, we have the identifications

$$A = \gamma - \delta\beta^{-1}\alpha = \beta^T - \delta\beta^{-1}\alpha, \tag{6.5.26}$$

$$B = \delta\beta^{-1}, \tag{6.5.27}$$

$$C = -\beta^{-1}\alpha, \tag{6.5.28}$$

$$D = \beta^{-1}. \tag{6.5.29}$$

At this point we remark that the relations (4.8.9) and (4.8.10) resemble the relations (5.14) and (5.15); and the relations (4.8.14) through (4.8.17) are identical to the relations (5.26) through (5.29). This is as it should be because linear maps are a special case of general maps. We also see from (5.29) that D must be invertible, $\det(D) \neq 0$, in accord with (5.5).

Finally, we must verify that M is symplectic. With the aid of (5.20) we find the relations

$$A^T = \gamma^T - \alpha^T(\beta^{-1})^T\delta^T = \beta - \alpha\gamma^{-1}\delta, \tag{6.5.30}$$

$$B^T = (\beta^{-1})^T\delta^T = \gamma^{-1}\delta, \tag{6.5.31}$$

$$C^T = -\alpha^T(\beta^{-1})^T = -\alpha\gamma^{-1}, \tag{6.5.32}$$

$$D^T = (\beta^{-1})^T = \gamma^{-1}. \tag{6.5.33}$$

Compute the various combinations of matrices that appear in (3.3.3) through (3.3.5). We find the results

$$A^T C = (\beta - \alpha\gamma^{-1}\delta)(-\beta^{-1}\alpha) = -\alpha + \alpha\gamma^{-1}\delta\beta^{-1}\alpha, \tag{6.5.34}$$

$$C^T A = -\alpha - \gamma^{-1}(\gamma - \delta\beta^{-1}\alpha) = -\alpha + \alpha\gamma^{-1}\delta\beta^{-1}\alpha, \tag{6.5.35}$$

$$B^T D = \gamma^{-1}\delta\beta^{-1}, \tag{6.5.36}$$

$$D^T B = \gamma^{-1}\delta\beta^{-1}, \tag{6.5.37}$$

$$A^T D = (\beta - \alpha\gamma^{-1}\delta)\beta^{-1} = I - \alpha\gamma^{-1}\delta\beta^{-1}, \tag{6.5.38}$$

$$C^T B = -\alpha\gamma^{-1}\delta\beta^{-1}. \tag{6.5.39}$$

By looking at these results we see that the relations (3.3.3) through (3.3.5) are satisfied, and therefore M is symplectic. Correspondingly, when the implicit relations (5.8) and (5.9) are solved to yield Z in terms of z, the result is a symplectic map $\mathcal{M}$.

6.5.1.3 Use of F_1

We have examined the use of F_2. As a second example, we will consider the use of F_1. The cases of F_3 and F_4 proceed similarly. For the case of F_1 the equations (5.4) have the implicit form

$$p_k = p_k(q, Q, t), \tag{6.5.40}$$

$$P_k = P_k(q, Q, t), \tag{6.5.41}$$

and have to be brought to the explicit form

$$Q_k = Q_k(q, p, t), \tag{6.5.42}$$

$$P_k = P_k(q, p, t). \tag{6.5.43}$$

Take differentials of both sides of (5.40) and (5.41), and use (5.4), to get the relations

$$\begin{aligned} dp_k &= \sum_\ell (\partial p_k/\partial q_\ell)dq_\ell + (\partial p_k/\partial Q_\ell)dQ_\ell \\ &= \sum_\ell (\partial^2 F_1/\partial q_k \partial q_\ell)dq_\ell + (\partial^2 F_1/\partial q_k \partial Q_\ell)dQ_\ell, \end{aligned} \tag{6.5.44}$$

$$\begin{aligned} dP_k &= \sum_\ell (\partial P_k/\partial q_\ell)dq_\ell + (\partial P_k/\partial Q_\ell)dQ_\ell \\ &= -\sum_\ell (\partial^2 F_1/\partial Q_k \partial q_\ell)dq_\ell - (\partial^2 F_1/\partial Q_k \partial Q_\ell)dQ_\ell. \end{aligned} \tag{6.5.45}$$

These relations can be written in the matrix form

$$dp = \alpha dq + \beta dQ, \tag{6.5.46}$$

$$dP = \gamma dq + \delta dQ, \tag{6.5.47}$$

where α through δ are the matrices

$$\alpha_{k\ell} = \partial p_\ell/\partial q_\ell = \partial^2 F_1/\partial q_k \partial q_\ell, \tag{6.5.48}$$

$$\beta_{k\ell} = \partial p_k/\partial Q_\ell = \partial^2 F_1/\partial q_k \partial Q_\ell, \tag{6.5.49}$$

$$\gamma_{k\ell} = \partial P_k/\partial q_\ell = -\partial^2 F_1/\partial Q_k \partial q_\ell, \tag{6.5.50}$$

$$\delta_{k\ell} = \partial P_k/\partial Q_\ell = -\partial^2 F_1/\partial Q_k \partial Q_\ell. \tag{6.5.51}$$

By inspection, these matrices have the properties

$$\alpha^T = \alpha\ ,\ \beta^T = -\gamma\ ,\ \delta^T = \delta. \tag{6.5.52}$$

Solve (5.46) for dQ to find the result

$$dQ = -\beta^{-1}\alpha dq + \beta^{-1} dp, \tag{6.5.53}$$

and insert this result in (5.47) to get the complementary relation

$$dP = (\gamma - \delta\beta^{-1}\alpha)dq + \delta\beta^{-1}dp. \tag{6.5.54}$$

Note that these manipulations require that β be invertible. [See (5.49) and the assumption made in (5.4).] By the inverse function theorem, this invertibility is equivalent to requiring that the first set of equations in (5.4), see (5.40), can be solved for the Q_ℓ to find $Q_\ell(q, p, t)$.

As before, write (5.53) and (5.54) in the compact matrix form (5.23) and employ (5.25). Comparison of (5.53) and (5.54) with (5.23) and (5.25) yields the relations

$$A = -\beta^{-1}\alpha, \tag{6.5.55}$$

$$B = \beta^{-1}, \tag{6.5.56}$$

$$C = \gamma - \delta\beta^{-1}\alpha, \tag{6.5.57}$$

$$D = \delta\beta^{-1}. \tag{6.5.58}$$

We see from (5.56) that B must be invertible, $\det(B) \neq 0$, in accord with (5.4).

Finally, we must verify that M is symplectic. With the aid of (5.52) we find the relations

$$A^T = -\alpha^T(\beta^{-1})^T = \alpha\gamma^{-1}, \tag{6.5.59}$$

$$B^T = (\beta^{-1})^T = -\gamma^{-1}, \tag{6.5.60}$$

$$C^T = \gamma^T - \alpha^T(\beta^{-1})^T\delta^T = -\beta + \alpha\gamma^{-1}\delta, \tag{6.5.61}$$

$$D^T = (\beta^{-1})^T\delta^T = -\gamma^{-1}\delta. \tag{6.5.62}$$

Compute the various combinations of matrices that appear in (3.3.3) through (3.3.5). We find the results

$$A^TC = \alpha\gamma^{-1}(\gamma - \delta\beta^{-1}\alpha) = \alpha - \alpha\gamma^{-1}\delta\beta^{-1}\alpha, \tag{6.5.63}$$

$$C^TA = (\beta - \alpha\gamma^{-1}\delta)\beta^{-1}\alpha = \alpha - \alpha\gamma^{-1}\delta\beta^{-1}\alpha, \tag{6.5.64}$$

$$B^TD = -\gamma^{-1}\delta\beta^{-1}, \tag{6.5.65}$$

$$D^TB = -\gamma^{-1}\delta\beta^{-1}, \tag{6.5.66}$$

$$A^TD = \alpha\gamma^{-1}\delta\beta^{-1}, \tag{6.5.67}$$

$$C^TB = (-\beta + \alpha\gamma^{-1}\delta)\beta^{-1} = -I + \alpha\gamma^{-1}\delta\beta^{-1}. \tag{6.5.68}$$

By looking at these results we see that the relations (3.3.3) through (3.3.5) are satisfied, and therefore M is symplectic. Correspondingly, when the implicit relations (5.40) and (5.41) are solved to yield Z in terms of z, the result is a symplectic map $\mathcal{M}$.

6.5.1.4 What Maps Can Be Produced by What F_j?

We have seen that the F_j produce symplectic maps, but now wonder what maps can be produced in this fashion. Here we will make a few observations. A more complete exploration of this question is made in Subsection 7.4.

Suppose $\mathcal{M}$ is the identity map so that

$$M = I^{2n}. \tag{6.5.69}$$

In this case

$$\det(A) = \det(D) = \det(I^n) = 1. \tag{6.5.70}$$

Therefore, according to (5.5) and (5.6), we expect the use of F_2 and F_3 to succeed. Indeed, it is easy to verify that $F_2(q,P)$ and $F_3(p,Q)$ given by

$$F_2(q,P) = \sum_k q_k P_k \tag{6.5.71}$$

and

$$F_3(p,Q) = -\sum_k p_k Q_k \tag{6.5.72}$$

do indeed produce the identity map. By contrast,

$$\det(B) = \det(C) = 0 \tag{6.5.73}$$

for the identity map. Therefore, according to (5.4) and (5.7), use of either F_1 or F_4 cannot produce the identity map; attempted use of either F_1 or F_4 fails.

What about the linear symplectic map $\mathcal{M}$ for which $M = J$? In this case we see from (3.1.1) that

$$\det(A) = \det(D) = 0 \tag{6.5.74}$$

and

$$\det(B) \neq 0 \text{ and } \det(C) \neq 0. \tag{6.5.75}$$

Examination of (5.4) through (5.7) shows that attempted use of F_2 and F_3 are expected to fail, and attempted use of F_1 and F_4 are expected to succeed. Indeed, it is easily verified that

$$F_1(q,Q,t) = \sum_k q_k Q_k \tag{6.5.76}$$

and

$$F_4(p,P,t) = \sum_k p_k P_k \tag{6.5.77}$$

produce $M = J$ when employed in (5.4) and (5.7), respectively.

What about the linear symplectic map $\mathcal{M}$ for which $M = R$ where R is the symplectic matrix given by (4.8.31)? If you worked Exercise 4.8.4, you verified that in this case all the submatrices A through D *fail* to have inverses. Therefore none of the yields/results listed in (5,4) through (5.7) can be realized if one attempts to produce this R using any F_j. It follows that one cannot produce this R using any of the F_j; attempted use of any of the F_j fails.

6.5.1.5 Differentials and Differential Forms associated with the F_j

Associated with each of the F_j are both a differential dF_j and a *differential* form which we will call ω_j. For example, we may write the differential

$$dF_1(q,Q,t) = \sum_k (\partial F_1/\partial q_k)dq_k + (\partial F_1/\partial Q_k)dQ_k. \tag{6.5.78}$$

Correspondingly, making use of the relations for p_k and P_k in (5.4), we define an associated differential form ω_1 by the rule

$$\omega_1 = \sum_k p_k dq_k - P_k dQ_k. \tag{6.5.79}$$

Note that ω_1 involves the $2n+2n = 4n$ variables z and Z. Similarly, there are the differentials and associated differential forms

$$dF_2(q,P,t) = \sum_k (\partial F_2/\partial q_k)dq_k + (\partial F_2/\partial P_k)dP_k, \tag{6.5.80}$$

$$\omega_2 = \sum_k p_k dq_k + Q_k dP_k; \tag{6.5.81}$$

$$dF_3(p,Q,t) = \sum_k (\partial F_3/\partial p_k)dp_k + (\partial F_3/\partial Q_k)dQ_k, \tag{6.5.82}$$

$$\omega_3 = \sum_k -q_k dp_k - P_k dQ_k; \tag{6.5.83}$$

$$dF_4(p,P,t) = \sum_k (\partial F_4/\partial p_k)dp_k + (\partial F_4/\partial P_k)dP_k, \tag{6.5.84}$$

$$\omega_4 = \sum_k -q_k dp_k + Q_k dP_k. \tag{6.5.85}$$

We have seen, if the Hessian of a given F_j is invertible, then there is an associated symplectic map which we will call $\mathcal{M}_j$, and the relevant $n \times n$ block in the associated Jacobian matrix M_j as given by (5.25) will be invertible. [In the case of F_1, for example, according to (5.4) the relevant block is the matrix B.] Conversely, given a symplectic map $\mathcal{M}$ and the desire of find an associated generating function F_j, and after verifying that the nature of $\mathcal{M}$ is such that all the remaining variables among the z and Z can be found in terms of the variables on which F_j is supposed to depend, then it can be shown that the differential form ω_j is exact in terms of these variables, and correspondingly the desired F_j can be constructed. Moreover, the requirement that all remaining variables can be found in terms of the variables on which F_j is supposed to depend is equivalent to assuming that the relevant $n \times n$ block in M is invertible. (In this case we say that form of the desired F_j is *compatible* with the nature of $\mathcal{M}$.) See, for example, Subsection 5.2.1 where F_2 is constructed from a knowledge of $\mathcal{M}$.

But what happens if the form of the desired F_j is *not* compatible with the nature of $\mathcal{M}$? Then an attempted construction of F_j will fail. Suppose, for example, that $\mathcal{M}$ is the identity map. In this case, since $Q_k = q_k$ and $P_k = p_k$, there are, according to (5.79), (5.81), (5.83), and (5.85), the results

$$\omega_1 = \sum_k p_k dq_k - P_k dQ_k = \sum_k p_k dq_k - p_k dq_k = 0, \tag{6.5.86}$$

$$\omega_2 = \sum_k p_k dq_k + Q_k dP_k = \sum_k p_k dq_k + q_k dp_k = d(\sum_k q_k p_k) = d(\sum_k q_k P_k), \tag{6.5.87}$$

$$\omega_3 = \sum_k -q_k dp_k - P_k dQ_k = \sum_k -q_k dp_k - p_k dq_k = d(-\sum_k p_k q_k) = d(-\sum_k p_k Q_k), \tag{6.5.88}$$

$$\omega_4 = \sum_k -q_k dp_k + Q_k dP_k = \sum_k -q_k dp_k + q_k dp_k = 0. \tag{6.5.89}$$

Note that (5.86) and (5.89) are in accord with the fact that attempted use of F_1 or F_4 fails for the identity map. Also, (5.87) and (5.71), and (5.88) and (5.72), are in agreement for the identity map. That is, $\omega_2 = dF_2$ and $\omega_3 = dF_3$.

6.5.2 Finding a Generating Function from a Map or a Generating Hamiltonian

We have seen that, modulo the invertibility of certain matrices, the mixed-variable generating functions F_1 through F_4 can be used to produce symplectic maps $\mathcal{M}$. What about the converse: given a symplectic map $\mathcal{M}$, can we find a mixed-variable generating function that produces it? Or, given the Hamiltonian H that generates a family of symplectic maps $\mathcal{M}(t)$, can we find an associated time-dependent generating function? We shall see that, again modulo the invertibility of certain matrices which amounts to the question of compatibility, the answer is *yes.*

6.5.2.1 Finding a Generating Function Directly from a Map

As an example, we will consider the problem of constructing $F_2(q, P, t)$ given a symplectic map $\mathcal{M}$. Begin by writing the relation (5.3) in the component form

$$Q_k = S_k(q, p, t), \tag{6.5.90}$$

$$P_k = T_k(q, p, t), \tag{6.5.91}$$

and assume that the S_k and T_k are known functions. Next assume that the relations (5.91) can be inverted to give the p_k as functions of q, P, and t,

$$p_k = p_k(q, P, t). \tag{6.5.92}$$

By the inverse function theorem, this inversion is possible if the Jacobian matrix

$$\partial P_k / \partial p_\ell = \partial T_k / \partial p_\ell \tag{6.5.93}$$

is invertible. Next substitute the relations (5.92) into (5.190) to obtain the Q_k as functions of q, P, and t,

$$Q_k = Q_k(q, P, t). \tag{6.5.94}$$

Now consider the differential form

$$\omega_2 = \sum_k (p_k dq_k + Q_k dP_k). \tag{6.5.95}$$

Recall (5.81). We shall soon see that the assumption that $\mathcal{M}$ is symplectic implies that this differential form is exact with regard to the variables q_k, P_k. Taking this assertion as granted, we may define a function $F_2(q, P, t)$ by the path integral

$$F_2(q, P, t) = \int^{q,P} \omega_2 = \int^{q,P} \sum_k [p_k(q', P', t)dq'_k + Q_k(q', P', t)dP'_k]. \tag{6.5.96}$$

By construction F_2 will have the properties

$$\partial F_2/\partial q_k = p_k(q, P, t), \tag{6.5.97}$$

$$\partial F_2/\partial P_k = Q_k(q, P, t), \tag{6.5.98}$$

and we see that the desired relations (5.5) have been obtained.

We still must show that ω_2 given by (5.95) is exact. According to Exercise 1.1, we must verify the relations (1.26). In the present context these relations take the form

$$\partial p_m/\partial q_n = \partial p_n/\partial q_m, \tag{6.5.99}$$

$$\partial Q_m/\partial q_n = \partial p_n/\partial P_m, \tag{6.5.100}$$

$$\partial p_m/\partial P_n = \partial Q_n/\partial q_m, \tag{6.5.101}$$

$$\partial Q_m/\partial P_n = \partial Q_n/\partial P_m. \tag{6.5.102}$$

Note that (5.100) and (5.101) say the same thing.

Take differentials of both sides of (5.90) and (5.91) and use the notation of (5.1), (5.2), (5.23), and (5.25) to find the relations

$$dQ = Adq + Bdp, \tag{6.5.103}$$

$$dP = Cdq + Ddp, \tag{6.5.104}$$

where A through D are the matrices

$$A_{k\ell} = \partial Q_k/\partial q_\ell = \partial S_k/\partial q_\ell \ , \ B_{k\ell} = \partial Q_k/\partial p_\ell = \partial S_k/\partial p_\ell, \tag{6.5.105}$$

$$C_{k\ell} = \partial P_k/\partial q_\ell = \partial T_k/\partial q_\ell \ , \ D_{k\ell} = \partial P_k/\partial p_\ell = \partial T_k/\partial p_\ell. \tag{6.5.106}$$

We now want to take q and P as independent variables. Solve (5.104) and (5.103) for dp and dQ in terms of dq and dP to find the results

$$dp = -D^{-1}Cdq + D^{-1}dP, \tag{6.5.107}$$

$$dQ = (A - BD^{-1}C)dq + BD^{-1}dP. \tag{6.5.108}$$

Note that in finding these results we assumed the existence of D^{-1}. But comparison of (5.93) and (5.106) shows that D is the Jacobian matrix whose invertibility has already been assumed. From (5.107) and (5.108) we obtain the results

$$\partial p_m/\partial q_n = (-D^{-1}C)_{mn}, \tag{6.5.109}$$

$$\partial p_m/\partial P_n = (D^{-1})_{mn}, \tag{6.5.110}$$

$$\partial Q_m/\partial q_n = (A - BD^{-1}C)_{mn}, \tag{6.5.111}$$

$$\partial Q_m/\partial P_n = (BD^{-1})_{mn}. \tag{6.5.112}$$

With these results before us, we see that establishing the relations (5.99) through (5.102) is equivalent to verifying the conjectures

$$(D^{-1}C) \overset{?}{=} (D^{-1}C)^T, \tag{6.5.113}$$

$$A - BD^{-1}C \overset{?}{=} (D^{-1})^T, \tag{6.5.114}$$

$$(BD^{-1}) \overset{?}{=} (BD^{-1})^T. \tag{6.5.115}$$

But, thanks to the symplectic condition, (5.113) is a consequence of (3.3.7), (5.115) is a consequence of (3.3.4), and (5.114) is a consequence of (3.3.8) and (3.37). Thus we have proved that ω_2 is an exact differential, and have verified that F_2 can be constructed using (5.96). Note that in this construction the time t played no role and, if present at all, appeared only as a parameter.

6.5.2.2 Finding a Generating Function from a Generating Hamiltonian

Given a Hamiltonian H, we know that integrating Hamilton's equations of motion produces a time-dependent symplectic map $\mathcal{M}(t)$. Conversely, given a time-dependent symplectic map $\mathcal{M}(t)$, we know that there is an underlying generating Hamiltonian H. Recall Subsection 4.2. Here we explore how the generating Hamiltonian H can be used to construct the $F_2(q, P, t)$ generating function associated with $\mathcal{M}(t)$. Similar constructions can be made for the F_1, F_3, and F_4 generating functions.

To see how F_2 can be constructed, it is convenient, in analogy with (1.7.9), to introduce the phase-space variables

$$\zeta = (\xi, \eta). \tag{6.5.116}$$

Here the ξ's play the role of coordinates and the η's are conjugate momenta. Let q, p be initial conditions at $t = t^i$, and let Q, P be the final conditions reached by following to time t the trajectories generated by $H(\zeta, t)$ starting with these initial conditions. We know that trajectories can be labeled by specifying either the initial conditions q, p or the final conditions Q, P. Assume that the trajectories are such that they can also be be labeled by specifying q and P. See Figure 5.1. This means that there are relations of the form

$$Q_j = Q_j(q, P, t), \tag{6.5.117}$$

$$p_j = p_j(q, P, t). \tag{6.5.118}$$

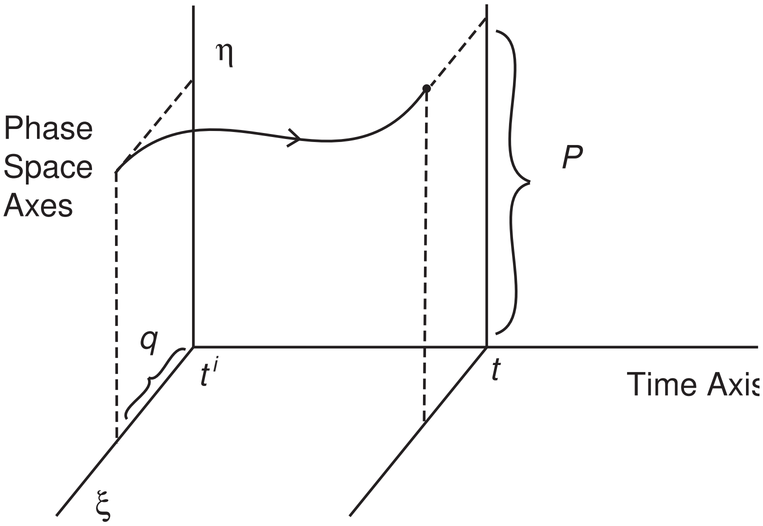


Figure 6.5.1: A trajectory of $H(\zeta, t)$ in the augmented ξ, η, t phase space having initial coordinates q and final momenta P.

With these assumptions in mind, define/construct the function F_2 by the rule

$$F_2(q, P, t) = \sum_k P_k Q_k - \int_{t^i}^{t} d\tau [(\sum_j \eta_j \dot{\xi}_j) - H(\zeta, \tau)]. \tag{6.5.119}$$

Here the integral on the right side is to be evaluated over the trajectory generated by H whose initial coordinates are q and final momenta are P. In actual practice, this trajectory may have to be found by some kind of *shooting* method: One integrates a variety of trajectories all having the same initial q and various initial p until one finds a trajectory that has the desired final momenta P. The search for this trajectory may be facilitated by also integrating the variational equations, see Exercise 1.4.6, to determine how changes in the initial conditions produce changes in the final conditions.

We will want to see how F_2 changes when changes are made in q, P, t. As a first step, let us study how the integral on the right side of (5.119) depends on the variables q, P. We make the definition

$$A(q, P, t) = \int_{t^i}^{t} d\tau [(\sum_j \eta_j \dot{\xi}_j) - H(\zeta, \tau)], \tag{6.5.120}$$

and recognize that A is the *action*. See (1.6.11). Define $\mathcal{A}$ by the rule

$$\mathcal{A}(\zeta, \dot{\zeta}, \tau) = (\sum_j \eta_j \dot{\xi}_j) - H(\zeta, \tau) \tag{6.5.121}$$

so that we may write

$$A(q, P, t) = \int_{t^i}^{t} \mathcal{A}(\zeta, \dot{\zeta}, \tau) d\tau. \tag{6.5.122}$$

Note that $\mathcal{A}$ is the Lagrangian L associated with the Hamiltonian H.

Changing the q's and P's changes the trajectory. Consequently, from variational calculus, we find that the change in A is given by the relation

$$\delta A = \int_{t^i}^{t} d\tau [\sum_j (\partial\mathcal{A}/\partial\xi_j)\delta\xi_j + (\partial\mathcal{A}/\partial\dot{\xi}_j)\delta\dot{\xi}_j + (\partial\mathcal{A}/\partial\eta_j)\delta\eta_j + (\partial\mathcal{A}/\partial\dot{\eta}_j)\delta\dot{\eta}_j]. \tag{6.5.123}$$

The integrand in (5.123) can be manipulated in the standard way to rewrite δA in the form

$$\begin{aligned}\delta A &= \int_{t^i}^{t} d\tau \{\sum_j [(\partial\mathcal{A}/\partial\xi_j) - (d/d\tau)(\partial\mathcal{A}/\partial\dot{\xi}_j)]\delta\xi_j \\ &+ \sum_j [(\partial\mathcal{A}/\partial\eta_j) - (d/d\tau)(\partial\mathcal{A}/\partial\dot{\eta}_j)]\delta\eta_j \\ &+ (d/d\tau)[\sum_j (\partial\mathcal{A}/\partial\dot{\xi}_j)\delta\xi_j + (\partial\mathcal{A}/\partial\dot{\eta}_j)\delta\eta_j]\}. \end{aligned} \tag{6.5.124}$$

For the various ingredients in the integrand of (5.124) we find the results

$$\partial\mathcal{A}/\partial\xi_j - (d/d\tau)(\partial\mathcal{A}/\partial\dot{\xi}_j) = -\partial H/\partial\xi_j - (d/d\tau)\eta_j = -\partial H/\partial\xi_j - \dot{\eta}_j = 0, \tag{6.5.125}$$

$$\partial\mathcal{A}/\partial\dot{\xi}_j = \eta_j, \tag{6.5.126}$$

$$\partial\mathcal{A}/\partial\dot{\eta}_j = 0, \tag{6.5.127}$$

$$\partial\mathcal{A}/\partial\eta_j - (d/d\tau)(\partial\mathcal{A}/\partial\dot{\eta}_j) = \partial\mathcal{A}/\partial\eta_j = \dot{\xi}_j - \partial H/\partial\eta_j = 0. \tag{6.5.128}$$

Here (5.127) follows from the fact that $\mathcal{A}$ does not actually depend on the $\dot{\eta}_j$. See (5.121). And (5.125) and (5.128) follow from the stipulation that the $\zeta(\tau)$ are trajectories of H. As a consequence of these results, δA becomes

$$\delta A = \int_{t^i}^{t} d\tau (d/d\tau)[\sum_j \eta_j\delta\xi_j] = [\sum_j \eta_j\delta\xi_j]|_{t^i}^{t} = \sum_j P_j\delta Q_j - p_j\delta q_j. \tag{6.5.129}$$

We are now ready to study F_2. In terms of the definition (5.120) the expression (5.119) for F_2 can be rewritten in the form

$$F_2(q,P,t) = -A(q,P,t) + \sum_k P_kQ_k. \tag{6.5.130}$$

It follows that the change in F_2 produced by changes in q, P is given by the relation

$$\begin{aligned}\delta F_2 &= -\delta A + \delta(\sum_k P_kQ_k) \\ &= \sum_j (-P_j\delta Q_j + p_j\delta q_j + P_j\delta Q_j + Q_j\delta P_j) \\ &= \sum_j (p_j\delta q_j + Q_j\delta P_j). \end{aligned} \tag{6.5.131}$$

Here we have used (5.129). Evidently (5.131) yields the relations

$$\partial F_2/\partial q_j = p_j, \quad \partial F_2/\partial P_j = Q_j, \tag{6.5.132}$$

which are the desired results (5.5).

As a final step, and in anticipation of results to be established in the next section, let us take the total time derivative of both sides of (5.130). From the chain rule we find the result

$$dF_2/dt = \partial F_2/\partial t + \sum_j (\partial F_2/\partial P_j)\dot{P}_j = \partial F_2/\partial t + \sum_j Q_j \dot{P}_j. \tag{6.5.133}$$

Here we have also used (5.132). For A as given by (5.120) we find the result

$$dA/dt = [(\sum_j \eta_j \dot{\xi}_j) - H(\zeta, \tau)]_{\tau = t} = (\sum_j P_j \dot{Q}_j) - H(Q, P, t). \tag{6.5.134}$$

Also, there is the simple result

$$(d/dt)(\sum_j P_j Q_j) = \sum_j (\dot{P}_j Q_j + P_j \dot{Q}_j). \tag{6.5.135}$$

It follows that the total time derivative of (5.130) is given by the relation

$$\begin{aligned} dF_2/dt = (d/dt)[-A + (\sum_j P_j Q_j)] &= [\sum_j (\dot{P}_j Q_j + P_j \dot{Q}_j - P_j \dot{Q}_j)] + H(Q, P, t) \\ &= (\sum_j Q_j \dot{P}_j) - H(Q, P, t). \end{aligned} \tag{6.5.136}$$

Comparison of (5.133) and (5.136) gives the final result

$$\partial F_2/\partial t = H(Q, P, t). \tag{6.5.137}$$

6.5.3 Finding the Generating Hamiltonian from a Generating Function; Hamilton-Jacobi Theory/Equations

If a generating function F_j is time dependent, then its use in the appropriate associated relation selected from (5.4) through (5.7) will produce a *family* of symplectic maps $\mathcal{M}(t)$. Thanks to the work of Section 6.4, we know that any family of symplectic maps is generated by a Hamiltonian. In this subsection we will find the Hamiltonian associated with a time dependent F_j.

6.5.3.1 Derivation

Consider, for example, the case where $F_2(q, P, t)$ is employed. Since our derivation will involve a flurry of partial differentiations with respect to various variables, it is convenient to introduce the notation

$$F_2(q, P, t; \ , \ , 1) = \partial F_2/\partial t, \tag{6.5.138}$$

$$F_2(q,P,t;k,\ ,1) = \partial^2 F_2/\partial q_k \partial t, \qquad (6.5.139)$$

$$F_2(q,P,t;k\ell,\ ,\ ) = \partial^2 F_2/\partial q_k \partial q_\ell, \qquad (6.5.140)$$

$$F_2(q,P,t;k,\ell,\ ) = \partial^2 F_2/\partial q_k \partial P_\ell. \qquad (6.5.141)$$

With this notation in mind, define the function $F_2^t(q,P,t)$ by the rule

$$F_2^t(q,P,t) = F_2(q,P,t;\ ,\ ,1). \qquad (6.5.142)$$

We know that use of the map produced by F_2 yields relations of the form (5.10) and (5.11). Moreover, since the map is symplectic, these relations can be inverted to yield relations of the form

$$q_k = q_k(Q,P,t), \qquad (6.5.143)$$

$$p_k = p_k(Q,P,t). \qquad (6.5.144)$$

Now substitute (5.143) into the first argument of (5.142) to produce the function $H_2(Q,P,t)$ defined by the rule

$$H_2(Q,P,t) = F_2^t(q(Q,P,t),P,t), \qquad (6.5.145)$$

which we write more compactly, but with less precision, as

$$H_2 = \partial F_2/\partial t. \qquad (6.5.146)$$

We claim that H_2 is the Hamiltonian that generates the family of maps $\mathcal{M}(t)$ produced by the use of $F_2(q,P,t)$.

To see that this claim is correct, write (5.5) in the form

$$p_k = F_2(q,P,t;k,\ ,\ ), \qquad (6.5.147)$$

$$Q_k = F_2(q,P,t;\ ,k,\ ). \qquad (6.5.148)$$

Now suppose the q,p are held *fixed*, and t is changed by an amount dt. So doing will change the Q,P by the amounts dQ,dP given by the relations

$$0 = dp_k = \sum_\ell F_2(q,P,t;k,\ell,\ )dP_\ell + F_2(q,P,t;k,\ ,1)dt, \qquad (6.5.149)$$

$$dQ_k = \sum_\ell F_2(q,P,t;\ ,\ell k,\ )dP_\ell + F_2(q,P,t;\ ,k,1)dt. \qquad (6.5.150)$$

Note that the zero on the left side of (5.149) indicates that the p_k remain fixed, as desired. Recall the matrices α and δ given in (5.16) and (5.19). In terms of these matrices (5.149) and (5.150) can be written in the form

$$0 = \sum_\ell \beta_{k\ell} dP_\ell + F_2(q,P,t;\ ,k,1)dt, \qquad (6.5.151)$$

$$dQ_k = \sum_\ell \delta_{k\ell} dP_\ell + F_2(q,P,t;\ ,k,1)dt. \qquad (6.5.152)$$

Solve (5.151) for the dP to find the result

$$dP_m = -dt \sum_n (\beta^{-1})_{mn} F_2(q, P, t; n, \ , 1). \tag{6.5.153}$$

Also, insert (5.153) into (5.152) to give an expression for the dQ,

$$dQ_m = dt[-\sum_n (\delta\beta^{-1})_{mn} F_2(q, P, t; n, \ , 1)] + dt F_2(q, P, t; \ , m, 1). \tag{6.5.154}$$

Finally, dividing through by dt gives the results

$$dQ_m/dt = -[\sum_n (\delta\beta^{-1})_{mn} F_2(q, P, t; n, \ , 1)] + F_2(q, P, t; \ , m, 1), \tag{6.5.155}$$

$$dP_m/dt = -\sum_n (\beta^{-1})_{mn} F_2(q, P, t; n, \ , 1). \tag{6.5.156}$$

Note that, as before, these manipulations require that β be invertible.

Next let us work out $(\partial H_2/\partial Q)$ and $(\partial H_2/\partial P)$. From (5.145) and (5.142) we find the result

$$(\partial H_2/\partial Q_m) = \sum_n F_2(q, P, t; n, \ , 1)(\partial q_n/\partial Q_m). \tag{6.5.157}$$

However, if we solve (5.15) and (5.14) for dq and dp, we find the relations

$$dq = \gamma^{-1} dQ - \gamma^{-1}\delta dP, \tag{6.5.158}$$

$$dp = \alpha\gamma^{-1} dQ + (\beta - \alpha\gamma^{-1}\delta) dP. \tag{6.5.159}$$

Note that, according to (5.20), the invertibility of γ is guaranteed by the invertibility of β. From (5.158) and (5.20) we find the relation

$$(\partial q_n/\partial Q_m) = (\gamma^{-1})_{nm} = (\beta^{-1})_{mn}. \tag{6.5.160}$$

Therefore (5.157) can also be written in the form

$$(\partial H_2/\partial Q_m) = \sum_n (\beta^{-1})_{mn} F_2(q, P, t; n, \ , 1). \tag{6.5.161}$$

For $(\partial H_2/\partial P_m)$ we find from (5.145) and (5.142) the more complicated result

$$(\partial H_2/\partial P_m) = [\sum_n F_2(q, P, t; n, \ , 1)(\partial q_n/\partial P_m)] + F_2(q, P, t; \ , m, 1). \tag{6.5.162}$$

From (5.158) and (5.20) we find the relation

$$(\partial q_n/\partial P_m) = -(\gamma^{-1}\delta)_{nm} = -[\delta^T(\gamma^{-1})^T]_{mn} = -(\delta\beta^{-1})_{mn}. \tag{6.5.163}$$

Therefore (5.162) can also be written in the form

$$(\partial H_2/\partial P_m) = -[\sum_n (\delta\beta^{-1})_{mn} F_2(q, P, t; n, \ , 1)] + F_2(q, P, t; \ , m, 1). \tag{6.5.164}$$

Now we are essentially done. Comparison of the right sides of (5.155) and (5.164) shows that they agree; and comparison of the right sides of (5.156) and (5.161) shows that they agree except for a minus sign. We therefore have demonstrated the desired results

$$dQ_m/dt = \partial H_2/\partial P_m, \tag{6.5.165}$$

$$dP_m/dt = -\partial H_2/\partial Q_m. \tag{6.5.166}$$

Finally we remark that similar calculations for all the F_j show (again after a transformation to the variables Q,P,t has been made) that there is the general result

$$H_j = \partial F_j/\partial t. \tag{6.5.167}$$

Note that (5.137) is a special case of (5.167). The relations (5.167) are closely related to the *Hamilton-Jacobi* equations. See the discussion below. We will revisit this subject in Subsection 7.3.

6.5.3.2 Transformation of Hamiltonians and Application to Hamilton-Jacobi Theory

Subsection 4.2 showed that any family of symplectic maps $\mathcal{N}(t)$ is Hamiltonian generated, and the associated Hamiltonian was called G. Subsection 4.4 described the transformation of an old Hamiltonian to a new Hamiltonian under the action of a symplectic map. Here we study the relation between the old and new Hamiltonians in the case that the symplectic map $\mathcal{N}$ arises from some specified mixed-variable generating function F_j, and apply the results to Hamilton-Jacobi theory for this case.

If we make the identification

$$Z = \bar{z}, \tag{6.5.168}$$

the relation (4.56) between old and new Hamiltonians can be rewritten in the form

$$K(Z;t) = H(z(Z,t);t) + G(Z;t). \tag{6.5.169}$$

In the special case that $\mathcal{N}(t)$ arises from the use of an F_j, we found in Subsection 5.3.1 that the associated generating Hamiltonian, which we called H_j, was given by the relation (5.167). Therefore, if we make the identification

$$G = H_j = \partial F_j/\partial t, \tag{6.5.170}$$

we see that (5.169) can be rewritten in the form

$$K(Z;t) = H(z(Z,t);t) + \partial F_j/\partial t \tag{6.5.171}$$

when $\mathcal{N}$ arises from the use of an F_j.

Suppose an $\mathcal{N}(t)$ can be found such that

$$K(Z;t) = 0. \tag{6.5.172}$$

This is, in principle, always possible because we can take the Z to be the initial conditions and take $\mathcal{N}(t)$ to be the symplectic map that transforms final conditions into initial conditions. If an F_j can be found such that $\mathcal{N}(t)$ arises from the use of this F_j, then combining (5.171) and (5.172) gives the Hamilton-Jacobi relation/equation

$$H(z(Z,t);t) + \partial F_j/\partial t = 0. \tag{6.5.173}$$

Exercises

6.5.1. Consider linear symplectic maps of the form (3.3.9), (3.3.10), (3.3.11), (3.10.16), and (3.10,19). Determine which generating functions F_j can be used in these cases, and find explicitly those that are applicable.

6.5.2. Consider the matrices (3.3.9) through (3.3.11). Show that they can all be produced by one of the mixed-variable generating functions F_1 through F_4, and hence any symplectic matrix can be produced by using a *sequence* of such generating functions.

6.5.3. We have seen that the matrix R given by (4.8.31) cannot be produced by any one of the mixed-variable generating functions F_1 through F_4. Refer to (4.8.27). Show that there are matrices M near R for which none of the matrices a through d are invertible and hence for these M the method of mixed-variable generating function symplectification using F_1 through F_4 fails.

6.5.4. Use the machinery of Section 4.2 to produce the relations (5.167).

6.5.5. In some situations, for example in passing from Cartesian to curvilinear coordinates in configuration (position) space, it is desirable to make configuration coordinate transformations of the kind

$$Q_k = f_k(q,t). \tag{6.5.174}$$

Transformations of this kind are called *Lagrange point transformations.* Here we assume that the relations (5.174) are invertible so that there are functions $g_k(Q,t)$ such that

$$q_k = g_k(Q,t). \tag{6.5.175}$$

If this change of variables is done in a canonical context, we would like to extend the configuration-space transformation (5.174) into a full phase-space transformation. The purpose of this exercise is to show that this extension can be done symplectically with the aid of the generating function F_2 given by

$$F_2(q,P,t) = \sum_{m=1}^{n} P_m f_m(q,t). \tag{6.5.176}$$

This symplectic extension is called a *lift* of the configuration coordinate transformation from configuration space to phase space.

Review Subsection 5.1. Show, with the aid of (5.5), that use of the F_2 given by (5.176) yields the desired relation (5.174). Find the matrices α through δ in this case and verify that the matrix β is invertible (as required) if, as has been assumed, (5.174) is invertible. You should find that

$$\alpha_{k\ell} = \partial^2 F_2/\partial q_k \partial q_\ell = \sum_{m=1}^{n} P_m \partial^2 f_m(q,t)/\partial q_k \partial q_\ell, \tag{6.5.177}$$

$$\beta_{k\ell} = \partial^2 F_2/\partial q_k \partial P_\ell = \partial f_\ell/\partial q_k, \tag{6.5.178}$$

$$\gamma_{k\ell} = \partial^2 F_2/\partial P_k \partial q_\ell = \partial f_k/\partial q_\ell = (\beta^T)_{k\ell}, \tag{6.5.179}$$

$$\delta_{k\ell} = \partial^2 F_2/\partial P_k \partial P_\ell = 0. \tag{6.5.180}$$

See (5.16) through (5.19).

Verify, using (5.5), that there is the relation

$$p_k = \sum_{m=1}^{n} P_m \beta_{km} = \sum_{m=1}^{n} \beta_{km} P_m, \tag{6.5.181}$$

which can be written in the compact matrix-vector form

$$p = \beta P. \tag{6.5.182}$$

It follows that there is the relation

$$P = \beta^{-1} p, \tag{6.5.183}$$

which specifies the transformed momenta associated with the transformed positions (5.174). Verify from (5.174) and (5.178) that there is the differential relation

$$dQ = \beta^T dq. \tag{6.5.184}$$

Canonical transformations given by relations of the form (5.174) and (5.183) are called *Mathieu* transformations, and (5.176) may be called a Mathieu generating function.

Verify that for Mathieu transformations the corresponding A through D matrices are given by the relations

$$A = \gamma - \delta\beta^{-1}\alpha = \gamma = \beta^T, \tag{6.5.185}$$

$$B = \delta\beta^{-1} = 0, \tag{6.5.186}$$

$$C = -\beta^{-1}\alpha, \tag{6.5.187}$$

$$D = \beta^{-1}. \tag{6.5.188}$$

According to Exercise 3.10.5, symplectic matrices with $B = 0$ form a subgroup of the symplectic group. Since the Jacobian of the product of two maps is the product of their Jacobians, it follows that Mathieu transformations form a subgroup of the group of symplectic maps. What is the nature of this subgroup? Evidently invertible Lagrange point transformations form a group which, assuming the underlying topology of configuration space to be Cartesian/Euclidean, is (under differentiability assumptions) the diffeomorphism group *Diff* $(\mathbb{R}^n)$. Thus, the subgroup of Mathieu transformations is isomorphic to the group *Diff* $(\mathbb{R}^n)$.

Suppose that the transformation (5.174) is in fact *linear* so that it can be written in the form

$$Q = Nq \tag{6.5.189}$$

where N is any real and invertible $n \times n$ matrix. That is, $N \in GL(n, \mathbb{R})$. Find the matrices α through δ and A through D in this case. Show, in particular, that in this case $\alpha = 0$ so that $C = 0$, and that corresponding to (5.189) there is the complementary relation

$$P = (N^T)^{-1} p. \tag{6.5.190}$$

Compare this result to (3.3.13). Verify that all Mathieu transformations for which a relation of the form (5.189) holds constitute a subgroup of the group of all Mathieu transformations, and this subgroup is isomorphic to $GL(n, \mathbb{R})$.

Suppose that the transformation (5.174) is in fact linear and *orthogonal* so that it can be written in the form

$$Q = Oq \tag{6.5.191}$$

where O is an orthogonal matrix. Show that in this case there is the complementary relation

$$P = Op. \tag{6.5.192}$$

Compare this result to (3.3.13) and the discussion of $SO(n, \mathbb{R})$ at the end of Section 7.2.2 and in Exercise 7.2.5. Verify that all Mathieu transformations for which a relation of the form (5.191) holds constitute a subgroup of the group of all Mathieu transformations, and this subgroup is isomorphic to $O(n, \mathbb{R})$.

6.5.6. Consider F_2 generating functions of the form

$$F_2(q, P, t) = -\chi(q, t) + \sum_{m=1}^{n} P_m q_m. \tag{6.5.193}$$

Show that these F_2 produce symplectic transformations of the form

$$Q_m = q_m, \tag{6.5.194}$$

$$P_m = p_m + \partial\chi/\partial q_m. \tag{6.5.195}$$

These symplectic transformations/maps are sometimes called *gauge* transformations because they arise naturally, for the case $n = 3$, in the context of charged-particle motion in electromagnetic fields.[17] Indeed, we have already seen in Exercise 2.8 that gauge transformations are symplectic maps.

Show that for gauge transformations the matrices α through δ and A through D are given by the relations

$$\alpha_{k\ell} = \partial^2 F_2/\partial q_k \partial q_\ell = -\partial^2\chi/\partial q_k \partial q_\ell, \tag{6.5.196}$$

$$\beta_{k\ell} = \partial^2 F_2/\partial q_k \partial P_\ell = \bar{\delta}_{k\ell}, \tag{6.5.197}$$

$$\gamma_{k\ell} = \partial^2 F_2/\partial P_k \partial q_\ell = \bar{\delta}_{k\ell}, \tag{6.5.198}$$

$$\delta_{k\ell} = \partial^2 F_2/\partial P_k \partial P_\ell = 0; \tag{6.5.199}$$

$$A = \gamma - \delta\beta^{-1}\alpha = \gamma = I, \tag{6.5.200}$$

$$B = \delta\beta^{-1} = 0, \tag{6.5.201}$$

$$C_{k\ell} = -(\beta^{-1}\alpha)_{k\ell} = -\alpha_{k\ell} = \partial^2\chi/\partial q_k \partial q_\ell, \tag{6.5.202}$$

$$D = \beta^{-1} = I. \tag{6.5.203}$$

Here we have used the symbol $\bar{\delta}_{k\ell}$ to denote the Kronecker delta.

[17] However note that the symplectic transformations given by (5.194) and (5.195) are defined for all n.

Compare the matrices A through D found above with those for (3.3.10). Show that gauge transformation symplectic maps form a subgroup of the set of all symplectic maps. Hint: See Exercise 3.10.1.

What is the nature of this subgroup? Define a symplectic map $\mathcal{M}$ by the rule

$$\mathcal{M} = \exp :\chi(q,t): . \tag{6.5.204}$$

Verify that the assertions

$$Q = \mathcal{M}q, \tag{6.5.205}$$

$$P = \mathcal{M}p \tag{6.5.206}$$

yield (5.194) and (5.195). Let $\chi(q,t)$ and $\chi'(q,t)$ be any two gauge functions. Evidently there is the relation

$$[\chi, \chi'] = 0. \tag{6.5.207}$$

It follows that the maps $\mathcal{M}$ and $\mathcal{M}'$ defined by (5.204) and

$$\mathcal{M}' = \exp :\chi'(q,t): \tag{6.5.208}$$

commute. Also, observe that functions of the form $\chi(q,t)$ comprise an infinite-dimensional vector space. Therefore the set of gauge transformations comprises an infinite-dimensional Abelian group.

6.6 Generating Functions Come from an Exact Differential

6.6.1 Overview

So far the discussion of generating functions has been relatively straight forward, but not particularly illuminating. Let us write (5.3) in the form

$$Z = \mathcal{M}z. \tag{6.6.1}$$

By this relation we mean that there are $2n$ functions $K_a(z,t)$ of the $2n$ variables z_b, and perhaps the time t, such that

$$Z_a = K_a(z,t). \tag{6.6.2}$$

That is, in general $2n$ functions are required to specify a map in $2n$ variables.

However in the last section we have seen that, with the use of any one of the generateng functions F_1 through F_4, all the required $2n$ functions come from a *single* master function, namely the generating function being employed. How does it happen that the information required to specify $2n$ functions can come from a single function? Presumably this occurs because in our case the $2n$ functions $K_a(z,t)$ are *not*, in fact, independent. Of course, in principle we know that they are not independent because of the assumption that $\mathcal{M}$ is symplectic. Apparently the symplectic condition is so stringent as to reduce the number of required functions down from $2n$ to a *single* function. In one sense this should not be too surprising, because we know that any family of symplectic maps $\mathcal{M}(t)$ is generated by a single function, namely the Hamiltonian. But that is an infinitesimal statement. How, precisely, could one have guessed that there were functions F_1 through F_4 that could be used in the manner (5.4) through (5.7) to manufacture symplectic maps? And can all symplectic maps be obtained in this fashion? Below we present a partial clue. Still deeper insight is presented in Subsection 7.1. There we will learn that the functions F_1 through F_4 are but 4 members of a $2n(4n+1)$ parameter family of generating functions, all of which can be used to manufacture symplectic maps. The final explanation is given in Subsection 7.2.

6.6.2 A Democratic Differential Form

6.6.2.1 Definition

Consider the differential form

$$\omega_d = (Z, JdZ) - (z, Jdz). \tag{6.6.3}$$

Note that ω_d involves all the $4n$ variables z and Z. It has the beauty that it treats the coordinates and momenta on an equal footing, and is "*democratic*" in its use of z and Z. Also, we will see in Subsection 7.2 that it arises in a natural way.[18]

6.6.2.2 The Democratic Differential Form Is Exact Iff $\mathcal{M}$ Is Symplectic

Suppose the Z's are viewed as functions of the z's by using (5.23) to write ω_d as

$$\omega_d(z) = (Z, JMdz) - (z, Jdz). \tag{6.6.4}$$

Then, if the $\mathcal{M}$ in (6.1) is a symplectic map, we will find that ω_d is *exact* with respect to the $2n$ variables z. Similarly, if the z's are viewed as functions of the Z's, ω_d can be rewritten as

$$\omega_d(Z) = (Z, JdZ) - (z, JM^{-1}dZ). \tag{6.6.5}$$

It can be shown that this form is also exact (with respect to the $2n$ variables Z) if $\mathcal{M}$ is a symplectic map. We will soon verify these claims by brute calculation. Subsection 7.2 will find the same results in an obvious way.

To see that ω_d is exact with respect to the variables z, observe that that it can be written more explicitly as

$$\begin{aligned}\omega_d(z) &= (Z, JMdz) - (z, Jdz) = (M^T J^T Z, dz) - (J^T z, dz)\\ &= \sum_b [(M^T J^T Z)_b - (J^T z)_b] dz_b.\end{aligned} \tag{6.6.6}$$

Upon comparing (6.6) with (1.22), we see that the coefficients $C_b(z,t)$ are given by the relation

$$C_b(z,t) = (M^T J^T Z)_b - (J^T z)_b. \tag{6.6.7}$$

[18] Despite its attractive appearance, the differential form ω_d given by (6.3) is not commonly employed by (and perhaps unfamiliar to some) other authors. Its existence, exactness, and utility were known, however, to Poincaré.

Note that there is a possible time dependence since $\mathcal{M}$ may depend on t. However, as before, t only plays the role of a parameter.

We must see if the conditions (1.26) are met. An easy computation gives for the second term in (6.7) the result

$$(\partial/\partial z_a)(J^T z)_b = (\partial/\partial z_a)\sum_c (J^T)_{bc} z_c = \sum_c (J^T)_{bc}\delta_{ac} = (J^T)_{ba} = J_{ab}. \tag{6.6.8}$$

Dealing with the first term in (6.7) is more complicated. We find the preliminary result

$$\begin{aligned}(\partial/\partial z_a)(M^T J^T Z)_b &= (\partial/\partial z_a)\sum_c (M^T J^T)_{bc} Z_c = \sum_c [(\partial/\partial z_a)(M^T J^T)_{bc}] Z_c \\ &+ \sum_c (M^T J^T)_{bc}(\partial/\partial z_a) Z_c. \end{aligned}\tag{6.6.9}$$

But, from (5.23), there is the relation

$$(\partial/\partial z_a) Z_c = M_{ca}. \tag{6.6.10}$$

It follows that for the second term on the right side of (6.9) there is the simplification

$$\sum_c (M^T J^T)_{bc}(\partial/\partial z_a) Z_c = \sum_c (M^T J^T)_{bc} M_{ca} = (M^T J^T M)_{ba} = (M^T J M)_{ab}. \tag{6.6.11}$$

For the first term on the right side of (6.9) there is the result

$$\begin{aligned}\sum_c [(\partial/\partial z_a)(M^T J^T)_{bc}] Z_c &= \sum_c Z_c (\partial/\partial z_a)(JM)_{cb} = \sum_{cd} Z_c J_{cd} (\partial/\partial z_a) M_{db} \\ &= \sum_{cd} Z_c J_{cd} (\partial^2 Z_d/\partial z_a \partial z_b). \end{aligned}\tag{6.6.12}$$

Here we have again used a variant of (6.10). Combining (6.8) (6.9), (6.11), and (6.12) gives the net result

$$(\partial/\partial z_a) C_b = [(M^T J M)_{ab} - J_{ab}] + [\sum_{cd} Z_c J_{cd} (\partial^2 Z_d/\partial z_a \partial z_b)]. \tag{6.6.13}$$

Here we have separated the right side of (6.13) into parts that are antisymmetric and symmetric under the interchange of a and b. It follows that there is the relation

$$(\partial/\partial z_a) C_b - (\partial/\partial z_b) C_a = 2[M^T J M - J]_{ab}. \tag{6.6.14}$$

Consequently the differential form $\omega_d(z)$ given by (6.4) is exact if, and only if, the map $\mathcal{M}$ is symplectic.

It can be shown in a similar way that the differential form $\omega_d(Z)$ given by (6.5) is exact if, and only if, the map $\mathcal{M}$ is symplectic.

6.6.3 Information about $\mathcal{M}$ Carried by the Democratic Form

Since (6.3) is exact, there is a function $F(z,t)$ such that

$$dF = \omega_d = (Z, JdZ) - (z, Jdz). \tag{6.6.15}$$

The function $F(z,t)$ may be called the *primitive* function associated with the differential form ω_d.[19]

How much information does $F(z,t)$ carry about $\mathcal{M}$? Put another way, by its definition, the differential form ω_d depends on $\mathcal{M}$. Are there possibly several maps $\mathcal{M}$ that produce the same differential form ω_d? According to (5.23) and (6.1) we may rewrite write (6.15) in the form

$$dF_{\mathcal{M}} = (\mathcal{M}z, JMdz) - (z, Jdz) \tag{6.6.16}$$

where have have appended the subscript $\mathcal{M}$ to F to indicate that F depends on $\mathcal{M}$. We will begin our exploration of this uniqueness question by considering various symplectic maps $\mathcal{M}$.

Suppose $\mathcal{M}$ is a member of the *inhomogeneous* symplectic group $ISp(2n,\mathbb{R})$. See Subsection 2.2 and Section 9.2. At this point it is convenient to use Lie-algebraic notation and tools. See Chapter 7 for details. Let f_1 be a first-degree polynomial such that

$$\exp(:f_1:)z = z + \delta. \tag{6.6.17}$$

We define a *translation* operator τ by writing

$$\tau = \exp(:f_1:)z \tag{6.6.18}$$

so that τ has the action

$$\tau z = z + \delta. \tag{6.6.19}$$

Also, let

$$\mathcal{R}_f = \exp(:f_2^c:)\exp(:f_2^a:),\ \text{etc.} \tag{6.6.20}$$

be a general linear symplectic map. It has the action

$$\mathcal{R}_f z = R_f z \tag{6.6.21}$$

where R_f is a general symplectic matrix. From the work of Section 9.2 we know that any element in $ISp(2n,\mathbb{R})$ can be written in the factored form

$$\mathcal{M}_f = \tau\mathcal{R} \tag{6.6.22}$$

and has the action

$$Z = \mathcal{M}_f z = R_f\delta + R_f z. \tag{6.6.23}$$

Let us employ this result in (6.16). We observe from (6.23) that there is the relation

$$M = R_f. \tag{6.6.24}$$

[19] In calculus parlance the terms antiderivative, primitive function, primitive integral, and indefinite integral are used interchangeably.

Therefore, in this case, (6.16) takes the form

$$dF_{\mathcal{M}_f} = (R_f\delta + R_f z, JR_f dz) - (z, Jdz) = (R_f\delta, JR_f dz) + (R_f z, JR_f dz) - (z, Jdz). \quad (6.6.25)$$

But, employing the symplectic condition for R_f yields the results

$$(R_f\delta, JR_f dz) = (\delta, R_f^T JR_f dz) = (\delta, Jdz), \quad (6.6.26)$$

$$(R_f z, JR_f dz) - (z, Jdz) = (z, R_f^T JR_f dz) - (z, Jdz) = (z, Jdz) - (z, Jdz) = 0. \quad (6.6.27)$$

Consequently, (6.25) becomes

$$dF_{\mathcal{M}_f} = (\delta, Jdz) \quad (6.6.28)$$

and therefore

$$F_{\mathcal{M}_f} = (\delta, Jz) + C \quad (6.6.29)$$

where C is an arbitrary additive constant.

What can we conclude from looking at (6.28)? First, suppose there is no translation part so that $\delta = 0$ and $\tau = \mathcal{I}$. Then we see from (6.22) that

$$\mathcal{M}_f = \mathcal{R}_f, \quad (6.6.30)$$

and it follows from (6.28) that

$$dF_{\mathcal{R}_f} = 0. \quad (6.6.31)$$

Consequently, up to an inconsequential additive constant, all *linear* symplectic maps produce the *same* primitive function and this primitive function may be taken to be *zero*. Second, (6.29) can be rewritten in the form

$$dF_{\mathcal{M}_f} = dF_{\tau\mathcal{R}_f} = dF_\tau \quad (6.6.32)$$

with

$$dF_\tau = (\delta, Jdz). \quad (6.6.33)$$

That is, in the case of $ISp(2n, \mathbb{R})$ and when the factorization (6.22) is employed, the differential form $dF_{\mathcal{M}_f}$ depends *only* on the *translation* part of $\mathcal{M}_f$.

The results obtained this far suggest the following exploration: Suppose $\mathcal{N}$ is any symplectic map.[20] We will write $\mathcal{N}$ in the Lie form

$$\mathcal{N} = \exp(: f_1 :)\exp(: f_2^c :)\exp(: f_2^a :)\exp(: f_3 :)\exp(: f_4 :)\cdots . \quad (6.6.34)$$

Now define a map $\mathcal{M}$ by writing

$$\mathcal{M} = \mathcal{N}\mathcal{L} \quad (6.6.35)$$

where $\mathcal{L}$ is a *linear* symplectic map with the action

$$\mathcal{L}z = Lz. \quad (6.6.36)$$

That is, $L \in Sp(2n, \mathbb{R})$. Let us consider $dF_{\mathcal{M}}$ in this case,

$$dF_{\mathcal{M}} = (\mathcal{M}z, JMdz) - (z, Jdz) \quad (6.6.37)$$

[20] That is, using the notation introduced in Subsection 2.2, assume only that $\mathcal{N} \in ISpM(2n, \mathbb{R})$.

where M is the Jacobian of $\mathcal{M}$.

Because $\mathcal{L}$ is a linear map, we may write

$$\mathcal{M}z = L\mathcal{N}z. \tag{6.6.38}$$

Also, by the chain rule, we know that M is given by the relation

$$M = LN \tag{6.6.39}$$

where N is the Jacobian of $\mathcal{N}$. Next make use of (6.38) and (6.39) and the symplectic condition to carry out the series of deductions

$$(\mathcal{M}z, JMdz) = (L\mathcal{N}z, JLNdz) = (\mathcal{N}z, L^T JLNdz) = (\mathcal{N}z, JNdz). \tag{6.6.40}$$

From (6.40) we conclude that

$$dF_{\mathcal{M}} = dF_{\mathcal{N}\mathcal{L}} = dF_{\mathcal{N}}. \tag{6.6.41}$$

[Note that (6.32) is a special case of (6.41).] Thus, up to a possible additive constant which is of no consequence, $F(z,t)$ is the *same* for all maps $\mathcal{M}$ obtained by *right* multiplication of $\mathcal{N}$ by linear symplectic maps $\mathcal{L}$. In coset language, F depends only on the left cosets $ISpM(2n,\mathbb{R})/Sp(2n,\mathbb{R})$. Recall Section 5.12 for a discussion of cosets.

6.6.4 Breaking the Degeneracy

Although the fact that many symplectic maps lead to the same F may seem alarming, we shall soon be adding other functions to F that will break this degeneracy. Let us rewrite (6.15) in terms of the q,p and the Q,P. It is easily verified that

$$(z, Jdz) = \sum_i (q_i dp_i - p_i dq_i) \text{ and } (Z, JdZ) = \sum_i (Q_i dP_i - P_i dQ_i). \tag{6.6.42}$$

It follows that

$$dF = \sum_i [(Q_i dP_i - P_i dQ_i) - (q_i dp_i - p_i dq_i)]. \tag{6.6.43}$$

This is a key result from which all the relations (5.4) through (5.7) follow.

6.6.4.1 F_2 Example

For example, suppose we define F_2 by the rule

$$F_2 = [F + \sum_i (p_i q_i + P_i Q_i)]/2. \tag{6.6.44}$$

Then we find the result

$$dF_2 = [dF + \sum_i (p_i dq_i + q_i dp_i + P_i dQ_i + Q_i dP_i)]/2 = \sum_i (p_i dq_i + Q_i dP_i). \tag{6.6.45}$$

Evidently the left side of (6.45) is an exact differential, and comparison with (5.71) provides an independent proof that the differential form ω_2 given by (5.81) is exact. Finally, from (6.45), we immediately derive the relations (5.5).

6.6.4.2 F_1 Example

As a second example, suppose we define F_1 by the rule

$$F_1 = [F + \sum_i (p_i q_i - P_i Q_i)]/2. \tag{6.6.46}$$

Then we find the result

$$dF_1 = [dF + \sum_i (p_i dq_i + q_i dp_i - P_i dQ_i - Q_i dP_i]/2 = \sum_i (p_i dq_i - P_i dQ_i). \tag{6.6.47}$$

From (6.47) it follows that

$$\partial F_1/\partial q_i = p_i, \quad \partial F_1/\partial Q_i = -P_i, \tag{6.6.48}$$

which are the relations (5.4). The relations (5.6) and (5.7) follow in a similar fashion. See Exercise 6.4.

6.6.4.3 Poincaré Generating Function

There are also other generating functions beside F_1 through F_4 that are less familiar. For example, consider the function F_+ defined by the rule

$$F_+ = F + (Z, Jz) = F + \sum_i (p_i Q_i - q_i P_i). \tag{6.6.49}$$

For F_+ we find the differential

$$\begin{aligned} dF_+ &= dF + \sum_i [(p_i dQ_i + Q_i dp_i) - (P_i dq_i + q_i dP_i)] \\ &= \sum_i [(Q_i dP_i - P_i dQ_i) - (q_i dp_i - p_i dq_i) + (p_i dQ_i + Q_i dp_i) - (P_i dq_i + q_i dP_i)] \\ &= \sum_i [(Q_i - q_i)(dp_i + dP_i) - (P_i - p_i)(dq_i + dQ_i)]. \end{aligned} \tag{6.6.50}$$

We may therefore write

$$F_+ = F_+[(q+Q), (p+P), t] \tag{6.6.51}$$

to obtain from (6.50) the relations

$$Q_i - q_i = \partial F_+/\partial(p_i + P_i), \tag{6.6.52}$$

$$P_i - p_i = -\partial F_+/\partial(q_i + Q_i). \tag{6.6.53}$$

The function F_+ is sometimes called a *Poincaré generating function.* It is more democratic then the functions F_1 through F_4 in the sense that it involves the old and new variables in a symmetric fashion. Indeed, introduce the quantities Σ and Δ by the rules

$$\Sigma = Z + z, \tag{6.6.54}$$

$$\Delta = Z - z. \tag{6.6.55}$$

Then for (6.51) we may write

$$F_+ = F_+(\Sigma, t). \tag{6.6.56}$$

Correspondingly, the relations (6.52) and (6.53) can be written in the compact form

$$\Delta = J\partial_\Sigma F_+|_{\Sigma = Z+z}. \tag{6.6.57}$$

Here, by employing J in (6.57), we have viewed $\Delta = \{(Q-q), (P-p)\}$ as composed of a canonical pair. We remark that if F_+ is a quadratic function of Σ, then the use of (6.57) leads to the Cayley transformation. See Exercise 6.5.

There is an important feature of the Poincaré generating function that is sometimes useful. A point Σ^c is called a *critical point* of F_+ if there is the result

$$\partial_\Sigma F_+|_{\Sigma=\Sigma^c} = 0. \tag{6.6.58}$$

From (6.57) we see that at a critical point $\Delta = 0$ so that

$$Z = \mathcal{M}z = z \tag{6.6.59}$$

and, by (6.54),

$$Z = z = \Sigma^c/2. \tag{6.6.60}$$

Thus, if we make the definition

$$z^f = \Sigma^c/2, \tag{6.6.61}$$

we see that z^f is a *fixed point* of $\mathcal{M}$. We conclude that critical points of F_+ correspond to fixed points of $\mathcal{M}$ and vice versa. This result can be useful in some circumstances because there are theorems (e.g. *Morse theory*) about critical points of smooth functions on various manifolds.

Exercises

6.6.1. Show that the differential form $\omega_d(Z)$ given by (6.5) is exact in terms of the variables Z.

6.6.2. Verify the claims (6.38) and (6.39).

6.6.3. Review Section 1.2.3. Define maps $\mathcal{S}$ and $\mathcal{R}(\phi)$ by the rules

$$\mathcal{S} = \exp(: q^3 :), \tag{6.6.62}$$

$$\mathcal{R}(\phi) = \exp[-(\phi/2) :p^2 + q^2:] \tag{6.6.63}$$

Let $\mathcal{M}$ be the map

$$\mathcal{M} = \mathcal{S}\mathcal{R}(\phi). \tag{6.6.64}$$

Verify that the maps (1.2.50) and (6.64) are related by a similarity transformation involving a map of the form $\mathcal{R}(\psi)$, which amounts to changing the observation point O.

Show that for the democratic differential form ω_d given by (6.3) there are the results

$$dF_{\mathcal{M}} = dF_{\mathcal{SR}(\phi)} = dF_{\mathcal{S}} =, \tag{6.6.65}$$

$$F_{\mathcal{M}} = . \tag{6.6.66}$$

Why is there no ϕ dependence?

6.6.4. Consider the functions F_3 and F_4 defined by the relations

$$F_3 = [F - \sum_i (p_i q_i + P_i Q_i)]/2, \tag{6.6.67}$$

$$F_4 = [F - \sum_i (p_i q_i - P_i Q_i)]/2. \tag{6.6.68}$$

where dF is the exact differential (6.15). Show that they satisfy (5.6) and (5.7). Compare (6.67) and (6.68) with (6.44) and (6.46). Observe that all these relations differ only in the signs assigned to the quantities $p_i q_i$ and $P_i Q_i$, and that the four possibilities yield the four functions F_1 through F_4.

6.6.5. Suppose that the Poincaré generating function F_+ is of the form

$$F_+(\Sigma) = (1/2)(\Sigma, W\Sigma) \tag{6.6.69}$$

where W is a symmetric matrix. It follows from (6.69) that in this case

$$\partial_\Sigma F_+ = W\Sigma. \tag{6.6.70}$$

Show that this F_+, when employed in (6.57), produces the result

$$\Delta = JW\Sigma, \tag{6.6.71}$$

which in turn yields the relation

$$Z - z = JW(Z + z). \tag{6.6.72}$$

Solve this relation for Z in terms of z to yield the linear relation

$$Z = (I - JW)^{-1}(I + JW)z. \tag{6.6.73}$$

Finally write the relation between Z and z in terms of a matrix M,

$$Z = Mz. \tag{6.6.74}$$

Comparison of (6.73) and (6.74) then gives the result

$$M = (I - JW)^{-1}(I + JW). \tag{6.6.75}$$

Observe that this result is the Cayley representation (3.12.5). It follows, as expected, that M is a symplectic matrix. Verify that the relation (6.75) can be inverted to give the result

$$W = -J(M - I)(M + I)^{-1} \tag{6.6.76}$$

in agreement with (3.12.19). Verify that there are the Cayley Möbius transformation relations

$$W = T_{\sigma}(M) \tag{6.6.77}$$

and

$$M = T_{\sigma^{-1}}(W) \tag{6.6.78}$$

with σ given by (5.13.11).

Suppose that F_+ is of the form

$$F_+(\Sigma) = (v, \Sigma) + (1/2)(\Sigma, W\Sigma) \tag{6.6.79}$$

where v is any vector. It follows from (6.79) that in this case

$$\partial_{\Sigma} F_+ = v + W\Sigma. \tag{6.6.80}$$

Show that this F_+, when employed in (6.57), produces the result

$$Z = Mz + (I - JW)^{-1} Jv. \tag{6.6.81}$$

6.6.6. Verify, using the methods of Subsection 5.1, that the Poincaré generating function F_+ when employed in (6.57) does indeed produce a symplectic map. Suppose that F_+ is time dependent so that its use produces a one-parameter family of symplectic maps. Find the associated generating Hamiltonian for this family. Hint: If you are stuck, see Subsection 7.3.1.

6.6.7. Find a generating function F_- analogous to F_+. That is, in (6.49), replace $+(Z, Jz)$ by $-(Z, Jz)$.

6.7 Plethora of Generating Functions

We have seen that there are the five generating function types F_1 through F_4 and F_+. We will now learn that (for a $2n$-dimensional phase space) there are an *infinite* number of generating functions types, of which the five cited above are but examples, that comprise a full $2n(4n + 1)$ parameter family. Indeed, there is a generating function type for each of the Darboux matrices/transformations of Section 5.13. And we know there is a distinct Darboux matrix corresponding to each element of $Sp(4n)$ whose dimension is $2n(4n + 1)$. See (5.13.28), (3.7.35), and Table 3.7.1. Thus, for example, in the simplest case of a two-dimensional phase space, there is a 10 parameter family of generating function types; and in the case of a six-dimensional phase space there is a 78 parameter family of generating function types. We will begin with a derivation of this result, and then follow the derivation with a discussion describing how the various results we have found all fit together.

6.7.1 Derivation

Consider again the differential form ω_d given by (6.3). If Z and z are $2n$ dimensional, let $\hat{Z}$ denote the $4n$ dimensional column vector *constructed* by appending the entries of z below those of Z,

$$\hat{Z} = (Z; z)^T. \tag{6.7.1}$$

With this notation, the differential form ω_d can be rewritten as

$$\omega_d = (Z, JdZ) - (z, Jdz) = (\hat{Z}, \tilde{J}^{4n} d\hat{Z}). \tag{6.7.2}$$

Here, as in section 5.13.3, we have used (5.13.16) to define $\tilde{J}^{4n}$.

Let α denote a $4n \times 4n$ invertible matrix. Use α to define new variables $\hat{U}$ by the rule

$$\hat{U} = \alpha \hat{Z}, \tag{6.7.3}$$

or

$$\hat{Z} = \alpha^{-1} \hat{U}. \tag{6.7.4}$$

With this change of variables the differential form on the right side of (7.2) becomes

$$(\hat{Z}, \tilde{J}^{4n} d\hat{Z}) = (\alpha^{-1}\hat{U}, \tilde{J}^{4n} \alpha^{-1} d\hat{U}) = (\hat{U}, (\alpha^{-1})^T \tilde{J}^{4n} (\alpha^{-1}) d\hat{U}). \tag{6.7.5}$$

Next, inspired by the discussion of Section 5.13, require that α satisfy the relation

$$(\alpha^{-1})^T \tilde{J}^{4n} (\alpha^{-1}) = J^{4n}. \tag{6.7.6}$$

That is, require that α be a Darboux transformation. Also, in analogy with (7.1), introduce $4n$ dimensional vectors $\hat{U}$ by letting U and u be the first $2n$ entries and last $2n$ entries of $\hat{U}$, respectively,

$$\hat{U} = (U; u)^T. \tag{6.7.7}$$

Upon combining (7.5) through (7.7), we find the result

$$(\hat{Z}, \tilde{J}^{4n} d\hat{Z}) = (\hat{U}, (\alpha^{-1})^T \tilde{J}^{4n} (\alpha^{-1}) d\hat{U}) = (\hat{U}, J^{4n} d\hat{U}) = (U, du) - (u, dU). \tag{6.7.8}$$

We know that $(\hat{Z}, \tilde{J}^{4n} d\hat{Z})$ is exact. See (6.15). Therefore, from the work so far, we have the relation

$$dF = (U, du) - (u, dU). \tag{6.7.9}$$

From F construct another function g by the rule

$$g = [F + (U, u)]/2. \tag{6.7.10}$$

Then, by this construction and the properties of F, g has the differential

$$dg = [dF + (U, du) + (u, dU)]/2 = [(U, du) - (u, dU) + (U, du) + (u, dU)]/2 = (U, du). \tag{6.7.11}$$

By the chain rule, we also have the relation

$$dg = \sum_a [(\partial g/\partial U_a)(dU_a) + (\partial g/\partial u_a)(du_a)]. \tag{6.7.12}$$

Upon comparing (7.11) and (7.12) we deduce the two relations

$$\partial g/\partial U_a = 0, \tag{6.7.13}$$

$$U_a = \partial g/\partial u_a. \tag{6.7.14}$$

The first of these states the remarkable result that g depends *only* on u,

$$g = g(u). \tag{6.7.15}$$

The second states that U and u are related by the *gradient* map $\mathcal{G}$ produced by the function g playing the role of a source function. More abstractly, we may write (7.14) in the form

$$U = \mathcal{G}u. \tag{6.7.16}$$

See Subsection 1.1.

When written in block form, (7.3) is equivalent to the two relations

$$U = A^\alpha Z + B^\alpha z, \tag{6.7.17}$$

$$u = C^\alpha Z + D^\alpha z. \tag{6.7.18}$$

When expanded in component form, (7.17) reads

$$U_a = \sum_b [(A^\alpha)_{ab} Z_b + (B^\alpha)_{ab} z_b], \tag{6.7.19}$$

and (7.18) reads

$$u_c = \sum_d [(C^\alpha)_{cd} Z_d + (D^\alpha)_{cd} z_d]. \tag{6.7.20}$$

In terms of these components, the relations (7.14) become

$$\sum_b [(A^\alpha)_{ab} Z_b + (B^\alpha)_{ab} z_b] = [\partial g/\partial u_a]|_{u=C^\alpha Z + D^\alpha z}. \tag{6.7.21}$$

In matrix-vector form they can be written more compactly as

$$A^\alpha Z + B^\alpha z = \partial_u g|_{u=C^\alpha Z + D^\alpha z}. \tag{6.7.22}$$

Equations (7.21) provide $2n$ *implicit* relations between Z and z. When made *explicit*, they produce the map $\mathcal{M}$ (which will soon be shown to be symplectic) with

$$Z = \mathcal{M}z. \tag{6.7.23}$$

Equations (7.21) relate the symplectic map $\mathcal{M}$ to the gradient map $\mathcal{G}$ associated with the source function g. Any such function produces a symplectic map, and we will call g, in association with the Darboux matrix α, the *generating* function that produces $\mathcal{M}$. Note that in principle, although we have not taken note of it until now, g could also depend on the time t,

$$g = g(u, t). \tag{6.7.24}$$

Here t should again be regarded as a parameter, and its presence in g when employed in (7.14) and (7.21) leads to a parameter dependent gradient map $\mathcal{G}(t)$ and a parameter dependent symplectic map $\mathcal{M}(t)$.

At this point we note that there is another way of viewing the relations (7.21). Suppose we pick a $2n$-vector u and, if $\mathcal{G}$ depends on t, also specify a value for t. Then, according to (7.16), we can determine $U(u,t)$ by the rule

$$U(u,t) = \mathcal{G}(t)u. \tag{6.7.25}$$

In view of (7.7), we have now also specified $\hat{U}(u,t)$. Indeed, we have the relation

$$\hat{U}(u,t) = (U(u,t);u)^T = (\mathcal{G}(t)u;u)^T. \tag{6.7.26}$$

Next, use the Darboux relation (7.4) to determine $\hat{Z}(u,t)$ by writing

$$\hat{Z}(u,t) = \alpha^{-1}\hat{U}(u,t). \tag{6.7.27}$$

When written in block form, (7.27) yields the relations

$$Z(u,t) = A^{\alpha^{-1}}\mathcal{G}(t)u + B^{\alpha^{-1}}u, \tag{6.7.28}$$

and

$$z(u,t) = C^{\alpha^{-1}}\mathcal{G}(t)u + D^{\alpha^{-1}}u. \tag{6.7.29}$$

We see that (7.28) and (7.29) specify the map $\mathcal{M}(t)$ in *parametric* form with u being a set of $2n$ parameters.

We still have to verify that the implicit relations (7.21), or the parametric relations (7.28) and (7.29), can be made explicit. That is, given the z's, we need to show that we can solve for the Z's, and vice versa. At this point one might ask if there is a choice of α such that the relation (7.21) is already explicit. Exercise 7.3 shows that this is impossible. There is no such α that also satisfies the Darboux requirement (5.13.20). Thus, we must begin with implicit relations. We also note that gradient map $\mathcal{G}(t)$ employed in (7.21) is explicit in form. Therefore, as we will next see, the implicit nature of (7.21) with regard to the z's and Z's is controlled primarily by the properties of the associated Darboux matrix α and its related Möbius transformations T_α and $T_{\alpha^{-1}}$.

Suppose we make small variations dz in the variables z thereby producing small variations dZ in the variables Z. Then, by the inverse function theorem, we must show that there is a relation of the form

$$dZ = Mdz \tag{6.7.30}$$

where the matrix M is invertible. From (1.5) we find that

$$dU = Gdu. \tag{6.7.31}$$

And, from (7.19) and (7.20), we have the results

$$dU = A^\alpha dZ + B^\alpha dz, \tag{6.7.32}$$

and

$$du = C^\alpha dZ + D^\alpha dz. \tag{6.7.33}$$

Combining (7.31) through (7.33) gives the series of results

$$A^\alpha dZ + B^\alpha dz = G(C^\alpha dZ + D^\alpha dz), \tag{6.7.34}$$

$$(A^\alpha - GC^\alpha)dZ = (GD^\alpha - B^\alpha)dz, \tag{6.7.35}$$

$$dZ = [(A^\alpha - GC^\alpha)^{-1}(GD^\alpha - B^\alpha)]dz. \tag{6.7.36}$$

Comparison of (7.30) and (7.36) gives the relation

$$M = (A^\alpha - GC^\alpha)^{-1}(GD^\alpha - B^\alpha). \tag{6.7.37}$$

But, by (5.11.23) with the substitutions $M \to \alpha$ and $U' \to G$, there is the identity

$$(A^\alpha - GC^\alpha)^{-1}(GD^\alpha - B^\alpha) = (A^{\alpha^{-1}}G + B^{\alpha^{-1}})(C^{\alpha^{-1}}G + D^{\alpha^{-1}})^{-1}. \tag{6.7.38}$$

Therefore we may also write

$$M = (A^{\alpha^{-1}}G + B^{\alpha^{-1}})(C^{\alpha^{-1}}G + D^{\alpha^{-1}})^{-1}. \tag{6.7.39}$$

We conclude that M and G are related by the Möbius transformation $T_{\alpha^{-1}}$,

$$M = T_{\alpha^{-1}}(G). \tag{6.7.40}$$

See also Exercise 7.4.

We already know that G is symmetric. See (1.7). Moreover, from the work of Section 5.13, we know that Möbius transformations of the form $T_{\alpha^{-1}}$ can be found such that (7.40) is well defined, and we know that these Möbius transformations send symmetric matrices into symplectic matrices. It follows that M is a symplectic matrix, and therefore M is invertible. We have shown that the implicit relations (7.21) can be made explicit so that we may indeed write (7.23). Moreover, since M is a symplectic matrix, we have also verified that $\mathcal{M}$ is a symplectic map.

The discourse so far described how, given the gradient map $\mathcal{G}$ associated with any source function g and a Darboux matrix α, we can construct a symplectic map $\mathcal{M}$ by use of (7.21). In the spirit of Subsection 5.2, suppose we are instead given $\mathcal{M}$ and we wish to construct g. Begin with the relation (7.18), which can be rewritten in the form

$$u = C^\alpha(\mathcal{M}z) + D^\alpha z. \tag{6.7.41}$$

Let us see if this relation can be solved for z. That is, given u, we want to use (7.41) to determine z. Taking differentials of both sides of (7.41) and using (7.30 gives the result

$$du = C^\alpha dZ + D^\alpha dz = C^\alpha M dz + D^\alpha dz = (C^\alpha M + D^\alpha)dz. \tag{6.7.42}$$

By the inverse function theorem, if the matrix $(C^\alpha M + D^\alpha)$ is invertible, we may solve (7.41) for z, in which case there is also the relation

$$dz = (C^\alpha M + D^\alpha)^{-1}du. \tag{6.7.43}$$

Next we observe that (7.17) can be rewritten in the form

$$U = A^\alpha(\mathcal{M}z) + B^\alpha z. \tag{6.7.44}$$

Since we have already found z as a function of u by solving (7.41), equation (7.44) enables us to find U as a function of u.

We claim that the map $\mathcal{G}$ that sends u to U is a gradient map if $\mathcal{M}$ is a symplectic map. Taking differentials of both sides of (7.44), and again using (7.30), give the result

$$dU = A^\alpha dZ + B^\alpha dz = A^\alpha M dz + B^\alpha dz = (A^\alpha M + B^\alpha)dz. \tag{6.7.45}$$

Next insert (7.43) into (7.45) to yield the result

$$dU = (A^\alpha M + B^\alpha)(C^\alpha M + D^\alpha)^{-1} du. \tag{6.7.46}$$

Now compare (7.31) and (7.46) to find the relation

$$G = (A^\alpha M + B^\alpha)(C^\alpha M + D^\alpha)^{-1}. \tag{6.7.47}$$

This result is evidently the Möbius relation

$$G = T_\alpha(M), \tag{6.7.48}$$

which is consistent with (7.40). Also, we know from Section 5.13 that G will be symmetric if M is symplectic, and consequently $\mathcal{G}$ is indeed a gradient map.

Finally, we may determine the source function g associated with $\mathcal{G}$ by performing the path integral

$$g(u,t) = \int^u \sum_a U(u',t)_a \, du'_a. \tag{6.7.49}$$

Here we have indicated that g may also depend on t if $\mathcal{M}$ depends on t.

We close this subsection by listing the Darboux matrices α associated with the familiar generating function types F_1 through F_4 and F_+, and the related γ in the representation $\alpha = \gamma\sigma$. They appear in the table below. Note the interesting fact that all these Darboux matrices are orthogonal. In Exercises 7.5 through 7.7 you, dear reader, will have the pleasure of spot checking that the use of these Darboux matrices in (7.21) reproduces the relations (5.4) through (5.7) and (6.57).

Table 6.7.1: Darboux Matrices α for the Generating Function types F_1 through F_4 and F_+.

$$\text{Here } \alpha = \begin{pmatrix} A^\alpha & B^\alpha \\ C^\alpha & D^\alpha \end{pmatrix} = \gamma\sigma. \tag{6.7.50}$$

$F_1(q, Q, t)$

$$p_k = \partial F_1/\partial q_k,\ P_k = -\partial F_1/\partial Q_k. \tag{6.7.51}$$

$$A^\alpha = \begin{pmatrix} 0 & 0 \\ 0 & -I^n \end{pmatrix},\ B^\alpha = \begin{pmatrix} 0 & I^n \\ 0 & 0 \end{pmatrix}, \tag{6.7.52}$$

$$C^\alpha = \begin{pmatrix} 0 & 0 \\ I^n & 0 \end{pmatrix},\ D^\alpha = \begin{pmatrix} I^n & 0 \\ 0 & 0 \end{pmatrix}. \tag{6.7.53}$$

$$\gamma = (1/\sqrt{2}) \begin{pmatrix} I^n & 0 & 0 & I^n \\ I^n & 0 & 0 & -I^n \\ 0 & -I^n & I^n & 0 \\ 0 & I^n & I^n & 0 \end{pmatrix}. \tag{6.7.54}$$

$F_2(q, P, t)$

$$p_k = \partial F_2/\partial q_k,\ Q_k = \partial F_2/\partial P_k. \tag{6.7.55}$$

$$A^\alpha = \begin{pmatrix} 0 & 0 \\ I^n & 0 \end{pmatrix},\ B^\alpha = \begin{pmatrix} 0 & I^n \\ 0 & 0 \end{pmatrix}, \tag{6.7.56}$$

$$C^\alpha = \begin{pmatrix} 0 & 0 \\ 0 & I^n \end{pmatrix},\ D^\alpha = \begin{pmatrix} I^n & 0 \\ 0 & 0 \end{pmatrix}. \tag{6.7.57}$$

$$\gamma = (1/\sqrt{2}) \begin{pmatrix} I^n & 0 & 0 & I^n \\ 0 & I^n & I^n & 0 \\ 0 & -I^n & I^n & 0 \\ -I^n & 0 & 0 & I^n \end{pmatrix}. \tag{6.7.58}$$

$F_3(p, Q, t)$

$$q_k = -\partial F_3/\partial p_k,\ P_k = -\partial F_3/\partial Q_k. \tag{6.7.59}$$

$$A^\alpha = \begin{pmatrix} 0 & 0 \\ 0 & -I^n \end{pmatrix},\ B^\alpha = \begin{pmatrix} -I^n & 0 \\ 0 & 0 \end{pmatrix}, \tag{6.7.60}$$

$$C^\alpha = \begin{pmatrix} 0 & 0 \\ I^n & 0 \end{pmatrix},\ D^\alpha = \begin{pmatrix} 0 & I^n \\ 0 & 0 \end{pmatrix}. \tag{6.7.61}$$

$$\gamma = (1/\sqrt{2}) \begin{pmatrix} 0 & I^n & -I^n & 0 \\ I^n & 0 & 0 & -I^n \\ I^n & 0 & 0 & I^n \\ 0 & I^n & I^n & 0 \end{pmatrix}. \tag{6.7.62}$$

Table 6.7.1 continued

$F_4(p,P,t)$

$$q_k = -\partial F_4/\partial p_k,\ Q_k = \partial F_4/\partial P_k. \tag{6.7.63}$$

$$A^\alpha = \begin{pmatrix} 0 & 0 \\ I^n & 0 \end{pmatrix},\ B^\alpha = \begin{pmatrix} -I^n & 0 \\ 0 & 0 \end{pmatrix}, \tag{6.7.64}$$

$$C^\alpha = \begin{pmatrix} 0 & 0 \\ 0 & I^n \end{pmatrix},\ D^\alpha = \begin{pmatrix} 0 & I^n \\ 0 & 0 \end{pmatrix}. \tag{6.7.65}$$

$$\gamma = (1/\sqrt{2}) \begin{pmatrix} 0 & I^n & -I^n & 0 \\ 0 & I^n & I^n & 0 \\ I^n & 0 & 0 & I^n \\ -I^n & 0 & 0 & I^n \end{pmatrix}. \tag{6.7.66}$$

$F_+(\Sigma,t)$

$$\Delta = J\partial_\Sigma F_+ \text{ where } \Sigma = Z + z \text{ and } \Delta = Z - z. \tag{6.7.67}$$

$$\alpha = \sigma = (1/\sqrt{2}) \begin{pmatrix} -J^{2n} & J^{2n} \\ I^{2n} & I^{2n} \end{pmatrix}. \tag{6.7.68}$$

$$\gamma = I^{4n}. \tag{6.7.69}$$

6.7.2 Discussion

6.7.2.1 Graphs

In this section, and in Sections 5.13 and 5.14, we introduced a $4n$-dimensional space even though the underlying entities of interest, namely symplectic matrices and symplectic maps, were associated with a $2n$-dimensional space. How could one have guessed that this would be a good thing to do? Of course, results are what count. But one way to look at the matter is in terms of *graphs*.

Suppose we wish to analyze a function $f(x)$ of a *single* real variable x. One way to do so is to introduce a *two*-dimensional space $\mathbb{R}^2$, with axes x and y, and then "darken" those points in $\mathbb{R}^2$ for which $y = f(x)$. The darkened points form the graph of f,

$$\text{graph of } f = \{\{x,y\} \in \mathbb{R}^2 \mid y = f(x)\}. \tag{6.7.70}$$

Moreover, the graph of f is a one-dimensional submanifold of $\mathbb{R}^2$. Let τ_1 be some parameter. Then we may also write

$$\text{graph of } f = \{\{x,y\} \in \mathbb{R}^2 \mid x = \tau_1,\ y = f(\tau_1) \text{ with } \tau_1 \in \mathbb{R}^1\}. \tag{6.7.71}$$

All these considerations are so commonplace that we hardly ever think about them, but they do involve a doubling of dimension.[21]

Now consider a $4n$-dimensional space with coordinates $\{Z_1 \cdots Z_{2n}\}$ and $\{z_1 \cdots z_{2n}\}$ or, equivalently, coordinates $\{\hat{Z}_1 \cdots \hat{Z}_{4n}\}$. Let $\mathcal{M}$ be the map (6.1). Then we can describe $\mathcal{M}$ in terms of a graph by writing

$$\text{graph of } \mathcal{M} = \{\hat{Z} \in \mathbb{R}^{4n} \mid Z_a = K_a(z) \;\text{ for }\; a = 1, 2n\}. \tag{6.7.72}$$

(Here we have suppressed the possible dependence of K on the parameter t.) We see that the construction (7.72) is completely analogous to (7.70). Moreover, the graph of $\mathcal{M}$ is a $2n$-dimensional submanifold of $\mathbb{R}^{4n}$. Let $\{\tau_1 \cdots \tau_{2n}\}$ be a set of $2n$ parameters. Then we may also write

$$\text{graph of } \mathcal{M} = \{\hat{Z} \in \mathbb{R}^{4n} \mid Z_a = K_a(\tau),\; z_a = \tau_a \;\text{ for }\; a = 1, 2n \;\text{ with }\; \tau \in \mathbb{R}^{2n}\}. \tag{6.7.73}$$

6.7.2.2 The Graph of $\mathcal{M}$ Is a $\tilde{J}^{4n}$ Lagrangian Submanifold

Let us find the tangent vectors to the graph of $\mathcal{M}$ (now regarded as a $2n$-dimensional submanifold in a $4n$-dimensional space). They describe how $\hat{Z}$ varies when τ is varied. Write τ in the form

$$\tau = \tau^0 + \sum_1^{2n} \lambda_i e^i. \tag{6.7.74}$$

Here the vectors e^i are the same as those introduced in Section 5.13, namely those that form the columns of I^{2n}. Employing this notation for τ, we define $2n$ vectors ζ^j tangent to the graph of $\mathcal{M}$ at the point $\hat{Z}(\tau^0)$ by writing the definition

$$\zeta^j(\tau^0) = \partial\hat{Z}/\partial\lambda_j|_{\lambda=0} = (\partial Z/\partial\lambda_j|_{\lambda=0}\;;\;\partial z/\partial\lambda_j|_{\lambda=0})^T. \tag{6.7.75}$$

As indicated, the tangent vectors ζ^j are of length $4n$ (have $4n$ entries) as is appropriate for a $4n$-dimensional space. The last $2n$ entries in each tangent vector ζ^j, those to the right of the semicolon in (7.75), are easy to find. From (7.73) and (7.74) we readily compute the result

$$\partial z/\partial\lambda_j|_{\lambda=0} = e^j. \tag{6.7.76}$$

The calculation of the first $2n$ entries is a bit more involved. From (7.73) (which contains the information that $z = \tau$) and (7.74) we find, in terms of components, the result

$$\partial Z_i/\partial\lambda_j|_{\lambda=0} = \partial Z_i/\partial z_j|_{z=\tau^0} = M_{ij}(\tau^0) = m_i^j. \tag{6.7.77}$$

Here we have used the notation (5.13.36), and $M(\tau^0)$ is the Jacobian matrix $M(z)$ for $\mathcal{M}$ evaluated at $z = \tau^0$.

We see from (7.75) through (7.77) that there is the relation

$$\zeta^j = (m^j; e^j)^T, \tag{6.7.78}$$

[21] The use of graphs to portray functions was invented by Descartes. He plotted x along the vertical axis and y along the horizontal axis. Newton turned this around to plot x along the horizontal axis and y along the vertical axis, and humankind have followed his convention ever since.

and recognize that the ζ^j are just the vectors u^j introduced in Section 5.13 and given by (5.13.42). We know that these vectors are $\tilde{J}^{4n}$ isotropic. Thus, we have shown that the tangent vectors of the graph of $\mathcal{M}$ are $\tilde{J}^{4n}$ isotropic at any point $\hat{Z}(\tau^0)$ in the submanifold, and therefore span a $\tilde{J}^{4n}$ Lagrangian plane at every such point. For this reason, the graph of $\mathcal{M}$ is entitled to be called a $\tilde{J}^{4n}$ Lagrangian submanifold.

6.7.2.3 The Graph of $\mathcal{G}$ Is a J^{4n} Lagrangian Submanifold

We can carry out a similar analysis for the graph of $\mathcal{G}$. The map $\mathcal{G}$ is defined by (7.7) and (7.14) through (7.16). Again let $\{\tau_1 \cdots \tau_{2n}\}$ be a set of $2n$ parameters. The graph of $\mathcal{G}$, as a $2n$-dimensional submanifold in $\mathbb{R}^{4n}$, is given by the definition

$$\text{graph of } \mathcal{G} = \{\hat{U} \in \mathbb{R}^{4n} \mid U_a = \partial g(\tau)/\partial \tau_a,\ u_a = \tau_a \text{ for } a = 1, 2n \text{ with } \tau \in \mathbb{R}^{2n}\}. \quad (6.7.79)$$

Again employing the notation (7.74), the graph of $\mathcal{G}$ will have $2n$ tangent vectors ν^j at the point $\hat{U}(\tau^0)$ given by the definition

$$\nu^j(\tau^0) = \partial\hat{U}/\partial\lambda_j|_{\lambda=0} = (\partial U/\partial\lambda_j|_{\lambda=0}\ ;\ \partial u/\partial\lambda_j|_{\lambda=0})^T. \quad (6.7.80)$$

The last $2n$ entries in each tangent vector ν^j, those to the right of the semicolon in (7.80), are calculated the same way as in the case of $\mathcal{M}$, and are therefore given by the relation

$$\partial u/\partial\lambda_j|_{\lambda=0} = e^j. \quad (6.7.81)$$

In terms of components, the first $2n$ entries are given by the relations

$$\partial U_i/\partial\lambda_j|_{\lambda=0} = \partial^2 g/\partial\lambda_i\partial\lambda_j|_{\lambda=0} = G_{ij}(\tau^0). \quad (6.7.82)$$

In analogy to the notation (5.13.60) and (5.13.61) employed in Section 5.13, define vectors w^j, each of length $2n$, by writing

$$w^j_i = G_{ij}. \quad (6.7.83)$$

Then, with this notation, the tangent vectors ν^j become

$$\nu^j = (w^j; e^j)^T. \quad (6.7.84)$$

Since G is a symmetric matrix, these vectors are completely analogous to the vectors v^j constructed in Section 5.13 and given by (5.13.62). Therefore we know that these vectors are J^{4n} isotropic. Thus, we have shown that the tangent vectors of the graph of $\mathcal{G}$ are J^{4n} isotropic at any point $\hat{U}(\tau^0)$ in the submanifold, and therefore span a J^{4n} lagrangian plane at every such point. Consequently the graph of $\mathcal{G}$ is entitled to be called a J^{4n} Lagrangian submanifold.

6.7.2.4 Relation between the Graphs of $\mathcal{M}$ and $\mathcal{G}$

We have learned that the $4n$-dimensional constructions used in Sections 5.13 through 5.15, and in this section, appear to be less ad hoc when one thinks in terms of graphs. The graph of $\mathcal{M}$ is a $\tilde{J}^{4n}$ Lagrangian submanifold, and the graph of $\mathcal{G}$ is a J^{4n} Lagrangian submanifold. Moreover, according to (7.3) and (7.4), these two submanifolds are mapped into each other by a Darboux transformation α. Or, put another way, they are the *same* submanifold, and this submanifold appears to be $\tilde{J}^{4n}$ Lagrangian when the coordinates $\hat{Z}$ are used and J^{4n} Lagrangian when the coordinates $\hat{U}$ are used.

6.7.2.5 Reason for the Term "Lagrangian"

To keep a promise, we still need to describe the origin of the term "Lagrangian" when applied to planes and submanifolds. It has to do with *Lagrange* brackets. Consider a $2n$-dimensional phase space with coordinates $(q;p)$ as in (1.7.9). Define an n-dimensional submanifold in this phase space (parameterized by the quantities $\tau_1, \cdots, \tau_n$) by writing $2n$ equations of the form

$$z_a = f_a(\tau) \tag{6.7.85}$$

where the f_a are any functions of the n variables τ. Next form the tangent vectors $\partial z/\partial\tau_i$. [These n vectors are assumed to be linearly independent since (7.85) is assumed to define an n-dimensional submanifold.] The Lagrange bracket $\{\tau_i, \tau_j\}$, which is a function of the variables τ, is defined by the rule

$$\{\tau_i, \tau_j\} = (\partial z/\partial\tau_i, J^{2n}\, \partial z/\partial\tau_j). \tag{6.7.86}$$

If we use the specific form for J^{2n} given by (3.1.1), we observe that the Lagrange bracket can also be written in what may be the more familiar text-book form

$$\{\tau_i, \tau_j\} = \sum_k (\partial q_k/\partial\tau_i)(\partial p_k/\partial\tau_j) - (\partial p_k/\partial\tau_i)(\partial q_k/\partial\tau_j). \tag{6.7.87}$$

From (7.86) we see that the tangent vectors $\partial z/\partial\tau_i$ for any fixed $\tau = \tau^0$ are J^{2n} isotropic if the Lagrange brackets $\{\tau_i, \tau_j\}$ all vanish (for $\tau = \tau^0$), and we say that the plane spanned by the tangent vectors $\partial z/\partial\tau_i$ is Lagrangian. Correspondingly, the submanifold given by (7.85) is Lagrangian if the Lagrange brackets $\{\tau_i, \tau_j\}$ all vanish for all values of τ. Similar nomenclature carries over to $2n$-dimensional planes and $2n$-dimensional submanifolds in $4n$-dimensional spaces and the use of J^{4n} or $\tilde{J}^{4n}$.

6.7.2.6 Closing Observation

We close this subsection with the observation that the family of maps produced by the generating/source function $g(u,t)$ and some Darboux matrix α does not necessarily pass through the identity map $\mathcal{I}$ for some value of t. For each value of t there will be a symmetric matrix $G(u,t)$ given by (1.6), and for this value of t the map $\mathcal{M}(t)$ will have a Jacobian matrix M given by (7.40). Suppose $\mathcal{M}(t) = \mathcal{I}$ when $t = t_0$. Then, since the Jacobian matrix of the identity map $\mathcal{I}$ is the identity matrix I, (7.40) becomes the the relation

$$I = T_{\alpha^{-1}}(G) \tag{6.7.88}$$

which requires that

$$G = G_0 \tag{6.7.89}$$

where

$$G_0 = T_\alpha(I) = (A^\alpha + B^\alpha)(C^\alpha + D^\alpha)^{-1}. \tag{6.7.90}$$

Corresponding, we must have

$$g(u, t_0) = (1/2)(u, G_0 u) + (v, u) + g_0 \tag{6.7.91}$$

where v is a fixed vector yet to be determined and g_0 is an immaterial constant.

To determine v, which will turn out to vanish, we need to find the $\mathcal{M}$ associated with the $g(u, t_0)$ given by (7.91). Partial differentiation of (7.91) gives the intermediate result

$$\partial g/\partial u_a = v_a + \sum_b (G_0)_{ab} u_b, \tag{6.7.92}$$

and employing this intermediate result in (7.21) gives the further result

$$A^\alpha Z + B^\alpha z = G_0(C^\alpha Z + D^\alpha z) + v. \tag{6.7.93}$$

Now solve (7.93) to find the result

$$\begin{aligned} Z &= (A^\alpha - G_0 C^\alpha)^{-1}(G_0 D^\alpha - B^\alpha)z + [(A^\alpha - G_0 C^\alpha)^{-1}]v \\ &= [(A^{\alpha^{-1}} G_0 + B^{\alpha^{-1}})(C^{\alpha^{-1}} G_0 + D^{\alpha^{-1}})^{-1}]z + [(A^\alpha - G_0 C^\alpha)^{-1}]v \\ &= [T_{\alpha^{-1}}(G_0)]z + [(A^\alpha - G_0 C^\alpha)^{-1}]v \\ &= z + [(A^\alpha - G_0 C^\alpha)^{-1}]v. \end{aligned} \tag{6.7.94}$$

Here we have used (7.38) and (7.88). We see that we must have $v = 0$ to achieve the identity map, in which case

$$g(u, t_0) = (1/2)(u, G_0 u) + g_0. \tag{6.7.95}$$

Now it may well happen that $g(u, t)$ is never of the form (7.95) for any value of t, in which case the family of maps $\mathcal{M}(t)$ never passes through the identity map. There is also a possible second obstacle. Note that G_0 as given by (7.90) is not defined if the matrix $(C^\alpha + D^\alpha)$ is not invertible,

$$\det(C^\alpha + D^\alpha) = 0. \tag{6.7.96}$$

Thus, there are Darboux matrices for which $\mathcal{M}(t)$ can never pass through the identity map. See, for example, the Darboux matrices associated with F_1 and F_4 given in Table 6.7.1.

To conclude this observation we note that, although we have verified that there are families of symplectic maps that never pass through the identity map, the symplectic maps associated with the Hamiltonian Cauchy initial value problem, see Section 1.3 and Subsection 4.1, pass through the identity map by definition because of the initial condition requirement (4.13).

6.7.2.7 Final Remark

Finally, we remark that there is no reason why the Darboux matrix α cannot also be taken to depend on t. We know that we can always write

$$\alpha = \gamma\sigma \tag{6.7.97}$$

where σ is the matrix (5.13.11) and γ is any matrix in the group $Sp(4n)$. We also know that the symplectic group is connected. See Section 5.9. Therefore the set of Darboux matrices is connected, and we may sensibly write

$$\alpha(t) = \gamma(t)\sigma. \tag{6.7.98}$$

for any path $\gamma(t)$ in $Sp(4n)$. If we now employ $g(u, t)$ and $\alpha(t)$ in (7.21), the result will again be a family of symplectic maps $\mathcal{M}(t)$.

6.7.3 Relating Source Functions and Generating Hamiltonians, Transformation of Hamiltonians, and Hamilton-Jacobi Theory/Equations

Suppose the source function g appearing in (7.24) does indeed depend on the time. Then its use in (7.21) produces a family of symplectic maps which, for our present notational purposes, we will call $\mathcal{N}(t)$ rather than $\mathcal{M}(t)$ and will have Jacobian $N(t)$. We know from Subsection 4.2 that any such family is Hamiltonian generated. Indeed, Subsection 5.3 determined this Hamiltonian for the case of $F_2(q,P,t)$, and (5.167) covers the cases F_1 through F_4. The relation between $g(u,t)$, the associated symplectic map $\mathcal{N}(t)$, and the associated generating Hamiltonian, which we will here call H^g, is part of Hamilton-Jacobi theory; and Subsection 5.3.2 describes examples of the Hamilton-Jacobi equation for the cases where $\mathcal{N}(t)$ arises from one of the F_j. In this subsection we will solve the general problem of finding the Hamiltonian H^g when $\mathcal{N}(t)$ arises from $g(u,t)$ and the use of some Darboux matrix α. We will also solve the inverse general problem of finding $g(u,t)$ in terms of H^g. Finally, we will relate these results to Hamilton-Jacobi Theory for the general problem. Our results will also be of use for the work of Chapter 34.

6.7.3.1 Finding the Generating Hamiltonian H^g from the Source Function g

Our discussion will be patterned after that of Subsection 5.3, so again we will have to deal with a variety of partial derivatives. Therefore we introduce the notation

$$\begin{aligned} &g(u,t;\ ,1) = \partial g/\partial t,\\ &g(u,t;a,1) = \partial^2 g/\partial u_a \partial t,\\ &g(u,t;ab,\ ) = \partial^2 g/\partial u_a \partial u_b. \end{aligned} \tag{6.7.99}$$

Employing this notation, define the function $g^t(u,t)$ by the rule

$$g^t(u,t) = g(u,t;\ ,1) \tag{6.7.100}$$

According to (7.18), u may be regarded as a function of $Z(t)$ and z. Also, according to (7.22) with the substitution $\mathcal{M} \rightarrow \mathcal{N}$, we may view z as being a function of t and $Z(t)$ by writing

$$z = \mathcal{N}^{-1}(t)Z. \tag{6.7.101}$$

Therefore, u may be regarded as a function of Z and t,

$$u = u(Z,t). \tag{6.7.102}$$

Now substitute (7.102) into (7.100) to define the function $H^g(Z,t)$ by the rule

$$H^g(Z,t) = g^t(u(Z,t),t). \tag{6.7.103}$$

We claim that $H^g(Z,t)$ is the Hamiltonian that generates the map $\mathcal{N}(t)$ produced by the use of the source function $g(u,t)$ and some Darboux matrix α. Here we assume that α is some *fixed* Darboux matrix, although it would be interesting to also entertain the possibility (7.98).

We will now seek to verify this claim about H^g. Suppose z is held fixed and t is increased by the amount dt. So doing will change Z by the amount dZ. Also, according to (7.18), u will experience a change that we will call du. Look at the relations (7.28) and (7.29). From (7.28) we conclude that

$$\begin{aligned} dZ_a &= [\sum_b (A^{\alpha^{-1}})_{ab} g(u,t;b,1)]dt \\ &+ \sum_{bc} (A^{\alpha^{-1}})_{ab} g(u,t;bc,\ )du_c \\ &+ \sum_b (B^{\alpha^{-1}})_{ab} du_b \\ &= [\sum_b (A^{\alpha^{-1}})_{ab} g(u,t;b,1)]dt \\ &+ \sum_b [A^{\alpha^{-1}} G(u,t) + B^{\alpha^{-1}}]_{ab} du_b. \end{aligned} \tag{6.7.104}$$

Here we have used (1.6). That is, here G is the Hessian matrix of g and the Jacobian matrix of $\mathcal{G}$. From (7.29), since z is to be held fixed, we conclude that

$$\begin{aligned} 0 = dz_a &= [\sum_b (C^{\alpha^{-1}})_{ab} g(u,t;b,1)]dt \\ &+ \sum_{bc} (C^{\alpha^{-1}})_{ab} g(u,t;bc,\ )du_c \\ &+ \sum_b (D^{\alpha^{-1}})_{ab} du_b \\ &= [\sum_b (C^{\alpha^{-1}})_{ab} g(u,t;b,1)]dt \\ &+ \sum_b [C^{\alpha^{-1}} G(u,t) + D^{\alpha^{-1}}]_{ab} du_b. \end{aligned} \tag{6.7.105}$$

Let us now eliminate du between (7.104) and (7.105). First solve (7.105) for du to find the result

$$du_b = -dt \sum_c \{[C^{\alpha^{-1}} G(u,t) + D^{\alpha^{-1}}]^{-1} C^{\alpha^{-1}}\}_{bc} g(u,t;c,1). \tag{6.7.106}$$

Now substitute (7.106) into (7.104) to obtain the result

$$\begin{aligned} dZ_a &= dt \sum_b (A^{\alpha^{-1}})_{ab} g(u,t;b,1) \\ &- dt \sum_b \{[A^{\alpha^{-1}} G(u,t) + B^{\alpha^{-1}}][C^{\alpha^{-1}} G(u,t) + D^{\alpha^{-1}}]^{-1} C^{\alpha^{-1}}\}_{ab} g(u,t;b,1) \\ &= dt \sum_b \{A^{\alpha^{-1}} - [A^{\alpha^{-1}} G(u,t) + B^{\alpha^{-1}}][C^{\alpha^{-1}} G(u,t) + D^{\alpha^{-1}}]^{-1} C^{\alpha^{-1}}\}_{ab} g(u,t;b,1). \end{aligned} \tag{6.7.107}$$

We also observe that

$$[A^{\alpha^{-1}}G + B^{\alpha^{-1}}][C^{\alpha^{-1}}G + D^{\alpha^{-1}}]^{-1} = T_{\alpha^{-1}}(G) = N. \tag{6.7.108}$$

Therefore, (7.107) can also be written as

$$dZ_a = dt \sum_b (A^{\alpha^{-1}} - NC^{\alpha^{-1}})_{ab} g(u, t; b, 1), \tag{6.7.109}$$

or

$$dZ_a/dt = \sum_c (A^{\alpha^{-1}} - NC^{\alpha^{-1}})_{ac} g(u, t; c, 1). \tag{6.7.110}$$

Note that here we have renamed the dummy summation index.

Next let us work out the quantities $\partial H^g(Z, t)/\partial Z_a$. From (7.100), (7.103), and the chain rule (and holding t fixed) we have the result

$$dH^g = \sum_a g(u, t; a, 1) du_a. \tag{6.7.111}$$

Also, use of (7.18) provides the relation

$$du_a = \sum_b [(C^\alpha)_{ab} dZ_b + (D^\alpha)_{ab} dz_b] \tag{6.7.112}$$

which, using (7.30) with the substitution $M \to N$, can be rewritten in the form

$$du_a = \sum_b [C^\alpha + D^\alpha(N^{-1})]_{ab} dZ_b. \tag{6.7.113}$$

When combined, (7.111) and (7.113 yield the relation

$$dH^g = \sum_{ab} g(u, t; a, 1)[C^\alpha + D^\alpha(N^{-1})]_{ab} dZ_b \tag{6.7.114}$$

from which we conclude that

$$\begin{aligned} \partial H^g/\partial Z_b &= \sum_a g(u, t; a, 1)[C^\alpha + D^\alpha(N^{-1})]_{ab} \\ &= \sum_c \{[C^\alpha + D^\alpha(N^{-1})]^T\}_{bc} g(u, t; c, 1). \end{aligned} \tag{6.7.115}$$

We are almost done. Multiply both sides of (7.115) by J_{ab} and sum over b to find the result

$$\sum_b J_{ab}(\partial H^g/\partial Z_b) = \sum_c \{J[C^\alpha + D^\alpha(N^{-1})]^T\}_{ac} g(u, t; c, 1). \tag{6.7.116}$$

Here again we have renamed the dummy summation index. We now claim that there is the relation

$$(A^{\alpha^{-1}} - NC^{\alpha^{-1}}) = J[C^\alpha + D^\alpha(N^{-1})]^T. \tag{6.7.117}$$

If so, then comparison of (7.110) and (7.116) gives the result

$$dZ_a/dt = \sum_b J_{ab}(\partial H^g/\partial Z_b), \tag{6.7.118}$$

which is the expected equations of motion set for Z when H^g is the Hamiltonian.

To complete the proof, we need to verify (7.117). Its right side can be rewritten as

$$J[C^\alpha + D^\alpha(N^{-1})]^T = J(C^\alpha)^T + J(N^{-1})^T(D^\alpha)^T. \tag{6.7.119}$$

According to (3.1.11), the symplectic condition for N gives the relation

$$(N^{-1})^T = -JNJ. \tag{6.7.120}$$

Therefore the right side of (7.117) can also be rewritten as

$$J[C^\alpha + D^\alpha(N^{-1})]^T = J(C^\alpha)^T - JJNJ(D^\alpha)^T = J(C^\alpha)^T + NJ(D^\alpha)^T. \tag{6.7.121}$$

Now employ the relations (5.13.100) and (5.13.102) to again rewrite the right side of (7.117) as

$$J[C^\alpha + D^\alpha(N^{-1})]^T = J(C^\alpha)^T + NJ(D^\alpha)^T = A^{\alpha^{-1}} - NC^{\alpha^{-1}}, \tag{6.7.122}$$

which, we see, agrees with the left side of (7.117). Therefore our claim is correct.

In summary, we have shown that in the general case the generating Hamiltonian $H^g(Z,t)$ for the family $\mathcal{N}(t)$ of symplectic maps produced by the source function $g(u,t)$ and the Darboux matrix α is given by the relation

$$H^g(Z,t) = [\partial g(u,t)/\partial t]|_{u=C^\alpha Z + D^\alpha(\mathcal{N}^{-1}Z)}. \tag{6.7.123}$$

We also observe that (5.167) is a special case of (7.123)

6.7.3.2 Finding the Source Function g from the Generating Hamiltonian H^g

In Subsection 5.2.2 we showed, as an example, how to find the source function F_2, which amounts to the choice (7.56) and (7.57) for the Darboux matrix α, in terms of an integral over a trajectory arising from the generating Hamiltonian which we there called H. See (5.119) and (5.130). The purpose of this subsection is to provide an analogous treatment of the general case: Given a Darboux matrix α and a generating Hamiltonian $H^g(\zeta,t)$, find the source function $g(u,t)$. Here, as in Subsection 5.2.2, it is convenient to employ the phase-space variables $\zeta = (\xi,\eta)$.

Let $(q,p) = z$ be initial conditions at $\tau = t^i$, and let $(Q,P) = Z$ be the final conditions reached by following to time $\tau = t$ the trajectories generated by $H^g(\zeta,\tau)$ starting with these initial conditions. We know that trajectories can be labeled by specifying either the initial conditions z or the final conditions Z. Assume that the trajectories are such that they can also be be labeled by specifying u as given by (7.18). See Figure 7.1. To do so will generally require a $2n$-dimensional search: Pick a $2n$-vector u. Begin by guessing z. Next follow the trajectory with initial conditions

$$\zeta(t^i) = z \tag{6.7.124}$$

and generated by $H^g(\zeta, \tau)$ to the time $\tau = t$ and set

$$Z = \zeta(t). \tag{6.7.125}$$

Now compute the quantity $(C^\alpha Z + D^\alpha z)$ and see if it equals u,

$$C^\alpha Z + D^\alpha z = u? \tag{6.7.126}$$

If it does, then the desired z (and Z) have been found. If not, guess again. In actual practice, this trajectory may have to be found by some kind of *shooting* method facilitated, perhaps, by a Newton's method search that involves also integrating the variational equations to determine how changes in the initial conditions produce changes in the final conditions.[22]

Observe that taking differentials of both sides of (7.18) and using (7.30) gives the result

$$du = C^\alpha dZ + D^\alpha dz = (C^\alpha N + D^\alpha)dz \tag{6.7.127}$$

from which it follows that

$$dz = (C^\alpha N + D^\alpha)^{-1} du. \tag{6.7.128}$$

Thus, if

$$\det(C^\alpha N + D^\alpha) \neq 0, \tag{6.7.129}$$

the quantity z (by the inverse function theorem) is indeed specified by the quantity u.

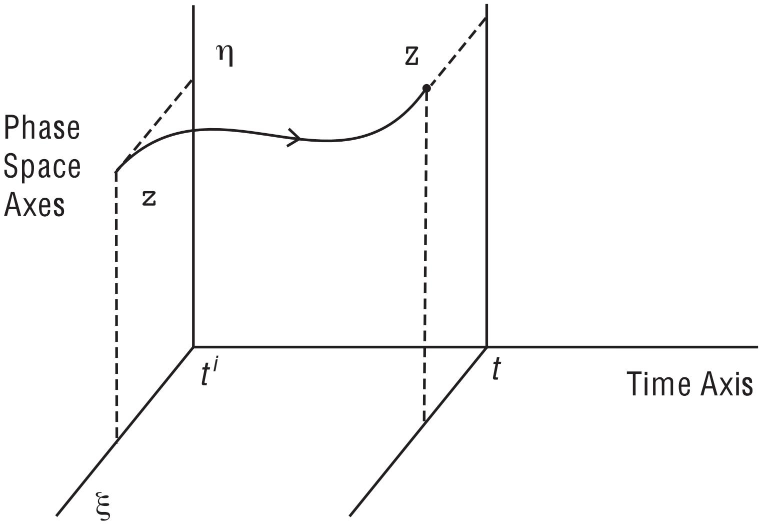


Figure 6.7.1: A trajectory of $H^g(\zeta, \tau)$ in the augmented $(\zeta, t) = (\xi, \eta; t)$ phase space. Given a Darboux matrix α, an initial time t^i, a final time t, and the $2n$-vector u, the initial condition $\zeta(t^i) = z$ is to be selected such that $C^\alpha Z + D^\alpha z = u$ where $\zeta(t) = Z$.

With these assumptions in mind, define the function $A'(u, t)$ by the rule

$$A'(u, t) = \int_{t^i}^{t} [(\zeta, J\dot{\zeta}) + 2H^g(\zeta, \tau)]d\tau. \tag{6.7.130}$$

[22] Note that although in general a $2n$-dimensional search is required, for some special α's, such as those for F_1 through F_4, only an n-dimensional search is required. See Subsection 5.2.2.

Here the integral on the right side is to be evaluated over the trajectory satisfying (7.126). We will want to see how $A'(u,t)$ changes when changes are made in u and/or t.

Write (7.130) in the form

$$A'(u,t) = \int_{t^i}^{t} \mathcal{A}'(\zeta,\dot{\zeta},\tau)d\tau \tag{6.7.131}$$

where

$$\mathcal{A}'(\zeta,\dot{\zeta},\tau) = (\sum_{cd} \zeta_c J_{cd}\dot{\zeta}_d) + 2H^g(\zeta,\tau). \tag{6.7.132}$$

Changing u (while holding t fixed) changes the initial and final conditions and the trajectory in between. Consequently, from variational calculus, we find that the change in $\mathcal{A}'$ is given by

$$\delta A' = \int_{t^i}^{t} d\tau\{\sum_a[(\partial\mathcal{A}'/\partial\zeta_a)\delta\zeta_a + (\partial\mathcal{A}'/\partial\dot{\zeta}_a)\delta\dot{\zeta}_a]\}. \tag{6.7.133}$$

The integrand in (7.133) can be manipulated in the standard way to rewrite $\delta A'$ in the form

$$\delta A' = \int_{t^i}^{t} d\tau\{\sum_a[(\partial\mathcal{A}'/\partial\zeta_a) - (d/d\tau)(\partial\mathcal{A}'/\partial\dot{\zeta}_a)]\delta\zeta_a + (d/d\tau)[\sum_a(\partial\mathcal{A}'/\partial\dot{\zeta}_a)\delta\zeta_a]\}. \tag{6.7.134}$$

For the various ingredients in the integrand of (7.134) we find the results

$$\partial\mathcal{A}'/\partial\zeta_a = (\sum_b J_{ab}\dot{\zeta}_b) + 2\partial H^g/\partial\zeta_a, \tag{6.7.135}$$

$$\partial\mathcal{A}'/\partial\dot{\zeta}_a = -\sum_b J_{ab}\zeta_b, \tag{6.7.136}$$

$$\sum_a(\partial\mathcal{A}'/\partial\dot{\zeta}_a)\delta\zeta_a = \sum_{ab}\zeta_a J_{ab}\delta\zeta_b = (\zeta, J\delta\zeta), \tag{6.7.137}$$

$$-(d/d\tau)(\partial\mathcal{A}'/\partial\dot{\zeta}_a) = \sum_b J_{ab}\dot{\zeta}_b, \tag{6.7.138}$$

$$[(\partial\mathcal{A}'/\partial\zeta_a) - (d/d\tau)(\partial\mathcal{A}'/\partial\dot{\zeta}_a)] = 2(\sum_b J_{ab}\dot{\zeta}_b) + 2(\partial H^g/\partial\zeta_a). \tag{6.7.139}$$

But, since ζ is assumed to be a trajectory for $H^g(\zeta,\tau)$, it satisfies Hamilton's equations

$$\dot{\zeta}_a = \sum_b J_{ab}(\partial H^g/\partial\zeta_b) \tag{6.7.140}$$

from which it follows that

$$[(\partial\mathcal{A}'/\partial\zeta_a) - (d/d\tau)(\partial\mathcal{A}'/\partial\dot{\zeta}_a)] = 2(\sum_b J_{ab}\dot{\zeta}_b) + 2(\partial H^g/\partial\zeta_a) = 0. \tag{6.7.141}$$

As a consequence of all these results, $\delta A'$ becomes

$$\delta A' = \int_{t^i}^{t} d\tau(d/d\tau)[(\zeta, J\delta\zeta)] = (\zeta, J\delta\zeta)|_{t^i}^{t} = (Z, JdZ) - (z, Jdz). \tag{6.7.142}$$

As a further step, we observe that the quantity (U, u) can be written in the form

$$(U, u) = (\hat{U}, S\hat{U}) \tag{6.7.143}$$

where S is the $4n \times 4n$ symmetric matrix

$$S = (1/2)\begin{pmatrix} 0 & I^{2n} \\ I^{2n} & 0 \end{pmatrix}. \tag{6.7.144}$$

[Recall the notation (7.1) and (7.7).] In terms of these quantities, and using (7.3), we may also write the relation

$$(U, u) = (\hat{U}, S\hat{U}) = (\alpha\hat{Z}, S\alpha\hat{Z}) = (\hat{Z}, \alpha^T S\alpha\hat{Z}). \tag{6.7.145}$$

We now have the tools to construct $g(u, t)$. It is defined by the rule

$$g(u, t) \overset{\text{def}}{=} [A'(u, t) + (\hat{Z}, \alpha^T S\alpha\hat{Z})]/2. \tag{6.7.146}$$

Our task is to verify that this g has the desired properties. We will first show that this g produces $\mathcal{N}$ according to the rule (7.21). Then we will show that it leads back to the specified Hamiltonian.

To see that this g produces $\mathcal{N}$, suppose t is held fixed and u is varied by an amount du. Then $\hat{Z}$ (that is, Z and z) will vary by an amount $d\hat{Z}$. Correspondingly, we find that g is changed by an amount δg with

$$\delta g = [\delta A'(u, t) + \delta(\hat{Z}, \alpha^T S\alpha\hat{Z})]/2. \tag{6.7.147}$$

Next employ (7.1) through (7.5) and (7.8) and (7.142) to rewrite $\delta A'$ in the form

$$\delta A' = (Z, JdZ) - (z, Jdz) = (\hat{Z}, \tilde{J}^{4n} d\hat{Z}) = (U, du) - (u, dU). \tag{6.7.148}$$

Also, we find that

$$\delta(\hat{Z}, \alpha^T S\alpha\hat{Z}) = 2(\hat{Z}, \alpha^T S\alpha d\hat{Z}) = 2(\alpha\hat{Z}, S\alpha d\hat{Z}) = 2(\hat{U}, Sd\hat{U}) = (U, du) + (u, dU). \tag{6.7.149}$$

Therefore we get the final relation

$$\delta g = [(U, du) - (u, dU) + (U, du) + (u, dU)]/2 = (U, du). \tag{6.7.150}$$

It follows that

$$U_a = \partial g(u, t)/\partial u_a, \tag{6.7.151}$$

which, in view of (7.19), is the desired result (7.21).

To check that this g in turn leads back to the specified Hamiltonian, let us take the total time derivative of both sides of (7.146). By ‘total’ we mean that the trajectory employed in computing A' should simply be extended in time, but otherwise unchanged. This means that z will not change, but Z and consequently also u will change. By the chain rule and using (7.151), we get for the left side of (7.146) the result

$$(d/dt)(\text{left side}) = dg/dt = \partial g/\partial t + \sum_a (\partial g/\partial u_a)(du_a/dt) = \partial g/\partial t + (U, \dot{u}). \tag{6.7.152}$$

For the right side of (7.146) we find

$$\begin{aligned}(d/dt)(\text{right side}) &= (d/dt)[A'(u,t) + (\hat{Z}, \alpha^T S \alpha \hat{Z})]/2 \\ &= (1/2)(d/dt)A'(u,t) + (1/2)(d/dt)(\hat{Z}, \alpha^T S \alpha \hat{Z}). \end{aligned} \quad (6.7.153)$$

The first term on the right side of (7.153) is easily evaluated using the fundamental theorem of calculus,

$$(1/2)(d/dt)[A'(u,t))] = (1/2)\mathcal{A}'(\zeta, \dot{\zeta}, \tau)|_{\tau=t} = (1/2)(Z, J\dot{Z}) + H^g(Z,t). \quad (6.7.154)$$

According our understanding that z should not change ($dz = 0$) there is also the result

$$(1/2)(z, J\dot{z}) = 0. \quad (6.7.155)$$

Therefore, using (7.155), (7.2), and (7.8), the relation (7.154) can also be written in the form

$$\begin{aligned}(1/2)(d/dt)[A'(u,t))] &= [(1/2)(Z, J\dot{Z}) - (1/2)(z, J\dot{z})] + H^g(Z,t) \\ &= (1/2)(\hat{Z}, \tilde{J}^{4n}(d/dt)\hat{Z}) + H^g(Z,t) \\ &= (1/2)[(U, \dot{u}) - (u, \dot{U})] + H^g(Z,t). \end{aligned} \quad (6.7.156)$$

Also, by (7.145), there is the simple result

$$(1/2)(d/dt)(\hat{Z}, \alpha^T S \alpha \hat{z}) = (1/2)(d/dt)(U, u) = (1/2)(\dot{U}, u) + (1/2)(U, \dot{u}). \quad (6.7.157)$$

Consequently the derivative of the right side of (7.146) can also be written as

$$\begin{aligned}(d/dt)(\text{right side}) &= (1/2)[(U, \dot{u}) - (u, \dot{U})] + H^g(Z,t) + (1/2)[(\dot{U}, u) + (U, \dot{u})] \\ &= (U, \dot{u}) + H^g(Z,t). \end{aligned}$$
$$(6.7.158)$$

Comparison of (7.152) and (7.158) now gives the final result

$$\partial g/\partial t = H^g(Z,t), \quad (6.7.159)$$

which is in agreement with (7.123).

6.7.3.3 Transformation of Hamiltonians and Application to Hamilton-Jacobi Theory in the General Case

Subsection 4.2 showed that any family of symplectic maps $\mathcal{N}(t)$ is Hamiltonian generated, and the associated Hamiltonian was called G. Subsection 4.4 described the transformation of an old Hamiltonian to a new Hamiltonian under the action of a symplectic map. And in Subsection 5.3.2 we found the the relation between the old and new Hamiltonians in the case that the symplectic map $\mathcal{N}$ arises from some specified mixed-variable generating function F_j, and applied the results to Hamilton-Jacobi theory for this case. Here we study the relation between the old and new Hamiltonian in the general case that the symplectic map $\mathcal{N}$ arises from some specified source function g and some specified Darboux matrix α, and apply the results to Hamilton-Jacobi theory for the general case.

We begin by recalling the relation (5.169), which we copy below,

$$K(Z;t) = H(z(Z,t);t) + G(Z;t), \tag{6.7.160}$$

and again remind ourselves that here G is the generating Hamiltonian for $\mathcal{N}$. In the general case that $\mathcal{N}$ arises from some specified source function g and some specified Darboux matrix α, we found in Subsections 7.3.1 and 7.3.2 that the associated generating Hamiltonian, which we there called H^g, was given by the relations (7.103) or (7.123) or (7.159). Therefore, if we make the identification

$$G = \partial g/\partial t, \tag{6.7.161}$$

we see that (7.160) can be rewritten in the form

$$K(Z;t) = H(z(Z,t);t) + \partial g/\partial t \tag{6.7.162}$$

when $\mathcal{N}$ arises from from some specified g, α pair. We have found the relation between the old and new Hamiltonians in the general case.

Suppose an $\mathcal{N}(t)$ can be found such that

$$K(Z;t) = 0. \tag{6.7.163}$$

This is, in principle, always possible because we can take the Z to be the initial conditions and take $\mathcal{N}(t)$ to be the symplectic map that transforms final conditions into initial conditions. If a g, α pair can be found such that $\mathcal{N}(t)$ arises from their use, then combining (7.162) and (7.163) gives the general Hamilton-Jacobi relation/equation

$$H(z(Z,t);t) + \partial g/\partial t = 0. \tag{6.7.164}$$

In the next subsection we will see that, at least locally, a g, α pair can be found for any $\mathcal{N}(t)$ such that $\mathcal{N}(t)$ arises from their use. Therefore there is always a suitable α such that the associated general Hamilton-Jacobi equation, at least locally, has a solution.

6.7.4 What Kind of Generating Function/Darboux Matrix Should We Choose?

6.7.4.1 Background

The relations (5.86) and (5.89) illustrated that attempted use of the generating functions F_1 and F_4 fails for the identity map. Here is another example of failure: Consider Mathieu transformations given by (5.174) and (5.183). Can they be obtained from an F_1 generating function? According to (5.4) we must have $\det(B) \neq 0$ for this to be possible. But we observe from (5.186) that $\det(B) = 0$ for any Mathieu transformation. Therefore, attempted construction of the desired F_1 must fail. Nevertheless, let us examine ω_1 in this case as given by (5.79). Using (5.79), (5.183), and (5.184) gives for Mathieu transformations the result

$$\begin{aligned}\omega_1 &= \sum_k p_k dq_k - P_k dQ_k = (p, dq) - (P, dQ) = (p, dq) - (\beta^{-1}p, \beta^T dq) \\ &= (p, dq) - (\beta\beta^{-1}p, dq) = (p, dq) - (p, dq) = 0.\end{aligned} \tag{6.7.165}$$

We see that ω_1 vanishes when evaluated for any Mathieu transformation, in accord with the fact that Mathieu transformations cannot be obtained from an F_1.

We also learned that the linear symplectic map described by the symplectic matrix R given by (4.8.31) cannot be obtained by use of any of the generating functions F_j. See the discussion in the paragraph below Equation (5.77). We will now verify that this troublesome R can be obtained using the Poincaré generating function F_+.

According to Exercise 6.6.5 the quadratic F_+ given by (6.69), when employed in the Poincaré recipe (6.57), produces the symplectic matrix M given by (6.75). Conversely, the symmetric matrix W specifying the quadratic F_+ is given in terms of M by the relation (6.76). Examination of (6.76) shows that W can be found if $(M+I)^{-1}$ exists or, equivalently, $\det(M+I) \neq 0$. That is, M must not have -1 as an eigenvalue. For the case at hand $M = R$ and therefore

$$M + I = R + I = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 2 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}. \tag{6.7.166}$$

Simple calculation yields the result

$$\det(M + I) = 8 \neq 0. \tag{6.7.167}$$

(More extensive calculation reveals that R has the eigenvalues $1, 1, i, -i$.) Therefore W is well defined by (6.76), and produces $M = R$ when employed in (6.57) and (6.75).

Here is a slightly different perspective on the general question: Suppose the relation of a symplectic map $\mathcal{M}$ to a gradient map $\mathcal{G}$ with the aid of a Darboux matrix α succeeds. Then, according to the work of Subsection 7.1, there must be the relations

$$G = T_\alpha(M), \tag{6.7.168}$$

and

$$M = T_{\alpha^{-1}}(G). \tag{6.7.169}$$

Here G is the Jacobian matrix of the gradient map $\mathcal{G}$, and accordingly must be a symmetric matrix; and M is the Jacobian matrix of the symplectic map $\mathcal{M}$, and accordingly must be a symplectic matrix. For the relation (7.168) to be a well defined Möbius transformation there must be the invertibility condition

$$\det(C^\alpha M + D^\alpha) \neq 0. \tag{6.7.170}$$

We also know from the work of Section 5.11.3 that if (7.170) is satisfied, then there is also the result

$$\det(C^{\alpha^{-1}} G + D^{\alpha^{-1}}) \neq 0, \tag{6.7.171}$$

and vice versa, so that the Möbius transformation (7.169) is also well defined. That is, there is the logical equivalence

$$\det(C^{\alpha^{-1}} G + D^{\alpha^{-1}}) \neq 0 \Leftrightarrow \det(C^\alpha M + D^\alpha) \neq 0. \tag{6.7.172}$$

To verify this claim, make the substitution $W \to G$ in (5.13.99).

Let us test the condition (7.170) for some of the examples we have already discussed: First suppose $M = I$ and we choose the Darboux matrix associated with F_2. Then we have the result

$$\det(C^\alpha M + D^\alpha) = \det(C^\alpha + D^\alpha). \tag{6.7.173}$$

But, from (7.57), we have the result

$$C^\alpha + D^\alpha = \begin{pmatrix} I^n & 0 \\ 0 & I^n \end{pmatrix}, \tag{6.7.174}$$

and therefore

$$\det(C^\alpha M + D^\alpha) = \det(C^\alpha + D^\alpha) = \det(I^{2n}) = 1 \neq 0. \tag{6.7.175}$$

Thus, we expect the use of an F_2 to succeed when $M = I$; and indeed F_2 as given by (5.71) does yield $M = I$.

Next suppose $M = J$ and we again choose the Darboux matrix associated with F_2. Then we have the result

$$C^\alpha M + D^\alpha = C^\alpha J + D^\alpha. \tag{6.7.176}$$

But, from (7.57), we find the results

$$C^\alpha J = \begin{pmatrix} 0 & 0 \\ 0 & I^n \end{pmatrix} \begin{pmatrix} 0 & I^n \\ -I^n & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ -I^n & 0 \end{pmatrix} \tag{6.7.177}$$

and

$$C^\alpha J + D^\alpha = \begin{pmatrix} I^n & 0 \\ -I^n & 0 \end{pmatrix}. \tag{6.7.178}$$

We see that in this case

$$\det(C^\alpha M + D^\alpha) = \det(C^\alpha J + D^\alpha) = 0. \tag{6.7.179}$$

It follows that attempted use of the Darboux matrix associated with F_2 must fail when $M = J$, as we already know from the work of Subsection 5.1.4.

To continue, suppose $M = J$ and we choose the Darboux matrix associated with F_1. Then we again have the result

$$C^\alpha M + D^\alpha = C^\alpha J + D^\alpha. \tag{6.7.180}$$

But now, from (7.53), we find the results

$$C^\alpha J = \begin{pmatrix} 0 & 0 \\ I^n & 0 \end{pmatrix} \begin{pmatrix} 0 & I^n \\ -I^n & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & I^n \end{pmatrix} \tag{6.7.181}$$

and

$$C^\alpha J + D^\alpha = \begin{pmatrix} I^n & 0 \\ 0 & I^n \end{pmatrix}. \tag{6.7.182}$$

We see that in this case

$$\det(C^\alpha M + D^\alpha) = \det(C^\alpha J + D^\alpha) = \det(I^{2n}) = 1 \neq 0. \tag{6.7.183}$$

Thus, we expect the use of an F_1 to succeed when $M = J$; and indeed F_1 as given by (5.76) does yield $M = J$.

Finally, to complete this set of examples, suppose again that $M = I$ but that we choose the Darboux matrix associated with F_1. Then we have the result

$$C^\alpha M + D^\alpha = C^\alpha + D^\alpha. \tag{6.7.184}$$

But, from (7.53), we find the result

$$C^\alpha + D^\alpha = \begin{pmatrix} I^n & 0 \\ I^n & 0 \end{pmatrix}. \tag{6.7.185}$$

We see that in this case

$$\det(C^\alpha M + D^\alpha) = \det(C^\alpha + D^\alpha) = 0. \tag{6.7.186}$$

It follows that attempted use of the Darboux matrix associated with F_1 must fail when $M = I$, as we already know from the work of Subsection 5.1.4.

6.7.4.2 The General case

What can be said in general? What we wish to examine is under what conditions there is a Darboux matrix α and a source/generating function $g(u)$ such that the implicit relation (7.21) can be made explicit to become the relation (7.23) with $\mathcal{M}$ being the desired map. We will first consider maps that have only a constant and a linear part, and then we will consider maps that have nonlinear parts.

6.7.4.2.1 Maps for the Inhomogeneous Symplectic Group $ISp(2n, \mathbb{R})$

Let us begin with maps that have only a constant and a linear part. These are the maps for the inhomogeneous symplectic group $ISp(2n, \mathbb{R})$. Our goal will be to find a Darboux matrix α and a source/generating function $g(u)$ such that their use produces maps of the form (2.10). As a first example, let us employ for the Darboux matrix α the $\tilde{\tilde{\beta}}$ given by (5.13.154) evaluated for $L = M$ and $V = 0$. So doing gives the result

$$\alpha = \tilde{\tilde{\beta}}|_{L=M,V=0} = (1/\sqrt{2}) \begin{pmatrix} -JM^{-1} & J \\ M^{-1} & I \end{pmatrix}. \tag{6.7.187}$$

For g make the choice

$$g(u) = (v, u) \tag{6.7.188}$$

where v is some vector yet to be determined. For this choice there is the relation

$$\partial_u g = v, \tag{6.7.189}$$

and use of (7.22) yields the result

$$(1/\sqrt{2})[-JM^{-1}Z + Jz) = v. \tag{6.7.190}$$

Upon solving (7.190) for Z we find the relation

$$Z = Mz + \sqrt{2}MJv. \tag{6.7.191}$$

This relation is equivalent to (2.10), after setting $Z = \bar{z}$, provided a v can be found such that

$$\sqrt{2}MJv = c, \tag{6.7.192}$$

for then (7.191) becomes

$$Z = Mz + c, \tag{6.7.193}$$

and our goal will have been accomplished. Finally, (7.192) can indeed be solved for v to yield the well defined relation

$$v = -(1/\sqrt{2})JM^{-1}c \tag{6.7.194}$$

because M is assumed to be symplectic and therefore invertible. Note that for this example the burden of producing M is borne *entirely* by the Darboux matrix, and the generating function provides only the translation part.

Next suppose we continue to use the Darboux matrix given by (7.187) but consider a more general generating function specified by the Ansatz

$$g(u) = (v, u) + (1/2)(u, Wu) \tag{6.7.195}$$

where v and W are to be determined. For this choice there is the relation

$$\partial_u g = v + Wu, \tag{6.7.196}$$

and use of (7.22) yields the result

$$(1/\sqrt{2})[-JM^{-1}Z + Jz) = v + W(1/\sqrt{2})(M^{-1}Z + z). \tag{6.7.197}$$

And solving (7.197) for Z gives the result

$$\begin{aligned} Z &= -(JM^{-1} + WM^{-1})^{-1}(W - J)z - (JM^{-1} + WM^{-1})^{-1}\sqrt{2}v \\ &= -M(J + W)^{-1}(W - J)z - M(J + W)^{-1}\sqrt{2}v \\ &= -M(J + W)^{-1}J^{-1}J(W - J)z - M(J + W)^{-1}J^{-1}J\sqrt{2}v \\ &= M(I - JW)^{-1}(I + JW)z + M(I - JW)^{-1}J\sqrt{2}v \end{aligned} \tag{6.7.198}$$

Let us define a matrix N by the rule

$$N = (I - JW)^{-1}(I + JW). \tag{6.7.199}$$

We observe from (7.199) that N is the Cayley transform of the symmetric matrix W, and therefore N is symplectic. Moreover, according to the work of Exercise 6.5, there are the relations

$$N = T_{\sigma^{-1}}(W) \tag{6.7.200}$$

and

$$W = T_\sigma(N) = -J(N - I)(N + I)^{-1}. \tag{6.7.201}$$

With the aid of the definition (7.199), the relation (7.198) can be rewritten in the form

$$Z = MNz + M(I - JW)^{-1}J\sqrt{2}v. \tag{6.7.202}$$

Yet a bit more can be accomplished by algebraic manipulation. From (7.199) we see that

$$N + I = (I - JW)^{-1}[(I + JW) + (I - JW)] = (I - JW)^{-1}(2I), \tag{6.7.203}$$

and therefore

$$(I - JW)^{-1} = (1/2)(N + I). \tag{6.7.204}$$

Consequently, (7.202) can also be rewritten in the pleasing form

$$Z = MNz + (1/\sqrt{2})M(N + I)Jv. \tag{6.7.205}$$

To continue this discussion of maps for $ISp(2n, \mathbb{R})$ using the Darboux matrix given by (7.187), let us make the definitions

$$M' = MN \text{ or } N = M^{-1}M' \tag{6.7.206}$$

and

$$c = (1/\sqrt{2})M(N + I)Jv \text{ or } v = -\sqrt{2}J(N + I)^{-1}M^{-1}c. \tag{6.7.207}$$

With these definitions we see that the relation between Z and z can be written in the final form

$$Z = M'z + c. \tag{6.7.208}$$

The general $ISp(2n, \mathbb{R})$ map has been obtained using the fixed Darboux matrix α given by (7.187) and the generating function g given by (7.195) subject only to the caveat that v and W be well defined. From the second form of (7.207) we see that the condition for v to be well defined is that $(N + I)^{-1}$ must exist. That is, there is the requirement

$$\det(N + I) \neq 0. \tag{6.7.209}$$

And from (7.201) we see that the same condition must hold for W to be well defined. We close the discussion of this example by noting that there are also the relations

$$W = T_{\alpha}(M') \tag{6.7.210}$$

and

$$M' = T_{\alpha^{-1}}(W). \tag{6.7.211}$$

See Exercise 7.2.

As a second example, suppose we use for the Darboux matrix α that given by (7.56) and (7.57), the Darboux matrix for the generating function F_2, and continue to employ a $g(u)$ of the form (7.195). In this case we find that use of (7.22) yields the result

$$Z = Mz + (A^{\alpha} - WC^{\alpha})^{-1}v \tag{6.7.212}$$

with M and W connected by the relation

$$M = T_{\alpha^{-1}}(W). \tag{6.7.213}$$

Again see Exercise 7.2.

The relation (7.213) has as its inverse the relation

$$W = T_\alpha(M) = (A^\alpha M + B^\alpha)(C^\alpha M + D^\alpha)^{-1}. \tag{6.7.214}$$

Evidently W is well defined in terms of M provided

$$\det(C^\alpha M + D^\alpha) \neq 0. \tag{6.7.215}$$

That is, a specified M can be obtained with the use of α as given by (7.56) and (7.57) and a generating function $g(u)$ provided M is such that (7.215) is satisfied. Let us write M in the block form (5.25). Employing this form gives the result

$$C^\alpha M + D^\alpha = \begin{pmatrix} I^n & 0 \\ C & D \end{pmatrix}, \tag{6.7.216}$$

from which it follows that

$$\det(C^\alpha M + D^\alpha) = \det D. \tag{6.7.217}$$

Therefore, for the use of the Darboux matrix α associated with F_2, the condition (7.215) becomes

$$\det D \neq 0, \tag{6.7.218}$$

in agreement with (5.5).

As a third example, suppose we use for α the Cayley Darboux matrix σ, the Darboux matrix associalted with F_+. See (7.68). Put another way, suppose we use for α the Darboux matrix $\tilde{\tilde{\beta}}$ given by (5.13.154) evaluated at $L = I$ and $V = 0$. So doing gives the result

$$\alpha = \tilde{\tilde{\beta}}|_{L=I,V=0} = (1/\sqrt{2}) \begin{pmatrix} -J & J \\ I & I \end{pmatrix} = \sigma. \tag{6.7.219}$$

The relations (7.212) through (7.215) continue to hold since they are true for any choice of α. But now (7.216) becomes

$$C^\sigma M + D^\sigma = (1/\sqrt{2})(M + I). \tag{6.7.220}$$

Correspondingly, the condition (2.15) becomes

$$\det(M + I) \neq 0. \tag{6.7.221}$$

That is, M must not have -1 as an eigenvalue, a requirement that we already have learned for the existence of a Cayley representation. Recall Section 3.12.

Let us summarize what we have learned about the use of Darboux matrices and generating functions for the case of $ISp(2n, \mathbb{R})$ maps. From the first example we have seen that for any symplectic matrix M there is an associated Darboux matrix α given by (7.187) such that the full burden of producing M is borne by the Darboux matrix and the generating function $g(u)$ is required only to produce the translation part. We may say that this Darboux matrix is *optimally compatible* with M. Moreover, $ISp(2n, \mathbb{R})$ maps of the form (7.208) with symplectic matrices M' of the form (7.206), can also be produced using the same Darboux matrix and a suitable $g(u)$ provided M' is sufficiently near M so that (7.209) is satisfied.

Conversely, it is attractive to conjecture that, for any Darboux matrix α, there is an $Sp(2n, \mathbb{R})$ [and also an $ISp(2n, \mathbb{R})$] map with symplectic matrix M for which M and α are incompatible. That is, given any Darboux matrix α, there is a symplectic matrix M such that

$$\det(C^{\alpha}M + D^{\alpha}) = 0. \tag{6.7.222}$$

Put another way, there is no fixed/universal Darboux matrix (generating function kind) that is compatible with all symplectic matrices. Correspondingly, there is no fixed/universal Darboux matrix (generating function kind) that is compatible with all symplectic maps. See Exercise 7.17. Examples 2 and 3 of this subsection, as well as examples at the beginning of this section, illustrate instances of incompatibility.

6.7.4.2.2 Maps with Nonlinear Parts

In Section 7.8 we will see that any symplectic map $\mathcal{M}$ can be written in the Lie form

$$\mathcal{M} = \exp(:f_1:)\mathcal{R}\mathcal{N} \tag{6.7.223}$$

where the factor $\exp(:f_1:)$ produces a translation, the factor

$$\mathcal{R} = \exp(:f_2^c:)\exp(:f_2^a:) \tag{6.7.224}$$

produces a linear transformation described by the symplectic matrix R, and $\mathcal{N}$ is the nonlinear map

$$\mathcal{N} = \exp(:f_3:)\exp(:f_4:)\cdots. \tag{6.7.225}$$

It can be shown that any such map $\mathcal{M}$ can be produced with the aid of a suitable Darboux matrix α and generating function $g(u)$ pair. However, if this is done, the choice of α will be constrained by the requirement that it be compatible with R. An alternative, which we will follow, is to represent just the nonlinear part $\mathcal{N}$ by a Darboux matrix-generating function pair. This allows for flexibility in the treatment of $\mathcal{N}$ and causes no undue problem in dealing with the $ISp(2n, \mathbb{R})$ part of $\mathcal{M}$, namely the $\exp(:f_1:)\mathcal{R}$ part, since it can be handled separately using the methods of Subsection 7.4.2.1 above and those of Chapter 9.

In the nonlinear case we would like the relation (7.22), and the ability to make it explicit, to hold over as large a phase-space region as possible. In this subsection we will see from some simple examples that the choice of α influences what can be achieved. These examples will also illustrate that the subject of nonlinear symplectic maps is very complicated. Therefore the discussion of this subsection will be limited. A fuller discussion will be undertaken in Chapter 34.

The complexity of nonlinear symplectic maps is already evident at the quadratic level in two variables. Consider the map

$$Q = q - 2qp + O(z^3), \tag{6.7.226}$$

$$P = p + p^2 + O(z^3). \tag{6.7.227}$$

It satisfies the relation

$$\begin{aligned}[Q,P] &= [q-2qp, p+p^2] + O(z^2) \\ &= [q,p] + [q,p^2] - 2[qp,p] - 2[qp,p^2] + O(z^2) \\ &= 1 + 2p - 2p + O(z^2) = 1 + O(z^2). \end{aligned} \tag{6.7.228}$$

Therefore the terms displayed in (7.226) and (7.227) constitute a *symplectic jet.* Indeed, they can be written in the form

$$Z = z + : f_3 : z + O(z^3) \tag{6.7.229}$$

with

$$f_3 = qp^2. \tag{6.7.230}$$

As described in Chapter 34, there are important instances in which a symplectic jet approximation to a symplectic map is *inadequate*. In these cases it is desirable to find an exactly symplectic map whose truncated Taylor expansion matches some specified symplectic jet. Such a symplectic map will be called a *symplectic completion* of the specified symplectic jet.

One way to symplectically complete the symplectic jet (7.229) is to write

$$Z = \mathcal{N} z \tag{6.7.231}$$

with

$$\mathcal{N} = \exp(: f_3 :). \tag{6.7.232}$$

We will call this *Lie symplectification*. So doing gives the result

$$Q = q(1-p)^2, \tag{6.7.233}$$

$$P = p/(1-p). \tag{6.7.234}$$

See Section 1.4.2. [Note that the Taylor expansion through second order of the map given by (7.233) and (7.234) agrees with that in (7.226) and (7.227).] We observe that the map given by (7.230) through (7.234) is analytic at the origin and has a pole on the surface $p = 1$.

We will next explore two examples of how symplectic completion can be achieved with the use of generating functions. In these examples we will again work with the symplectic jet given by (7.226) and (7.227). Let us begin with the use of an F_2 generating function. Suppose we make the Ansatz

$$F_2(q,P) = qP - qP^2. \tag{6.7.235}$$

Use of this Ansatz in (5.5) produces the implicit equations

$$p = \partial F_2/\partial q = P - P^2, \tag{6.7.236}$$

$$Q = \partial F_2/\partial P = q - 2qP. \tag{6.7.237}$$

Since these equtions are quadratic, they can be solved exactly, and we will do so shortly. First, however, let us find the first few terms in the Taylor expansions of $Q(q,p)$ and $P(q,p)$ in powers of q and p. Rewrite (7.236) in the form

$$P = p + P^2. \tag{6.7.238}$$

Now we can expand Q and P in powers of q and p by iteration of (7.237) and (7.238). In lowest approximation, they have the solution

$$Q = q + O(z^2), \tag{6.7.239}$$

$$P = p + O(z^2). \tag{6.7.240}$$

Now substitute (7.239) and (7.240) into (7.237) and (7.238) to get the improved solution

$$Q = q - 2qp + O(z^3), \tag{6.7.241}$$

$$P = p + p^2 + O(z^3). \tag{6.7.242}$$

We have verified that the use of the generating function Ansatz (7.235) produces a (symplectic) map whose Taylor expansion through second order yields the jet map given by (7.226) and (7.227).

Let us now solve (7.236) and (7.237) to find $Q(q, p)$ and $P(q, p)$ exactly. Solving (7.236) for P gives the result

$$P = (1/2)[1 - (1 - 4p)^{1/2}], \tag{6.7.243}$$

and substituting (7.243) into (7.237) gives the complementary result

$$Q = q(1 - 4p)^{1/2}. \tag{6.7.244}$$

[Note that the Taylor expansion through second order of the map given by (7.243) and (7.244) agrees with that in (7.226) and (7.227).] We see that the map given by (7.243) and (7.244) is analytic at the origin and has a branch point on the surface $p = 1/4$.

What happens if we use an F_+ generating function instead of F_2? According to (7.68) this amounts to choosing the Darboux matrix α to be σ. Also, according to (7.219) and the previous discussion of compatibility, in choosing α to be σ we have chosen α to be the Darboux matrix that is optimally compatible with the symplectic matrix I. Finally, the linear part of the map given by (7.226) and (7.227) is indeed the identity matrix I.

In the two-dimensional case the variable u appearing in (7.22) will have the components

$$u = \{u_1; u_2\}^T. \tag{6.7.245}$$

Make the Ansatz

$$g(u) = -(\sqrt{2}/4)u_1(u_2)^2. \tag{6.7.246}$$

In this case

$$\partial_u g = \{-(\sqrt{2}/4)(u_2)^2; -(\sqrt{2}/2)u_1u_2\}^T, \tag{6.7.247}$$

and use of (7.22) with the Darboux matrix α given by (7.68) yields the implicit equations

$$(1/\sqrt{2})(-JZ + Jz) = \{-(\sqrt{2}/4)(u_2)^2; -(\sqrt{2}/2)u_1u_2\}^T|_{u=(1/\sqrt{2})(Z+z)}. \tag{6.7.248}$$

The equations (7.248) can be rewritten in the vector form

$$\begin{aligned} Z - z &= (\sqrt{2})J\{-(\sqrt{2}/4)(u_2)^2; -(\sqrt{2}/2)u_1u_2\}^T|_{u=(1/\sqrt{2})(Z+z)} \\ &= J\{-(1/4)(P + p)^2; -(1/2)(Q + q)(P + p)\}^T \\ &= \{-(1/2)(Q + q)(P + p); +(1/4)(P + p)^2\}^T. \end{aligned} \tag{6.7.249}$$

Finally, the vector form (7.249 is equivalent to the component equations

$$Q - q = -(1/2)(Q + q)(P + p), \tag{6.7.250}$$

$$P - p = (1/4)(P + p)^2. \tag{6.7.251}$$

The component equations (7.250) and (7.251) are quadratic and can be solved explicitly, which we will do shortly. But first let us seek Taylor expansions for $Q(q, p)$ and $P(q, p)$. Rewrite (7.250) and (7.251) in the forms

$$Q = q - (1/2)(Q + q)(P + p), \tag{6.7.252}$$

$$P = p + (1/4)(P + p)^2 \tag{6.7.253}$$

and iterate them once and then once again to find the results

$$Q = q + O(z^2), \tag{6.7.254}$$

$$P = p + O(z^2); \tag{6.7.255}$$

$$Q = q - 2qp + O(z^3), \tag{6.7.256}$$

$$P = p + p^2 + O(z^3). \tag{6.7.257}$$

We see that (7.256) and (7.257) agree with (7.226) and (7.227) as desired.

Let us now solve (7.252) and (7.253) to find $Q(q, p)$ and $P(q, p)$ exactly. Solving (7.253) for P gives the result

$$P = 2 - p - 2(1 - 2p)^{1/2}, \tag{6.7.258}$$

and substituting (7.258) into (7.252) and solving for Q gives the complementary result

$$Q = q(1 - 2p)^{1/2}/[2 - (1 - 2p)^{1/2}]. \tag{6.7.259}$$

[Note that the Taylor expansion through second order of the map given by (7.258) and (7.259) agrees with that in (7.226) and (7.227).] We see that the map given by (7.258) and (7.259) is analytic at the origin and has a branch point on the surface $p = 1/2$.

6.7.4.2.3 Concluding Discussion

We have studied how the choice of a Darboux matrix-generating function pair affects the representation of maps that have only constant and linear terms, namely $ISp(2n, \mathbb{R})$ maps, and also how the choice affects nonlinear maps. For the $ISp(2n, \mathbb{R})$ case, because of its relative simplicity, the treatment was essentially complete. By contrast, the nonlinear case is much more complicated. Below are some questions/observations that naturally arise about the nonlinear case along with partial responses:

- How did we know to make the generating function Ansätze (7.235) and (7.246)? Chapter 34 describes how, for any choice of Darboux matrix α, the Lie f_n and the generating function g_n are related.

- Why are the resulting maps given by (7.233) and (7.234), by (7.243) and (7.244), and by (7.258) and (7.259) all different even though they symplectify (symplectically complete) the same two-jet given by (7.226) and (7.227)? As found in Chapter 34, even if all the f_n vanish beyond some n value $n_{\max}$, the same will not be true for the associated g_n. Had these high-order g_n (generally infinite in number) been retained, the resulting maps would agree.

- Because only g_3 (and perhaps g_2) generating functions were involved in our examples, the implicit equations produced by their use were quadratic, and therefore could be solved exactly. What can be done, as occurs for more realistic cases, when the implicit equations are higher order? Chapter 34 describes various iterative methods, including Newton's method, for efficiently solving the implicit equations numerically.

- For the f_n case studied, namely that given by (7.230), it was found that the use of an F_2 generating function produced a map that was singular at $p = 1/4$, and the use of an F_+ generating function produced a map that was singular at $p = 1/2$. We have seen that in the linear case an F_+ generating function is more compatible with the identity symplectic matrix I than is an F_2 generating function, and for nonlinear maps of the form (7.225) the linear part is the identity map. Is it significant that the use of an F_+ generating function produced a map with a *larger* domain of analyticity than the use of an F_2 generating function?

  Let us review/reconsider when the Möbius transformations relating gradient and symplectic maps succeed or fail. We recall that G is the Jacobian of the gradient map $\mathcal{G}$ and M is the Jacobian of the symplectic map $\mathcal{M}$. Let α be the Darboux matrix that produces the Möbius transformations transformations T_α and $T_{\alpha^{-1}}$. According to (7.48) and (7.40) there are the complementary relations

$$G = T_\alpha(M) = (A^\alpha M + B^\alpha)(C^\alpha M + D^\alpha)^{-1} \tag{6.7.260}$$

  and

$$M = T_{\alpha^{-1}}(G) = (A^{\alpha^{-1}} G + B^{\alpha^{-1}})(C^{\alpha^{-1}} G + D^{\alpha^{-1}})^{-1}. \tag{6.7.261}$$

  For (7.260) to be well defined we must have

$$\det(C^\alpha M + D^\alpha) \neq 0, \tag{6.7.262}$$

  and for (7.261) to be well defined we must have

$$\det(C^{\alpha^{-1}} G + D^{\alpha^{-1}}) \neq 0. \tag{6.7.263}$$

  Recall that the conditions (7.262) and (7.263) are logically equivalent. See (7.172). Therefore (7.260) is well defined if (7.261) is well defined, and vice versa.

  Conversely, we expect that determination of G and therefore construction of the associated generating function $g(u)$ will fail if

$$\det(C^\alpha M + D^\alpha) = 0 : \text{ Condition for incompatibility of } \alpha \text{ and } M. \tag{6.7.264}$$

We have already seen examples of this incompatibility. Moreover, we expect that the construction of a satisfactory $\mathcal{M}$ from a generating function $g(u)$ using (7.22) will fail if

$$\det(C^{\alpha^{-1}}G + D^{\alpha^{-1}}) = 0 : \text{ Condition for map construction failure.} \tag{6.7.265}$$

Put another way, complementary to the logical equivalence (7.172), there are the logical implications

$$\det(C^{\alpha}M + D^{\alpha}) = 0 \;\Rightarrow\; G \text{ and hence } \mathcal{G} \text{ and } g \text{ are not defined,} \tag{6.7.266}$$

$$\det(C^{\alpha^{-1}}G + D^{\alpha^{-1}}) = 0 \;\Rightarrow\; M \text{ and hence } \mathcal{M} \text{ are not defined.} \tag{6.7.267}$$

We will see, for our examples, that the appearance of singularities is related to map construction failure, the condition (7.265).

Consider first the use of F_2. We could treat this case by employing the condition (7.265) with α being the Darboux matrix associated with F_2. See Exercise 7.18. Instead, and equivalently as shown in Exercise 7.19, we already know that use of F_2 assumes that

$$\det(\partial^2 F_2/\partial q_k \partial P_\ell) \neq 0. \tag{6.7.268}$$

See (5.5). For the case (7.235) there is the result

$$\partial^2 F_2/\partial q \partial P = 1 - 2P, \tag{6.7.269}$$

and (7.268) fails when

$$P = 1/2. \tag{6.7.270}$$

From (7.243) we see that (7.270) implies that

$$p = 1/4, \tag{6.7.271}$$

the value for which the map given by (7.243) and (7.244) has a singularity!

Next consider the use of F_+. For the case of F_+, the Darboux matrix is σ. See (7.68). Correspondingly, the condition for map construction failure becomes

$$\det(C^{\sigma^{-1}}G + D^{\sigma^{-1}}) = 0. \tag{6.7.272}$$

From (5.13.12) we see that (7.272) has the specific form

$$\det(-JG + I) = 0. \tag{6.7.273}$$

Since G is the Jacobian of the gradient map $\mathcal{G}$, see (1.6), it follows that in our two-dimensional case G is given by the matrix

$$G = \begin{pmatrix} \partial^2 g/\partial u_1 \partial u_1 & \partial^2 g/\partial u_1 \partial u_2 \\ \partial^2 g/\partial u_2 \partial u_1 & \partial^2 g/\partial u_2 \partial u_2 \end{pmatrix} = \begin{pmatrix} 0 & -(\sqrt{2}/2)u_2 \\ -(\sqrt{2}/2)u_2 & -(\sqrt{2}/2)u_1 \end{pmatrix}. \tag{6.7.274}$$

Here we have used (7.246). Consequently, we find that

$$-JG+I=\begin{pmatrix} 1+(\sqrt{2}/2)u_2 & (\sqrt{2}/2)u_1 \\ 0 & 1-(\sqrt{2}/2)u_2 \end{pmatrix}, \tag{6.7.275}$$

and the condition (7.273) yields the relation

$$1-(1/2)(u_2)^2=0 \tag{6.7.276}$$

with the solution

$$u_2=\pm\sqrt{2}. \tag{6.7.277}$$

Also, when $\alpha=\sigma$, (7.18) becomes

$$u=C^\sigma Z+D^\sigma z=(1/\sqrt{2})(Z+z)=(1/\sqrt{2})\{Q+q;P+p\} \tag{6.7.278}$$

so that

$$u_2=(1/\sqrt{2})(P+p). \tag{6.7.279}$$

See (7.68). Employing (7.277) with a + sign converts (7.279) to the relation

$$P+p=2. \tag{6.7.280}$$

Finally, inserting (7.280) into (7.258) yields the result

$$p=1/2, \tag{6.7.281}$$

the value for which the map given by (7.258) and (7.259) has a singularity!

Let us summarize what has been learned about generating function symplectification: Suppose one is given a symplectic jet of the form

$$Z_a=z_a+\sum_{bc}T_{abc}z_bz_c+\sum_{bcd}U_{abcd}z_bz_cz_d+\cdots+O(z^{n_{\max}+1}). \tag{6.7.282}$$

That is, only the terms through degree $n_{\max}$ are given. Select a Darboux matrix α that is compatible with the identity matrix I. From the coefficients in the jet (7.282) and the selected α one can construct a unique polynomial generating function (that will depend on α) of the form

$$g=\sum_{n=2}^{n=n_{\max}+1}g_n \tag{6.7.283}$$

such that its use in (7.22) reproduces the jet (7.282). Then the (symplectic) map obtained by making explicit the implicit relations (7.22), call it $\mathcal{M}(\alpha,g)$, will be analytic about the origin and may be expected to have singularities when (7.265) holds. If we select $\alpha=\sigma$, the Darboux matrix that is optimally compatible with the symplectic matrix I, then $g_2=0$. Moreover, based on our examples, we anticipate that $\mathcal{M}(\sigma,g)$ will have optimal analytic properties.

- Lie symplectification (symplectic completion), that given by (7.232), appears from our first example to be superior from the perspective of the size of the analyticity domain because it produced a symplectic map that is singular on the surface $p = 1$. But Lie symplectification can be carried out analytically only in a few special cases, and its implementation by numerical methods requires the summation of a very large (generally infinite) number of terms. Recall the definition (1.2.44). Therefore its use is generally impractical when rapid computation is required.

- For a symplectification (symplectic completion) procedure that produces from a symplectic jet a symplectic map having *no* singularities (save at infinity), see Section 34.2.4.

Exercises

6.7.1. Relate the discussion surrounding (4.8.21) through (4.8.26) in Section 4.8 to the relations (7.55) through (7.58) in Subsection 7.1.

6.7.2. Suppose g is a source function of the form

$$g(u) = (1/2)(u, Wu) \tag{6.7.284}$$

where W is a symmetric $2n \times 2n$ matrix. It follows that in this case

$$\partial_u g = Wu. \tag{6.7.285}$$

Show that the $\mathcal{M}$ produced by this g using (7.21) is linear and is described by the matrix

$$M = T_{\alpha^{-1}}(W). \tag{6.7.286}$$

To do so, use (7.38) with the substitution $G \to W$.

Suppose g is of the form

$$g(u) = (v, u) + (1/2)(u, Wu) \tag{6.7.287}$$

where v is any vector. It follows that in this case

$$\partial_u g = v + Wu. \tag{6.7.288}$$

Show that now there is the relation

$$Z = Mz + (A^\alpha - WC^\alpha)^{-1}v. \tag{6.7.289}$$

Verify that

$$M = T_{\alpha^{-1}}(W) = (A^{\alpha^{-1}}W + B^{\alpha^{-1}})(C^{\alpha^{-1}}W + D^{\alpha^{-1}})^{-1} \tag{6.7.290}$$

is well defined providing

$$\det(C^{\alpha^{-1}}W + D^{\alpha^{-1}}) \neq 0. \tag{6.7.291}$$

But, for (7.289) to make sense, we must also have

$$\det(A^\alpha - WC^\alpha) \neq 0. \tag{6.7.292}$$

Is this an additional requirement? Show that it is not. Select from the chain of inferences (5.11.42) the inference

$$\det(C^M U + D^M) \neq 0 \;\Leftrightarrow\; \det(UC^{M^{-1}} - A^{M^{-1}}) \neq 0 \tag{6.7.293}$$

and verify that one may make the substitutions $M \to \alpha^{-1}$ and $U \to W$ to obtain the logical equivalence

$$\det(C^{\alpha^{-1}} W + D^{\alpha^{-1}}) \neq 0 \Leftrightarrow \det(A^\alpha - WC^\alpha) \neq 0. \tag{6.7.294}$$

Therefore (7.292) is a consequence of (7.291), and vice versa.

Verify the correctness of the discussion involving (7.88) through (7.96).

6.7.3. If (7.21) or (7.22) is explicit, α must have the property $B^\alpha = C^\alpha = 0$. See (7.17) and (7.18). Compute γ for such an α using the relation $\gamma = \alpha\sigma^{-1}$ with σ^{-1} given by (5.13.12). Show that the γ so obtained cannot satisfy the symplectic conditions (3.3.3) and (3.3.4). Therefore α cannot be a Darboux matrix.

6.7.4. Verify that the parametric representation of $\mathcal{M}$ given by (7.28) and (7.29) yields (7.39) and (7.40) directly.

6.7.5. The purpose of this exercise is to verify the F_1 contents of Table 7.1. Verify that the α given by (7.52) and (7.53) is a Darboux matrix, is also orthogonal, and its use reproduces the equations (7.51) when employed in (7.21). Verify that γ given by (7.54) satisfies $\alpha = \gamma\sigma$ and is J^{4n} symplectic.

Hint: Show that, for the α given by (7.52) and (7.53), the relations (7.17) and (7.18) take the form

$$U_i = p_i \text{ for } i = 1 \text{ to } n, \tag{6.7.295}$$

$$U_i = -P_{i-n} \text{ for } i = n+1 \text{ to } 2n, \tag{6.7.296}$$

$$u_i = q_i \text{ for } i = 1 \text{ to } n, \tag{6.7.297}$$

$$u_i = Q_{i-n} \text{ for } i = n+1 \text{ to } 2n. \tag{6.7.298}$$

Suppose we partition u into two parts, each of length/dimension n, by writing

$$u = (v; w). \tag{6.7.299}$$

Then the relations (7.297) and (7.298) become

$$v = q, \quad w = Q. \tag{6.7.300}$$

Thus, if we use the partition (7.299), we may write

$$g(u, t) = g(v; w, t). \tag{6.7.301}$$

Show that there is the relation

$$g(v; w, t) = F_1(v, w, t). \tag{6.7.302}$$

6.7.6. The purpose of this exercise is to verify the F_2 contents of Table 7.1. Verify that the α given by (7.56) and (7.57) is a Darboux matrix, is also orthogonal, and its use reproduces the equations (7.55) when employed in (7.21). Verify that γ given by (7.58) satisfies $\alpha = \gamma\sigma$ and is J^{4n} symplectic.
Hint: Show that, for the α given by (7.56) and (7.57), the relations (7.17) and (7.18) take the form

$$U_i = p_i \text{ for } i = 1 \text{ to } n, \tag{6.7.303}$$

$$U_i = Q_{i-n} \text{ for } i = n+1 \text{ to } 2n, \tag{6.7.304}$$

$$u_i = q_i \text{ for } i = 1 \text{ to } n, \tag{6.7.305}$$

$$u_i = P_{i-n} \text{ for } i = n+1 \text{ to } 2n. \tag{6.7.306}$$

Suppose we partition u into two parts, each of length/dimension n, by writing

$$u = (v; w). \tag{6.7.307}$$

Then the relations (7.305) and (7.306) become

$$v = q, \quad w = P. \tag{6.7.308}$$

Thus, if we use the partition (7.307), we may write

$$g(u,t) = g(v; w, t). \tag{6.7.309}$$

Show that there is the relation

$$g(v; w, t) = F_2(v, w, t). \tag{6.7.310}$$

6.7.7. The purpose of this exercise is to verify and work with the F_+ contents of Table 7.1. Your task is to show that use of the α given by (7.68) reproduces the equations (7.67) when employed in (7.21) or (7.22).

First show that, for the α given by (7.68), the relations (7.17) and (7.18) take the form

$$U = A^\sigma Z + B^\sigma z = -(1/\sqrt{2})J\Delta, \tag{6.7.311}$$

$$u = C^\sigma Z + D^\sigma z = (1/\sqrt{2})\Sigma. \tag{6.7.312}$$

Note that (7.312) can be rewritten as

$$\Sigma = \sqrt{2}u. \tag{6.7.313}$$

Next show that the relation (7.22) takes the form

$$-(1/\sqrt{2})J\Delta = \partial_u g|_{u=(1/\sqrt{2})\Sigma}, \tag{6.7.314}$$

from which it follows that

$$\Delta = \sqrt{2}J\partial_u g|_{u=(1/\sqrt{2})\Sigma}. \tag{6.7.315}$$

Let us compare this result with the Poincaré generating function result (7.67) which reads

$$\Delta = J\partial_{\Sigma}F_{+}. \tag{6.7.316}$$

Verify that (7.315 and (7.316) are equivalent when there is the relation

$$g(u,t) = (1/2)F_{+}(\Sigma,t) = (1/2)F_{+}(u\sqrt{2},t). \tag{6.7.317}$$

To do so, apply the chain rule to (7.317) to show that

$$\partial g/\partial u_a = (1/2)\sum_b(\partial F_{+}/\partial\Sigma_b)(\partial\Sigma_b/\partial u_a). \tag{6.7.318}$$

But, by (7.313), verify that there is the relation

$$\partial\Sigma_b/\partial u_a = \sqrt{2}\delta_{ba}. \tag{6.7.319}$$

Verify that therefore (7.318) can be rewritten in the component form

$$\partial g/\partial u_a = (1/\sqrt{2})\partial F_{+}/\partial\Sigma_a \tag{6.7.320}$$

or, more compactly, in the vector form

$$\partial_u g = (1/\sqrt{2})\partial_{\Sigma}F_{+}. \tag{6.7.321}$$

Finally, employ (7.321) in (7.315) to find the result

$$\Delta = \sqrt{2}J\partial_u g = \sqrt{2}J(1/\sqrt{2})\partial_{\Sigma}F_{+} = J\partial_{\Sigma}F_{+}, \tag{6.7.322}$$

which agrees with (7.316).

As a simple example, suppose g is of the form

$$g(u) = (v',u) + (1/2)(u,W'u) \tag{6.7.323}$$

where v' is any vector and W' is any symmetric matrix. Then

$$\partial_u g = v' + W'u. \tag{6.7.324}$$

Show that in this case there is the relation

$$\begin{aligned} Z &= Mz + (A^{\sigma} - W'C^{\sigma})^{-1}v' = Mz + [-(1/\sqrt{2})J - (1/\sqrt{2})W']^{-1}v' \\ &= Mz - \sqrt{2}(J+W')^{-1}v' = Mz - \sqrt{2}(J - JJW')^{-1}v' \\ &= Mz - \sqrt{2}(I - JW')^{-1}J^{-1}v' = Mz + \sqrt{2}(I - JW')^{-1}Jv' \end{aligned} \tag{6.7.325}$$

with

$$M = T_{\sigma^{-1}}(W') = (I - JW')^{-1}(I + JW'). \tag{6.7.326}$$

Compare this result with that given by (6.75) in the case that F_{+} is of the form (6.69). Verify that (6.81) and (7.326) agree provided

$$W' = W, \tag{6.7.327}$$

and

$$v' = (1/\sqrt{2})v. \tag{6.7.328}$$

Is this result consistent with (7.317)?

6.7.8. Verify that (7.87) is equivalent to (7.86).

6.7.9. In a $2n$-dimensional phase space consider the n-dimensional submanifold parameterized by the equations

$$q_i = \tau_i, \tag{6.7.329}$$

$$p_i = p_i^0. \tag{6.7.330}$$

Show that this submanifold is J^{2n} Lagrangian. Consider the submanifold parameterized by the equations

$$q_i = \tau_i, \tag{6.7.331}$$

$$p_i = \tau_i. \tag{6.7.332}$$

What can be said about it? What can be said about the submanifold parameterized by the equations

$$q_i = \tau_i, \tag{6.7.333}$$

$$p_i = \partial f(\tau)/\partial \tau_i, \tag{6.7.334}$$

where f is any function of τ?

6.7.10. Verify that the graph of $\mathcal{M}$ is a $\tilde{J}^{4n}$ Lagrangian submanifold using the parametric form of $\mathcal{M}$ given by (7.28) and (7.29). Hint: Write that

$$\text{graph of } \mathcal{M} = \{\hat{Z} \in \mathbb{R}^{4n} \mid \hat{Z} = \alpha^{-1}\hat{U},\ \ \hat{U} = (U;u)^T = (\mathcal{G}u;u)^T \text{ with } \ u \in \mathbb{R}^{2n}\}. \tag{6.7.335}$$

Make the Ansatz

$$u = u^0 + \sum_1^{2n} \lambda_i e^i. \tag{6.7.336}$$

Show that the tangent vectors ζ^j to the graph of $\mathcal{M}$ are given by the relations

$$\zeta^j(u^0) = \partial\hat{Z}/\partial\lambda_j|_{\lambda=0} = \alpha^{-1}\partial\hat{U}/\partial\lambda_j|_{\lambda=0} = \alpha^{-1}\nu^j(u^0). \tag{6.7.337}$$

Verify that these tangent vectors are $\tilde{J}^{4n}$ isotropic by showing that

$$(\zeta^j, \tilde{J}^{4n}\zeta^k) = (\alpha^{-1}\nu^j, \tilde{J}^{4n}\alpha^{-1}\nu^k) = (\nu^j, (\alpha^{-1})^T\tilde{J}^{4n}\alpha^{-1}\nu^k) = (\nu^j, J^{4n}\nu^k) = 0. \tag{6.7.338}$$

6.7.11. Suppose all the tangent vectors of a submanifold are mutually isotropic at each point in the submanifold. Such a submanifold is called isotropic or *null*. Show that in a $2n$-dimensional phase space the largest dimension a null submanifold can have is n. Thus a Lagrangian submanifold has the largest possible dimension for a null submanifold. For simplicity, work with J^{2n} isotropic submanifolds.

6.7.12. Suppose, in the relation between symplectic and gradient maps, we wish to arrange to have the identity map $\mathcal{I}$ correspond to the case of a zero (or constant) source function g, and vice versa. What can be said about the Darboux matrix α in this case? Is it unique? Far from it.

If g is zero or constant, we must have

$$\mathcal{G} = 0 \tag{6.7.339}$$

so that

$$G = 0. \tag{6.7.340}$$

Show that, since the linear part of the identity map is the identity matrix I, the relation (7.40) then becomes

$$I = T_{\alpha^{-1}}(0), \tag{6.7.341}$$

or, equivalently,

$$T_\alpha(I) = 0. \tag{6.7.342}$$

Employ the factorization (7.50), namely

$$\alpha = \gamma\sigma, \tag{6.7.343}$$

so that (7.342) becomes

$$T_{\gamma\sigma}(I) = 0. \tag{6.7.344}$$

Show, from the group property of Möbius transformations and (5.14.2), that

$$T_{\gamma\sigma}(I) = T_\gamma(T_\sigma(I)) = T_\gamma(0), \tag{6.7.345}$$

and conclude that

$$T_\gamma(0) = 0. \tag{6.7.346}$$

Review the discussions at the ends of Sections 5.12.7 and 5.13.9.3, and show that one must have

$$\gamma \in H(4n, \mathbb{R}). \tag{6.7.347}$$

Finally, verify that (7.339), (7.343), and (7.347), when employed in (7.28) and (7.29), yield the identity map,

$$Z = z. \tag{6.7.348}$$

6.7.13. The discussion at the end of Subsection 7.4.1 examined what the conditions on M were for W as given by (6.76) to be well defined. What are the conditions on W for M as given by (6.75) to be well defined? Compute W for the case $M = R$ with R given by (4.8.31). Verify that this W satisfies the conditions for M as given by (6.75) to be well defined.

6.7.14. Observe that A as given by (5.120) and A' as given by (7.130) are different. From our previous discussion we know that they both take extrema on the trajectories generated by H. However, A' treats the coordinates ξ and momenta η on an equal footing while A does not. Nevertheless, they are related. Show that

$$A = -(1/2)A' + (1/2)\sum_i (Q_iP_i - q_ip_i). \tag{6.7.349}$$

6.7.15. From (6.15) and (7.142) we see that F and A' are related. Show that if $\mathcal{M}(t^i, t^f)$ is a symplectic map generated by the Hamiltonian $H(\zeta, t)$ using t^i and t^f as initial and final times, then the F of (6.15) is given by the relation

$$F(z, t^i, t^f) = 2\int_{t^i,z}^{t^f,Z(z)} [(\zeta, J\dot{\zeta})/2 + H(\zeta, t)]dt. \tag{6.7.350}$$

Here the integral is to be evaluated for the trajectory of H satisfying (7.124).

Consider each of the three cases

$$H = (k, \zeta) \tag{6.7.351}$$

where k is a constant vector,

$$H = (1/2)(\zeta, S\zeta) \tag{6.7.352}$$

where S is a constant symmetric matrix, and

$$H = (1/3)(\zeta_1)^3. \tag{6.7.353}$$

In each case find $\mathcal{M}$, verify that (6.3) when viewed as a function z is an exact differential, and find an F such that (6.15) is satisfied.

Next suppose that H is of the form

$$H(\zeta) = h_m(\zeta) \tag{6.7.354}$$

where $h_m(\zeta)$ is a homogeneous polynomial of degree m in the variables ζ_a. Show that in this case

$$F(z, t^i, t^f) = -(m-2)(t^f - t^i)h_m(z). \tag{6.7.355}$$

Note that if H is quadratic, then F vanishes. Because quadratic Hamiltonians generate linear symplectic maps, this result is consistent with the earlier discussion of the fact that F is the same for all linear symplectic maps.

Finally, if

$$H(\zeta) = \sum_m h_m(\zeta), \tag{6.7.356}$$

show that

$$F(z, t^i, t^f) = -\int_{t^i}^{t^f} dt \, \sum_m (m-2)h_m(\zeta). \tag{6.7.357}$$

6.7.16. For the map given by (7.233) and (7.234) and the map given by (7.243) and (7.244) show by direct computation/evaluation that

$$[Q, P] = 1, \tag{6.7.358}$$

thereby verifying that these maps are symplectic.

6.7.17. Recall the *incompatibility* condition (7.264), which also appears in (7.222). To free up some symbols for subsequent different use, let us rewrite (7.264) in the form

$$\det(C^\beta M' + D^\beta) = 0. \tag{6.7.359}$$

Our goal is to show that, given any Darboux matrix β, there is a symplectic matrix M' such that the incompatibility condition (7.359) holds.

According to Section 5.13.9.4, the most general Darboux matrix β can be written in the form (5.13.148). Verify that, consequently, there are the relations

$$C^\beta = (1/\sqrt{2})(A^T)^{-1}(-CJ + I)L^{-1} \tag{6.7.360}$$

and

$$D^\beta = (1/\sqrt{2})(A^T)^{-1}(CJ+I); \tag{6.7.361}$$

and therefore (7.359) amounts to the requirement

$$\det[(1/\sqrt{2})(A^T)^{-1}(-CJ+I)L^{-1}M' + (1/\sqrt{2})(A^T)^{-1}(CJ+I)] = 0. \tag{6.7.362}$$

Verify, since A is assumed to be invertible, that the requirement (7.362) is equivalent to the requirement

$$\det[(-CJ+I)L^{-1}M' + (CJ+I)] = 0. \tag{6.7.363}$$

Let us now write M' in the factorized form

$$M' = -LK \tag{6.7.364}$$

where K is yet to be determined. Verify that employing this M' in the argument of (7.363) produces the result

$$\begin{aligned}(-CJ+I)L^{-1}M' + (CJ+I) &= -(-CJ+I)L^{-1}LK + (I+CJ)\\ &= -(-CJ+I)K + (I+CJ).\end{aligned} \tag{6.7.365}$$

Suppose we require that

$$(-CJ+I)L^{-1}M' + (CJ+I) = -(-CJ+I)K + (I+CJ) = 0. \tag{6.7.366}$$

Then (7.363) is automatically satisfied, and accordingly we should examine the implication for K of the requirement

$$-(-CJ+I)K + (I+CJ) = 0. \tag{6.7.367}$$

To do so, begin by verifying the manipulation

$$CJ = JJ^{-1}CJ = JW \tag{6.7.368}$$

with

$$W = J^{-1}CJ = -JCJ. \tag{6.7.369}$$

By assumption C is symmetric. Verify from (7.369) that therefore W is also symmetric. Consequently the requirement (7.367) can also be written in the form

$$-(I-JW)K + (I+JW) = 0. \tag{6.7.370}$$

with W being symmetric.

Suppose (7.370) can be solved for K. This is possible if the *invertibility* condition

$$\det(I-JW) \neq 0 \tag{6.7.371}$$

holds, and doing so gives the result

$$K = (I-JW)^{-1}(I+JW). \tag{6.7.372}$$

Observe that (7.372) is a Cayley representation, and therefore K is symplectic. See (3.12.5). Also, by assumption, L is symplectic, and therefore $-L$ is symplectic. It follows from the group property of symplectic matrices that M', given by the product (7.364), is symplectic. We have achieved our goal in the generic subcase (7.371). We have found, for the invertible subccase, a symplectic matrix M' such that the incompatibility condition (7.359) holds.

To complete our discussion, we must also explore what can be said when the generic condition (7.371) does not hold and instead there is the *singular/noninvertibility* condition

$$\det(I - JW) = 0. \tag{6.7.373}$$

This appears to be a more difficult subcase. But, considerable progress can be made using group theory.

First verify that there is the logical implication

$$\det(I - JW) = 0 \Leftrightarrow \det(I + JW) = 0 \tag{6.7.374}$$

To see this, check the equality chain

$$\begin{aligned} \det(I - JW) &= \det[(I - JW)^T] = \det(I + WJ) \\ &= \det[J(I + WJ)J^{-1}] = \det(I + JW). \end{aligned} \tag{6.7.375}$$

Next observe that JW is a Hamiltonian matrix. See Sections 3.7.2 and 3.7.3. Let H be any Hamiltonian matrix and N be any symplectic matrix. Verify that $\bar{H}$ defined by

$$\bar{H} = NHN^{-1} \tag{6.7.376}$$

is also a Hamiltonian matrix. To do this, set up and employ some Lie-algebraic machinery. Let A and B be any two matrices of the same dimension. Associated with A introduce an operator $\#A\#$ that maps matrices to matrices by the rule

$$\#A\#B = \{A, B\}. \tag{6.7.377}$$

Note that $\#A\#$ is essentially the *adjoint* operator associated with A. See the discussion in Sections 3.7.7, 5.3, and 8.1 where something similar is described. Next, it can be verified that

$$\begin{aligned} \exp(A)B\exp(-A) &= \exp(\#A\#)B = B + \#A\#B + (1/2!)(\#A\#)^2B + \cdots \\ &= B + \{A, B\} + (1/2!)\{A, \{A, B\}\} + (1/3!)\{A, \{A, \{A, B\}\}\} + \cdots . \end{aligned} \tag{6.7.378}$$

See the discussion in Section 8.1 where again something similar is described. Now write N in the factored form

$$N = \exp(JS^a)\exp(JS^c). \tag{6.7.379}$$

See (3.8.26). Show, using (7.378) and (7.379), that there is the result

$$\begin{aligned} \bar{H} &= NHN^{-1} = \exp(JS^a)\exp(JS^c)H\exp(-JS^c)\exp(-JS^a) \\ &= \exp(JS^a)[\exp(\#JS^c\#)H]\exp(-JS^a) = \exp(\#JS^a\#)[\exp(\#JS^c\#)H]. \end{aligned} \tag{6.7.380}$$

Observe that JS^c is a Hamiltonian matrix, and recall that Hamiltonian matrices form the Lie algebra $sp(2n, \mathbb{R})$. It follows from (7.378) that $\exp(\#JS^c\#)H$ is also a Hamiltonian matrix. And, again by analogous reasoning, it follows that $\exp(\#JS^a\#)[\exp(\#JS^c\#)H]$ is also a Hamiltonian matrix, thereby verifying (7.376).

Show, as a consequence of (7.376), that there are the results

$$N(I \pm JW)N^{-1} = (I \pm J\hat{W}) \tag{6.7.381}$$

where $\hat{W}$ is symmetric iff W is symmetric. Indeed, if we write

$$NJWN^{-1} = J\hat{W}, \tag{6.7.382}$$

verify that

$$\hat{W} = -JNJWN^{-1} = (N^T)^{-1}WN^{-1} = (N^{-1})^T WN^{-1}. \tag{6.7.383}$$

To do so, verify that the symplectic condition

$$N^T JN = J \tag{6.7.384}$$

can be rewritten in the form

$$JNJ = -(N^T)^{-1}. \tag{6.7.385}$$

If two symmetric matrices W and $\hat{W}$ are connected by a relation of the form

$$\hat{W} = (N^{-1})^T WN^{-1} \tag{6.7.386}$$

where N is symplectic, then we write

$$\hat{W} \sim W. \tag{6.7.387}$$

Verify that $\sim$ is an equivalence relation. See Exercise (5.12.7). Finally, suppose further that the noninvertibility condition (7.373) holds. Show that there are the results

$$\det(I \pm J\hat{W}) = \det[N(I \pm JW)N^{-1}] = \det(I \pm JW) = 0. \tag{6.7.388}$$

Thus, the "sandwiching" operation described by (7.381) preserves noninvertibility.

The stage is now set to study the incompatibility requirement (7.363) in the noninvertible subcase. Show, using (7.364), (7.365), and (7.368), that (7.363) is equivalent to the requirement

$$\det[-(I - JW)K + (I + JW)] = 0. \tag{6.7.389}$$

Suppose (7.369) holds. Then it is also true that

$$\det\{N[-(I - JW)K + (I + JW)]N^{-1}\} = 0, \tag{6.7.390}$$

and vice versa. Verify that matrix manipulation gives the result

$$\begin{aligned} N[-(I - JW)K + (I + JW)]N^{-1} &= -N(I - JW)N^{-1}NKN^{-1} + N(I + JW)N^{-1} \\ &= -(I - J\hat{W})\check{K} + (I + J\hat{W}) \end{aligned} \tag{6.7.391}$$

where

$$\check{K} = NKN^{-1}. \tag{6.7.392}$$

Note that $\check{K}$ will be symplectic iff K is symplectic. If two symplectic matrices $\check{K}$ and K are connected by a relation of the form (7.392), then we write

$$\check{K} \approx K. \tag{6.7.393}$$

Verify that $\approx$ is also an equivalence relation.

What you have shown is that there is the logical implication

$$\det[-(I - JW)K + (I + JW)] = 0 \Leftrightarrow \det[-(I - J\hat{W})\check{K} + (I + J\hat{W})] = 0. \tag{6.7.394}$$

Therefore, the incompatibility condition is a *class* condition. If it holds for the W, K pair, then it also holds for the $\hat{W}, \check{K}$ pair, and vice versa.

A possible strategy now comes into view. Suppose we partition the set of symmetric matrices W into equivalence classes using the equivalence relation $\sim$ and select a normal form W^{norm} for each equivalence class. Suppose each normal form is sufficiently simple that we can construct a corresponding matrix $\bar{K}$ such that the W^{norm},$\bar{K}$ pair is incompatible. Then we will have proved, also in the noninvertible subcase, that for every choice of a Darboux matrix β there exists a symplectic matrix M' such that (7.359) holds.

Let us see how this strategy works in the case of a two-dimensional phase space so that J, W, and K are 2×2 matrices. In the two-dimensional case W has the general form

$$W = \begin{pmatrix} c & a \\ a & b \end{pmatrix}. \tag{6.7.395}$$

Verify that in this case JW takes the form

$$JW = \begin{pmatrix} a & b \\ -c & -a \end{pmatrix} \tag{6.7.396}$$

and $I \pm JW$ take the forms

$$I \pm JW = \begin{pmatrix} 1 \pm a & \pm b \\ \mp c & 1 \mp a \end{pmatrix}. \tag{6.7.397}$$

Next show that

$$\det(I \pm JW) = 1 - a^2 + bc. \tag{6.7.398}$$

Also, verify that from (7.382) that $\det(JW)$ is a class function. That is,

$$\det(J\hat{W}) = \det(JW). \tag{6.7.399}$$

For the parameterization (7.395) we find that

$$d \overset{\text{def}}{=} \det(JW) = [\det(J)][\det(W)] = \det(W) = -a^2 + bc, \tag{6.7.400}$$

Verify that, for the case of a two-dimensional phase space, there is the relation

$$\det(I \pm JW) = 1 + d. \tag{6.7.401}$$

Finally, from the noninvertibility condition (7.373), verify that we are interested in any equivalence class for which

$$d = -1. \tag{6.7.402}$$

Serendipitously, the normal forms for 2 × 2 symmetric matrices are given in Section 32.2.2.1 in the context of finding normal forms for second-order polynomials. There δ plays the role of d,

$$\delta = d, \tag{6.7.403}$$

and we find that a possible normal form is given by

$$W^{\text{norm}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{6.7.404}$$

See (32.2.41) and (32.2.45). Correspondingly, verify that there are the results

$$JW^{\text{norm}} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{6.7.405}$$

$$I + JW^{\text{norm}} = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}, \tag{6.7.406}$$

$$I - JW^{\text{norm}} = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}. \tag{6.7.407}$$

Let us evaluate the argument of the right side of (7.394) when

$$\hat{W} = W^{\text{norm}} \tag{6.7.408}$$

and we make the inspired (and symplectic) choice

$$\check{K} = \bar{K} = J. \tag{6.7.409}$$

Show that so doing gives the result

$$\begin{aligned} -(I - JW^{\text{norm}})\bar{K} + (I + JW^{\text{norm}}) &= \begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} + \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix} + \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 0 \\ 2 & 0 \end{pmatrix}. \end{aligned} \tag{6.7.410}$$

Evidently the matrix on the far right end of (7.410) has determinant 0. It follows that the W^{norm},$\bar{K}$ pair given by (7.408) and (7.409) is incompatible. Consequently we have shown, for the case of a two-dimensional phase space, that for every choice of a Darboux matrix β there exists a symplectic matrix M' such that the incompatibility condition (7.359) holds even in the noninvertible subcase.

What can be said about higher dimensional phase-space cases? W have already treated the invertible subcase for any number of dimensions. What remains is to treat the noninvertible subcase for dimensions four and higher. We can try to proceed in analogy to the two-dimensional phase-space case. In principle normal forms are known for symmetric matrices in any (even) number of dimensions. See Sections 32.2.2.2 and 32.2.2.3. Therefore some of the necessary tools are available to proceed. But we will not pursue the question further in this exercise.

6.7.18. Review Exercise 7.17. The machinery associated with (7.377) through (7.380) and used to prove (7.376) is elegant, but not really necessary. Verify that (7.376) can also be proved directly using (7.382) through (7.385).

6.7.19. Review Exercise 6.7.6. The purpose of this exercise is to show, for an F_2 generating function, that use of the failure condition (7.265) yields the result

$$\det(\partial^2 F_2/\partial q_k \partial P_\ell) = 0. \tag{6.7.411}$$

Verify that, according to (7.265), (5.13.102), (5.13.103), (6.7.56), and (6.7.57), the relevant matrices in this case are

$$C^{\alpha^{-1}} = -J^{2n}(D^\alpha)^T = \begin{pmatrix} 0 & 0 \\ I^n & 0 \end{pmatrix} \tag{6.7.412}$$

and

$$D^{\alpha^{-1}} = J^{2n}(B^\alpha)^T = \begin{pmatrix} I^n & 0 \\ 0 & 0 \end{pmatrix}. \tag{6.7.413}$$

Write G in the block form

$$G = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \tag{6.7.414}$$

and verify that

$$C^{\alpha^{-1}} G + D^{\alpha^{-1}} = \begin{pmatrix} I^n & 0 \\ A & B \end{pmatrix}, \tag{6.7.415}$$

and therefore in this case (7.265) becomes

$$\det(C^{\alpha^{-1}} G + D^{\alpha^{-1}}) = \det(B) = 0. \tag{6.7.416}$$

Recall that G is the Hessian of g. Verify from the work of Exercise 6.7.6 that

$$B_{k\ell} = \partial^2 F_2/\partial q_k \partial P_\ell, \tag{6.7.417}$$

and that therefore (7.416) implies (7.411).

6.8 Symplectic Invariants

We have seen that Hamiltonian flows produce symplectic maps, and that essentially any symplectic map can be produced by a Hamiltonian flow. Therefore, a fundamental problem is to understand the action of symplectic maps on phase space. Is the action completely general, or are there restrictions? If there are restrictions, what is their nature, and are there any associated invariants? This is a difficult and only partially understood subject. Indeed, it is still a matter of intensive theoretical research in the general setting of symplectic geometry and symplectic topology. It is also of great practical interest. Consider, for example, the field of Accelerator Physics. Suppose some charged-particle source produces a collection of low-energy particles described by some initial distribution function. Imagine these particles are now acted upon over time by some combination of electric and magnetic fields to accelerate them to high energy. What is the final distribution function under the assumption that interactions among the particles, e.g. space-charge effects, are ignored? By a suitable choice of electric and magnetic fields can it be made anything one desires, or are there restrictions? In this section we will partially explore some elementary aspects of this subject.

6.8.1 Liouville's Theorem

Consider some $2n$-dimensional region R^i_{2n} of phase space that will be referred to as an *initial* region. Suppose some symplectic map $\mathcal{M}$ acts on phase space, and in so doing sends R^i_{2n} to some region R^f_{2n} that will be referred to as a *final* region.[23] Let V^i be the volume of the initial region,

$$V^i = \int_{R^i_{2n}} dz^i_1 \cdots dz^i_{2n}, \tag{6.8.1}$$

and let V^f be the volume of the final region,

$$V^f = \int_{R^f_{2n}} dz^f_1 \cdots dz^f_{2n}. \tag{6.8.2}$$

Here the z^i are coordinates for the initial region R^i_{2n}, and the z^f are coordinates for the final region R^f_{2n}. They are related by the map $\mathcal{M}$

$$z^f = \mathcal{M} z^i, \tag{6.8.3}$$

and correspondingly their differentials are related by M,

$$dz^f = M dz^i. \tag{6.8.4}$$

It follows from the standard rules for changing variables of integration that the volume V^f is also given by the relation

$$V^f = \int_{R^i_{2n}} |\det(M)| dz^i_1 \cdots dz^i_{2n}. \tag{6.8.5}$$

But, since M is a symplectic matrix, it must have determinant $+1$. Therefore, comparison of (8.5) and (8.1) shows that the two volumes V^f and V^i are the same,

$$V^f = V^i. \tag{6.8.6}$$

[23] Here, and elsewhere, we assume that the mapping from R^i_{2n} to R^f_{2n} is a bijection so that $\mathcal{M}^{-1}$ exists and both $\mathcal{M}$ and $\mathcal{M}^{-1}$ are single valued. By the existence and uniqueness theorems, this will certainly be the case if $\mathcal{M}$ is the result of integrating some set of differential equations. In that case $\mathcal{M}^{-1}$ is found by integrating backwards in time.

Symplectic maps preserve volume in phase space. This result is called *Liouville's* theorem. Note that in the 2-dimensional case, "volume" is simply area. Therefore, in two dimensions, area (and orientation) preserving maps are symplectic maps, and vice versa.

There is a slightly different phrasing of Liouville's theorem that is also worth mentioning. Consider an ensemble of *noninteracting* systems with each member of the ensemble governed by the same Hamiltonian $H(z, t)$. At some initial instant t^i, let each member of the ensemble be characterized by a point in phase space corresponding to its initial condtions. Suppose, further, that all the points of the ensemble at the initial instant t^i occupy a certain region R^i_{2n} of phase space. Now follow all the trajectories of the members of the ensemble through augmented phase space to some later instant t^f. The members of the ensemble will then occupy some final region R^f_{2n} of phase space. Since the map produced by this Hamiltonian flow is symplectic, we have the relation (8.6). The volume in phase space occupied by the ensemble remains constant. Also, by construction, the number of ensemble points in V^f and V^i is the same. Therefore, since V^f equals V^i, one may also say that the *density* of points in phase space (the number of points divided by the volume they occupy) is preserved by Hamiltonian flows. The collection of ensemble points moves about in phase space (and augmented phase space) like an *incompressible* fluid. In the context of Accelerator Physics, this result means that the density of (assumed noninteracting) beam particles in phase space after acceleration can never exceed (and, in fact, must equal) their initial phase-space density at the source provided their motions are all governed by the same Hamiltonian. That is, particles cannot be concentrated in phase space by solely Hamiltonian means.

6.8.2 Gromov's Nonsqueezing Theorem and the Symplectic Camel

According to Liouville's theorem, if an initial region R^i_{2n} of phase space is sent into a final region R^f_{2n} of phase space under the action of a symplectic map $\mathcal{M}$, then these two regions must have the same volume. One might wonder about the converse: Given two regions of phase space having the same volume, is there a symplectic map that sends one into the other? The answer is *yes* in the case of two-dimensional phase space ($n = 1$), and, as will be done in Chapter 33, it is fairly easy to show that the answer is *no* in the case of four or more phase-space dimensions ($n > 1$) if one is restricted to *linear* symplectic maps. But what about the far more complicated case where nonlinear symplectic maps are allowed? The answer to this question was unknown until 1985 when *Gromov* announced his famous *nonsqueezing theorem* and its application to the *symplectic camel.*[24] The proof of his theorem is beyond the scope of this text and is part of the deep new field of *symplectic topology*. However, it is easy to state and understand its contents.

In the spirit of the theoretical physicist who instructed the farmer to first consider a spherical cow, the mathematician Gromov considered a spherical region in phase space, the

[24] "It is easier for a camel to go through the eye of a needle than for a rich man to enter into the kingdom of God", a saying of Jesus as quoted in Matthew 19, Mark 10, and Luke 18.

symplectic ball $B^{2n}(r)$ of radius r given by the relation

$$B^{2n}(r) = \{z \in R^{2n} \mid \sum_{j=1}^{n}(p_j^2 + q_j^2) \leq r^2\}. \tag{6.8.7}$$

(This ball is called *symplectic* because its definition involves the p_j as well as the q_j.) It is an easy calculation to show that $B^{2n}(r)$ has a finite volume $V(r)$ given by the relation

$$V(r) = r^{2n}\pi^n/[n\Gamma(n)] = r^{2n}\pi^n/n!. \tag{6.8.8}$$

Gromov also considered a symplectic cylinder $C_1^{2n}(r')$ of radius r' given by the relation

$$C_1^{2n}(r') = B_1^2(r') \times R^{2n-2}. \tag{6.8.9}$$

Here $B_1^2(r')$ is the set

$$(p_1^2 + q_1^2) \leq (r')^2 \tag{6.8.10}$$

and, according to (8.9), the remaining variables q_j and p_j for $j > 1$ are allowed to range from minus to plus infinity,

$$q_j \in (-\infty, +\infty),\ p_j \in (-\infty, +\infty) \text{ for } j \in [2, n]. \tag{6.8.11}$$

Evidently, because of (8.11), $C_1^{2n}(r')$ has infinite volume. We now ask if there is a symplectic map $\mathcal{M}$, possibly nonlinear, such that when $\mathcal{M}$ is applied to $B^{2n}(r)$ the resulting region lies within (is *embedded* in) $C_1^{2n}(r')$,

$$\mathcal{M}B^{2n}(r) \subset C_1^{2n}(r')? \tag{6.8.12}$$

Put another way, if we regard $B^{2n}(r)$ as a "symplectic camel", can this camel be squeezed into the cylinder $C_1^{2n}(r')$ under the action of some symplectic map? Liouville would not object because the volume of the cylinder, being infinite, would certainly exceed the volume of the camel. However, Gromov showed that there was no symplectic $\mathcal{M}$ that would map the camel to a region lying within the cylinder unless the radius of the cylinder equaled or exceeded that of the camel (ball),

$$r' \geq r. \tag{6.8.13}$$

A related question, more akin to passing a camel through the eye of a needle, is this: Suppose there is a camel on one side of a wall, and this wall has a hole in it. Is there a continuous family of symplectic maps $\mathcal{M}(\tau)$ such that $\mathcal{M}(0)$ is the identity map $\mathcal{I}$ and the map $\mathcal{M}(1)$ has the property that when it acts on the camel the result is a camel on the other side of the wall? Moreover, is it the case that all points obtained by letting $\mathcal{M}(\tau)$ act on the camel (for $\tau \in [0, 1]$) lie either outside the wall or within the hole in the wall?

Because we are working in dimension four or higher where our intuition may easily fail, let us phrase the question more precisely in mathematical terms. We define the wall W to be the hyperplane $q_1 = 0$,

$$W = \{z \in R^{2n} \mid q_1 = 0\}. \tag{6.8.14}$$

We define the hole in the wall, $H(r')$, to be the set

$$H(r') = \{z \in R^{2n} \mid q_1 = 0 \text{ and } \sum_{j=1}^{n}(p_j^2 + q_j^2) \leq (r')^2\}. \tag{6.8.15}$$

As for the symplectic camel, we define the two sets $B_+^{2n}(r,a)$ and $B_-^{2n}(r,a)$, with $a > r$, by the rules

$$B_+^{2n}(r,a) = \{z \in R^{2n} \mid (q_1 - a)^2 + p_1^2 + \sum_{j=2}^{n}(p_j^2 + q_j^2) \leq r^2\}, \tag{6.8.16}$$

$$B_-^{2n}(r,a) = \{z \in R^{2n} \mid (q_1 + a)^2 + p_1^2 + \sum_{j=2}^{n}(p_j^2 + q_j^2) \leq r^2\}. \tag{6.8.17}$$

Evidently $B_+^{2n}(r,a)$ is a camel centered around the point given by $q_1 = a$ with all remaining coordinates being zero, and $B_-^{2n}(r,a)$ is a camel centered around the point given by $q_1 = -a$ with all remaining coordinates being zero. And since we have assumed $r < a$, no part of either camel is in contact with the wall. Therefore $B_+^{2n}(r,a)$ is a camel located on the side of the wall W with $q_1 > 0$, and $B_-^{2n}(r,a)$ is a camel located on the side of the wall W with $q_1 < 0$.

Now suppose the camel is smaller than the hole in the wall, $r < r'$. Then it is easy to see that the camel can be moved through the hole from one side of the wall to the other by a simple translation along the q_1 axis of the form (6.2.9). Employing notation to be introduced in Section 7.7, we may then write $\mathcal{M}(\tau)$ in the form

$$\mathcal{M}(\tau) = \exp(2\tau a :p_1:). \tag{6.8.18}$$

It easily verified that there are the relations

$$\mathcal{M}(0)B_+^{2n}(r,a) = B_+^{2n}(r,a), \tag{6.8.19}$$

$$\mathcal{M}(1)B_+^{2n}(r,a) = B_-^{2n}(r,a). \tag{6.8.20}$$

Moreover all the points given by

$$\mathcal{M}(\tau)B_+^{2n}(r,a) = B_+^{2n}(r,a(1-2\tau)) \tag{6.8.21}$$

with $q_1 = 0$ satisfy

$$(a(1-2\tau))^2 + p_1^2 + \sum_{j=2}^{n}(p_j^2 + q_j^2) \leq r^2. \tag{6.8.22}$$

From (8.22) we conclude that either $q_1 \neq 0$ or

$$q_1 = 0 \text{ and } \sum_{j=1}^{n}(p_j^2 + q_j^2) \leq r^2 - (a(1-2\tau))^2 \leq (r')^2, \tag{6.8.23}$$

and therefore all points of the camel are either off the wall ($q_1 \neq 0$) or are within the hole $H(r')$ as the camel passes through the wall under the action of $\mathcal{M}(\tau)$. We have moved the camel from the side with $q_1 > 0$ to the side with $q_1 < 0$.

What happens in the more interesting case where the camel is larger than the hole, $r > r'$? In that case Gromov has shown that there is *no* continuous family of symplectic maps $\mathcal{M}(\tau)$ satisfying (8.19) and (8.20) without some points of $\mathcal{M}(\tau)B_+^{2n}(r,a)$ lying in the wall and outside the hole for some intermediate τ values. Thus for a symplectic camel to pass through the eye of a needle under the action of a continuous family of symplectic maps, the eye of the needle must be larger than the camel. By contrast, if one is allowed to use maps that are simply volume preserving but not symplectic, it is easy to see that one can pass the camel through the eye of any needle no matter how large the camel is or how small the eye of the needle is. For example, one may first stretch and thin the camel by pulling along her tail in the $+q_1$ direction while holding her nose fixed. (We assume the camel is eyeing the eye with some trepidation, and we plan to pass her through head first.) While increasing her length in the q_1 direction, we appropriately compress her in all other directions so that her volume remains unchanged. Then this thinned camel may be safely passed through the eye of the needle. Finally, the camel can be brought back to her original shape by holding her hind quarters fixed, pushing on her nose thereby compressing her q_1 dimension, and letting her other dimensions expand to their original values.

The discussion so far has been concerned with the 'spherical' camel $B^{2n}(r)$ given by (8.7). It can be extended to the case of a *general elliptic* camel $E^{2n}(r)$. By a general elliptic camel we mean the set defined by the rule

$$E^{2n}(r) = \{z \in R^{2n} \mid (z, Sz) \leq r^2\} \tag{6.8.24}$$

where S is a positive-definite matrix. Suppose we make the symplectic change of variables

$$z = AZ \text{ or } Z = A^{-1}z \tag{6.8.25}$$

where A is a symplectic matrix. Then there is the relation

$$(z, Sz) = (AZ, SAZ) = (Z, A^T SAZ). \tag{6.8.26}$$

As will be seen in Chapter 33, if S is positive definite, there is always a symplectic A such that

$$A^T SA = S_\lambda \tag{6.8.27}$$

where S_λ is a diagonal matrix with pair-wise degenerate positive entries. (S_λ is called the Williamson diagonal or normal form of S.) In the 4×4 case, for example, S_λ has the form

$$S_\lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & 0 & \lambda_2 \end{pmatrix}. \tag{6.8.28}$$

Correspondingly, and in the case of general dimension, there is the relation

$$(z, Sz) = (Z, S_\lambda Z) = \sum_{j=1}^{n} \lambda_j (P_j^2 + Q_j^2). \tag{6.8.29}$$

Here, without loss of generality, we may select A such the λ_j are ordered in the fashion

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0. \tag{6.8.30}$$

Motivated by this result, we will define a *normal-form elliptic* camel $E^{2n}_{\text{nf}}(r,\lambda)$ by the relation

$$E^{2n}_{\text{nf}}(r,\lambda) = \{z \in R^{2n} \mid \sum_{j=1}^{n} \lambda_j(p_j^2 + q_j^2) \leq r^2\}. \tag{6.8.31}$$

We conclude that the general elliptic camel can be transformed into a normal-form elliptic camel by a linear symplectic map. But now there is a generalization of the nonsqueezing result for the spherical camel to the case of a normal-form elliptic camel. It states that a normal-form elliptic camel cannot be imbedded in the cylinder $C_1^{2n}(r')$ by a symplectic map unless

$$r' \geq r/\sqrt{\lambda_1}. \tag{6.8.32}$$

Similarly, the normal-form elliptic camel cannot be passed through the hole $H(r')$ by a family of symplectic maps unless (8.32) holds.

Evidently the nonsqueezing theorem and the symplectic camel results, which are examples of the general subject of symplectic *capacities*, have important applications to Accelerator Physics. The nonsqueezing theorem has implications for the feasibility of *emittance trading* (the hope that one might be able to concentrate particles in some phase-space plane at the expense of possible dilution in other planes), and the symplectic camel results bear on problems of linear and nonlinear beam transport. Of course one would like to have analogous results for camels and needle eyes with more general shapes than the simple spherical and elliptical and cylindrical shapes assumed in this section. Also, even if all of a camel cannot be squeezed into, say, some cylinder or some other camel, what fraction of the camel can be so squeezed, and how? Some important results have been found in these directions. See the references to Symplectic Geometry and Topology given at the end of this chapter.[25] The study of such matters is still in its infancy. And even when such results have been obtained and should possibly useful nonlinear symplectic maps be found, there will still be the problem of designing beamline elements and sequences of beamline elements to realize the desired symplectic maps.[26] Clearly, in this area as in so many others, there is still much to be learned about the effects of nonlinear maps and how to achieve, exploit, or mitigate them.

[25] There are many surprises. For example, when $r' < r$ so that according to Gromov 100% of the spherical camel cannot be embedded in the cylinder, nevertheless any fraction less than 100% can be embedded. Moreover, the construction of such embeddings is very complicated, and in some cases only an existence proof is available. Apparently there will always be some points whose images are outside the cylinder, and perhaps quite far outside the cylinder, whose measure can be made as small as one might desire.

[26] In the case of Accelerator Physics the maps will arise from Hamiltonian flows and will be analytic. With heroic effort it could probably be possible to achieve maps of the form
$\mathcal{M} = \exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_3:)\exp(:f_4:)\exp(:f_5:)\exp(:f_6:)\cdots\exp(:g_1:)$
with the $g_1, f_2^a, f_2^c, f_3, f_4, f_5, f_6$ being any desired homogeneous polynomials and the f_m with $m > 6$ being small. (See Section 7.7 for the meaning of this notation.) In these considerations questions of differentiability might be important and the distinction between $ISpM(2n,\mathbb{R})$ $[= Symp(n)]$ and $Ham(n)$ might be relevant.

Finally, we close this section with a related consideration. Suppose $f(z)$ and $g(z)$ are two phase-space distributions. It would be nice to know whether or not two different phase-space distributions could be sent into each other by a symplectic map. For example, as already asked earlier, given the phase-space distribution coming out of some ion source or electron gun, is there any possible collection of beamline elements that would transform this distribution into some desired distribution at the end of some beamline or accelerator complex? Mathematically stated, one would like to decompose phase-space distributions into equivalence classes. This too is a deep question about which little is known.

6.8.3 Poincaré Integral Invariants

The volume invariant of Liouville's theorem is actually the last in a hierarchy of invariants called the *Poincaré integral invariants.* The first invariant in the series consists of a certain 2-dimensional integral over a 2-dimensional submanifold in phase space. The next consists of a 4-dimensional integral over a 4-dimensional submanifold, etc. The last consists of a $2n$ dimensional integral, which is just the volume of Liouville's theorem.

A complete and proper discussion of all the Poincaré invariants requires the use of the *exterior calculus of differential forms.* However, the first in the series of invariants is easily discussed using ordinary calculus and the fundamental symplectic 2-form $(\delta z, Jdz)$ introduced earlier, and we will do so shortly. At this point it is worth noting that, in constructing the general higher-order Poincaré invariants, the exterior calculus of differential forms is used to fabricate general $2m$-forms for $m = 2, 3, \cdots, n$ using as the *only* building block the fundamental symplectic 2-form. The invariance of all these forms, including the last of the hierarchy $(m = n)$, which is simply the volume element, follows from the invariance (1.19) of the fundamental symplectic 2-form. This invariance is in turn equivalent to the symplectic condition (1.12). Thus, the symplectic condition is really the fundamental condition from which everything else follows. To the author's knowledge, the utility of the $2m$-forms for the intermediate m values $2 \leq m < n$ is an open question. Finally it is worth remarking that it is the symplectic structure at the classical level of mechanics that makes possible the uncertainty principle at the quantum level.

Let R_2^i be some *initial* 2-dimensional submanifold in phase space. To be more precise, we construct it as follows. We imagine a 2-dimensional Euclidean space with coordinates α,β, and consider a domain Γ_2 in this space. See Figure 8.1 below. We map Γ_2 into R_2^i with the aid of $2n$ relations of the form

$$z_a^i = g_a(\alpha, \beta). \tag{6.8.33}$$

That is, the functions $g_1 \cdots g_{2n}$ specify the mapping of Γ_2 into R_2^i.

Next we will define a certain integral I_2^i over R_2^i. Subdivide Γ_2 into N rectangles with each rectangle having sides $d\alpha$ and $d\beta$. This subdivision of Γ_2 will produce a corresponding subdivision of R_2^i into "parallelograms" with sides dz^i and δz^i. Here dz^i is the vector formed using (8.33) when only α is allowed to vary,

$$dz_a^i = (\partial g_a/\partial \alpha)d\alpha \tag{6.8.34}$$

or, in vector notation,

$$dz^i = \partial_\alpha g d\alpha. \tag{6.8.35}$$

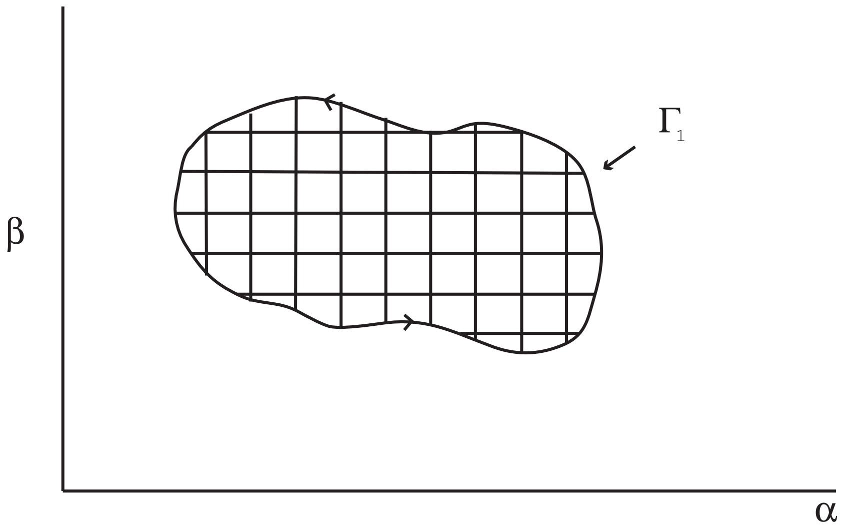


Figure 6.8.1: The domain Γ_2 in α,β space. Also shown is its subdivision into rectangles of sides $d\alpha$, $d\beta$ and its boundary Γ_1.

Similarly, δz^i is the vector formed when only β is allowed to vary,

$$\delta z^i = \partial_\beta g d\beta. \tag{6.8.36}$$

Now, for each parallelogram in R_2^i, compute the quantity $(\delta z^i, Jdz^i)$. By using the relations (8.35) and (8.36) we find the result

$$(\delta z^i, Jdz^i) = (\partial_\beta g, J\partial_\alpha g)d\alpha d\beta. \tag{6.8.37}$$

The left side of (8.37) is the result of evaluating a 2-form in phase space for a small parallelogram in R_2^i. The right side of (8.37) is a 2-form in α,β space. It is called the *pullback* (into α,β space) of the form in phase space.[27] We may summarize the situation as follows: The functions g_a provide a mapping of Γ_2 in α,β space into R_2^i in phase space, with small rectangles in Γ_2 mapped into small parallelograms in R_2^i. On phase space there is a 2-form, namely $(\delta z^i, Jdz^i)$, which induces the 2-form $\{(\partial_\beta g, J\partial_\alpha g)d\alpha d\beta\}$ back in the original α,β space. See Exercise 8.3. We remark that in this terminology the integrand in the right side of (4.47) is the pullback to the τ parameter space of the 1-form (4.41) in z space.

As a last step, form a Riemann sum over all parallelograms in R_2^i and a corresponding Riemann sum over all rectangles in Γ_2. Upon continually refining the subdivision of Γ_2 by letting N go to infinity, we obtain the integrals and the relation

$$I_2^i = \int_{R_2^i} (\delta z^i, Jdz^i) = \int_{\Gamma_2} (\partial_\beta g, J\partial_\alpha g)d\alpha d\beta. \tag{6.8.38}$$

[27] Why is the 2-form on the right side of (8.37) called a pullback? The relations (8.33) provide a mapping from points in Γ_2 to points in phase space. Suppose we regard this as a mapping in the *forward* direction. It is sometimes described by saying that points in Γ_2 are *pushed forward* into points in phase space. By contrast, the relation (8.37) begins with a 2-form in phase space and yields a 2-form *back* in Γ_2 space. If points may be regarded as being pushed forward, then associated forms may be regarded as being related by pulling back.

Put another way, the integral over Γ_2 on the right side of (8.38) is well defined; and, based on (8.37), defines what is meant by the integral I_2^i of the 2-form $(\delta z^i, Jdz^i)$ over R_2^i. [We remark that it can be shown, as desired, that the value of the right side of (8.38) is independent of the choice of parameterization.]

Now suppose some symplectic map $\mathcal{M}$ sends the points in R_2^i to some other 2-dimensional submanifold R_2^f according to the rule (6.3). For this submanifold we can compute the associated integral

$$I_2^f = \int_{R_2^f} (\delta z^f, Jdz^f). \tag{6.8.39}$$

With the aid of (8.4) and its counterpart for δz, this integral can be pulled back from the final phase space to the initial phase space, and then puled back further to α,β space to give the result

$$I_2^f = \int_{R_2^f} (\delta z^f, Jdz^f) = \int_{R_2^i} (M\delta z^i, JMdz^i) = \int_{\Gamma_2} (M\partial_\beta g, JM\partial_\alpha g) d\alpha d\beta. \tag{6.8.40}$$

Here we have used the relations (8.4) and (8.34) to write

$$dz_a^f = (Mdz^i)_a = \sum_c M_{ac} dz_c^i = \sum_c M_{ac}(\partial g_c/\partial\alpha)d\alpha = (M\partial_\alpha g)_a d\alpha, \tag{6.8.41}$$

and similarly for δz^f. But, from the symplectic condition, we have the result

$$(M\partial_\beta g, JM\partial_\alpha g) = (\partial_\beta g, M^T JM\partial_\alpha g) = (\partial_\beta g, J\partial_\alpha g). \tag{6.8.42}$$

See also (1.19). It follows that

$$I_2^f = \int_{\gamma_2} (M\partial_\beta g, JM\partial_\alpha g) d\alpha d\beta = \int (\partial_\beta g, J\partial_\alpha g) d\alpha d\beta = I_2^i. \tag{6.8.43}$$

Under the action of a symplectic map, the 2-dimensional integral based on the fundamental symplectic 2-form is conserved,

$$I_2^f = I_2^i. \tag{6.8.44}$$

Finally, if we wish, we may associate the points on R_2^i with the members of some ensemble at some initial time t^i. Assuming that the members at the ensemble are governed by some Hamiltonian $H(z,t)$, we may follow, as before, the trajectories of the members of the ensemble through augmented phase space to some later instant t^f when they terminate on R_2^f. Since H generates a symplectic map, we have the relation (8.44). Integrals over sums of projected signed areas are conserved. Recall Exercise 1.2.

6.8.4 Connection between Surface and Line Integrals

There is an intimate connection between the 2-form $(\delta z, Jdz)$ and the differential form (1-form)

$$(z, Jdz). \tag{6.8.45}$$

Moreover, we will learn that this connection and the relation (8.44) are the inspiration for the differential form (6.3).

Let Γ_1 be the *boundary* of Γ_2 as illustrated in Figure 8.1. View Γ_1 as a closed path in α,β space, and parameterize it using the parameter $\tau \in [0,1]$ by introducing functions $\alpha(\tau)$ and $\beta(\tau)$. Under the mapping (8.33) there is an associated closed phase-space path in R_2^i, call it R_1^i, given by the relations

$$z_a^i(\tau) = g_a(\alpha(\tau), \beta(\tau)). \tag{6.8.46}$$

By construction, R_1^i is the boundary of R_2^i. Now form the closed phase-space path integral

$$I_1^i = \int_{R_1^i} (z^i, Jdz^i). \tag{6.8.47}$$

To be more explicit, take the differential of (8.46) to find the result

$$dz_a^i(\tau) = (\partial g_a/\partial\alpha)d\alpha + (\partial g_a/\partial\beta)d\beta \tag{6.8.48}$$

or, in vector notation,

$$dz^i = \partial_\alpha g d\alpha + \partial_\beta g d\beta. \tag{6.8.49}$$

These results enable us to write the relations

$$(z^i, Jdz^i) = (z^i, J\partial_\alpha g)d\alpha + (z^i, J\partial_\beta g)d\beta, \tag{6.8.50}$$

$$\begin{aligned} I_1^i = \int_{R_1^i} (z^i, Jdz^i) &= \int_{\Gamma_1} [(z^i, J\partial_\alpha g)d\alpha + (z^i, J\partial_\beta g)d\beta] \\ &= \int_0^1 d\tau[(z^i, J\partial_\alpha g)(d\alpha/d\tau) + (z^i, J\partial_\beta g)(d\beta/d\tau)]. \end{aligned} \tag{6.8.51}$$

Observe that the left side of (8.50) is a differential 1-form in phase space, and the right side is a differential 1-form in α,β space. In analogy to our earlier discussion, the differential form in α,β space is the pullback of the 1-form in phase space. And the integrand on the far right side of (8.51) is a differential 1-form in τ space that is a pullback from α, β space, and the pullback of the pullback from phase space.

Introduce the functions C_α and C_β by the rules

$$C_\alpha(\alpha, \beta) = (z^i, J\partial_\alpha g) = (g, J\partial_\alpha g), \tag{6.8.52}$$

$$C_\beta(\alpha, \beta) = (z^i, J\partial_\beta g) = (g, J\partial_\beta g). \tag{6.8.53}$$

They allow us to write the integral over the closed path Γ_1 in the more compact form

$$I_1^i = \int_{\Gamma_1} [(z^i, J\partial_\alpha g)d\alpha + (z^i, J\partial_\beta g)d\beta] = \int_{\Gamma_1} C_\alpha d\alpha + C_\beta d\beta. \tag{6.8.54}$$

Now apply Green's (or Stokes') theorem to convert the path integral over Γ_1 to a surface integral over Γ_2. Doing so gives the result

$$I_1^i = \int_{\Gamma_1} C_\alpha d\alpha + C_\beta d\beta = \int_{\Gamma_2} (\partial C_\beta/\partial\alpha - \partial C_\alpha/\partial\beta)d\alpha d\beta. \tag{6.8.55}$$

However, from (8.52) and (8.53) we find the results

$$\partial C_\beta/\partial\alpha = (\partial_\alpha g, J\partial_\beta g) + (g, J\partial_\alpha\partial_\beta g), \tag{6.8.56}$$

$$\partial C_\alpha/\partial\beta = (\partial_\beta g, J\partial_\alpha g) + (g, J\partial_\beta\partial_\alpha g). \tag{6.8.57}$$

It follows from the symmetry of mixed partials and the antisymmetry of J that there is the relation

$$\partial C_\beta/\partial\alpha - \partial C_\alpha/\partial\beta = 2(\partial_\alpha g, J\partial_\beta g). \tag{6.8.58}$$

Consequently, (8.55) can be rewriten in the form

$$I_1^i = 2\int_{\Gamma_2}(\partial_\alpha g, J\partial_\beta g)d\alpha d\beta. \tag{6.8.59}$$

Finally, upon comparing (8.38) and (8.59), we find the key result

$$I_1^i = -2I_2^i. \tag{6.8.60}$$

Note that this result is completely general in that it holds for any surface R_2^i and its boundary R_1^i. We remark that one of the features of the exterior calculus of differential forms, see the beginning of Subsection 6.8.3, is that it incorporates the Poincaré lemma of Exercise 1.1 and its generalizations in a systematic way so that relations like (8.60) become routinely obvious.

Now suppose, as in Subsection 6.8.2, that the symplectic map $\mathcal{M}$ sends R_2^i to R_2^f according to the rule (8.3). It will then send R_1^i to R_1^f where R_1^f is the boundary of R_2^f. Let I_1^f be the result of integrating the differential (z^f, Jdz^f) over the path R_1^f,

$$I_1^f = \int_{R_1^f}(z^f, Jdz^f). \tag{6.8.61}$$

Based on the result just found, there is the relation

$$I_1^f = -2I_2^f, \tag{6.8.62}$$

no matter what the nature of $\mathcal{M}$ is save that it be differentiable and invertible. (Given R_2^i and R_1^i, R_2^f and R_1^f must be well defined.) But if $\mathcal{M}$ is symplectic, then (8.44) must hold, and we conclude from (8.60) and (8.62) that there must also be the relation

$$I_1^f = I_1^i. \tag{6.8.63}$$

When written out in full, the relation (8.63) reads

$$\int_{R_1^f}(z^f, Jdz^f) = \int_{R_1^i}(z^i, Jdz^i). \tag{6.8.64}$$

By using (8.4) to change variables, the integral on the left side of (8.64) can be rewritten as

$$\int_{R_1^f}(z^f, Jdz^f) = \int_{R_1^i}(z^f, JMdz^i). \tag{6.8.65}$$

Now combine (8.64) and (8.65) to find the result

$$\int_{R_1^i} [(z^f, JMdz^i) - (z^i, Jdz^i)] = 0. \tag{6.8.66}$$

We know that a necessary and sufficient condition for this result to hold for any closed path R_1^i in phase space is that the differential form

$$(z^f, JMdz^i) - (z^i, Jdz^i) \tag{6.8.67}$$

be exact. But, in slightly different notation (identify z^i with z and z^f with Z), this is the differential form (6.4). And we know that a necessary and sufficient condition for this form to be exact is that $\mathcal{M}$ be a symplectic map. What we have found is that (8.64) or (8.66) holding for all closed paths is a necessary and sufficient condition for $\mathcal{M}$ to be a symplectic map.

We close this subsection with some comments. Recall the relation (6.30). There is also the simple result

$$d(\sum_j p_j q_j) = \sum_j (p_j dq_j + q_j dp_j). \tag{6.8.68}$$

Combining (6.30) and (8.68) gives the relation

$$\sum_j p_j dq_j = -[(z, Jdz) - d(\sum_j p_j q_j)]/2. \tag{6.8.69}$$

Since by definition the quantity $d(\sum_j p_j q_j)$ is an exact differential, there must be the result

$$\int_{R_1} d(\sum_j p_j q_j) = 0 \tag{6.8.70}$$

for any closed phase-space path R_1. Therefore from (8.69) and (8.70) we have the general relation

$$\int_{R_1} \sum_j p_j dq_j = -(1/2) \int_{R_1} (z, Jdz) = -(1/2) I_1 \tag{6.8.71}$$

for any closed phase-space path R_1. It follows that (8.64) can also be written as

$$\int_{R_1^f} \sum_j P_j dQ_j = \int_{R_1^i} \sum_j p_j dq_j. \tag{6.8.72}$$

This relation, although it does not appear to treat the coordinates and momenta on an equal footing, is still true whenever the Q, P and the q, p are related by a symplectic map $\mathcal{M}$, and frequently occurs in the literature. [An integral quantity of the form appearing on the left (or right) side of (8.72) is sometimes called a *circulation* because, if the q_i are regarded as the coordinates of a "position" vector $\mathbf{r}$ and the p_i are regarded as being proportional to the coordinates of a "velocity" vector $\mathbf{v}$, the integrand is of the form $\mathbf{v}{\cdot}d\mathbf{r}$.] Evidently

(8.72) holding for all closed paths is also a necessary and sufficient condition for $\mathcal{M}$ to be symplectic. Finally, we note that combining (8.60) and (8.71) gives the relation

$$\int_{R_1} \sum_j p_j dq_j = I_2 = \int_{R_2} (\delta z, Jdz) \tag{6.8.73}$$

for any phase-space surface R_2 whose boundary is the closed phase-space path R_1.[28]

6.8.5 Poincaré-Cartan Integral Invariant

Suppose we are given a family of symplectic maps $\mathcal{N}(t)$. Then we know there is an associated generating Hamiltonian $H(z,t)$. Recall Theorem 4.2. Alternatively, given a Hamiltonian, we know from Theorem 4.1 that it generates a family of symplectic maps. In this context, let C^i be a closed path in the associated $(2n+1)$ dimensional augmented phase space. See Figure 8.2. Specifically, for a parameter $\tau \in [0,1]$, we describe C^i by $(2n+1)$ relations of the form

$$z_a^i(\tau) = g_a(\tau), \tag{6.8.74}$$

$$t^i(\tau) = g_{2n+1}(\tau). \tag{6.8.75}$$

Note that different points of C^i may have different values of t.

View each point of C^i as an initial condition. For each point on C^i launch a trajectory governed by the Hamiltonian $H(z,t)$, and follow this trajectory to some final time t^f. Allow this time to vary from trajectory to trajectory by specifying yet one more relation of the form

$$t^f(\tau) = g_{2n+2}(\tau). \tag{6.8.76}$$

So doing produces a set of final conditions that constitutes another closed path C^f in augmented phase space. [Note that all the functions appearing on the right sides of (8.74) through (8.76) are assumed to be periodic in τ with period 1.] Put another way, C^i and C^f are any two augmented phase-space paths that surround a common bundle of phase-space trajectories produced by H.

For augmented phase space consider the differential form

$$(\sum_j p_j dq_j) - Hdt. \tag{6.8.77}$$

Then, according to Poincaré and Cartan, there is the path-integral relation

$$\int_{C^f} [(\sum_j p_j dq_j) - Hdt] = \int_{C^i} [(\sum_j p_j dq_j) - Hdt]. \tag{6.8.78}$$

Note that in the special case that t is constant on both C^i and C^f, (8.78) reduces to (8.72).

[28] We remark that sometimes the differential form $\sum_j p_j dq_j$ is called the *Liouville form*. By the same token, the differential form (8.45) could be called the *Poincaré form*. There does not seem to be any name for the differential form $\sum_j q_j dp_j$. Like the Liouville form in (8.72) and the Poincaré form in (8.64), it too is "invariant" under the action of symplectic maps. See Exercise 8.7.

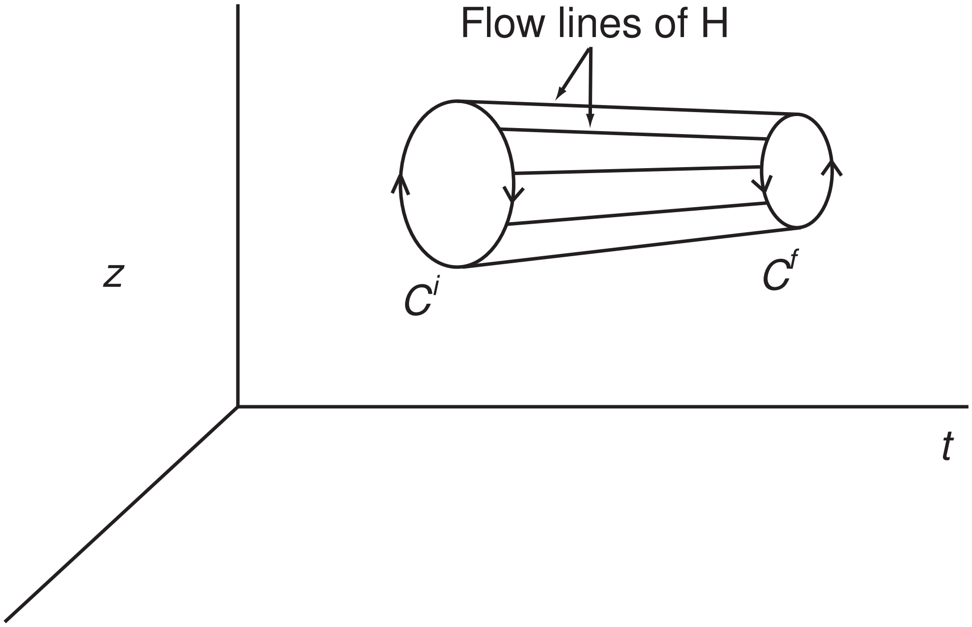


Figure 6.8.2: The closed paths C^i and C^f in augmented phase space and the trajectories that join them.

There are several ways to prove the Poincaré-Cartan relation. Our proof will use variational calculus. The trajectories originating on C^i and terminating on C^f form a two-dimensional surface in augmented phase space that is topologically equivalent to a cylinder. Indeed, points on this surface can be viewed as the image of a two-dimensional parameter space region described by τ and t with

$$\tau \in [0, 1], \tag{6.8.79}$$

$$t \in [g_{2n+1}(\tau), g_{2n+2}(\tau)], \tag{6.8.80}$$

and the understanding that the lines $\tau = 0$ and $\tau = 1$ are to be identified. Introduce for the integral on the right side of (8.78) the short-hand notation

$$\int_{C^i} **, \tag{6.8.81}$$

and similarly for the integral on the left side. Also, let

$$\int_{-C^f} ** \tag{6.8.82}$$

denote the integral on the left side of (8.78) with the path traversed in the opposite sense. With these understandings, (8.78) can be rewritten in the form

$$\int_{C^i} ** + \int_{-C^f} ** = 0. \tag{6.8.83}$$

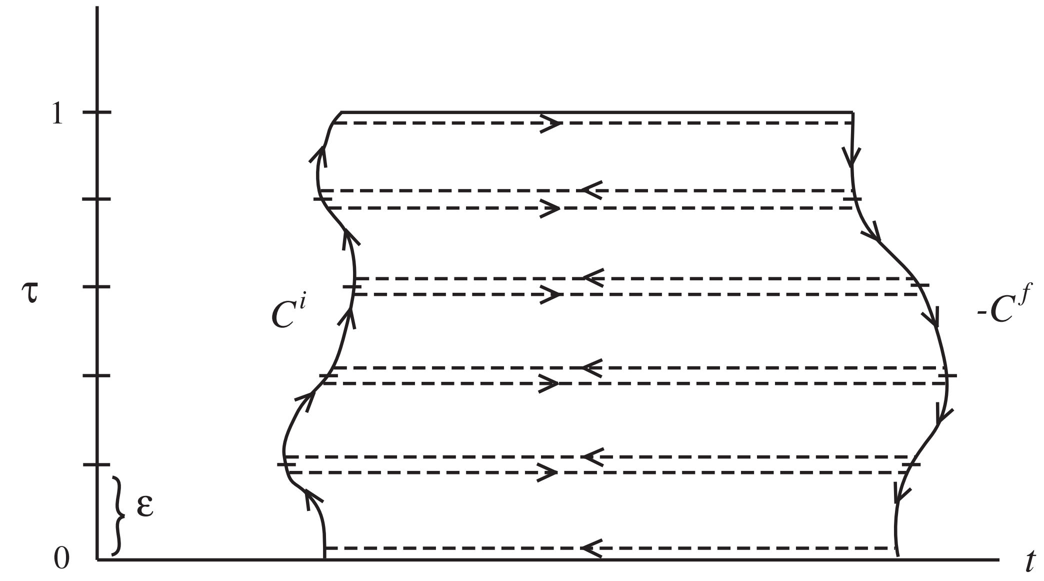


Figure 6.8.3: The t, τ parameter space. The left and right boundaries are the curves $t^i(\tau)$ and $t^f(\tau)$, and their augmented phase-space images are the paths C^i and C^f. Also shown as dashed lines are pairs of parameter-space paths traversed in opposite directions whose images are augmented phase-space trajectories traversed in opposite directions. Note that the lines $\tau = 0$ and $\tau = 1$ have the same image in augmented phase space.

The paths C^i and $-C^f$ are the images of the left and right boundaries of the parameter-space region. See Figure 8.3.

Divide the τ interval (8.79) into N equal pieces of size $\epsilon = 1/N$. For each subdivision consider pairs of parameter-space paths of constant τ traversed in opposite directions. See Figure 8.3. By construction, their images in augmented phase-space are trajectories for the Hamiltonian H traversed forward and backward in time. Imagine integrating the differential form (8.77) over these pairs of augmented phased-space trajectories. So doing will give a null net result because, by construction, the integrals so produced cancel in pairs. Add these self-canceling path integrals to those occurring in (8.83). Evidently the sum of integrals thus obtained can be reorganized into a sum of integrations over N thin loops ℓ_j,

$$\int_{C^i} ** + \int_{-C^f} ** + \text{canceling integral pairs } = \sum_{j=1}^{N} \int_{\ell_j} ** . \tag{6.8.84}$$

See Figures 8.4 and 8.5.

Now consider an individual loop. It can be viewed as a sum of top and bottom halves. See Figure 8.6. Each half can in turn be viewed as the result of deforming (in parameter space) a line of constant τ. See Figure 8.7. Note that the image of a line of constant τ in parameter space is a trajectory for the Hamiltonian H in augmented phase space.

Observe that by definition the sum of a path integral over a trajectory of H and its reverse, see Figure 8.7, cancel. It follows that the integral over any loop ℓ_j is the sum of the

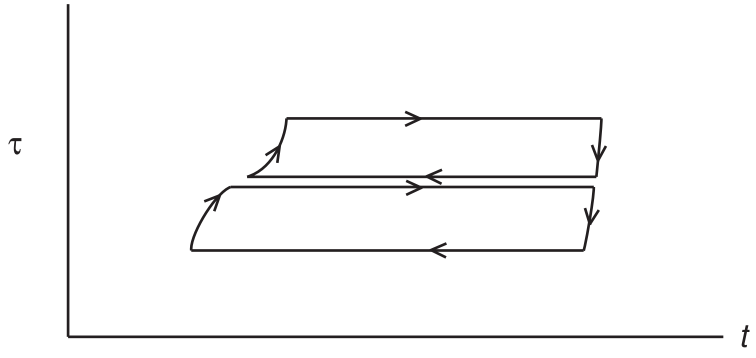


Figure 6.8.4: Two adjacent loops in parameter space.

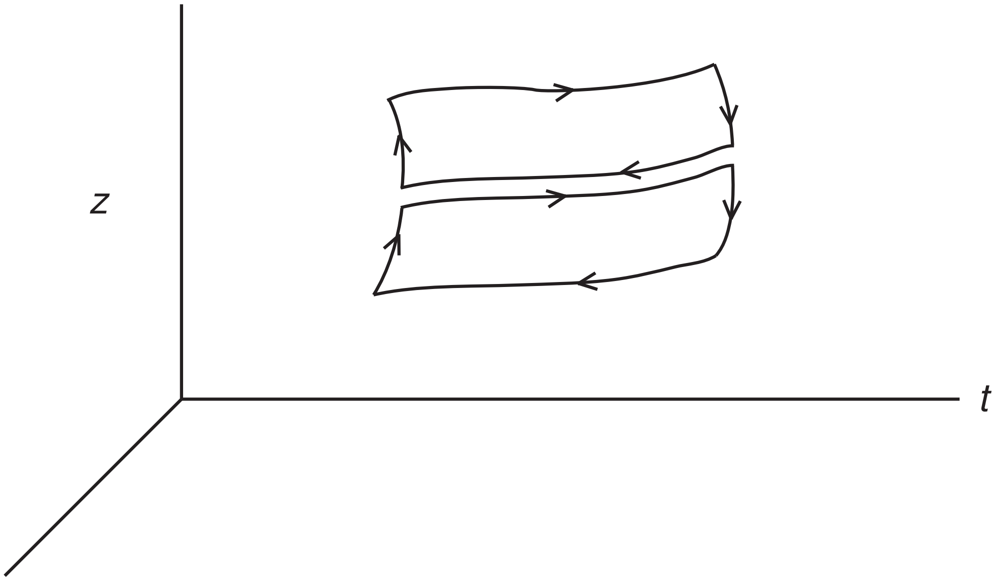


Figure 6.8.5: The loops in augmented phase space corresponding to the two parameter-space loops of Figure 8.4. Note that the long sides of the loops are trajectories for the Hamiltonian H, and the short sides are pieces of C^i and C^f.

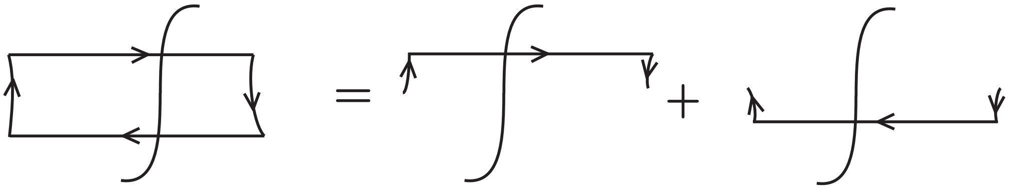


Figure 6.8.6: The integral over a loop is the sum of integrals over top and bottom halves.

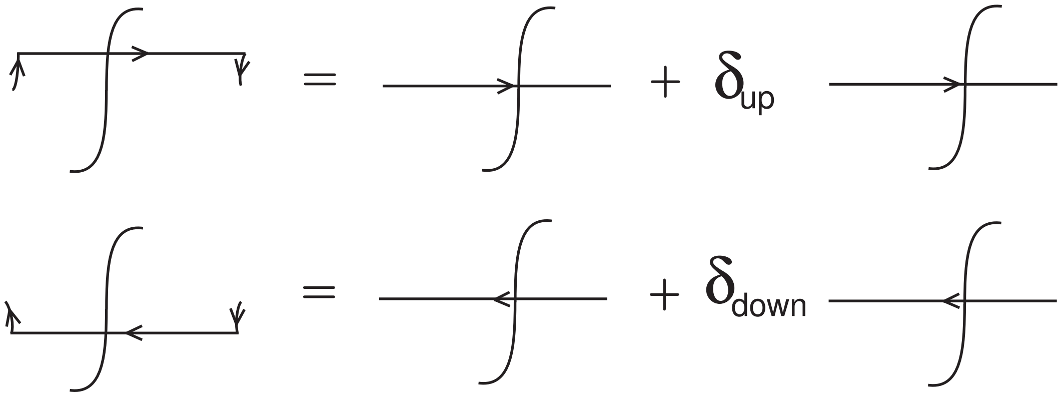


Figure 6.8.7: The integral over a half loop is the integral over a trajectory of H or its reverse plus the change in the integral resulting from deforming this path.

"up" and "down" variations about a trajectory of H. See Figures 8.6 and 8.7.

$$\int_{\ell_j} ** = \delta_{\text{up}} \int ** + \delta_{\text{down}} \int ** \,. \tag{6.8.85}$$

But, from Hamilton's (modified) principle, we know that the functional formed by integrating (8.77) over paths in augmented phase space has an extremum on paths that are trajectories of H. See Exercise 8.8 for details. Therefore each term on the right side of (8.85) vanishes through terms of order ϵ, and we have the result

$$\int_{\ell_j} ** = 0 + O(\epsilon^2). \tag{6.8.86}$$

Insert this result into (8.84) to find the relation

$$\int_{C^i} ** + \int_{-C^f} ** = 0 + O(N\epsilon^2). \tag{6.8.87}$$

Now let $N \to \infty$ and, correspondingly, $\epsilon \to 0$. Then $N\epsilon^2 \to 0$, and (8.87) becomes the desired relation (8.83) or (8.78).

Exercises

6.8.1. Suppose a "burst" of protons is injected into a uniform electric field $\boldsymbol{E} = E_0 \boldsymbol{e}_z$. Assume the burst is initially concentrated at x and $y = 0$ and v_x and $v_y = 0$, but is uniformly spread in z and v_z about the values $z = 0$ and $v_z = v_z^0$ within intervals $\pm\Delta z$ and $\pm\Delta v_z$. Thus the problem is essentially that of one-dimensional motion along the z axis. The initial distribution is shown schematically in Figure 8.8. Find the distribution at later times, and verify Liouville's theorem. Do not assume that Δz and Δv_z are infinitesimal. Neglect Coulomb interactions between particles.

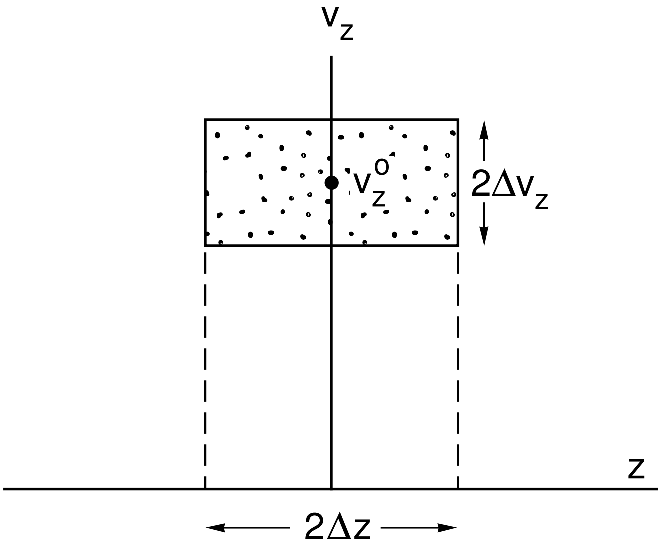


Figure 6.8.8: Initial phase-space distribution for Exercise 8.1.

6.8.2. Problem about Liouville's theorem and divergence theorem and how density transforms.

6.8.3. Exercise showing that, in the case of electromagnetic fields, Liouville's theorem also holds in terms of spatial coordinates and mechanical momenta.

6.8.4. Verify (8.8).

6.8.5. Construct a nonsymplectic but volume preserving family of maps $\mathcal{N}(\tau)$ that will send any symplectic camel through the eye of any needle.

6.8.6. Show that for elliptic camels there is the result

$$\text{Volume of } E^{2n}(r) = \text{Volume of } E^{2n}_{\text{nf}}(r,\lambda) = r^{2n}\pi^n/[(n!)(\lambda_1\lambda_2\lambda_3\cdots\lambda_n)]. \tag{6.8.88}$$

Show the impossibility of sending a symplectic cigar into a symplectic ball of the same volume using a symplectic map.

6.8.7. Show that (8.72) can also be written as

$$\int_{R_1^f}\sum_j Q_j dP_j = \int_{R_1^i}\sum_j q_j dp_j. \tag{6.8.89}$$

Consider the differential form

$$-[(z, Jdz) - d(\lambda\sum_j p_j q_j)]/2 \tag{6.8.90}$$

where λ is a parameter. Evaluate this form for the cases $\lambda = -1, 0, 1$. Show that it is invariant for all λ.

6.8.8. Refer to Exercise 6.2. Show that the Poincaré-Cartan relation (8.78) can also be written in the more democratic form

$$\int_{C^f}[(z,Jdz)/2+H(z,t)dt]=\int_{C^i}[(z,Jdz)/2+H(z,t)dt]. \tag{6.8.91}$$

6.8.9. The observant reader may object that, in deriving the Poincaré-Cartan invariant of Section 6.8.5, we invoked Hamilton's modified principle (1.6.11) and (1.6.12) in an unusual way because we employed paths in augmented phase space along which the time t may possibly both increase and decrease. So, some special explanation is required. To take into account the possibility of this more general case, suppose the path in (1.6.11) is parameterized by considering the q_i, the p_i, and t itself to be functions of some parameter σ where $\sigma\in[0,1]$. Introduce the notation

$$q_i'=dq_i/d\sigma,\ p_i'=dp_i/d\sigma,\ t'=dt/d\sigma. \tag{6.8.92}$$

Verify that (1.6.11) can be rewritten in the form

$$\mathcal{A}=\int_0^1 d\sigma A \tag{6.8.93}$$

where

$$A(q,q',p,p',t,t')=\sum_i p_iq_i'-Ht'. \tag{6.8.94}$$

Show, employing the usual variational calculus machinery, that the variation in $\mathcal{A}$ for fixed end points q,p,t is given by the relation

$$\begin{aligned}\delta\mathcal{A} &= \int_0^1 d\sigma\{\sum_i[(d/d\tau)(\partial A/\partial q_i')-\partial A/\partial q_i]\delta q_i+\sum_i[(d/d\tau)(\partial A/\partial p_i')-\partial A/\partial p_i]\delta p_i\\ &+ [(d/d\tau)(\partial A/\partial t')-\partial A/\partial t]\delta t\}+O(\epsilon^2).\end{aligned} \tag{6.8.95}$$

Next show that the various partial derivatives in (8.95) are given by the relations

$$\partial A/\partial q_i'=p_i, \qquad \partial A/\partial q_i=-t'(\partial H/\partial q_i), \tag{6.8.96}$$

$$\partial A/\partial p_i'=0, \qquad \partial A/\partial p_i=q_i'-t'(\partial H/\partial p_i), \tag{6.8.97}$$

$$\partial A/\partial t'=-H, \qquad \partial A/\partial t=-t'(\partial H/\partial t). \tag{6.8.98}$$

From these results, and Hamilton's equations of motion (1.5.11) augmented by (1.5.14), verify the following conclusions about the terms appearing in the integrand of (8.95):

$$[(d/d\tau)(\partial A/\partial q_i')-\partial A/\partial q_i]=dp_i/d\tau+t'(\partial H/\partial q_i)=p_i'-t'\dot{p}_i=0, \tag{6.8.99}$$

$$[(d/d\tau)(\partial A/\partial p_i')-\partial A/\partial p_i]=t'(\partial H/\partial p_i)-q_i'=t'\dot{q}_i-q_i'=0, \tag{6.8.100}$$

$$[(d/d\tau)(\partial A/\partial t')-\partial A/\partial t]=-dH/d\tau+t'(\partial H/\partial t)=-dH/d\tau+t'(dH/dt)=0. \tag{6.8.101}$$

We see that in the general case, as claimed, the variation in $\mathcal{A}$ about a trajectory is given by the relation

$$\delta\mathcal{A}=0+O(\epsilon^2). \tag{6.8.102}$$

6.9 Poincaré Surface of Section and Poincaré Return Maps

In Section 6.4.1 we saw that Hamiltonian flows between two times t^i and t^f generated symplectic maps. In this section we will study two generalizations of this result. The first, the *Poincaré surface of section map*, is related to the concept of using a coordinate as an independent variable. For an application of Poincaré surface of section maps see Section 21.7.2.

The second is related to long-term behavior. Recall that Section 1.4.3 illustrated how the determination of the long-term behavior of a periodically driven system could be reduced to the study of the behavior of a certain map, the stroboscopic map, under repeated iteration. For some Hamiltonian problems a similar simplification can be obtained by the use of a *Poincaré return map*. How this can be done will be a second generalization.

Poincaré maps may have other uses as well.

6.9.1 Poincaré Surface of Section Maps

Consider the case of conservative Hamiltonian flows in $2n$ dimensional phase space. That is, we assume $\partial H/\partial t = 0$. In this case we know that H is an integral of motion. Let g and h be two phase-space functions and let S^g and S^h be two $(2n-2)$ dimensional submanifolds in phase space defined by the equations

$$S^g: \; H(z) = \mathcal{E} \text{ and } g(z) = 0, \tag{6.9.1}$$

$$S^h: \; H(z) = \mathcal{E} \text{ and } h(z) = 0. \tag{6.9.2}$$

Note that each of the equations in (9.1) defines a $(2n-1)$ dimensional submanifold. For their intersection to define a $(2n-2)$ dimensional submanifold S^g, the gradients $\partial_z H$ and $\partial_z g$ must not be colinear. The analogous condition must also hold for S^h.

Next assume that S^g is *transverse* to the flow generated by H. What does this mean? Suppose z is some point in S^g. Then, we want z to leave S^g under both the foward and backward time evolution generated by H. Under time evolution the change in z is given by

$$dz = (J\partial_z H)dt. \tag{6.9.3}$$

In order for z to leave S^g, the quantity dz must have some component in at least one of the directions $\partial_z H$ and $\partial_z g$. Suppose we require that dz have some component in the direction of $\partial_z g$,

$$(\partial_z g, dz) \neq 0. \tag{6.9.4}$$

In view of (9.3), this requirement is equivalent to the condition

$$(\partial_z g, J\partial_z H) = [g, H] \neq 0. \tag{6.9.5}$$

We also observe that

$$(\partial_z H, dz) = (\partial_z H, J\partial_z H)dt = 0 \tag{6.9.6}$$

due to the antisymmetry of J. Therefore, (9.5) is a necessary and sufficient condition for z to leave S^g. Finally, we note that (9.5) guarantees that $\partial_z H$ and $\partial_z g$ cannot be colinear (proportional). Thus (9.5) is a necessary and sufficient condition both for S^g to be defined by (9.1) and for the flow to cross S^g. The surface S^g is said to be a *surface of section* for the flow generated by H.

Suppose S^h is also a surface of section. Suppose further that for some region R^g_{2n-2} of S^g the points $z \in R^g_{2n-2}$ have the property that their phase-space trajectories generated by H, when followed sufficiently forward in time, arrive at some region R^h_{2n-2} in S^h. Note that the interval of time required for this to occur may vary from trajectory to trajectory. Also, since H does not depend on the time, without loss of generality we may assume that all trajectories are launched at some common initial time $t = t^i$. See Figure 9.1. Then, by this operation, we have produced a mapping $\mathcal{M}$, called a *Poincaré surface of section map*, that sends R^g_{2n-2} to R^h_{2n-2}. Moreover, $\mathcal{M}$ is invertible since, given any point in R^h_{2n-2}, we can always follow trajectories backward in time until they reach R^g_{2n-2}.

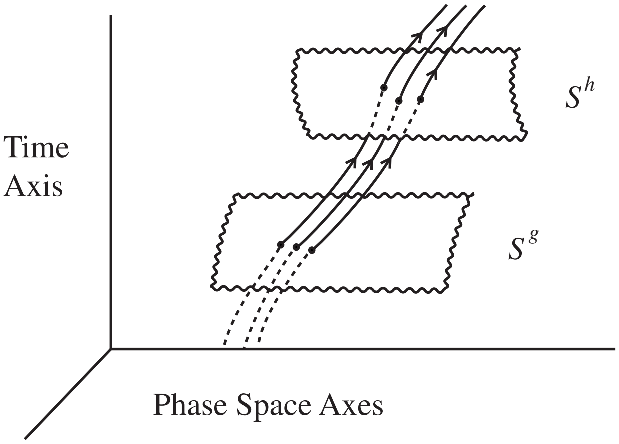


Figure 6.9.1: Two surfaces of section in augmented phase space. Trajectories leaving S^g are assumed to eventually enter and cross S^h, perhaps at different times.

Let R^g_1 be any closed path in R^g_{2n-2}, and let R^h_1 be its image in R^h_{2n-2} under the action of $\mathcal{M}$. Then, since R^g_1 and R^h_1 are related by following trajectories generated by H, the Poincaré-Cartan relation (8.78) takes form

$$\int_{R^g_1} [(\sum_j p_j dq_j) - H dt] = \int_{R^h_1} [(\sum_j p_j dq_j) - H dt]. \tag{6.9.7}$$

Since we assumed all trajectories on R^g_1 were launched at $t = t^i$, we have $dt = 0$ for the integral on the left side of (9.7). Therefore, we have the result

$$\int_{R^g_1} [(\sum_j p_j dq_j) - H dt] = \int_{R^g_1} \sum_j p_j dq_j. \tag{6.9.8}$$

Also, since all trajectories lie on the surface $H = \mathcal{E}$, see (9.1) and (9.2), we have the result

$$\int_{R^h_1} (-H) dt = -\mathcal{E} \int_{R^h_1} dt = 0 \tag{6.9.9}$$

because R_1^h is a closed curve.[29] Therefore we have the result

$$\int_{R_1^h}[(\sum_j p_j dq_j) - Hdt] = \int_{R_1^h}\sum_j p_j dq_j. \qquad (6.9.10)$$

It follows, for a Poincaré map, that

$$\int_{R_1^g}\sum_j p_j dq_j = \int_{R_1^h}\sum_j p_j dq_j. \qquad (6.9.11)$$

In some cases still more can be said. Suppose it can be arranged, perhaps by a suitable choice of variables, that $g(z)$ and $h(z)$ take the form

$$g(z) = 0 \rightarrow q_1 = \alpha, \qquad (6.9.12)$$

$$h(z) = 0 \rightarrow q_1 = \beta, \qquad (6.9.13)$$

when α and β are certain constants. Then we have $dq_1 = 0$ on R_1^g and R_1^h, and (9.11) becomes the relation

$$\int_{R_1^g}\sum_{j=2}^{n} p_j dq_j = \int_{R_1^h}\sum_{j=2}^{n} p_j dq_j. \qquad (6.9.14)$$

Let z be an initial condition in R_{2n-2}^g. We know the value of q_1 from (9.12). Suppose $q_2, p_2 \cdots q_n, p_n$ are selected to lie in R_{2n-2}^g. Then p_1 can be determined, perhaps up to a sign, from the condition $H = \mathcal{E}$. The sign ambiguity can be resolved by requiring that the trajectory launched from R_{2n-2}^g reach R_{2n-2}^h when traced forward in time. Thus, we may assume that points in R_{2n-2}^g (and R_{2n-2}^h) are described by the $(2n-2)$ coordinates $q_2, p_2 \cdots q_n, p_n$; and the Poincaré map $\mathcal{M}$ acts on this $(2n-2)$ dimensional space. Finally, from the results of Section 6.8.4, the relation (9.14) implies that the Poincaré map $\mathcal{M}$ is a symplectic map on this $(2n-2)$ dimensional space.

6.9.2 Poincaré Return Maps

Many Hamiltonian flows of physical interest have the property that they repeatedly re-enter some region of phase space. For example, in a *Penning* trap or a mirror machine, particles repeatedly return to some midplane region. In a circular accelerator or storage ring, particles repeatedly pass through any given beam-line element. In celestial and galactic dynamics, trajectories sufficiently close to a periodic trajectory nearly repeat themselves.

For such systems there are surfaces of section that are crossed repeatedly by a bundle of trajectories, and such a surface can be used to define a *Poincaré return* map. Let S^g be such a surface of section, and let R_{2n-2}^g be some region in S^g. For any point $z \in R_{2n-2}^g$ suppose the trajectory launched with these initial conditions returns to S^g. Then, by following these trajectories, we obtain a mapping of S^g onto itself,

$$\mathcal{M}: \; S^g \rightarrow S^g. \qquad (6.9.15)$$

[29] Note that we could have used the same argument to deduce (9.8) without the assumption that all points on S^g are launched with the same times t^i.

Moreover, like the case of a stroboscopic map, the long-term behavior of such a system can be found by studying the repeated action of $\mathcal{M}$. See (1.4.34). Finally, if coordinates can be selected so that (9.12) holds, the map $\mathcal{M}$ is symplectic.

Consider, for example, a circular accelerator or storage ring as shown schematically in Figure 1.2.6. At the point O we may introduce Cartesian coordinates as in Figure 1.6.1 so that particle trajectories repeatedly cross the plane $z = 0$ as they go around the ring. For the Hamiltonian we will use H^{eff} as given by (1.6.34). In particular near O we will employ the conjugate coordinate pairs (z, p_z), (x, p_x), (y, p_y), and (t, p_t), and the independent time-like variable τ. By construction H^{eff} is conserved, and we may restrict our attention to trajectories for which $H^{\text{eff}} = 0$, in which case (1.6.5) holds. Given the values of (z, p_z), (x, p_x), (y, p_y), and (t, p_t) in the plane $z = 0$, we can find p_z as in (1.6.6). Starting with these initial conditions, we follow a trajectory until it again crosses $z = 0$. In this way we find a mapping $\mathcal{M}$ of the surface of section into itself,

$$\mathcal{M}: \ (x, p_x), (y, p_y), (t, p_t) \to (\bar{x}, \bar{p}_x), (\bar{y}, \bar{p}_y), (\bar{t}, \bar{p}_t). \tag{6.9.16}$$

Moreover, we have the relation

$$\int_{R_1} (p_x dx + p_y dy + p_t dt) = \int_{\bar{R}_1} (p_x dx + p_y dy + p_t dt) \tag{6.9.17}$$

for any closed path R_1 in the phase-space surface $z = 0$ and its image $\bar{R}_1$ under the action of $\mathcal{M}$. Therefore, $\mathcal{M}$ is a symplectic map. Finally, determining the long-term behavior of trajectories in the ring is equivalent to determining the effect of the repeated action of $\mathcal{M}$ on points in the surface of section.

6.10 Overview and Preview

We have studied symplectic maps and have seen their intimate connection with Hamiltonian dynamics. Thus, a key goal is to be able to produce, manipulate, and apply symplectic maps.

We have also learned that symplectic maps can be produced using mixed-variable generating functions. However, while often useful, this method has some disadvantages. As we have seen, the relations between old and new variables are initially implicit, and must be made explicit. This fact makes it difficult to apply, multiply, and invert symplectic maps specified in terms of generating functions.

In subsequent chapters we will learn that symplectic maps can also be produced using Lie transformations. This approach has the advantage of being explicit. Moreover, we will develop tools for inverting, multiplying, and otherwise manipulating symplectic maps in Lie form. Finally, the use of Lie methods yields physical insight and facilitates high-order perturbation theory.

Bibliography

Legendre Transormations

(See also the results of Googling “Legendre transformation”.)

[1] V.I. Arnold, *Mathematical Methods of Classical Mechanics*, Second Edition, page 61, Springer Verlag (1989).

[2] R. K. P. Zia, E. F. Redish, and S. R. McKay, “Making Sense of the Legendre Transform”, *American Journal of Physics* **77**, 614-622 (2009).

Transformation (Generating) Functions

(See also the Variational Calculus section of the Bibliography for Chapter 1.)

[3] A. Wintner, *the Analytical Foundations of Celestial Mechanics*, Princeton University Press (1947).

[4] C. Carathéodory, *Calculus of Variations and Partial Differential Equations of the First Order, Parts I and II*, Holden-Day (1965).

[5] H. Goldstein, *Classical Mechanics*, Addison-Wesley (1980).

[6] J.V. Jose and E.J. Salatan, *Classical Dynamics: A Contemporary Approach*, Cambridge University Press (1998).

[7] H. Poincaré, *New Methods of Celestial Mechanics* (American Institute of Physics), New York, (1893/1993), Vol. 3, chap. 28, section 319.

[8] V.I. Arnold, *Mathematical Methods of Classical Mechanics*, Second Edition, Springer Verlag (1989).

[9] R. Abraham and J. Marsden, *Foundations of Mechanics*, American Mathematical Society (2008).

[10] S. H. Benton, *The Hamilton-Jacobi Equation: A Global Approach*, Academic Press (1977).

[11] Feng Kang, Wu Hua-mo, Qin Meng-shao, and Wang Dao-liu, “Construction of Canonical Difference Schemes for Hamiltonian Formalism via Generating Functions”, *Journal of Computational Mathematics* **11**, p. 71 (1989).

[12] Feng Kang, "The Calculus of Generating Functions and the Formal Energy for Hamiltonian Algorithms", *Journal of Computational Mathematics* **16**, p. 481 (1998).

[13] Feng Kang and Mengzhao Qin, *Symplectic Geometric Algorithms for Hamiltonian Systems*, Zhejiang Publishing and Springer-Verlag (2010).

[14] A. Weinstein, "Symplectic Manifolds and their Lagrangian Submanifolds", *Advances in Math.* **6**, 329 (1971).

[15] A. Weinstein, "The Invariance of Poincaré's Generating Function for Canonical Transformations", *Invent. Math.* **16**, 202 (1972).

[16] A. Weinstein, "Lagrangian Submanifolds and Hamiltonian Systems", *Ann. of Math.* **98**, 377 (1973).

[17] A. Weinstein, "Normal Modes for Nonlinear Hamiltonian Systems", *Invent. Math.* **20**, 47 (1973).

[18] A. Weinstein, *Lectures on symplectic manifolds*, CBMS Reg. Conf. Series in Math. 29, American Mathematical Society, Providence, RI (1977).

[19] A. Weinstein, "Symplectic Geometry", *Bulletin Amer. Math. Soc.* (N.S.) **5**, 1-13 (1981).

[20] J. Amiet and P. Huguenin, "Generating functions of canonical maps", *Helv. Phys. Acta* **53**, 377 (1980).

[21] A. Ozorio de Almeida, "On the Symplectically Invariant Variational Principle and Generating Functions", *Proc. R. Soc. Lond. A* **431**, 403 (1990).

[22] M. Sewell and I. Roulstone, "Anatomy of the Canonical Transformation", *Phil. Trans. R. Soc. Lond. A* **345**, 577 (1993).

[23] H. Gzyl, *Hamiltonian Flows and Evolution Semigroups*, Longman Group (1990).

[24] B. Erdélyi, "Symplectic Approximation of Hamiltonian Flows and Accurate Simulation of Fringe Field Effects", Michigan State University Physics and Astronomy Department Ph.D. Thesis (2001).

[25] B. Erdélyi and M. Berz, "Optimal Symplectic Approximation of Hamiltonian Flows", *Physical Review Letters* **87**, 114302 (2001).

[26] B. Erdélyi and M. Berz, "Local Theory and Applications of Extended Generating Functions", *International Journal of Pure and Applied Mathematics* **11**, 241 (2004).

[27] Alex Haro, "The Primitive Function of an Exact Symplectomorphism", *Nonlinearity* **13**, 1483-1500, (2000).

Symplectic Geometry and Topology

[28] M. Gromov, "Pseudo-holomorphic curves in symplectic manifolds", *Invent. Math.* **82**, 307-347, (1985).

[29] H. Hofer and E. Zehnder, *Symplectic Invariants and Hamiltonian Dynamics*, Birkhäuser (1994).

[30] D. McDuff and D. Salamon, *Introduction to Symplectic Topology*, Clarendon Press (1995).

[31] D. Salamon, Edit., *Symplectic Geometry*, London Mathematical Society Lecture Note Series 192, Cambridge University Press (1993).

[32] A. T. Fomenko, *Symplectic Geometry*, Gordon and Breach (1995).

[33] A. Crumeyrolle and J. Grifone, Edit., *Symplectic Geometry*, Pitman (1983).

[34] L. Polterovich, *The Geometry of the Group of Symplectic Diffeomorphisms*, Birkhäuser (2000).

[35] R. Berndt, *An Introduction to Symplectic Geometry*, American Mathematical Society (2001).

[36] A. Banyaga, *The Structure of Classical Diffeomorphism Groups*, Kluwer Academic Publishers (1997).

[37] B. Khesin and R. Wendt, *The Geometry of Infinite-Dimensional Groups*, Springer (2009).

[38] V.I. Arnold, "Symplectic geometry and topology", *J. of Math. Phys.* **41**, 6, 3307 (2000).

[39] V.I. Arnold, *Dynamical Systems IV, Symplectic Geometry and its Applications*, Springer-Verlag (1990).

[40] M.A. de Gosson, *The principles of Newtonian and quantum mechanics: the need for Planck's constant, h*, Imperial College Press, London (2001).

[41] M.A. de Gosson, "The symplectic camel and phase space quantization", *J. Phys. A: Math. Gen.* **34** p. 10085 (2001).

[42] M.A. de Gosson, "The 'symplectic camel principle' and semiclassical mechanics", *J. Phys A: Math. Gen.* **35**, p. 6825 (2002).

[43] M.A. de Gosson, "Symplectically covariant Schrödinger equation in phase space", *J. Phys A: Math. Gen.* **38**, p. 9263 (2005).

[44] M.A. de Gosson, "Uncertainty Principle, Phase-Space Ellipsoids, and Weyl Calculus", *Operator Theory: Advances and Applications*, Vol. 164, p. 121, Birkhäuser Verlag (2006).

[45] M.A. de Gosson, *Symplectic geometry and quantum mechanics*, Birkhäuser Verlag (2006).

[46] M.A. de Gosson and F. Luef, "Symplectic capacities and the geometry of uncertainty: The irruption of symplectic topology in classical and quantum mechanics", *Physics Reports* 484, 131-179, Elsevier (2009).

[47] F. Schlenk, *Embedding Problems in Symplectic Geometry*, de Gruyter Expositions in Mathematics, vol. 40, Berlin (2005).

[48] L. Traynor, Book Review of *Embedding Problems in Symplectic Geometry* by F. Schlenk, *Bulletin of the American Mathematical Society* **43**, p. 593 (2006).

[49] Y. Eliashberg and L. Traynor, Editors, *Symplectic Geometry and Topology*, American Mathematical Society (1999).

[50] K. Cieliebak, H. Hofer, J. Latschev, and F. Schlenk, "Quantitative symplectic geometry", 22 May 2006, arXiv:math.SG/0506191 v1 10 June 2005.

[51] A. Cannas da Silva, *Lectures on Symplectic Geometry*, Corrected second printing, Springer Verlag (2008).

[52] Y. Long, *Index Theory for Symplectic Paths with Applications*, Birkhäuser (2002).

[53] V. Guillmen and S. Sternberg, *Symplectic Techniques in Physics*, Cambridge (1984).

[54] V. Guillmen and S. Sternberg, *Geometric Asymptotics*, American Mathematical Society (1977).

[55] P. Libermann and C-M. Marle, *Symplectic Geometry and Analytical Mechanics*, D. Reidel (1987).

Differential Manifolds and Forms

[56] H.M. Edwards, *Advanced Calculus: A Differential Forms Approach*, Birkhäuser (1994).

[57] H. Flanders, *Differential Forms with Applications to the Physical Sciences*, Academic Press (1963).

[58] B. Schultz, *Geometrical Methods of Mathematical Physics*, Cambridge University Press (1980).

[59] M. Spivak, *A Comprehensive Introduction to Differential Geometry*, Vols. 1-5, Publish or Perish (1999).

[60] M. Schreiber, *Differential Forms, A Heuristic Introduction*, Springer-Verlag (1977).

[61] R. Courant and F. John, *Introduction to Calculus and Analysis*, Vol. I, Vol. II/1, Vol. II/2, Springer-Verlag (1998, 1999, 2000).

[62] R. Abraham, J. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications*, Springer-Verlag (1988).

[63] J. Marsden and T. Ratiu, *Introduction to Mechanics and Symmetry*, Second Edition, Springer (1999).

[64] J. Marsden, R. Montgomery, and T. Ratiu, *Reduction, Symmetry , and Phases in Mechanics*, American Mathematical Society (1990).

[65] C. J. Isham, *Modern Differential Geometry for Physicists*, Second Edition, World Scientific (1999).

[66] T. Frankel, *The Geometry of Physics*, Second Edition, Cambridge University Press (2004).

[67] S. Lang, *Fundamentals of Differential Geometry*, Springer-Verlag (1999).

[68] W. Rudin, *Principles of Mathematical Analysis*, Third Edition, Mc-Graw-Hill (1976).

[69] J. Dieudonné, *Treatise on Analysis*, Volumes 10-III and 10-IV in the series Pure and Applied Mathematics, Academic Press (1972 and 1974).

[70] P. Iglesias-Zemmour, *Diffeology*, (2010), `http://math.huji.ac.il/~piz/documents/Diffeology.pdf`.

Poincaré Maps

[71] 3-body problem

[72] Galactic Dynamics

[73] A.J. Dragt, "Trapped Orbits in a Magnetic Dipole Field", *Rev. Geophys.* **3**, p. 255-298 (1965).

[74] A.J. Dragt and J.M. Finn, "Insolubility of Trapped Particle Motion in a Magnetic Dipole Field", *J. Geophys. Res.* **81** p. 2327-2340 (1976).

[75] A.J. Dragt and J.M. Finn, "Normal Form for Mirror Machine Hamiltonians", *J. Math. Physics* **20**, p. 2649-2660 (1979).

The Galilean Group and Group Contraction/Deformation

[76] Theodore Jacobson provided the ideas for Exercise 6.2.5.

[77] E. Inönü and E.P. Wigner, "On the contraction of groups and their representations", *Proc. Nat. Acad. Sci. U.S.A.* **39**, 510-524 (1953).

[78] J. Lõhmus, E. Paal, and L. Sorgsepp, *Nonassociative Algebras in Physics*, Hadronic Press (1994).

Chapter 7

Lie Transformations and Symplectic Maps

Chapter 6 showed that there is an intimate connection between symplectic maps and Hamiltonian flows, and showed how symplectic maps could be produced (in implicit form) with the aid of mixed-variable generating functions. This chapter explores how Lie transformations can be used for the same purpose, and how their use produces symplectic maps in explicit form. It also displays how the group of all symplectic maps is a Lie group whose Lie algebra is the Poisson bracket Lie algebra of all phase-space functions.

7.1 Production of Symplectic Maps

Let $f(z,t)$ be any dynamical variable, and let $\exp(:f(z,t):)$ be the Lie transformation associated with f. (Here, as in Section 6.1, the time t simply plays the role of a parameter.) This Lie transformation can be used to define a map $\mathcal{M}$ that produces new variables $\overline{z}(z,t)$ by the rule

$$\overline{z}_a(z,t) = \exp(:f(z,t):)z_a \quad , \quad a = 1, 2, \cdots 2n. \tag{7.1.1}$$

The relations (1.1) can also be expressed more compactly by writing

$$\overline{z} = \mathcal{M}z, \tag{7.1.2}$$

$$\mathcal{M} = \exp(:f:). \tag{7.1.3}$$

Note that in writing (1.1) we have indicated explicitly the arguments of f. Generally these arguments will be omitted for simplicity of notation. However, it is always important to keep in mind what these arguments are, and they should and will always be stated explicitly whenever there is any possibility for confusion.

Consider the Poisson brackets of the various $\overline{z}$'s with each other. Using the definition (1.1), the isomorphism condition (5.4.14), and (5.4.22), we find the result

$$\begin{aligned}
[\overline{z}_a, \overline{z}_b]_z &= [\exp(:f:)z_a, \exp(:f:)z_b]_z \\
&= \exp(:f:)[z_a, z_b]_z \\
&= \exp(:f:)J_{ab} = J_{ab}.
\end{aligned} \tag{7.1.4}$$

It follows from (1.4) that $\mathcal{M}$ is a symplectic map! What has been shown is that every Lie transformation may be viewed as a symplectic map. Consequently, Lie transformations produce an endless supply of symplectic maps. And, unlike the case for mixed-variable generating functions (see Section 6.5.1), these maps are immediately in explicit form. Finally, we see from (5.4.15) and its generalization that products of Lie transformations also produce symplectic maps.

Consider the map $\mathcal{M}(\lambda)$ depending on the parameter λ and defined by the relation

$$\mathcal{M}(\lambda) = \exp(\lambda :f:). \tag{7.1.5}$$

This map produces the transformation

$$\begin{aligned} \bar{z}(z,t;\lambda) &= \exp(\lambda :f:)z \\ &= z + \lambda :f: z + (\lambda^2/2!) :f:^2 z + \cdots . \end{aligned} \tag{7.1.6}$$

Evidently we have the relations

$$\mathcal{M}(0) = \mathcal{I} = \text{identity map}, \tag{7.1.7}$$

$$\mathcal{M}(1) = \mathcal{M}. \tag{7.1.8}$$

Next let $\mathcal{M}(\lambda_1)$ and $\mathcal{M}(\lambda_2)$ be two maps of the form (1.5) corresponding to the λ values λ_1 and λ_2, respectively. Consider the product map given by the relation

$$\mathcal{M}(\lambda_1)\mathcal{M}(\lambda_2) = \exp(\lambda_1 :f:)\exp(\lambda_2 :f:). \tag{7.1.9}$$

Because Lie operators are linear operators, their behavior is in many ways analogous to that of matrices. In particular, observe that we may attempt to combine the exponents appearing on the right side of (1.9) into a single exponent using the Baker-Campbell-Hausdorff (BCH) series. See (3.7.33) and (3.7.34). According to (5.3.14), the Lie operators $\lambda_1 :f:$ and $\lambda_2 :f:$ commute. Consequently, the exponents in (1.9) simply add to give the result

$$\begin{aligned} \mathcal{M}(\lambda_1)\mathcal{M}(\lambda_2) &= \exp(\lambda_1 :f:)\exp(\lambda_2 :f:) \\ &= \exp(\lambda_1 :f: + \lambda_2 :f:) \\ &= \exp((\lambda_1+\lambda_2) :f:) \\ &= \mathcal{M}(\lambda_1+\lambda_2). \end{aligned} \tag{7.1.10}$$

Section 6.2 showed that the set of all symplectic maps forms a group. The relation (1.10) shows that the subset of symplectic maps given by (1.5) forms a one-parameter subgroup of symplectic maps. Moreover (1.10), together with (1.7) and (1.8), shows that any Lie transformation lies on a one-parameter subgroup of symplectic maps. That is, any Lie transformation is continuously connected to the identity map $\mathcal{I}$ by a path whose points are all elements of some common subgroup of symplectic maps.

We have seen that Lie transformations produce symplectic maps that act on the phase-space variables z. According to (5.4.11), Lie transformations also act on general functions. Let $g^{\text{old}}(z,t)$ be any function of the phase-space variables z and perhaps the time t. Then the Lie transformation (1.3), using (5.4.11), produces a *new* function $g^{\text{new}}(z,t)$ according to the rule

$$\begin{aligned} g^{\text{new}}(z,t) &= \mathcal{M}g^{\text{old}}(z,t) = \exp(:f:)g^{\text{old}}(z,t) \\ &= g^{\text{old}}(\exp(:f:)z,t) = g^{\text{old}}(\overline{z}(z,t),t) \\ &= g^{\text{old}}(\mathcal{M}z,t). \end{aligned} \tag{7.1.11}$$

Note that the relation (1.11) is analogous to (6.3.6).

The symplectic map defined by (1.3) has the particular property that f is an *invariant function* for the map. That is, there is the relation

$$f(\overline{z},t) = f(z,t)\ , \quad \text{or} \quad f^{\text{new}}(z,t) = f^{\text{old}}(z,t). \tag{7.1.12}$$

To see the truth of this assertion, apply (5.4.11) to the case where $g = f$. We find, using the notation of (1.1), the result

$$\exp(:f:)f(z,t) = f(\overline{z},t). \tag{7.1.13}$$

However, using the expression (5.4.2), we also obtain the result

$$\begin{aligned} \exp(:f:)f(z,t) &= f + [f,f] + [f[f,f]]/2! + \cdots \\ &= f(z,t) \end{aligned} \tag{7.1.14}$$

since the Poisson bracket $[f,f]$ is zero by the antisymmetry condition. Comparison of (1.13) and (1.14) shows that (1.12) is indeed correct. Note again that in all these calculations, the time t plays no essential role and may be regarded simply as a parameter.

Suppose the symplectic map $\exp(-:f(z,t):)$ is applied to both sides of (1.1). We find the result

$$\exp(-:f:)\overline{z}_a = \exp(-:f:)\exp(:f:)z_a. \tag{7.1.15}$$

Consider first the problem of evaluating the right side of (1.15). Observe that the Lie operators $:f:$ and $-:f:$ commute. Consequently, the exponents on the right side of (1.15) can be added to give the result

$$\exp(-:f:)\exp(:f:) = \exp(:0:) = \mathcal{I}. \tag{7.1.16}$$

Correspondingly, when read from right to left, (1.15) may be rewritten in the form

$$z_a = \exp(-:f:)\overline{z}_a. \tag{7.1.17}$$

The right side of (1.17), which is the left side of (1.15), can be viewed in two ways. First, both f and the $\overline{z}_a$ can be regarded as functions of z (and perhaps the time t), and all indicated Poisson brackets are to be taken with respect to the variables z. The result of these operations is simply to produce the functions z_a as indicated by the left side of (1.17). Alternatively, f may be viewed as a function of $\overline{z}$ by writing the relation

$$f(z,t) = f^*(\overline{z},t) = f(z(\overline{z},t),t). \tag{7.1.18}$$

See (6.3.3) and (6.3.5). Then, thanks to the preservation of Poisson brackets under symplectic maps as expressed by (6.3.11) and (6.3.21), the relation (1.17) can be written in the form

$$z_a(\overline{z}, t) = \exp(- : f^*(\overline{z}, t) :)\overline{z}_a \tag{7.1.19}$$

where now all Poisson brackets are to be taken with respect to the variables $\overline{z}$. However, the invariance condition (1.12) can be written in the form

$$f^*(\overline{z}, t) = f(z, t) = f(\overline{z}, t). \tag{7.1.20}$$

Consequently, (1.17) can also be written in the final form

$$z_a(\overline{z}, t) = \exp(- : f(\overline{z}, t) :)\overline{z}_a \tag{7.1.21}$$

where all Poisson brackets are to be taken with respect to the variables $\overline{z}$. What has been shown is that if $\mathcal{M}$ is given by the relations (1.1) through (1.3), then, when due regard is taken for the variables involved, the inverse relation (1.21) can be written in the compact form

$$z = \mathcal{M}^{-1}\overline{z} \tag{7.1.22}$$

with

$$\mathcal{M}^{-1} = \exp(- : f :). \tag{7.1.23}$$

Exercises

7.1.1. Verify in detail the steps leading from (1.15) to (1.23).

7.1.2. Suppose f and g are two phase-space functions in involution. That is,

$$[f, g] = 0. \tag{7.1.24}$$

Show from the power series definition (5.4.1) that in this case there is the relation

$$\exp(: f :)\exp(: g :) = \exp(: f + g :). \tag{7.1.25}$$

See Exercise 3.7.11.

7.1.3. Consider the map of Exercise 5.4.6 written in the form

$$\overline{q}(q, p) = \exp(: f :)q = q(1 - \lambda p)^2, \tag{7.1.26}$$

$$\overline{p}(q, p) = \exp(: f :)p = p/(1 - \lambda p). \tag{7.1.27}$$

Verify by direct computation that $[\overline{q}, \overline{p}]_z = 1$. Verify that $f = \lambda qp^2$ is an invariant function, that is $f(\overline{q}, \overline{p}) = f(q, p)$. Solve (1.26) and (1.27) for $q(\overline{q}, \overline{p}), p(\overline{q}, \overline{p})$. Verify by direct computation that $[q, p]_{\overline{z}} = 1$. Verify directly that $\mathcal{M}^{-1}$ is given by (1.23).

7.1.4. Repeat Exercise 1.3 for the f of Exercise 5.4.5.

7.2 Realization of the Group $Sp(2n)$ and Its Subgroups

7.2.1 Realization of General Group Element

Let $\mathcal{M}$ be the map given by the relation

$$\mathcal{M}: \ z \rightarrow \overline{z} = Mz, \tag{7.2.1}$$

where M is a symplectic matrix. Then, according to Exercise 6.2.1, $\mathcal{M}$ is a symplectic map. Since M is a symplectic matrix, it can be written in the form

$$M = PO = \exp(JS^a)\exp(JS^c). \tag{7.2.2}$$

See (3.8.1) and (3.8.24).

Define a quadratic polynomial f_2^a in terms of the matrix S^a appearing in the decomposition (2.2) by the relation

$$f_2^a = -(1/2)\sum_{de} S^a_{de} z_d z_e. \tag{7.2.3}$$

Now consider the Lie operator $: f_2^a :$. Suppose this Lie operator acts on the various z's. We find the result

$$\begin{aligned} : f_2^a : z_b &= -(1/2)\sum_{de} S^a_{de}[z_d z_e, z_b] \\ &= -(1/2)\sum_{de} S^a_{de}\{[z_d, z_b]z_e + [z_e, z_b]z_d\} \\ &= -(1/2)\sum_{de} S^a_{de}\{J_{db}z_e + J_{eb}z_d\} \\ &= \sum_d (JS^a)_{bd} z_d. \end{aligned} \tag{7.2.4}$$

Here use has been made of the antisymmetry of J and the symmetry of S^a. Using matrix and vector notation, (2.4) can also be written in the more compact form

$$: f_2^a : z = (JS^a)z. \tag{7.2.5}$$

From this form it is easy to see that there is the general relation

$$: f_2^a :^m z = (JS^a)^m z. \tag{7.2.6}$$

Finally, it follows from (2.6) that we also have the relation

$$\exp(: f_2^a :)z = \exp(JS^a)z = Pz. \tag{7.2.7}$$

In a similar way, define a quadratic polynomial f_2^c in terms of the matrix S^c appearing in (2.2),

$$f_2^c = -(1/2)\sum_{de} S^c_{de} z_d z_e. \tag{7.2.8}$$

Correspondingly, we have the relation

$$\exp(:f_2^c:)z = \exp(JS^c)z = Oz. \tag{7.2.9}$$

Now define a symplectic map $\mathcal{M}$ by the relation

$$\mathcal{M} = \exp(:f_2^c:)\exp(:f_2^a:). \tag{7.2.10}$$

Here $\mathcal{M}$ is intended to act on the phase-space variables z, and both f_2^a and f_2^c are functions of z. We find, using (2.7) and (2.9), the result

$$\begin{aligned} \mathcal{M}z_b &= \exp(:f_2^c:)\exp(:f_2^a:)z_b \\ &= \exp(:f_2^c:)\sum_d P_{bd}z_d \\ &= \sum_d P_{bd}\exp(:f_2^c:)z_d \\ &= \sum_{de} P_{bd}O_{de}z_e = (Mz)_b. \end{aligned} \tag{7.2.11}$$

This result may be written in the more compact form

$$\bar{z} = \mathcal{M}z = Mz. \tag{7.2.12}$$

Notice two things. First, we have shown that any linear symplectic transformation of the form (2.1) can be realized as the product of two Lie transformations. Second, comparison of (2.2) and (2.10) shows that the corresponding factors appear in opposite order. That is, when Lie transformations all involve the *same* phase-space variables, they act from *left to right*. This particular feature of Lie transformations will be explored in greater detail in Section 8.3. There it will also be explained why the difference in sign between relations such as (5.5.1) and (2.3) is not arbitrary. The reader will soon come to realize that Lie transformations lead lives of their own, and possess many unexpected properties.

7.2.2 Realization of Various Subgroups

We next employ Lie transformations to study various aspects of subgroups of $Sp(2n)$. We begin with symplectic matrices of the form (3.3.9). As shown in Section (3.10), such matrices are related to matrices S of the form (3.10.2). Correspondingly, let f_2^B be the quadratic polynomial given by the relation

$$f_2^B = -(1/2)\sum_{de} S_{de}z_dz_e = -(1/2)\sum_{jk} B_{jk}p_jp_k. \tag{7.2.13}$$

Then we have the relation

$$\bar{z} = \exp(:f_2^B:)z = Mz, \tag{7.2.14}$$

with M given by (3.3.9) or (3.10.5). We have learned that the subgroup of symplectic matrices of the form (3.3.9) is produced by Lie transformations whose Lie operators arise from monomials of the form p_jp_k. Evidently, monomials of this form are mutually in involution. See (5.5.14). Correspondingly, the subgroup is Abelian, as has already been seen earlier.

In a similar fashion, it is easily checked that the subgroup of symplectic matrices of the form (3.3.10) is generated by the Lie operator $: f_2^C :$ given by the relation

$$f_2^C = -(1/2)\sum_{de} S_{de} z_d z_e = (1/2)\sum_{jk} C_{jk} q_j q_k \tag{7.2.15}$$

with S given by (3.10.7). That is, we have the relation

$$\overline{z} = \exp(: f_2^C :) z = Mz \tag{7.2.16}$$

with M given by (3.3.10). We have learned that Lie transformations arising from monomials of the form q_jq_k produce the subgroup of symplectic matrices of the form (3.3.10). Monomials of the form q_jq_k are also mutually in involution, and the corresponding subgroup is again Abelian. See (5.5.15).

Next consider the subgroup of matrices of the form (3.3.11). Let f_2 be the quadratic polynomial defined by the relation

$$\begin{aligned} f_2 &= -(1/2)\sum_{de} S_{de} z_d z_e = -(1/2)(z, Sz) \\ &= -(1/2)[(q, a^T p) + (p, aq)] = -(q, a^T p) \\ &= -\sum_{jk} a^T_{jk} q_j p_k. \end{aligned} \tag{7.2.17}$$

Here S is given by (3.10.13). Then, for matrices M of the form (3.3.11) and sufficiently near the identity, we have the relation

$$\overline{z} = \exp(: f_2 :) z = Mz \tag{7.2.18}$$

with f_2 given by (2.17). We have learned that Lie transformations arising from monomials of the form q_jp_k produce symplectic matrices of the form (3.3.11). It is easily verified that the set of monomials of the form q_jp_k forms a Lie algebra under the Poisson bracket operation. See (5.5.16). Correspondingly, matrices of the form (3.3.11) constitute a group. As we saw in Section 3.10, this group is $GL(n, \mathbb{R})$.

As a special case of (2.18), consider the Lie transformation $\exp(: -\lambda q_\ell p_\ell :)$ where ℓ is some integer satisfying $0 \le \ell \le n$, and λ is a parameter. Then for $j \ne \ell$ we have the relations

$$\begin{aligned} \overline{q}_j &= \exp(: -\lambda q_\ell p_\ell :) q_j = q_j, \\ \overline{p}_j &= \exp(: -\lambda q_\ell p_\ell :) p_j = p_j. \end{aligned} \tag{7.2.19}$$

And for $j = \ell$ we find the result

$$\begin{aligned} \overline{q}_j &= \exp(: -\lambda q_\ell p_\ell :) q_\ell = (e^{\lambda}) q_\ell, \\ \overline{p}_j &= \exp(: -\lambda q_\ell p_\ell :) p_\ell = (e^{-\lambda}) p_\ell. \end{aligned} \tag{7.2.20}$$

See Exercise 5.4.4. We conclude that $\exp(:-\lambda q_\ell p_\ell:)$ *scales* q_ℓ and p_ℓ by the (positive) factors e^λ and $e^{-\lambda}$, respectively, and leaves the remaining q_j and p_j untouched.

Consider next Lie transformations corresponding to the quadratic polynomials f_2^ℓ given by the definition

$$f_2^\ell = -(1/2)\theta_\ell(q_\ell^2 + p_\ell^2). \tag{7.2.21}$$

Then for $j \neq \ell$ we have the relations

$$\bar{q}_j = \exp(:f_2^\ell:)q_j = q_j,$$

$$\bar{p}_j = \exp(:f_2^\ell:)p_j = p_j. \tag{7.2.22}$$

And for $j = \ell$ we find the results

$$\bar{q}_\ell = \exp(:f_2^\ell:)q_\ell = q_\ell\cos\theta_\ell + p_\ell\sin\theta_\ell,$$

$$\bar{p}_\ell = \exp(:f_2^\ell:)p_\ell = -q_\ell\sin\theta_\ell + p_\ell\cos\theta_\ell. \tag{7.2.23}$$

See Exercise 5.4.5. We conclude that in this case $\exp(:f_2^\ell:)$ produces a *rotation* by angle θ_ℓ in the q_ℓ, p_ℓ plane. Because these two variables are conjugate, such a rotation is sometimes referred to as a *phase advance*.

At this point we remark that there is a correspondence between phase advances and the maximal $Sp(2n,\mathbb{R})$ torus described at the end of Section 3.9. From (2.22) and (2.23) we find the result

$$\exp(:f_2^1 + f_2^2 + \cdots + f_2^n:)z_a = \sum_b [N(\theta_1, \theta_2, \cdots \theta_n)]_{ab} z_b \tag{7.2.24}$$

where N is given by (3.8.85), or (3.5.60) and (3.5.61). Here we have used the ordering (3.2.4).

Finally, consider Lie transformations corresponding to the quadratic polynomials f_2^{jk} given by the definition

$$f_2^{jk} = \theta_{jk}(q_j p_k - q_k p_j). \tag{7.2.25}$$

It is easily verified that the set of such polynomials is closed under the Poisson bracket operation, and thus constitutes a Lie algebra. Furthermore, for $\ell \neq j, k$ we have the evident result

$$\bar{q}_\ell = \exp(:f_2^{jk}:)q_\ell = q_\ell,$$

$$\bar{p}_\ell = \exp(:f_2^{jk}:)p_\ell = p_\ell. \tag{7.2.26}$$

Also, explicit calculation gives the results

$$\bar{q}_j = \exp(:f_2^{jk}:)q_j = q_j\cos\theta_{jk} + q_k\sin\theta_{jk},$$

$$\bar{q}_k = \exp(:f_2^{jk}:)q_k = -q_j\sin\theta_{jk} + q_k\cos\theta_{jk},$$

$$\bar{p}_j = \exp(:f_2^{jk}:)p_j = p_j\cos\theta_{jk} + p_k\sin\theta_{jk},$$

$$\bar{p}_k = \exp(:f_2^{jk}:)p_k = -p_j\sin\theta_{jk} + p_k\cos\theta_{jk} \tag{7.2.27}$$

We conclude that in this case exp(: f_2^{jk} :) produces a rotation by angle θ_{jk} in the q_j, q_k plane, and simultaneously, the same rotation in the p_j, p_k plane. These rotations provide a realization of the special orthogonal group $SO(n, \mathbb{R})$. See Exercise 2.5.

Finally, it can be verified that the f_2^{ℓ} given by (2.21) and the f_2^{jk} given by (2.25) all correspond to matrices S^c that commute with J. Consequently, the transformations given by (2.22), (2.23) and (2.26), (2.27) are all in the $U(n)$ subgroup of $Sp(2n)$.

7.2.3 Another Proof of Transitive Action of $Sp(2n)$ on Phase Space

Near the end of Section 3.6 we showed in effect that if $\tilde{z}^i$ and $\tilde{z}^f$ are *any* two points in phase space *distinct* from the origin, then there is a *linear* symplectic map of the form (2.1) such that

$$\tilde{z}^f = M\tilde{z}^i. \tag{7.2.28}$$

See (3.6.115). To recapitulate, with the exception of the origin, any point in phase space can be sent into any other point by a linear symplectic transformation. (The origin is obviously sent into itself.) Following the terminology of Section 5.12, we say that, with the exception of the origin, $Sp(2n)$ acts *transitively* on phase space.

We will now provide another proof of this result using a series of constructive steps that have some instructive merit. First, suppose that $\tilde{z}^i$ is not the origin. Perform successive phase advances of the form (2.22), (2.23) to remove all "p" type components from $\tilde{z}^i$. Next perform a rotation of the form (2.26), (2.27) in the $(n-1), n$ plane to remove any q_n component. In so doing, no p_n component is produced. Thus both p_n and q_n components have been removed. Next perform a rotation in the $(n-2), (n-1)$ plane to remove any q_{n-1} component, etc. The net result of a sequence of such rotations is that all components have been transformed to zero save for the q_1 component. Also, this component cannot be zero, because transformations of the form (2.22), (2.23) and (2.26), (2.27) evidently preserve the inner product (z, z), and this quantity cannot vanish if $\tilde{z}^i$ is not the origin. Moreover, the q_1 component can be taken to be positive. [If it is not, simply increase θ_{12} by π. See (2.27).] Finally, apply a scaling transformation of the form (2.19), (2.20) with $\ell = 1$ to transform the q_1 component so that it has the numerical value 1. Since all the transformations just described are linear symplectic maps, and linear symplectic maps form a group, it follows that there is a symplectic matrix M^i such that

$$M^i\tilde{z}^i = z^1. \tag{7.2.29}$$

Here z^1 is a vector (phase-space point) whose q_1 component is 1, and all others are zero,

$$z^1_a = \delta_{a1}. \tag{7.2.30}$$

By an analogous argument, there is also a symplectic matrix M^f such that

$$M^f\tilde{z}^f = z^1. \tag{7.2.31}$$

Upon combining (2.30) and (2.28), we get the result

$$\tilde{z}^f = (M^f)^{-1}z^1 = (M^f)^{-1}M^i\tilde{z}^i. \tag{7.2.32}$$

That is, the advertised result (2.27) is correct with M given by the relation

$$M = (M^f)^{-1}M^i. \tag{7.2.33}$$

Note that M as given by (2.33) is again symplectic as a consequence of the group property.

Introduce the term *punctured phase space* to refer to the set of all points in phase space with the exception of the origin. We have learned that $Sp(2n, \mathbb{R})$ acts transitively on punctured phase space. Consequently, according to the discussion in Section 5.12, punctured phase space is a homogeneous space with respect to $Sp(2n, \mathbb{R})$, and therefore must be a coset space of $Sp(2n, \mathbb{R})$ with respect to one of its subgroups. What is this subgroup? See Exercise 7.4 in Section 7.7.

Exercises

7.2.1. Verify the relations (2.4) through (2.7).

7.2.2. Verify (2.17) and (2.18).

7.2.3. Verify (2.19) and (2.20).

7.2.4. Verify (2.22) and (2.23).

7.2.5. The orthogonal group $O(n, \mathbb{R})$ is defined by the set of real $n \times n$ matrices satisfying (5.10.13). Show that such matrices do indeed form a group. See Exercise (3.7.24). Show that (5.10.13) implies the relation

$$\det O = \pm 1.$$

Orthogonal matrices with determinant +1 are called *proper*. Show that proper $O(n, \mathbb{R})$ matrices form a subgroup of $O(n, \mathbb{R})$. Recall that this subgroup is $SO(n, \mathbb{R})$, the special orthogonal group. Show that the set of $O(n, \mathbb{R})$ matrices with determinant -1 (called *improper* orthogonal) does not form a subgroup, and is disconnected from $SO(n, \mathbb{R})$. Show that any matrix of the form (3.3.11) with

$$A = D = O, \tag{7.2.34}$$

and O orthogonal, is symplectic. Recall Exercise 6.5.2.

If O is special (proper) real orthogonal, then it can be written in the form

$$O = \exp(F) \tag{7.2.35}$$

where F is $n \times n$, real, and antisymmetric,

$$F^T = -F. \tag{7.2.36}$$

Show that M as given by (3.3.11) and (2.34) has the form

$$M = \exp(JS^c) \tag{7.2.37}$$

where JS^c is the matrix

$$JS^c = \begin{pmatrix} F & 0 \\ 0 & F \end{pmatrix}. \tag{7.2.38}$$

Show that S^c is given by the relation

$$S^c = \begin{pmatrix} 0 & -F \\ F & 0 \end{pmatrix}. \tag{7.2.39}$$

Show that S^c is symmetric and, as the notation suggests, commutes with J. See (3.9.6). Use (2.8) and (2.39) to derive the result

$$\begin{aligned} f_2^c &= -(1/2)(z, S^c z) = (q, Fp) \\ &= \sum_{jk} F_{jk} q_j p_k \\ &= (1/2) \sum_{jk} F_{jk}(q_j p_k - q_k p_j). \end{aligned} \tag{7.2.40}$$

7.2.6. Show that the set of polynomials of the form (2.25) or (2.40) constitutes a Lie algebra under the Poisson bracket operation. This Lie algebra is $so(n, \mathbb{R})$, the Lie algebra of $SO(n, \mathbb{R})$.

7.2.7. Verify (2.26) and (2.27).

7.2.8. Verify that the transformations (2.22), (2.23) and (2.26), (2.27) preserve the inner product (w, z).

7.2.9. Show that for any (square) matrix G there is the identity

$$\exp(G) = \cosh G + \sinh G. \tag{7.2.41}$$

Using (3.1.3) and the series expansion for cosh and sinh, verify the relation

$$\exp(\lambda J) = I \cos\lambda + J \sin\lambda. \tag{7.2.42}$$

Find quadratic polynomials f_2 such that $\mathcal{M}$ given by $\mathcal{M} = \exp(:f_2:)$ satisfies (2.1) with $M = \pm J$ and $M = \pm I$.

7.2.10. Review Exercises 6.2.6 and 6.2.7. There we learned that Lorentz transformations are symplectic maps, and in fact are linear symplectic maps. Therefore, based on the work of this section, we suspect that they can be written as Lie transformations. Lorentz transformations consist of rotations about the x, y, z axes and velocity transformations (sometimes called *boosts*) along these axes. (We remark that the factorization of Lorentz transformations into rotations and boosts arises naturally in a polar decomposition of the Lorentz group.) The relations (2.27) show that rotations can be written as Lie transformations. We want to show that the same is true of boosts. For simplicity we will consider boosts along the z axis. Boosts along the other axes, and in arbitrary directions, can be treated analogously.

Verify that the quantities β, γ defined by (6.2.54) and (6.2.55) satisfy the relation

$$\gamma^2 - \gamma^2\beta^2 = 1. \tag{7.2.43}$$

Therefore, we can define a quantity χ called the *rapidity* such that

$$\sinh\chi = \beta\gamma, \tag{7.2.44}$$

$$\cosh\chi = \gamma. \tag{7.2.45}$$

With this notation, show that (6.2.58) and (6.2.59), and their momentum counterparts, can be written on the form

$$\tilde{x}^3 = x^3\cosh\chi + x^4\sinh\chi, \tag{7.2.46}$$

$$\tilde{x}^4 = x^3\sinh\chi + x^4\cosh\chi, \tag{7.2.47}$$

$$\tilde{p}^3 = p^3\cosh\chi + p^4\sinh\chi, \tag{7.2.48}$$

$$\tilde{p}^4 = p^3\sinh\chi + p^4\cosh\chi. \tag{7.2.49}$$

Using the metric tensor $\bar{g}$ given by (1.6.75), show that (2.48) and (2.49) can be rewritten as

$$\tilde{p}_3 = p_3\cosh\chi - p_4\sinh\chi, \tag{7.2.50}$$

$$\tilde{p}_4 = -p_3\sinh\chi + p_4\cosh\chi. \tag{7.2.51}$$

Let f_2 be the quadratic polynomial defined by the relation

$$f_2 = -\chi(x^3p_4 + x^4p_3). \tag{7.2.52}$$

Verify the relations

$$:f_2:x^3 = \chi x^4, \tag{7.2.53}$$

$$:f_2:x^4 = \chi x^3, \tag{7.2.54}$$

$$:f_2:p_3 = -\chi p_4, \tag{7.2.55}$$

$$:f_2:p_4 = -\chi p_3. \tag{7.2.56}$$

Finally, by summing the relevant infinite series, show that

$$\tilde{x}^3 = \exp(:f_2:)\,x^3, \tag{7.2.57}$$

$$\tilde{x}^4 = \exp(:f_2:)\,x^4, \tag{7.2.58}$$

$$\tilde{p}_3 = \exp(:f_2:)\,p_3, \tag{7.2.59}$$

$$\tilde{p}_4 = \exp(:f_2:)\,p_4. \tag{7.2.60}$$

7.2.11. We have seen that, apart from the origin, $Sp(2n)$ acts transitively on the $2n$-dimensional Euclidean space E^{2n}. Does $O(2n,\mathbb{R})$ act transitively on E^{2n}?

7.2.12. Consider the 3-dimensional *isotropic* harmonic oscillator described by the Hamiltonian

$$H = \sum_1^3 p_j^2/(2m) + (k/2)q_j^2. \tag{7.2.61}$$

Show that there is a linear canonical transformation that brings H to the form

$$H = (\omega/2) \sum_1^3 (p_j^2 + q_j^2). \tag{7.2.62}$$

See Exercises (5.4.4) and (6.4.3). Consider the set of all linear canonical transformations that leaves H invariant. Show that these transformations form a group isomorphic to $U(3)$. See Section (5.8). Show that there is an even larger group of linear and nonlinear canonical transformations that leaves H invariant.

7.2.13. Equation (5.11.39) provides the partial Iwasawa decomposition for any element in the group $Sp(2n, \mathbb{R})$. The purpose of this exercise is to find the corresponding decomposition of the Lie algebra $sp(2n, \mathbb{R})$. From (5.11.39) we see that we must study the elements in the Lie algebra associated with $M(Z)$ given by (5.11.18), and the elements in the Lie algebra associated with $M(m)$ given by (3.9.19). The case of $M(m)$ has already been discussed, and is realized in terms of Lie transformations by symplectic maps of the form $\exp(: f_2^c :)$. Thus, the associated elements in the Lie algebra $sp(2n, \mathbb{R})$ are the polynomials f_2^c when the Poisson bracket realization is used, the Lie operators $: f_2^c :$ when the Lie operator realization is used, and the matrices of the form JS^c when the matrix realization is used. We now turn to the case of $M(Z)$. According to (5.11.18) it can be written as the product of two factors. Consider first the second factor. It can be written in the form (5.11.41). Show that matrices of this form are equivalent to those given by (3.3.9). That is, show that any real symmetric B can be written in the form

$$B = Y^{-1/2} X Y^{-1/2} \tag{7.2.63}$$

with X real symmetric, and Y real symmetric and positive definite. Thus, this case has already been treated in the discussion surrounding (2.13) and (2.14). Finally, consider the first factor in (4.11.18). It can be written in the form given by (5.11.43) and (5.11.44). Show that this case is a special case of (3.3.11) with A symmetric. Do symplectic matrices of the form (3.3.11) with A symmetric form a subgroup? According to Exercise 5.11.9, $\log(Y)$ is real and symmetric. Thus, show that this case is a special case of that treated in the discussion surrounding (2.17). Specifically, show in this case that the matrix a appearing in (2.17) is real and symmetric. Consider the Lie algebra $sp(2, \mathbb{R})$. According to Section 5.6, it has a Poisson bracket realization in terms of the basis polynomials b^0, f, and g. See (5.6.6), (5.6.11), and (5.6.12). Show that the partial Iwasawa basis for $sp(2, \mathbb{R})$ is given by the polynomials b^0, $p^2 = b^0 - f$, and $qp = g$. Show that the partial Iwasawa basis for the quadratic polynomial realization of $sp(4, \mathbb{R})$ is given by the polynomials b^0,b^1, b^2, b^3; $p_1^2 = (1/2)(b^0 + b^3 + f^1 - g^2)$, $p_1 p_2 = (1/2)(b^1 - f^3)$, $p_2^2 = (1/2)(b^0 - b^3 - f^1 - g^2)$; $q_1 p_1 = -(1/2)(f^2 + g^1)$, $q_2 p_2 = (1/2)(g^1 - f^2)$, $q_1 p_2 + q_2 p_1 = g^3$. See Section 5.7.

7.2.14. The *center* of a group G consists of those elements of G that commute with all elements of G. Show that the center of a group forms a subgroup of G. Show that the center of $Sp(2n,\mathbb{R})$ consists of the elements $\pm I$. What is the center of $Sp(2n,\mathbb{C})$? What is the center of $U(n)$? What is the center of $SU(n)$? What is the center of $O(n,\mathbb{R})$?

7.2.15. Refer to Exercise 4.5.4. Show that static symplectic matrices form a group. Show that this group is generated by quadratic polynomials f_2 that obey $\partial f_2/\partial t = 0$.

7.3 Invariant Scalar Product

In Section 5.8, in the context of describing representations of $su(3)$, the need for a suitable scalar product was mentioned. In this section we will introduce a particularly convenient scalar product. The choice of scalar product is not unique, and the whole matter is discussed in greater detail in Appendix G.

7.3.1 Definition of Scalar Product

For simplicity, we will treat the case of $sp(6)$. From this treatment it will be easy to read off the results for the general case $sp(2n)$. Let $G(\mu;\nu)$ denote the general monomial defined by the relation

$$G(\mu;\nu) = (\mu_1!\nu_1!\mu_2!\nu_2!\mu_3!\nu_3!)^{-1/2}p_1^{\mu_1}q_1^{\nu_1}p_2^{\mu_2}q_2^{\nu_2}p_3^{\mu_3}q_3^{\nu_3}. \tag{7.3.1}$$

It is evident that the $G(\mu;\nu)$ form a basis for the set of all phase-space functions.

For reasons that will become clear shortly, let us pause to consider the Lie operators associated with the quadratic polynomials $q_1^2, p_1^2, q_1q_2, p_1p_2$,
q_1p_1, and q_1p_2. Explicit calculation gives the relations

$$:q_1^2:G(\mu_1\mu_2\mu_3;\nu_1\nu_2\nu_3) = 2\sqrt{\mu_1(\nu_1+1)}G(\mu_1-1,\mu_2,\mu_3;\nu_1+1,\nu_2,\nu_3), \tag{7.3.2}$$

$$:p_1^2:G(\mu_1\mu_2\mu_3;\nu_1\nu_2\nu_3) = -2\sqrt{\nu_1(\mu_1+1)}G(\mu_1+1,\mu_2,\mu_3;\nu_1-1,\nu_2,\nu_3), \tag{7.3.3}$$

$$\begin{aligned}:q_1q_2:G(\mu_1\mu_2\mu_3;\nu_1\nu_2\nu_3) &= \sqrt{\mu_1(\nu_2+1)}G(\mu_1-1,\mu_2,\mu_3;\nu_1,\nu_2+1,\nu_3)\\ &+ \sqrt{\mu_2(\nu_1+1)}G(\mu_1,\mu_2-1,\mu_3;\nu_1+1,\nu_2,\nu_3),\end{aligned} \tag{7.3.4}$$

$$\begin{aligned}:p_1p_2:G(\mu_1\mu_2\mu_3;\nu_1\nu_2\nu_3) &= -\sqrt{\nu_1(\mu_2+1)}G(\mu_1,\mu_2+1,\mu_3;\nu_1-1,\nu_2,\nu_3)\\ &- \sqrt{\nu_2(\mu_1+1)}G(\mu_1+1,\mu_2,\mu_3;\nu_1,\nu_2-1,\nu_3),\end{aligned} \tag{7.3.5}$$

$$\begin{aligned}:q_1p_2:G(\mu_1\mu_2\mu_3;\nu_1\nu_2\nu_3) &= \sqrt{\mu_1(\mu_2+1)}G(\mu_1-1,\mu_2+1,\mu_3;\nu_1\nu_2\nu_3)\\ &- \sqrt{\nu_2(\nu_1+1)}G(\mu_1\mu_2\mu_3;\nu_1+1,\nu_2-1,\nu_3).\end{aligned} \tag{7.3.6}$$

$$:q_1p_1:G(\mu_1\mu_2\mu_3;\nu_1\nu_2\nu_3) = (\mu_1-\nu_1)G(\mu_1\mu_2\mu_3;\nu_1\nu_2\nu_3). \tag{7.3.7}$$

With this detour behind us, define a scalar product among the basis elements $G(\mu;\nu)$ by the rule

$$\langle G(\mu';\nu'), G(\mu;\nu)\rangle = \delta_{\mu'\mu}\delta_{\nu'\nu}. \tag{7.3.8}$$

Here we use the short-hand notation

$$\delta_{\mu'\mu} = \delta_{\mu_1'\mu_1}\delta_{\mu_2'\mu_2}\delta_{\mu_3'\mu_3}\ , \text{ etc.} \tag{7.3.9}$$

That is, the basis elements are defined to be an orthonormal set. Note that although the notation is the same, this scalar product is not to be confused with that introduced for phase-space vectors in Section 3.5.

It is easily verified that the rule (3.8) induces a positive-definite scalar product among the set of all phase-space functions. Let f and g be any two (possibly complex) functions. Make the expansions

$$f = \sum_{\mu\nu} f_{\mu\nu} p_1^{\mu_1} q_1^{\nu_1} p_2^{\mu_2} q_2^{\nu_2} p_3^{\mu_3} q_3^{\nu_3}, \tag{7.3.10}$$

$$g = \sum_{\mu\nu} g_{\mu\nu} p_1^{\mu_1} q_1^{\nu_1} p_2^{\mu_2} q_2^{\nu_2} p_3^{\mu_3} q_3^{\nu_3}. \tag{7.3.11}$$

Then we have the relation

$$\langle f, g\rangle = \sum_{\mu\nu} \overline{f}_{\mu\nu} g_{\mu\nu} \mu_1!\nu_1!\mu_2!\nu_2!\mu_3!\nu_3!. \tag{7.3.12}$$

Examination of (3.12) shows that there is another equivalent way of defining the scalar product. Let ∂_z denote the set of partial differentiation operators, $\partial_z = (\partial/\partial z_1, \partial/\partial z_2, \cdots \partial/\partial z_6)$. Then we also have the result

$$\langle f, g\rangle = \overline{f}(\partial_z) g(z)|_{z=0} = g(\partial_z)\overline{f}(z)|_{z=0}. \tag{7.3.13}$$

There is a corollary that will be of later use. Let h be any phase-space function. Then, from (3.13), we find the result

$$\langle hf, g\rangle = \langle f, \overline{h}(\partial_z) g\rangle. \tag{7.3.14}$$

We close this subsection with the remark that in the definition of the scalar product given by (3.1) and (3.8) it was convenient to treat the q"s and $p's$ separately. For a somewhat different notation that treats them on the same footing, see Exercise 3.23.

7.3.2 Definition of Hermitian Conjugate

Given a scalar product and any linear operator O, the Hermitian conjugate $O^\dagger$ is defined by the relation

$$\langle f, O^\dagger g\rangle = \langle Of, g\rangle. \tag{7.3.15}$$

The virtue of the scalar product (3.8) is that the Lie operators associated with quadratic polynomials have particularly simple Hermitian conjugates. From the relations (3.2) through (3.7) and their generalizations to all $z_a z_b$ pairs, and the definition (3.15), we find the pleasing results

$$: q_j q_k :^\dagger = - : p_j p_k :, \tag{7.3.16}$$

$$: p_j p_k :^\dagger = - : q_j q_k :, \tag{7.3.17}$$

$$: q_j p_k :^\dagger =: q_k p_j : . \tag{7.3.18}$$

Indeed, let $\mathcal{L}_{ab}$ be any vector field of the form

$$\mathcal{L}_{ab} = z_a(\partial/\partial z_b). \tag{7.3.19}$$

See Section 5.3. Then we find the result

$$(\mathcal{L}_{ab})^\dagger = \mathcal{L}_{ba}. \tag{7.3.20}$$

Consider the quadratic polynomials b^0 through b^8 defined by (5.8.5). It is easily verified from their definitions, and the relations (3.16) through (3.18), that their associated Lie operators are anti-Hermitian,

$$: b^j :^\dagger = - : b^j : . \tag{7.3.21}$$

Also, from (5.8.50), (5.8.51), and (3.21), we have the relations

$$: c^j :^\dagger =: c^j :, \tag{7.3.22}$$

$$: r(\boldsymbol{\mu}) :^\dagger =: r(-\boldsymbol{\mu}) : . \tag{7.3.23}$$

Thus the $: c^j :$ are Hermitian as desired for the construction of a representation theory. Finally, the Lie operators associated with the quadratic polynomials f^j and g^j given by (5.8.63) are Hermitian,

$$: f^j :^\dagger =: f^j :, \tag{7.3.24}$$

$$: g^j :^\dagger =: g^j : . \tag{7.3.25}$$

Suppose f_2^c is a real quadratic polynomial defined in terms of a real matrix S^c as in (2.8). Then, in the case of a 6-dimensional phase space, such an f_2^c can be written as a linear combination of the polynomials b^0 through b^8, with real coefficients. Correspondingly, the Lie operator associated with f_2^c is anti-Hermitian,

$$: f_2^c :^\dagger = - : f_2^c : . \tag{7.3.26}$$

Let $\mathcal{M}$ be the symplectic map associated with f_2^c,

$$\mathcal{M} = \exp(: f_2^c :). \tag{7.3.27}$$

Then, from (3.26), we find the result

$$\mathcal{M}^\dagger = \exp(: f_2^c :^\dagger) = \exp(- : f_2^c :) = \mathcal{M}^{-1}. \tag{7.3.28}$$

It follows that $\mathcal{M}$ is *unitary* with respect to the scalar product (3.12). That is, we have the relation

$$\langle \mathcal{M}f, \mathcal{M}g \rangle = \langle f, \mathcal{M}^\dagger \mathcal{M}g \rangle = \langle f, g \rangle. \tag{7.3.29}$$

We already know that Lie transformations of the form (3.27) are a realization of the group $U(3)$. From this perspective, the relation (3.28) indicates that the scalar product defined by (3.12) is *invariant* under $U(3)$.

Remarkably, the scalar product (3.12) is in fact invariant under the full group $USp(6)$. Suppose f_2^a is a real quadratic polynomial defined in terms of a real matrix S^a as in (2.3). Then it is easily verified from (5.8.43), (3.24), and (3.25) that the Lie operator $: f_2^a :$ is Hermitian,

$$: f_2^a :^\dagger =: f_2^a : . \tag{7.3.30}$$

Let $\mathcal{M}$ be any (complex) symplectic map of the form

$$\mathcal{M} = \exp(: f_2^c :) \exp(i : f_2^a :). \tag{7.3.31}$$

Then, from (3.26) and (3.30), we find the result

$$\begin{aligned} \mathcal{M}^\dagger &= \exp(-i : f_2^a :^\dagger) \exp(: f_2^c :^\dagger) \\ &= \exp(-i : f_2^a :) \exp(- : f_2^c :) \\ &= \mathcal{M}^{-1}. \end{aligned} \tag{7.3.32}$$

It follows as before that $\mathcal{M}$ is unitary with respect to the scalar product (3.12). Moreover, we know from Section 5.10 and (2.10) that symplectic maps of the form (3.31) are a realization of the group $USp(6)$. Thus, the scalar product (3.12) is invariant under $USp(6)$.

One might wonder if it is possible to define a scalar product that would be invariant under all of the *real* symplectic group $Sp(6, \mathbb{R})$. The answer is no. Let $\mathcal{M}$ be the symplectic map associated with the monomial $\lambda q_1 p_1$ with λ real,

$$\mathcal{M} = \exp(\lambda : q_1 p_1 :). \tag{7.3.33}$$

From (3.7) we have the result

$$\mathcal{M} G(\mu; \nu) = \exp[\lambda(\mu_1 - \nu_1)] G(\mu; \nu). \tag{7.3.34}$$

It follows that for any definition of the scalar product, there is the relation

$$\langle \mathcal{M} G(\mu; \nu), \mathcal{M} G(\mu; \nu) \rangle = \exp[2\lambda(\mu_1 - \nu_1)] \langle G(\mu; \nu), G(\mu; \nu) \rangle. \tag{7.3.35}$$

We conclude that if the scalar product is such that the elements $G(\mu; \nu)$ are normalizable, then this scalar product cannot be invariant under all of $Sp(6, \mathbb{R})$. What we are observing here is a consequence of the fact that the group $Sp(6, \mathbb{R})$ is not compact. It can be shown that a noncompact group cannot have finite-dimensional unitary representations. Note that any $\mathcal{M}$ of the form (2.10), when acting on a homogeneous polynomial, preserves the degree of that polynomial. See Lemma 7.6.3. Consequently, any realization of $Sp(6, \mathbb{R})$ associated with a polynomial basis must be finite dimensional, and thus, by the comment above, cannot be unitary. Finally, we remark that $U(3)$ is the largest compact subgroup of $Sp(6, \mathbb{R})$, and therefore is the largest subgroup for which we can hope to obtain finite-dimensional unitary representations.

While the relations (3.2) through (3.7) are fresh in the mind, we take this opportunity to observe that any Lie operator of the form $: f_2 :$ is traceless. Correspondingly, according to (3.7.56), any map $\mathcal{M}$ of the form

$$\mathcal{M} = \exp(: f_2 :) \tag{7.3.36}$$

has determinant $+1$. These statements may seem somewhat surprising since both $: f_2 :$ and $\mathcal{M}$ given by (3.36) may be viewed as infinite dimensional matrices in the sense that they are both linear operators that act on infinite dimensional vector spaces. We therefore have to be more precise.

We have already noted that any $\mathcal{M}$ of the form (2.10) and therefore also of the form (3.36), when acting on a homogeneous polynomial, preserves the degree of that polynomial. We capitalize on this fact by slightly changing the notation for the monomials $G(\mu,\nu)$ defined by (3.1). Specifically, we denote the same monomials by the symbols G_r^m where m now denotes the degree of the monomial, and r is some index that labels the various possibilities for the exponents μ_i and ν_j subject to their sum being equal to m,

$$\sum(\mu_i+\nu_i)=m, \tag{7.3.37}$$

$$r=\{\mu_i,\nu_j\}. \tag{7.3.38}$$

With this notation, the scalar product (3.8) takes the form

$$\langle G_{r'}^{m'},G_r^m\rangle=\delta_{m'm}\delta_{r'r}. \tag{7.3.39}$$

Let us pause for a moment to make the selection of the index r more specific. Consider all monomials of degree m in d variables. Let $N(m,d)$ be the number of such monomials. Combinatoric considerations (see Exercises 3.9 through 3.13) show that N is given by the relation

$$N(m,d)=\left(\begin{array}{c} m+d-1 \\ m \end{array}\right)=\frac{(m+d-1)!}{m!(d-1)!}. \tag{7.3.40}$$

Table 3.1 below shows values of $N(m,d)$ for various values of m for the case of 6 dimensional phase space ($d=6$), and for other values of d that may be of interest later. Consequently, for $d=6$ and each value of m, we may take for the index r the integers running from $r=1$ through $r=N(m,6)$. More sophisticated indexing schemes are described in Section 27.2.

7.3.3 Matrices Associated with Quadratic Lie Generators

We now return to our main discussion. As is evident from (3.2) through (3.7), any quantity of the form $:f_2:G_r^m$ must be a homogeneous polynomial of degree m. See also Lemma 7.6.3. Thus, we must have a result of the form

$$:f_2:G_r^m=\sum_{r'}F_{r'r}^m G_{r'}^m \tag{7.3.41}$$

where the $F_{r'r}^m$ are coefficients yet to be determined. In fact, from (3.39) and (3.41) we have the result

$$F_{r'r}^m=\langle G_{r'}^m,:f_2:G_r^m\rangle. \tag{7.3.42}$$

Let $\mathcal{P}_m$ denote the space of all homogeneous polynomials of degree m. We know its dimension is $N(m,6)$. What we have made explicit is that the general $:f_2:$ sends $\mathcal{P}_m$ into itself. Indeed, the action of $:f_2:$ on $\mathcal{P}_m$ for each value of m is described by the $N(m,6)\times N(m,6)$ matrix F^m given by (3.41) and (3.42). Let us compute the matrices corresponding to powers of $:f_2:$. From (3.41) we find the result

$$\begin{aligned} :f_2:^2 G_r^m &= \sum_{r'}F_{r'r}^m :f_2:G_{r'}^m=\sum_{r'r''}F_{r'r}^m F_{r''r'}^m G_{r''}^m \\ &= \sum_{r''}(\sum_{r'}F_{r''r'}^m F_{r'r}^m)G_{r''}^m=\sum_{r''}[(F^m)^2]_{r''r}G_{r''}^m. \end{aligned} \tag{7.3.43}$$

Table 7.3.1: Number of monomials of degree m in various numbers of variables.

m	$N(m,4)$	$N(m,5)$	$N(m,6)$	$N(m,7)$	$N(m,8)$	$N(m,9)$	$N(m,10)$	$N(m,11)$
0	1	1	1	1	1	1	1	1
1	4	5	6	7	8	9	10	11
2	10	15	21	28	36	45	55	66
3	20	35	56	84	120	165	220	286
4	35	70	126	210	330	495	715	1001
5	56	126	252	462	792	1287	2002	3003
6	84	210	462	924	1716	3003	5005	8008
7	120	330	792	1716	3432	6435	11440	19448
8	165	495	1287	3003	6435	12870	24310	43758
9	220	715	2002	5005	11440	24310	48620	92378
10	286	1001	3003	8008	19448	43758	92378	184756
11	364	1365	4368	12376	31824	75582	167960	352716
12	455	1820	6188	18564	50388	125970	293930	646646

It follows that the matrix corresponding to the action of $:f_2:^2$ on $\mathcal{P}_m$ is $(F^m)^2$. Similarly, the matrix corresponding to the action of $:f_2:^\ell$ on $\mathcal{P}_m$ is $(F^m)^\ell$. Correspondingly, it follows that the action of $\mathcal{M} = \exp(:f_2:)$ on $\mathcal{P}_m$ is given by a relation of the form

$$\mathcal{M}G_r^m = \sum_{r'} M_{r'r}^m G_{r'}^m, \tag{7.3.44}$$

and the matrices M^m are related to the F^m by the equations

$$M^m = \exp(F^m). \tag{7.3.45}$$

Let us now examine the form of F^m using (3.2) through (3.7) and (3.42). We see from (3.2) through (3.6) and (3.42) that the matrices F^m associated with any $:f_2:$ made from monomials of the form q_i^2, p_i^2, q_iq_j, p_ip_j, q_ip_j have no diagonal entries. Thus, all such F^m must be traceless. The only monomials that can produce diagonal entries in the F^m are of the form q_ip_i. But, from (3.7), we see that these entries are either zero or occur in positive and negative pairs. For example, let α and β be any two positive integers. Then, referring to (3.7), for the case $\mu_1 = \alpha$ and $\nu_1 = \beta$ there must also be the case $\mu_1 = \beta$ and $\nu_1 = \alpha$. We conclude that all the F^m must be traceless,

$$\text{tr } F^m = 0. \tag{7.3.46}$$

Correspondingly, the M^m given by (3.44) and (3.45) must have unit determinant,

$$\det M^m = 1. \tag{7.3.47}$$

We close this subsection with a final remark. The relations (3.26) and (3.30) show that $: f_2 :^\dagger$ is a Lie operator for any f_2. This need not be the case for $: f_m :^\dagger$ with $m > 2$. See Exercise 3.21.

Exercises

7.3.1. Verify the relations (3.2) through (3.7).

7.3.2. Verify the relations (3.13) and (3.14), and (3.16) through (3.20).

7.3.3. Verify the relations (3.21) through (3.23).

7.3.4. Verify the relations (3.24) and (3.25).

7.3.5. Verify the relations (3.26) and (3.30).

7.3.6. Suppose that the scalar product (3.8) is generalized to the case of a $2n$ dimensional phase space in the obvious way. See Exercise 3.23. Show that the relations (3.26) and (3.30) still hold. It follows that this scalar product is invariant under $U(n)$ and $USp(2n)$. Suggestion: Let f_2 and f_2^* be quadratic polynomials defined by the equations

$$f_2 = -(1/2)(z, Sz), \tag{7.3.48}$$

$$f_2^*(z) = f_2(Jz) = -(1/2)(Jz, SJz). \tag{7.3.49}$$

Here S is a real symmetric matrix. Prove the relation

$$: f_2 :^\dagger = - : f_2^* : . \tag{7.3.50}$$

7.3.7. Show that f_2 and f_2^* as defined by (3.48) and (3.49) are connected by the relation

$$\exp[-(\pi/2) : b^0 :] f_2 = f_2^*. \tag{7.3.51}$$

7.3.8. Verify the scalar product relations

$$\langle z_a z_b, z_c z_d \rangle = \delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc}. \tag{7.3.52}$$

For two *quadratic* polynomials f and g written in the forms (5.5.1) and (5.5.2), verify that

$$\langle f, g \rangle = (1/2) \text{ tr } (S^f S^g). \tag{7.3.53}$$

7.3.9. Review Section 5.9.3 and Exercise 3.8 above. Let

$$\mathcal{R} = \exp(:f_2^c:) \tag{7.3.54}$$

be the map corresponding to the matrix R given by (5.9.28). Show that all of the $U(n)$ subgroup is covered by elements of the form (3.54) when

$$\langle f_2^c, f_2^c \rangle = (1/2)\mathrm{tr}[(S^c)^2] \leq n\pi^2. \tag{7.3.55}$$

7.3.10. Consider the Lie algebra $sp(2)$. Show that b^0 as given by (5.6.6) has a squared norm of 1. That is, $\langle b^0, b^0 \rangle = 1$. Show that f and g given by (5.6.11) and (5.6.12) also have a squared norm of 1. Show that all these $sp(2)$ elements are mutually orthogonal. Consider the Lie algebra $sp(4)$. Show that b^0 through b^3 as given by (5.7.4) have a squared norm of 2. Show that the f^j and the g^j given by (5.7.30) and (5.7.31) also have a squared norm of 2. Show that all these $sp(4)$ elements are mutually orthogonal. Consider the Lie algebra $sp(6)$. Show that b^0 as given by (5.8.5) has a squared norm of 3. Show that b^1 through b^8 as given by (5.8.5) have a squared norm of 2. Show that h^1, h^3, and h^5 as given by (5.8.37) have a squared norm of 2. Show that h^2, h^4, and h^6 have a squared norm of 4. Carry out an analogous computation for the $sp(6)$ elements $\overline{h}^j$. Show that all these $sp(6)$ elements are mutually orthogonal. The group-theoretical reason for this orthogonality is that these $sp(6)$ elements either have different $su(3)$ weights or belong to different $su(3)$ representations. Show that all f_2^a are orthogonal to all f_2^c. For some purposes we may wish to renormalize the $sp(6)$ elements so that they all have the same norm. As shown in Chapter 27, in a suitable Cartan basis all the basis elements are orthonormal.

7.3.11. Prove (3.40). Hint: First show that the binomial coefficients obey the recursion relations

$$\binom{n+1}{m} = \binom{n}{m} + \binom{n}{m-1}, \tag{7.3.56}$$

and hence

$$\binom{n+1}{m} = \binom{n}{m} + \binom{n-1}{m-1} + \cdots + \binom{n-m}{0}, \tag{7.3.57}$$

and the identity

$$\binom{n}{m} = \binom{n}{n-m}. \tag{7.3.58}$$

Next, by definition, $N(m,d)$ is the number of monomials of degree m in d variables. Verify the relations

$$N(m,1) = 1, \tag{7.3.59}$$

$$N(m,2) = m+1. \tag{7.3.60}$$

In fact, show that $N(m,d)$ satisfies the recursion relation

$$N(m,d+1) = N(m,d) + N(m-1,d) + N(m-2,d) + \cdots + N(0,d) = \sum_{j=0}^{m} N(j,d). \tag{7.3.61}$$

(Let z_{d+1} be the extra variable that is added to pass from d to $d+1$ variables. Then the monomials of degree m in $d+1$ variables may be partitioned into subsets that contain z_{d+1} to the zero power, z_{d+1} to the first power, z_{d+1} to the second power, etc.) Finally, show that $N(m,d)$ as given by (3.40) satisfies the recursion relation (3.59) and the initial condition (3.57), and verify that these facts specify $N(m,d)$ uniquely.

7.3.12. The fact that, according to (3.40), $N(m,d)$ is given simply by a binomial coefficient suggests that there should be a simple proof of this result. There is: Given n things, the number of combinations of these n things taken ℓ at a time is specified by the binomial coefficient

$$_nC_\ell = C[n,\ell] = n \text{ choose } \ell = \left(\begin{array}{c} n \\ \ell \end{array} \right) = \frac{n!}{\ell!(n-\ell)!}. \tag{7.3.62}$$

On a sheet of paper lay out $(m+d+1)$ spaces as shown below, and number them from 1 to $(m+d+1)$. Place a vertical bar "|" in the first and last spaces.

$$\begin{array}{cccccc} \underline{\qquad | \qquad} & \underline{\qquad\qquad} & \underline{\qquad\qquad} & \cdots & \underline{\qquad\qquad} & \underline{\qquad | \qquad} \\ 1 & 2 & 3 & & (m+d) & (m+d+1) \end{array}$$

After this construction $(m+d-1)$ empty spaces remain. Select m of these spaces, and place an "X" in each. The number of ways of doing this is given by the binomial coefficient

$$C[(m+d-1),m] = \left(\begin{array}{c} m+d-1 \\ m \end{array} \right). \tag{7.3.63}$$

For example, suppose $m=4$ and $d=3$. Then one way of placing the X's is shown below.

$$\begin{array}{cccccccc} \underline{\ |\ } & \underline{\quad} & \underline{X} & \underline{X} & \underline{X} & \underline{\quad} & \underline{X} & \underline{\ |\ } \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array}$$

Next, put vertical bars in the remaining empty spaces. Doing so for the $m=4$ and $d=3$ example just cited yields this picture.

$$\begin{array}{cccccccc} \underline{\ |\ } & \underline{\ |\ } & \underline{X} & \underline{X} & \underline{X} & \underline{\ |\ } & \underline{X} & \underline{\ |\ } \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array}$$

Evidently these are $(d+1)$ vertical bars after this is done. Now regard the $(d+1)$ vertical bars as representing the walls of d cells, and count the number of X's in each cell. In the case shown above, and proceeding from left to right, these counts are 0,3,1, respectively. Consider the monomial $z_1^{j_1} z_2^{j_2} \cdots z_d^{j_d}$ subject to the homogeneity condition

$$j_1 + j_2 + \cdots + j_d = m. \tag{7.3.64}$$

We may regard the numbers $j_1, j_2, \cdots j_d$ as possible cell counts for the cells 1, 2, $\cdots d$ since our construction automatically satisfies (3.62). We now see that (3.58) is the number of ways of selecting d non-negative integers $j_1, j_2, \cdots j_d$ such that (3.62) is satisfied. It follows that (3.40) is correct. Verify the argument just given for several examples.

7.3.13. There is another way to derive (3.40). Define a *composition* of m into d parts to be a representation of the form

$$m = j_1 + j_2 + \cdots + j_d, \tag{7.3.65}$$

where $j_1, j_2, \cdots j_d$ are d non-negative integers, and the order of the summands is significant. Evidently $N(m, d)$ is the number of compositions for a given m and d. Suppose we have several power series, and we want to find their product. How can we calculate the coefficients of the product series? Consider, for example, the product

$$(\sum a_i x^i)(\sum b_j x^j)(\sum c_k x^k) = \sum d_\ell x^\ell. \tag{7.3.66}$$

Then we have the relation

$$d_\ell = \sum_{i+j+k=\ell} a_i b_j c_k. \tag{7.3.67}$$

On the right side of (3.65) there will be a term for each composition of ℓ into 3 parts. Consider, by inspiration, the power series

$$f(x) = 1 + x + x^2 + \cdots . \tag{7.3.68}$$

Verify the relation

$$[f(x)]^d = \sum_{j_1=0}^{\infty} \sum_{j_2=0}^{\infty} \cdots \sum_{j_d=0}^{\infty} x^{j_1+j_2+\cdots+j_d}. \tag{7.3.69}$$

Collect terms with the same exponent, and show that the exponent x^m occurs $N(m, d)$ times. Thus, verify the relation

$$[f(x)]^d = \sum_{\ell=0}^{\infty} N(\ell, d) x^\ell. \tag{7.3.70}$$

You have shown that the functions $g(x; d)$ defined by the relations

$$g(x; d) = [f(x)]^d \tag{7.3.71}$$

are the *generating functions* (for each value of d) for the quantities $N(\ell, d)$. Next verify the relations

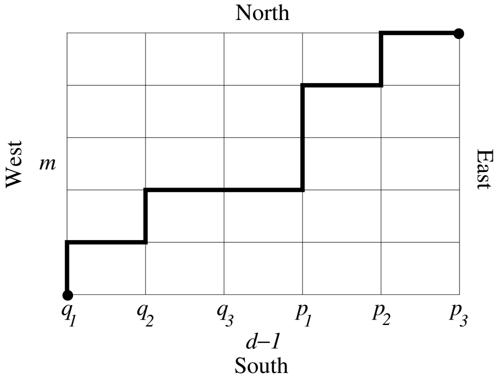
$$[f(x)]^d = (1 - x)^{-d}, \tag{7.3.72}$$

and, by the binomial theorem,

$$(1 - x)^{-d} = \sum_{\ell=0}^{\infty} \left(\begin{array}{c} \ell + d - 1 \\ \ell \end{array} \right) x^\ell. \tag{7.3.73}$$

Now compare (3.68) through (3.71) to prove (3.40). We remark that apparently *de Moivre* was the first person to view collections of numbers or functions as coefficients in the power series of some master function. Laplace subsequently championed the use of such master functions, for which he coined the term *generating functions*.

7.3.14. Here is yet another way to derive (3.40). Suppose a walker in Manhattan wants to go from **A** to **B** (see the picture below where **A** is taken to be at the lower left corner and **B** at the upper right corner). While walking, he is thinking of the problem of finding the number of all the distinct monomials $N(m, d)$ of degree m in d variables. He soon realizes that $N(m, d)$ equals the number of different paths he can walk through from **A** to **B** [1], provided that the number of blocks in the East-West direction is $d - 1$ and the number of blocks in the South-North direction is m (in the picture $m = 5$, $d = 6$). He can easily associate a path with a monomial. First he labels each street going North with a variable name. Before leaving **A** he sets to zero the exponents of all the variables of a monomial. Then he increases by one the exponent of any variable in the monomial each time he goes North by one block along the corresponding street. The monomial he is left with when he reaches **B** is the one associated with the particular path he has gone through. For example, the path shown in the picture represents the monomial $q_1 q_2 p_1^2 p_2$.



How many different paths can he walk through? In his way from **A** to **B** he has to decide if he wants to go East or North at the corner of each block. He has to take $m + (d - 1)$ decisions. The only constraint is that, overall, he needs to choose to go North (N) m times, and go East (E) $d - 1$ times. With each path we can associate a sequence of N's and E's. The following table represents the path shown in the picture.

N	E	N	E	E	N	N	E	N	E

The problem is equivalent to finding the number of all possible rearrangements of such sequences. Each sequence has $m + (d - 1)$ slots; the symbol N has to appear m times, while the symbol E appears $d - 1$ times. Therefore the number of all the possible rearrangements is given by the relation

$$N(m, d) = \frac{[m + (d - 1)]!}{m!(d - 1)!}.$$

7.3.15. Show from (3.40) or (3.59) that $N(m, d)$ can be generated by the recursion relation

$$N(m, d) = N(m, d - 1) + N(m - 1, d) \tag{7.3.74}$$

with the boundary conditions

$$N(m, 1) = 1, \tag{7.3.75}$$

[1] Only those paths that minimize the walking distance count. He never walks south or west.

$$N(0,d)=1 \text{ or } N(1,d)=d. \tag{7.3.76}$$

7.3.16. From (3.37) with $d=6$ we find the results $N(0,6)=1$, $N(1,6)=6$, $N(2,6)=21$, $N(3,6)=56$, etc. See Table 3.1. Equations (5.8.24) and (5.8.28) describe how homogeneous polynomials f_ℓ of degree ℓ in 6 variables can be decomposed into irreducible representations of $su(3)$. Equation (5.8.18) gives the dimension of these representations. Show by explicit calculation that the dimension counts for both sides of (5.8.24) match for the cases $\ell=0,2,4$. Do the same for (5.8.28) for the cases $\ell=1$ and 3. Can you show that the dimension counts agree for general ℓ?

7.3.17. Verify that the matrix corresponding to the action of $:f_2:^\ell$ on $\mathcal{P}_m$ is $(F^m)^\ell$. Derive (3.45) from (3.36), (3.41), and (3.44).

7.3.18. Show that the matrices JS^a, JS^c and P,O given by (2.4), (2.7) etc. are special cases of the matrices F^m and M^m, respectively. What is m in this case?

7.3.19. Strictly speaking, what has been shown in the text is that all matrices M^m arising from the $\mathcal{M}$ given by (3.36) must satisfy (3.47). Show that (3.47) also holds for matrices M^m arising from $\mathcal{M}$ of the form (2.10).

7.3.20. Show that the matrices F^m associated with Lie operators of the form $:f_2^a:$ are real and symmetric. Show that the matrices F^m associated with Lie operators of the form $:f_2^c:$ are real and antisymmetric.

7.3.21. The dimension of $sp(2n)$ is given by (3.7.35). Compare this dimension with $N(2,2n)$ as given by (3.40), and explain why these two numbers must agree.

7.3.22. Let q,p be coordinates in a two-dimensional phase space. Show that $:q^3:^\dagger$ is not a derivation, and therefore not a vector field. Hint: Evaluate its action on q and q^2.

7.3.23. The purpose of this exercise is to introduce a somewhat different notation for the scalar product of Subsection 3.1 with the aim of treating the $q's$ and $p's$ more democratically. For monomials introduce the notation

$$z^k = z_1^{k_1} z_2^{k_2} \cdots z_{2n}^{k_{2n}}, \tag{7.3.77}$$

$$\delta_{kk'} = \delta_{k_1k_1'}\delta_{k_2k_2'}\cdots\delta_{k_{2n}k_{2n}'}, \tag{7.3.78}$$

$$k! = k_1!k_2!\cdots k_{2n}!. \tag{7.3.79}$$

In terms of this notation, show that the scalar product of Subsection 3.1 is given by the relation

$$\langle z^k, z^{k'}\rangle = \delta_{kk'}k!. \tag{7.3.80}$$

Show that the scalar product based on (3.80) is positive definite. That is, verify that

$$\langle f, f\rangle = 0 \tag{7.3.81}$$

when $f=0$, and

$$\langle f, f\rangle > 0 \tag{7.3.82}$$

otherwise.

7.3.24. The purpose of this exercise is to explore the relation between the Lie algebras $sp(2,\mathbb{R})$, $sl(2,\mathbb{R})$, $su(2)$, and $usp(2)$. From the work of Exercise 3.7.29 we know that $sl(n,\mathbb{R})$ and $su(n)$ are equivalent over the the complex field and therefore, as a particular case, $sl(2,\mathbb{R})$ and $su(2)$ are equivalent over the complex field. Also, from Section 5.10.1, we know that $sp(2n,\mathbb{R})$ and $usp(2n)$ are equivalent over the complex field and therefore, as a particular case, $sp(2,\mathbb{R})$ and $usp(2)$ are equivalent over the complex field. In fact, $sp(2,\mathbb{R})$ and $usp(2)$ are the same. See Exercise 5.10.17. There remains the case of $sp(2,\mathbb{R})$ and $su(2)$. Your task in this exercise is to verify that they are equivalent over the complex field and to explore various features of this equivalence. In Exercise 3.25 you will have the privilege of studying how $sp(2,\mathbb{R})$ and $su(2)$ are related to $s\ell(2,\mathbb{C})$.

Suppose that the quantities B_α, for $\alpha = 1, 2,$ or 3, are *any* set of matrices or operators that obey the $sp(2,\mathbb{R})$ commutation rules given by (3.7.69) through (3.7.71). Define associated matrices/operators B'_α by the (change of basis) rules

$$B'_1 = -iB_1, \tag{7.3.83}$$

$$B'_2 = -B_2, \tag{7.3.84}$$

$$B'_3 = -iB_3. \tag{7.3.85}$$

Verify that the B'_α obey the commutation rules

$$\{B'_1, B'_2\} = B'_3, \tag{7.3.86}$$

$$\{B'_2, B'_3\} = B'_1, \tag{7.3.87}$$

$$\{B'_3, B'_1\} = B'_2, \tag{7.3.88}$$

iff the B_α obey the $sp(2,\mathbb{R})$ commutation rules given by (3.7.69) through (3.7.71). But, according to Equations (3.7.173) and (3.7.174) of Exercise 3.7.31, the commutation rules (3.86) through (3.88) are those for $su(2)$. Thus $sp(2,\mathbb{R})$ and $su(2)$ are indeed equivalent over the complex field. [Note that the relations (3.83) through (3.85) do in fact involve complex coefficients.]

Suppose the quantities B_α *are*, in fact, the 2×2 matrices that appear on the right sides of (3.7.66) through (3.7.68). Show that in this case the matrices B'_α are the matrices given by the relations

$$B'_\alpha = K^\alpha. \tag{7.3.89}$$

Recall the definitions (3.7.169) through (3.7.171).

What are the polynomials b_α and b'_α associated with the B_α and the B'_α, and what are the properties of their associated Lie operators? Based on (5.6.6), (5.6.11), and (5.6.12), and (3.7.66) through (3.7.68), verify that

$$b_1 = (1/2)f = (1/4)(p^2 - q^2), \tag{7.3.90}$$

$$b_2 = (1/2)b^0 = (1/4)(p^2 + q^2), \tag{7.3.91}$$

$$b_3 = (1/2)g = (1/2)qp. \tag{7.3.92}$$

Based on (3.83) through (3.85) and (3.89) through (3.91), verify that

$$b_1' = -ib_1 = -i(1/2)f = -i(1/4)(p^2 - q^2), \tag{7.3.93}$$

$$b_2' = -b_2 = -(1/2)b^0 = -(1/4)(p^2 + q^2), \tag{7.3.94}$$

$$b_3' = -ib_3 = -i(1/2)g = -i(1/2)qp. \tag{7.3.95}$$

Verify the Poisson bracket relations

$$[b_1, b_2] = -b_3, \tag{7.3.96}$$

$$[b_2, b_3] = -b_1, \tag{7.3.97}$$

$$[b_3, b_1] = b_2. \tag{7.3.98}$$

They are to be expected from (3.7.69) through (3.7.71) which, as we have already seen, are a variant of the commutation rules for $sp(2, \mathbb{R})$. Verify, as expected from (3.86) through (3.88), that there are the Poisson bracket relations

$$[b_1', b_2'] = b_3', \tag{7.3.99}$$

$$[b_2', b_3'] = b_1', \tag{7.3.100}$$

$$[b_3', b_1'] = b_2', \tag{7.3.101}$$

which are a variant of the commutation rules for $su(2)$. Recall (3.16) through (3.18). Verify the conjugacy relations

$$: b_1 :^\dagger =: b_1 :, \tag{7.3.102}$$

$$: b_2 :^\dagger = - : b_2 :, \tag{7.3.103}$$

$$: b_3 :^\dagger =: b_3 :; \tag{7.3.104}$$

$$: b_\alpha' :^\dagger = - : b_\alpha' : . \tag{7.3.105}$$

Thus $: b_2 :$ is anti-Hermitian, $: b_1 :$ and $: b_3 :$ are Hermitian, and the $: b_\alpha' :$ are anti-Hermitian.

It is also useful to introduce another basis for $sp(2, \mathbb{R})$. Let $\ell_\pm$ and ℓ_0 be the monomials

$$\ell_+ = -(1/2)(f + b^0) = -(1/2)p^2, \tag{7.3.106}$$

$$\ell_- = -(1/2)(f - b^0) = (1/2)q^2, \tag{7.3.107}$$

$$\ell_0 = (1/2)g = (1/2)qp. \tag{7.3.108}$$

(Note that the $\ell_\pm$ are linear combinations of b^0 and f with *real* coefficients.) Define associated Lie operators by the rules

$$\mathcal{L}_+ =: \ell_+ :, \tag{7.3.109}$$

$$\mathcal{L}_0 =: \ell_0 :, \tag{7.3.110}$$

$$\mathcal{L}_- =: \ell_- : . \tag{7.3.111}$$

Using (7.3.16) through (7.3.18) verify that

$$(\mathcal{L}_+)^\dagger =: -p^2/2 :^\dagger =: q^2/2 := \mathcal{L}_-, \tag{7.3.112}$$

$$(\mathcal{L}_0)^\dagger =: (1/2)qp :^\dagger =: (1/2) : qp := \mathcal{L}_0, \tag{7.3.113}$$

$$(\mathcal{L}_-)^\dagger =: (1/2)q^2/2 :^\dagger =: -(1/2)p^2/2 := \mathcal{L}_+. \tag{7.3.114}$$

Also, let $L_\pm$ and L_0 be the matrices associated with $\mathcal{L}_\pm$ and $\mathcal{L}_\prime$. Verify that they are given by the relations

$$L_+ = -(1/2)(F + B^0) = -\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = -(1/2)(\sigma^1 + i\sigma^2), \tag{7.3.115}$$

$$L_0 = (1/2)G = (1/2)\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = (1/2)\sigma^3, \tag{7.3.116}$$

$$L_- = -(1/2)(F - B^0) = -\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = -(1/2)(\sigma^1 - i\sigma^2). \tag{7.3.117}$$

Verify that Lie operators (3.108) through (3.110) obey the commutation rules

$$\{\mathcal{L}_+, \mathcal{L}_-\} = 2\mathcal{L}_0, \tag{7.3.118}$$

$$\{\mathcal{L}_0, \mathcal{L}_+\} = \mathcal{L}_+, \tag{7.3.119}$$

$$\{\mathcal{L}_0, \mathcal{L}_-\} = -\mathcal{L}_-. \tag{7.3.120}$$

Since the $\ell_\pm$ and ℓ_0 are real polynomials and the $\mathcal{L}_\pm$ and $\mathcal{L}_0$ are the Lie operators associated with these polynomials, the commutation rules (3.117) through (3.119) are a variant of those for $sp(2,\mathbb{R})$. Verify, as expected by construction, that $L_\pm$ and L_0 obey the same commutation rules. Finally, we observe that they are also the commutation rules for $su(2)$ in its raising and lowering operator form. See Sections 27.1 and 27.2. That is, by a change of basis involving complex coefficients, the commutation rules for the usual form of $su(2)$ can be brought to the $sp(2,\mathbb{R})$ commutation rules (3.117) through (3.119). This change in form is another instance of the equivalence of $sp(2,\mathbb{R})$ and $su(2)$ under a change of basis involving complex coefficients.

7.3.25. Review Exercise 3.24. The purpose of this exercise is to show that $sp(2,\mathbb{R})$ and $su(2)$ are subalgebras of $s\ell(2,\mathbb{C})$, to see how they fit within $s\ell(2,\mathbb{C})$, and to make analogous statements about the relations between the corresponding groups $Sp(2,\mathbb{R})$, $SU(2)$, and $SL(2,\mathbb{C})$.

The group $SL(n,\mathbb{C})$ is the set of $n \times n$ matrices with entries drawn from the field $\mathbb{C}$ and having determinant $+1$. Let B_1 through B_3 be the matrices given by (3.7.66) through (3.7.68) and let $\boldsymbol{\gamma}$ be a three-component vector with entries drawn from $\mathbb{C}$,

$$\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3). \tag{7.3.121}$$

According to Exercise 3.1.3, a 2×2 matrix is symplectic iff its determinant is $+1$. And, according to Exercise 3.7.10, the exponent for the exponential form of any matrix (which always exists in some neighborhood of the identity) has trace 0 iff the the matrix has determinant $+1$. Thus, in agreement with Exercise 3.7.26, $s\ell(n,\mathbb{C})$, the Lie algebra of $SL(n,\mathbb{C})$, consists of all $n \times n$ matrices with entries drawn from the field $\mathbb{C}$ and having trace 0. In particular, $s\ell(2,\mathbb{C})$ consists of all 2×2 matrices with entries drawn from the field $\mathbb{C}$ and having trace 0. Verify, therefore, that $s\ell(2,\mathbb{C})$ consists of matrices of the form

$$(1/2)\begin{pmatrix} \gamma_3 & \gamma_1+\gamma_2 \\ \gamma_1-\gamma_2 & -\gamma_3 \end{pmatrix} = \boldsymbol{\gamma}\cdot\boldsymbol{B} \tag{7.3.122}$$

where

$$\boldsymbol{\gamma}\cdot\boldsymbol{B} = \sum_j \gamma_j B_j. \tag{7.3.123}$$

Let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ be three-component vectors with entries drawn from $\mathbb{R}$,

$$\boldsymbol{\alpha} = (\alpha_1,\alpha_2,\alpha_3), \tag{7.3.124}$$

$$\boldsymbol{\beta} = (\beta_1,\beta_2,\beta_3). \tag{7.3.125}$$

Decompose $\boldsymbol{\gamma}$ into real and imaginary parts by writing

$$\boldsymbol{\gamma} = \boldsymbol{\alpha} + i\boldsymbol{\beta}. \tag{7.3.126}$$

Verify that $sp(2,\mathbb{R})$ consists of the matrices (3.121) evaluated with

$$\boldsymbol{\beta} = 0 \text{ and } \boldsymbol{\alpha} \text{ unrestricted.} \tag{7.3.127}$$

Consequently, in the $sp(2,\mathbb{R})$ case, the $s\ell(2,\mathbb{C})$ matrix (3.121) takes the form

$$(1/2)\begin{pmatrix} \alpha_3 & \alpha_1+\alpha_2 \\ \alpha_1-\alpha_2 & -\alpha_3 \end{pmatrix} = \boldsymbol{\alpha}\cdot\boldsymbol{B}. \tag{7.3.128}$$

What is required for the matrix in (3.121) to be in $su(2)$? Show that it must be anti-Hermitian. What does this property require of the vector $\boldsymbol{\gamma}$? Verify that

$$(1/2)\begin{pmatrix} \gamma_3 & \gamma_1+\gamma_2 \\ \gamma_1-\gamma_2 & -\gamma_3 \end{pmatrix}^{\dagger} = \sum_j \bar{\gamma}_j B_j^{\dagger} = \bar{\gamma}_1 B_1 - \bar{\gamma}_2 B_2 + \bar{\gamma}_3 B_3. \tag{7.3.129}$$

[Recall that B_1 and B_3 are Hermitian and B_2 is anti-Hermitian. See (3.7.66) through (3.7.68).] Show that the matrix in (3.121) is anti-Hermitian iff

$$\bar{\gamma}_1 = -\gamma_1, \tag{7.3.130}$$

$$\bar{\gamma}_2 = \gamma_2, \tag{7.3.131}$$

$$\bar{\gamma}_3 = -\gamma_3. \tag{7.3.132}$$

Verify it follows that $su(2)$ consists of the matrices (3.121) evaluated with

$$\alpha_1 = 0 \text{ and } \beta_1 \text{ is unrestricted,} \tag{7.3.133}$$

$$\beta_2 = 0 \text{ and } \alpha_2 \text{ is unrestricted,} \tag{7.3.134}$$

$$\alpha_3 = 0 \text{ and } \beta_3 \text{ is unrestricted.} \tag{7.3.135}$$

Consequently, show that in the $su(2)$ case, the $s\ell(2,\mathbb{C})$ matrix (3.121) takes the form

$$\begin{aligned}(1/2)\begin{pmatrix} i\beta_3 & i\beta_1+\alpha_2 \\ i\beta_1-\alpha_2 & -i\beta_3 \end{pmatrix} &= \gamma_1 B_1+\gamma_2 B_2+\gamma_3 B_3 \\ &= i\beta_1 B_1+\alpha_2 B_2+i\beta_3 B_3 \\ &= -\beta_1 B_1'-\alpha_2 B_2'-\beta_3 B_3' \\ &= -\beta_1 K^1-\alpha_2 K^2-\beta_3 K^3. \end{aligned} \tag{7.3.136}$$

Recall the relations (3.7.169) through (3.7.171) and (7.3.86) through (7.3.89).

Is it possible for the matrix (3.122) to be in *both* $sp(2,\mathbb{R})$ and $su(2)$? Upon comparing the conditions (3.127) and the conditions (3.133) through (3.135), show that the answer is *yes* provided that

$$\boldsymbol{\beta}=0,\ \alpha_1=\alpha_3=0,\ \text{and } \alpha_2 \text{ is unrestricted.} \tag{7.3.137}$$

Thus $sp(2,\mathbb{R})$ and $su(2)$ have in common the one-dimensional Lie algebra (over the real field $\mathbb{R}$) spanned by B_2.

You have shown that $sp(2,\mathbb{R})$ and $su(2)$ are both subalgebras of $s\ell(2,\mathbb{C})$. Moreover, you have shown that these two subalgebras are equivalent under the (complex) change of basis given by (7.3.83) through (7.3.85).

So far you have been treating Lie algebras. What can be said about the associated Lie groups $Sp(2,\mathbb{R})$, $SU(2)$, and $SL(2,\mathbb{C})$?

We begin with the case corresponding to the one-dimensional Lie algebra spanned by B_2, which is easy. The matrix B_2 is real and antisymmetric. See (3.7.67). Show, therefore, that elements of the form $\exp(\theta B_2)$, with θ real, comprise an $SO(2,\mathbb{R})$ group. See (3.7.93) or (5.9.12). Note that this $SO(2,\mathbb{R})$ group is in both both $Sp(2,\mathbb{R})$ and $SU(2)$.

To continue, we know much about $Sp(2,\mathbb{R})$ and, in particular, that it has the topology $E^2\times T^1$. See (5.9.10) in Section 5.9.1.[2] Similarly, we know much about $SU(2)$ and, in particular, that it has the topology S^3. See Exercise 5.10.13. What remains is to examine the case of $SL(2,\mathbb{C})$.

Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be three-vectors with real entries. Employ the notation

$$\boldsymbol{a}\cdot\boldsymbol{\sigma}=\sum_j a_j\sigma^j,\ \text{etc.} \tag{7.3.138}$$

Verify that the matrix $\boldsymbol{a}\cdot\boldsymbol{\sigma}$ is Hermitian and the matrix $i\boldsymbol{b}\cdot\boldsymbol{\sigma}$ is anti-Hermitian. Show that any matrix of the form $\exp[(1/2)\boldsymbol{a}\cdot\boldsymbol{\sigma}]$ is Hermitian and positive definite. Show that any matrix of the form $\exp[(1/2)i\boldsymbol{b}\cdot\boldsymbol{\sigma}]$ is in $SU(2)$. Show that any element M in $SL(2,\mathbb{C})$ can be written in the form

$$M=\exp[(1/2)\boldsymbol{a}\cdot\boldsymbol{\sigma}]\exp[(1/2)i\boldsymbol{b}\cdot\boldsymbol{\sigma}] \tag{7.3.139}$$

where both factors are unique. (This result is the complex analog of orthogonal polar decomposition. See Section 4.2.) What is the topology of $SL(2,\mathbb{C})$? We know that the second factor in (3.139), because it is in $SU(2)$, has the topology S^3. Show that the first factor has the topology of E^3. It follows that $SL(2,\mathbb{C})$ has the topology $E^3\times S^3$.

[2] Observe that the T^1 in $Sp(2,\mathbb{R})$ is the $SO(2,\mathbb{R})$ just described.

Using the parameterization (3.139), how does one obtain the two subgroups $SU(2)$ and $Sp(2,\mathbb{R})$? Show that $SU(2)$ consists of the matrices (3.139) evaluated with

$$\boldsymbol{a} = 0 \text{ and } \boldsymbol{b} \text{ unrestricted.} \tag{7.3.140}$$

Show that $Sp(2,\mathbb{R})$ consists of the matrices (3.139) evaluated with

$$a_1 \text{ and } a_3 \text{ unrestricted, } a_2 = 0;\ b_1 = b_3 = 0,\ b_2 \text{ unrestricted.} \tag{7.3.141}$$

In summary, $SL(2,\mathbb{C})$ has both $SU(2)$ and $Sp(2,\mathbb{R})$ subgroups.

Finally, show that the $SO(2,\mathbb{R})$ that is in both $SU(2)$ and $Sp(2,\mathbb{R})$ consists of the matrices (3.139) evaluated with

$$\boldsymbol{a} = 0;\ b_1 = b_3 = 0,\ b_2 \text{ unrestricted.} \tag{7.3.142}$$

7.3.26. Review Exercise 6.2.6. The purpose of this exercise is to derive the result (6.2.49) from the requirement (6.2.48). Before doing so, let us check that the *finite* interval defined by (6.4.27) is consistent with the *infinitesimal* interval given by (1.6.46). Verify that

$$D^2(x+dx,x) = (x+dx-x)\cdot(x+dx-x) = (dx)\cdot(dx) = ds^2. \tag{7.3.143}$$

Now move on to the main purpose of this exercise. Our goal is to show that any transformation

$$x' = f(x) \tag{7.3.144}$$

that satisfies

$$D^2(x',y') = D^2(x,y) \tag{7.3.145}$$

for all pairs of vectors x, y must be of the form

$$f(x) = x' = \Lambda x + d \tag{7.3.146}$$

where d is a fixed vector and Λ is a fixed matrix. (The converse of this assertion has already been treated in Exercise 6.2.6.)

Begin by defining d to be the image of the origin,

$$d = f(0), \tag{7.3.147}$$

and define a new transformation $h(x)$ by the rule

$$h(x) = f(x) - d. \tag{7.3.148}$$

Verify that h maps the origin into itself,

$$h(0) = 0, \tag{7.3.149}$$

and also satisfies

$$D^2\{h(x),h(y)\} = D^2(x,y). \tag{7.3.150}$$

Verify that explicit evaluation of both sides of (3.150) gives the result

$$h(x) \cdot h(x) - 2h(x) \cdot h(y) + h(y) \cdot h(y) = x \cdot x - 2x \cdot y + y \cdot y. \tag{7.3.151}$$

Show that setting $y = 0$ in (3.151) gives the result

$$h(x) \cdot h(x) = x \cdot x. \tag{7.3.152}$$

Show that combining (3.151) and (3.152) yields the result

$$h(x) \cdot h(y) = x \cdot y. \tag{7.3.153}$$

Let e^1 to e^4 be the points/vectors

$$e^1 = (1000),\ e^2 = (0100),\ \text{etc.} \tag{7.3.154}$$

Define points/vectors c^j by the rule

$$c^j = h(e^j). \tag{7.3.155}$$

Show that using (3.153) and (3.155) gives the result

$$c^i \cdot c^j = h(e^i) \cdot h(e^j) = e^i \cdot e^j = g^{ij}. \tag{7.3.156}$$

Prove, therefore, that the vectors c^j are linearly independent and can be used as a basis set. Now define a matrix Λ by the rule

$$c^j = \Lambda e^j \ \Leftrightarrow\ \Lambda e^j = c^j, \tag{7.3.157}$$

and show that this definition results in the explicit relation

$$\Lambda^{ij} = (e^i, \Lambda e^j) = (e^i, c^j) \tag{7.3.158}$$

where $(*, *)$ denotes the usual/ordinary scalar product.

Let x be an arbitrary point having the expansion

$$x = \sum_j \xi^j e^j, \tag{7.3.159}$$

and set

$$x'' = h(x). \tag{7.3.160}$$

Show, since the c^j form a basis, that one may write

$$x'' = \sum_j g^{jj} \{c^j \cdot x''\} c^j. \tag{7.3.161}$$

Using (3.153), (3.155), (3.159), and (3.160), verify that

$$c^j \cdot x'' = h(e^j) \cdot h(x) = e^j \cdot x = g^{jj} \xi^j. \tag{7.3.162}$$

Show that combining this information with (3.157), (3.160), and (3.161) yields the result

$$h(x) = x'' = \sum_j (g^{jj})^2 \xi^j c^j = \sum_j \xi^j \Lambda e^j = \Lambda x. \tag{7.3.163}$$

Finally, verify that going back to (3.148) gives the advertised result

$$f(x) = h(x) + d = \Lambda x + d. \tag{7.3.164}$$

7.3.27. The purpose of this exercise is to study the Lie algebra of the Lorentz group and the Lie/exponential representation of group elements. But, before doing so, we digress to observe that the Lorentz group has four separate components, only one of which contains the identity element.

To see that the Lorentz group has four separate components, begin by verifying that the matrices g (spatial inversion), $-g$ (temporal inversion), and $-I$ (total inversion), as well as I, are all Lorentz transformations. Next, starting with the relation (6.2.50), verify the line of reasoning

$$\begin{aligned}&\Lambda^T g\Lambda = g \Rightarrow \det(\Lambda^T g\Lambda) = \det(g) \Rightarrow \det(\Lambda^T)\det(g)\det(\Lambda) = \det(g) \Rightarrow \\ &\det(\Lambda^T)\det(\Lambda) = 1 \Rightarrow [\det(\Lambda)]^2 = 1 \Rightarrow \det(\Lambda) = \pm 1. \end{aligned} \tag{7.3.165}$$

Since the determinant of a matrix is a continuous function of its entries, the last relation in (3.165) shows that the Lorentz transformations with determinant +1 are separated in matrix space from those with determinant −1. Also we know that the matrix g is a Lorentz transformation, and is easily verified to have $\det(g) = -1$. Therefore, if any Lorentz transformation matrix with determinant +1 is multiplied by g, show that the result will be a Lorentz transformation matrix with determinant −1. Finally, verify the line of reasoning

$$\begin{aligned}&\Lambda^T g\Lambda = g \Rightarrow (\Lambda^T g\Lambda)^{44} = g^{44} \Rightarrow (\Lambda^{44})^2 - \sum_{\mu=1}^{3}(\Lambda^{\mu 4})^2 = 1 \Rightarrow \\ &|\Lambda^{44}| \geq 1 \;\Leftrightarrow\; \Lambda^{44} \geq 1 \text{ or } \Lambda^{44} \leq -1. \end{aligned} \tag{7.3.166}$$

Evidently, among the four possibilities embraced by the last relations in (3.165) and (3.166), only the component with

$$\det(\Lambda) = 1 \text{ and } \Lambda^{44} \geq 1 \tag{7.3.167}$$

can contain the identity element. Verify that this component is a subgroup, and the other three are not. Verify that all the elements in the other components can be obtained by multiplying the elements in the identity component by g or $-g$ or $-I$.

With this digression behind us, we turn to studying the Lie structure of the identity component of the Lorentz group. Suppose Λ is sufficiently near the identity matrix I so that it can be written in the form

$$\Lambda = \exp(\epsilon S) = I + \epsilon S + O(\epsilon^2) \tag{7.3.168}$$

where ϵ is a small parameter and S is a matrix to be determined. Show that inserting (3.168) into (6.2.50) and equating powers of ϵ yields the condition

$$S^T g + gS = 0 \Leftrightarrow \tag{7.3.169}$$

$$S^T = -gSg. \tag{7.3.170}$$

Verify that the condition (3.170) is also sufficient for the Λ given by (3.168) to satisfy (6.2.50) exactly. Verify that matrices S that satisfy (3.170) form a Lie algebra.

Let us pause at this point to see how the tilde Lie algebraic conjugacy operator defined by $\tilde{\mathcal{C}}$ in Exercise 3.7.36 applies to elements in the Lorentz group Lie algebra. Show from the definition (3.7.219) and (3.170) that

$$\tilde{\mathcal{C}}(S) = -S^T = gSg = gSg^{-1}. \tag{7.3.171}$$

(Here we have used the fact that $g = g^{-1}$). Evidently, for the Lorentz group Lie algebra, this conjugate representation is equivalent to the original representation. Note that Lorentz transformation matrices Λ act on four-vectors, and four-vectors carry the representation $\Gamma(1/2, 1/2)$. See Exercise 7.3.29. You have shown that this representation is self conjugate under the tilde operation.

What happens if we instead use the conjugacy operators $\breve{\mathcal{C}}$ and $\grave{\mathcal{C}}$? Verify that

$$\breve{\mathcal{C}}(S) = \bar{S} = S \tag{7.3.172}$$

and

$$\grave{\mathcal{C}}(S) = -S^\dagger = -S^T = gSg = gSg^{-1} \tag{7.3.173}$$

because S is a real matrix. It follows that, for the Lorentz group Lie algebra, the conjugate representation is equivalent to the original representation no matter what conjugacy operator is used.

Now let us work out the consequences of (3.170) in detail. Begin by computing some matrix elements. Verify the following line of reasoning for diagonal elements:

$$\begin{aligned} &S^T = -gSg \Rightarrow (S^T)^{jj} = -\sum_{k\ell} g^{jk} S^{k\ell} g^{\ell j} \Rightarrow S^{jj} = -S^{jj} \\ &\Rightarrow \text{all diagonal elements of } S \text{ vanish.} \end{aligned} \tag{7.3.174}$$

Verify the following line of reasoning for $j4$ and $4j$ elements:

$$S^T = -gSg \Rightarrow (S^T)^{j4} = -\sum_{k\ell} g^{jk} S^{k\ell} g^{\ell 4} \Rightarrow S^{4j} = -g^{jj} S^{j4} \Rightarrow S^{4j} = S^{j4} \text{ for } j \neq 4. \tag{7.3.175}$$

Verify, consequently, that S can be written in the form

$$S = \begin{pmatrix} A & \boldsymbol{a} \\ \boldsymbol{a}^T & 0 \end{pmatrix} \tag{7.3.176}$$

where A is a 3×3 matrix and $\boldsymbol{a}$ is a three-component vector. Now take (3.170) into account once again. Show that it implies the matrix relation

$$\begin{pmatrix} A^T & \boldsymbol{a} \\ \boldsymbol{a}^T & 0 \end{pmatrix} = -\begin{pmatrix} -I & \boldsymbol{o} \\ \boldsymbol{o} & 1 \end{pmatrix} \begin{pmatrix} A & \boldsymbol{a} \\ \boldsymbol{a}^T & 0 \end{pmatrix} \begin{pmatrix} -I & \boldsymbol{o} \\ \boldsymbol{o} & 1 \end{pmatrix} \tag{7.3.177}$$

where $\boldsymbol{o}$ is a three-component vector all of whose entries are zero. Verify that carrying out the matrix multiplications appearing on the right side of (3.177) yields the final result

$$\begin{pmatrix} A^T & \boldsymbol{a} \\ \boldsymbol{a}^T & 0 \end{pmatrix} = \begin{pmatrix} -A & \boldsymbol{a} \\ \boldsymbol{a}^T & 0 \end{pmatrix}. \tag{7.3.178}$$

Consequently A must be antisymmetric,

$$A^T = -A, \tag{7.3.179}$$

and $\boldsymbol{a}$ can be any three-component vector.

We are ready to set up a convenient (and pleasing) basis for the matrices S. In analogy to (3.7.178) through (3.7.180) define in the present context matrices L^1 through L^3 by the rules

$$L^1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{7.3.180}$$

$$L^2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{7.3.181}$$

$$L^3 = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{7.3.182}$$

Also, define matrices N^1 through N^3 by the rules

$$N^1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \tag{7.3.183}$$

$$N^2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \tag{7.3.184}$$

$$N^3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{7.3.185}$$

Verify that these matrices form a basis for the vector space of matrices S that satisfy (3.176) with the condition (3.179). Thus, the Lie algebra of such matrices is six dimensional. Indeed, verify that there are the commutation rules

$$\{L^j, L^k\} = \sum_\ell \epsilon_{jk\ell} L^\ell, \tag{7.3.186}$$

$$\{L^j, N^k\} = \sum_\ell \epsilon_{jk\ell} N^\ell, \tag{7.3.187}$$

$$\{N^j, N^k\} = -\sum_\ell \epsilon_{jk\ell} L^\ell. \tag{7.3.188}$$

Finally, observe that the L^j are sent into themselves under the tilde operation (3.171), and the N^j are sent into their negatives. Glancing at (3.186) through (3.188), we see that the tilded basis elements satisfy the same commutation relations as the original elements, as expected.

Consider matrices Λ of the form

$$\Lambda(\lambda, \boldsymbol{m}; \theta, \boldsymbol{n}) = \exp(\lambda \boldsymbol{m} \cdot \boldsymbol{N}) \exp(\theta \boldsymbol{n} \cdot \boldsymbol{L}) \tag{7.3.189}$$

were $\boldsymbol{m}$ and $\boldsymbol{n}$ are unit vectors and

$$\boldsymbol{m} \cdot \boldsymbol{N} = \sum_j m_j N^j, \text{ etc.} \tag{7.3.190}$$

Verify that $\boldsymbol{m} \cdot \boldsymbol{N}$ is Hermitian and $\boldsymbol{n} \cdot \boldsymbol{L}$ is anti-Hermitian, and therefore (3.189) is a polar decomposition. Show that every Λ in the identity component of the Lorentz group can be uniquely written in this form. Show that all Λ in all four components of the Lorentz group can be uniquely written in the form.

$$\Lambda(\lambda, \boldsymbol{m}; \theta, \boldsymbol{n}; r, s) = g^r(-g)^s \exp(\lambda \boldsymbol{m} \cdot \boldsymbol{N}) \exp(\theta \boldsymbol{n} \cdot \boldsymbol{L}) \text{ with } r = 0,1 \text{ and } s = 0,1. \tag{7.3.191}$$

To simplify nomenclature, from here on we will refer to the identity component of the Lorentz group simply as the Lorentz group.

Evidently, as follows from the work of Exercise 3.7.31, the factor $\exp(\theta \boldsymbol{n} \cdot \boldsymbol{L})$ in (3.189) produces spatial rotations by angle θ about the axis $\boldsymbol{n}$. What can be said about the factor $\exp(\lambda \boldsymbol{m} \cdot \boldsymbol{N})$? Your next task is to verify that it produces velocity transformations along the $\boldsymbol{m}$ axis and to find the relation between λ and the magnitude of the velocity.

Begin with the case $\boldsymbol{m} = \boldsymbol{e}_3$, in which case we are interested in the effect of $\exp(\lambda N^3)$. Verify that

$$(N^3)^2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{7.3.192}$$

and

$$(N^3)^3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} = N^3. \tag{7.3.193}$$

Show, therefore, that there is the intermediate result

$$\begin{aligned} \exp(\lambda N^3) &= I + \lambda N^3 + \lambda^2 (N^3)^2/2! + \lambda^3 (N^3)^3/3! + \cdots \\ &= I + N^3 \sinh(\lambda) + (N^3)^2[\cosh(\lambda) - 1]. \end{aligned} \tag{7.3.194}$$

Verify that employing (3.185) and (3.192) in (3.194) gives the final matrix result

$$\exp(\lambda N^3) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cosh(\lambda) & \sinh(\lambda) \\ 0 & 0 & \sinh(\lambda) & \cosh(\lambda) \end{pmatrix}. \tag{7.3.195}$$

For extra credit, evaluate $\boldsymbol{m} \cdot \boldsymbol{N}$ and its second and third powers. Show that

$$(\boldsymbol{m} \cdot \boldsymbol{N})^3 = \boldsymbol{m} \cdot \boldsymbol{N}. \tag{7.3.196}$$

Use the results you find to generalize (3.194) and to evaluate $\exp(\lambda \boldsymbol{m} \cdot \boldsymbol{N})$ for arbitrary unit vector $\boldsymbol{m}$.

To continue, review Exercise 6.2.7 and, in particular, the result (6.2.53). How are (6.2.53) and (3.195) related? Let τ be a parameter and consider the world line

$$x^1 = x^2 = x^3 = 0,\ t = \tau \tag{7.3.197}$$

so that

$$x(\tau) = \{0, 0, 0, c\tau\}. \tag{7.3.198}$$

This is the world line for a particle at rest at the spatial origin.[3] Suppose $\bar{x}$ is the result of applying $\exp(\lambda N^3)$ to x. Verify that according to (3.195) there is the result

$$\bar{x}(\tau) = \exp(\lambda N^3) x(\tau) = \{0, 0, c\sinh(\lambda)\tau, c\cosh(\lambda)\tau\} \tag{7.3.199}$$

so that

$$\bar{x}^3 = c\sinh(\lambda)\tau \tag{7.3.200}$$

and

$$c\bar{t} = c\cosh(\lambda)\tau \Leftrightarrow \bar{t} = \cosh(\lambda)\tau. \tag{7.3.201}$$

Using (3.200) and (3.201) verify that, after the transformation/boost (3.195), the particle will be moving along the $+\boldsymbol{e}_3$ axis with velocity v given by

$$v = d\bar{x}_3/d\bar{t} = c\tanh(\lambda). \tag{7.3.202}$$

Using the definition

$$v = (v/c)c = \beta c \Leftrightarrow \beta = v/c, \tag{7.3.203}$$

verify that

$$\beta = \tanh(\lambda). \tag{7.3.204}$$

The quantity

$$\lambda = \tanh^{-1}(\beta) \tag{7.3.205}$$

is called the *rapidity*. At this point we may make a remark about sign choices. Observe that had we replaced the matrices defining the N^j by their negatives in (3.183) through (3.185), the commutation rules (3.186) through (3.188) would be unchanged. However, a minus sign would then appear in (3.202). Since it seems desirable for a positive rapidity to result in a positive velocity, the sign choice we have made seems to be the more natural.

Here is an occasion for two brief interludes: For the first, review Exercise 3.7.37 and suppose the tilde group element conjugacy relation defined by the operator $\tilde{\mathcal{D}}$ is applied to the Lorentz group element Λ given by (3.189). Show that

$$\tilde{\mathcal{D}}[\Lambda(\lambda, \boldsymbol{m}; \theta, \boldsymbol{n})] = \exp(-\lambda \boldsymbol{m} \cdot \boldsymbol{N}) \exp(\theta \boldsymbol{n} \cdot \boldsymbol{L}) = \Lambda(-\lambda, \boldsymbol{m}; \theta, \boldsymbol{n}). \tag{7.3.206}$$

[3] Note that for a particle to be possibly at rest in some inertial frame it must have finite mass.

Evidently the effect of $\tilde{\mathcal{D}}$ is to change the sign of the rapidity and therefore the sign of the boost velocity.

For the second interlude, suppose two velocity transformations with rapidities λ_1 and λ_2 are made successively in the same direction. Then, from the group property

$$exp(\lambda_1 \boldsymbol{m} \cdot \boldsymbol{N}) exp(\lambda_2 \boldsymbol{m} \cdot \boldsymbol{N}) = exp[(\lambda_1 + \lambda_2) \boldsymbol{m} \cdot \boldsymbol{N}], \tag{7.3.207}$$

we see that rapidities add. Show, as a result, that there is the relation

$$\beta_3 = \tanh(\lambda_3) = \tanh(\lambda_1 + \lambda_2). \tag{7.3.208}$$

Verify the chain of reasoning

$$\begin{aligned} \beta_3 &= \tanh(\lambda_1 + \lambda_2) = [\tanh(\lambda_1) + \tanh(\lambda_2)]/[1 + \tanh(\lambda_1)\tanh(\lambda_2)] \\ &= (\beta_1 + \beta_2)/(1 + \beta_1\beta_2). \end{aligned} \tag{7.3.209}$$

This is the relativistic law for the addition of parallel velocities.

We also take this opportunity to remark that in general velocity transformations do not commute because the right side of (3.188) does not vanish, but rather contains rotation generators. Consequently a sequence of velocity transformations can produce a rotation, called a *Wigner rotation*. This failure to commute, which is a relativistic phenomena, is the origin of *Thomas precession*.

To return to the main theme, verify using (3.204) that

$$\begin{aligned} \gamma &= 1/(1-\beta^2)^{1/2} = 1/[1 - \tanh^2(\lambda)]^{1/2} = \cosh(\lambda)/[\cosh^2(\lambda) - \sinh^2(\lambda)]^{1/2} \\ &= \cosh(\lambda). \end{aligned} \tag{7.3.210}$$

Also verify that

$$\beta\gamma = \tanh(\lambda)\cosh(\lambda) = \sinh(\lambda). \tag{7.3.211}$$

Consequently, (3.195) can be rewritten in the form

$$\Lambda(\lambda, \boldsymbol{e}_3; 0, \boldsymbol{n}) = \exp(\lambda N^3) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cosh(\lambda) & \sinh(\lambda) \\ 0 & 0 & \sinh(\lambda) & \cosh(\lambda) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \gamma & \beta\gamma \\ 0 & 0 & \beta\gamma & \gamma \end{pmatrix} \tag{7.3.212}$$

in agreement with (6.2.53).

7.3.28. Review Exercise 3.27. We have seen that the commutation rules (3.186) are those for $so(3,\mathbb{R})$ and they generate spatial rotations, $SO(3,\mathbb{R})$ transformations, described by Lorentz transformation group elements of the form $\exp(\theta \boldsymbol{n} \cdot \boldsymbol{L})$. Are there other subgroups of the Lorentz group?

Curiously, and generally not discussed in the literature, the Lorentz group also has $Sp(2,\mathbb{R})$ subgroups. Consider, for example, the generators N^3, N^1, and L^2. Verify from (3.187) and (3.188) that they obey the commutation rules

$$\{L^2, N^3\} = N^1, \tag{7.3.213}$$

$$\{L^2, N^1\} = -N^3, \tag{7.3.214}$$

$$\{N^3, N^1\} = -L^2, \tag{7.3.215}$$

and therefore generate a subalgebra. Next look at the $sp(2, \mathbb{R})$ commutation rules (3.7.69) through (3.7.71). Verify, under the correspondences

$$B_1 \leftrightarrow -N^1, \tag{7.3.216}$$

$$B_2 \leftrightarrow -L^2, \tag{7.3.217}$$

$$B_3 \leftrightarrow -N^3, \tag{7.3.218}$$

which taken together behave like a change of basis using *real* coefficients, that the commutation relations (3.213) through (3.215) are the same as the $sp(2, \mathbb{R})$ commutation rules (3.7.69) through (3.7.71).[4] Verify, therefore, that Lorentz transformations of the form

$$\Lambda(\lambda_3, \lambda_1, \theta) = \exp(\lambda_3 N^3 + \lambda_1 N^1) \exp(\theta L^2) \tag{7.3.219}$$

form a group that is isomorphic to the group $Sp(2, \mathbb{R})$.

How could one have guessed that the Lorentz group would have both $SO(3, \mathbb{R})$ and $Sp(2, \mathbb{R})$ subgroups? According to Exercise 3.25 the group $SL(2, \mathbb{C})$ has both $SU(2)$ and $Sp(2, \mathbb{R})$ subgroups. In Exercise 8.2.14 you will learn that the Lorentz group is homomorphic to $SL(2, \mathbb{C})$. In fact, $SL(2, \mathbb{C})$ is the covering group for the Lorentz group. Armed with this knowledge, one would expect that the Lorentz group would have both $SO(3, \mathbb{R})$ and $Sp(2, \mathbb{R})$ subgroups.

Finally, let Λ^f be any *fixed* Lorentz group element. Verify that all Lorentz group elements Λ' of the form

$$\Lambda' = \Lambda^f \exp(\lambda_1 N^3 + \lambda_2 N^1) \exp(\theta L^2) (\Lambda^f)^{-1} \tag{7.3.220}$$

also comprise an $Sp(2, \mathbb{R})$ subgroup. Thus the Lorentz group has many $Sp(2, \mathbb{R})$ subgroups depending on the choice of Λ^f. Make a similar argument for the case of $SO(3, \mathbb{R})$.

7.3.29. The purpose of this exercise is to describe some representations of the Lorentz group. We begin with the observation that the Lorentz group is *not* compact. Indeed, for example, looking at (3.208) we see that λ can be arbitrarily large thereby yielding matrices Λ that are arbitrarily far from the origin in matrix space. It follows that the Lorentz group does not have any *finite* dimensional *unitary* representations because unitary matrices are bounded in matrix space.[5] However, the Lorentz group does have nonunitary finite dimensional representations. They are useful for constructing classical and quantum fields including the Higgs fields, Dirac fields for leptons (neutrinos, electrons $\cdots$) and quarks, vector boson fields (gluons, photons, $W^{\pm}$, Z^0), the graviton field, and more. Some facts about these finite dimensional representations are the subject of this exercise.

[4] Making the correspondences (3.216) through (3.218) was facilitated by the common appearance of Pauli matrices in (3.7.66) through (3.7.68), (5.6.7), (5.6.13), (5.6.14), and (7.3.236) through (7.3.241).

[5] The Lorentz group does have *infinite* dimensional unitary representations. Their discussion is beyond the scope of this book.

Let $\check{L}^j$ and $\check{N}^j$ be a set of matrices/operators that obey the Lorentz group Lie algebra commutation rules (3.186) through (3.188). Define related matrices/operators A^j and B^j by the rules

$$A^j = (\check{L}^j + i\check{N}^j)/2, \tag{7.3.221}$$

$$B^j = (\check{L}^j - i\check{N}^j)/2, \tag{7.3.222}$$

Note that these definitions are essentially a change of basis with coefficients drawn from the complex field $\mathbb{C}$. Verify, from (3.186) through (3.188), that the A^j and B^k obey the commutation rules

$$\{A^j, B^k\} = 0, \tag{7.3.223}$$

$$\{A^j, A^k\} = \sum_\ell \epsilon_{jk\ell} A^\ell, \tag{7.3.224}$$

$$\{B^j, B^k\} = \sum_\ell \epsilon_{jk\ell} B^\ell. \tag{7.3.225}$$

That is, the A^j and B^k commute, and separately obey the commutation rules for $su(2)$. You have shown that, over the complex field, the Lie algebra of the Lorentz group is equivalent to $su(2) \oplus su(2)$, the direct sum of two commuting $su(2)$ Lie algebras. It follows that the Lorentz group Lie algebra, unlike the classical and exceptional Lie algebras listed in Table 3.7.2, is *not* a simple Lie algebra.[6] In retrospect, we should have already known this. Recall Exercise 3.7.40.

As is familiar from their occurrence in Quantum Mechanics, the representations of $su(2)$ are labelled by a quantity j that can take on the values

$$j = 0,\ 1/2,\ 1,\ 3/2,\ 2\ \cdots\ . \tag{7.3.226}$$

We also recall that the dimension of an $su(2)$ representation is given by the quantity $(2j+1)$. Since the Lie algebra of the Lorentz group is equivalent to $su(2) \oplus su(2)$, we will label a representation of the Lorentz group Lie algebra by the symbol $\Gamma(j_1, j_2)$ where j_1 and j_2 are the j values associated with the A_k and B_k Lie algebras, respectively.[7] Verify that the dimension of $\Gamma(j_1, j_2)$ is given by the relation

$$\dim\, \Gamma(j_1, j_2) = (2j_1 + 1)(2j_2 + 1). \tag{7.3.227}$$

What is the representation in the case that the matrices $\check{L}^j$ and $\check{N}^j$ are the matrices given by (3.180) through (3.185)? This will be the representation carried by the matrices Λ given by (3.189) acting on four-vectors. To answer this question we may compute the $su(2)$ Casimir operators given by $\boldsymbol{A} \cdot \boldsymbol{A}$ and $\boldsymbol{B} \cdot \boldsymbol{B}$ where

$$\boldsymbol{A} \cdot \boldsymbol{A} = \sum_j (A^j)^2, \text{ etc.} \tag{7.3.228}$$

[6] It is, however, semisimple since it is the direct sum of two commuting $su(2)$ Lie algebras and $su(2)$ is simple.

[7] The symbol $D(j_1, j_2)$ is also frequently used to denote a representation of the Lorentz group or its Lie algebra.

(For a discussion of Casimir operators see Exercise 3.7.31 and Section 27.11.) Verify from (3.221) that

$$\sum_j (A^j)^2 = (1/4)\sum_j [(L^j)^2 + i(L^jN^j + N^jL^j) - (N^j)^2]. \tag{7.3.229}$$

Show that the following results hold:

$$\sum_j (L^j)^2 = -2\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{7.3.230}$$

[see (3.7.215)],

$$L^jN^j = N^jL^j = 0, \tag{7.3.231}$$

$$\sum_j (N^j)^2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}. \tag{7.3.232}$$

Show, therefore, that

$$\sum_j (A^j)^2 = (1/4)\sum_j [(L^j)^2 - (N^j)^2] = -(3/4)I, \tag{7.3.233}$$

and that the same result holds for $\boldsymbol{B}\cdot\boldsymbol{B}$. Since the $su(2)$ quadratic Casimir operator has the value $-j(j+1)$ it follows that in the four-vector case there are the relations

$$j_1 = j_2 = 1/2. \tag{7.3.234}$$

You have shown that four-vectors carry the representation $\Gamma(1/2,1/2)$. Note, from the work of Exercise 3.27, we know that $\Gamma(1/2,1/2)$ is self conjugate under any of the conjugacy operations. This fact is customarily expressed by writing

$$\bar{\Gamma}(1/2,1/2) = \Gamma(1/2,1/2) \tag{7.3.235}$$

where here, to be precise, the bar represents any of the conjugacy operations. We also observe from (3.189) that, since the generators given by (3.180) through (3.185) are real, the matrices Λ are real. Therefore, we say that the representation $\Gamma(1/2,1/2)$ is real.

It can be shown that Dirac 4-spinors (to be defined subsequently) carry the representation $\Gamma(0,1/2) \oplus \Gamma(1/2,0)$, and antisymmetric tensors such as the electromagnetic field tensor $F^{\mu\nu}$ carry the representation $\Gamma(0,1) \oplus \Gamma(1,0)$. (See Exercises 8.2.17 and 3.33.) It can be shown that these representations are also real. Verify that four-vectors, Dirac 4-spinors, and antisymmetric tensors have the expected dimensions of 4, 4, and 6, respectively. Verify, using (6.2.51), that the metric tensor $g^{\mu\nu}$ carries the representation $\Gamma(0,0)$. Show that the same is true of the completely antisymmetric (*Levi-Civita*) tensor/symbol $\epsilon^{\alpha\beta\gamma\delta}$.

7.3.30. Review Exercise 3.7.26. One of its purposes was to show that $SL(n,\mathbb{C})$, the set of all $n \times n$ complex matrices with determinant $+1$, forms a group; and to verify that $s\ell(n,\mathbb{C})$, the set of all $n \times n$ complex matrices with trace 0, is its Lie algebra. The purpose of this exercise is to study in some detail (the simplest case) $SL(2,\mathbb{C})$ and its Lie algebra $s\ell(2,\mathbb{C})$.

To begin, define matrices $\hat{L}^j$ and $\hat{N}^j$ by the rules

$$\hat{L}^1 = K^1 = (-i/2)\sigma^1 = (-i/2)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{7.3.236}$$

$$\hat{L}^2 = K^2 = (-i/2)\sigma^2 = (-i/2)\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} = (-1/2)\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{7.3.237}$$

$$\hat{L}^3 = K^3 = (-i/2)\sigma^3 = (-i/2)\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}; \tag{7.3.238}$$

$$\hat{N}^1 = (1/2)\sigma^1 = (1/2)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{7.3.239}$$

$$\hat{N}^2 = (1/2)\sigma^2 = (1/2)\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \tag{7.3.240}$$

$$\hat{N}^3 = (1/2)\sigma^3 = (1/2)\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{7.3.241}$$

See Exercise 3.7.31 and (3.7.169) through (3.7.171). Note that

$$\hat{N}^j = i\hat{L}^j. \tag{7.3.242}$$

Verify that the $\hat{L}^j$ and $\hat{N}^j$ form a basis for $s\ell(2,\mathbb{C})$ and obey the commutation rules

$$\{\hat{L}^j, \hat{L}^k\} = \sum_\ell \epsilon_{jk\ell}\hat{L}^\ell, \tag{7.3.243}$$

$$\{\hat{L}^j, \hat{N}^k\} = \sum_\ell \epsilon_{jk\ell}\hat{N}^\ell, \tag{7.3.244}$$

$$\{\hat{N}^j, \hat{N}^k\} = -\sum_\ell \epsilon_{jk\ell}\hat{L}^\ell. \tag{7.3.245}$$

Observe that the structure constants in (3.243) through (3.245) are the same as those in (3.183) through (3.185). Therefore the Lie algebra $s\ell(2,\mathbb{C})$ is the *same* as the Lie algebra of the Lorentz group.

Consider $SL(2,\mathbb{C})$ matrices $\hat{\Lambda}$ of the form

$$\hat{\Lambda}(\lambda, \boldsymbol{m}; \theta, \boldsymbol{n}) = \exp(\lambda \boldsymbol{m} \cdot \hat{\boldsymbol{N}}) \exp(\theta \boldsymbol{n} \cdot \hat{\boldsymbol{L}}) \tag{7.3.246}$$

were $\boldsymbol{m}$ and $\boldsymbol{n}$ are unit vectors and

$$\boldsymbol{m} \cdot \hat{\boldsymbol{N}} = \sum_j m_j \hat{N}^j, \text{ etc.} \tag{7.3.247}$$

Verify that $\boldsymbol{m} \cdot \hat{\boldsymbol{N}}$ is Hermitian and $\boldsymbol{n} \cdot \hat{\boldsymbol{L}}$ is anti-Hermitian, and therefore (3.246) is a polar decomposition. Show that every matrix in $SL(2, \mathbb{C})$ can be uniquely written in this form. See Exercise 4.2.5.

You have shown that the Lie algebra $s\ell(2, \mathbb{C})$ is the *same* as the Lie algebra of the Lorentz group. Verify, moreover, that there are the analogous polar decompositions (3.246) and (3.189). We therefore expect that there is an intimate connection between the Lorentz group and the group $SL(2, \mathbb{C})$. This connection is explored in Exercise 8.2.14 where it is shown that $SL(2, \mathbb{C})$ is the covering group of the Lorentz group.

At this point it is possible to make three remarks: For the first, review Exercise 3.7.36. Suppose the grave Lie algebraic conjugacy operator defined by $\grave{\mathcal{C}}$ is applied to the elements $\hat{L}^j$ and $\hat{N}^j$ that comprise a basis for the $s\ell(2, \mathbb{C})$ Lie algebra. Show from the definition (3.7.225) and (3.236) through (3.241) that there are the results

$$\grave{\mathcal{C}}(\hat{L}^j) = \grave{\hat{L}}^j = -(\hat{L}^j)^\dagger = \hat{L}^j, \tag{7.3.248}$$

$$\grave{\mathcal{C}}(\hat{N}^j) = \grave{\hat{N}}^j = -(\hat{N}^j)^\dagger = -\hat{N}^j. \tag{7.3.249}$$

Evidently the $\hat{L}^j$ are left in peace and the $\hat{N}^j$ change sign. Verify, as expected, that these transformed elements obey the same commutation rules (3.243) through (3.245) as the original elements.

Can the $\grave{\hat{L}}^j$,$\grave{\hat{N}}^j$ be related to the $\hat{L}^j$,$\hat{N}^j$ by a similarity transformation as in (3.7.218)? Suppose we assume so. Then there will be the relations

$$\grave{\hat{L}}^j = E\hat{L}^j E^{-1}, \tag{7.3.250}$$

$$\grave{\hat{N}}^j = E\hat{N}^j E^{-1}. \tag{7.3.251}$$

Verify that combining (3.248) through (3.251) yields the relations

$$\hat{L}^j = E\hat{L}^j E^{-1}, \tag{7.3.252}$$

$$\hat{N}^j = -E\hat{N}^j E^{-1}. \tag{7.3.253}$$

But, from (3.242) and (3.253), we conclude that

$$i\hat{L}^j = -iE\hat{L}^j E^{-1} \Rightarrow \hat{L}^j = -E\hat{L}^j E^{-1}. \tag{7.3.254}$$

Observe that (3.252) and the far right side of (3.254) disagree! It follows there is *no* E for which (3.250) and (3.251) hold. Therefore the representations of $s\ell(2, \mathbb{C})$ provided by the $\hat{L}^j$,$\hat{N}^j$ and the $\grave{\hat{L}}^j$,$\grave{\hat{N}}^j$ are *not* equivalent.

For the second remark, review Exercise 3.7.37. Suppose the grave group element conjugacy relation defined by the operator $\grave{\mathcal{D}}$ is applied to the $SL(2, \mathbb{C})$ group elements $\hat{\Lambda}$ given by (3.246). Show that

$$\grave{\mathcal{D}}[\hat{\Lambda}(\lambda, \boldsymbol{m}; \theta, \boldsymbol{n})] = \exp(-\lambda \boldsymbol{m} \cdot \hat{\boldsymbol{N}}) \exp(\theta \boldsymbol{n} \cdot \hat{\boldsymbol{L}}) = \hat{\Lambda}(-\lambda, \boldsymbol{m}; \theta, \boldsymbol{n}). \tag{7.3.255}$$

Evidently the effect of $\grave{\mathcal{D}}$ is to change the sign of the $SL(2, \mathbb{C})$ analog of the rapidity and therefore the sign of the $SL(2, \mathbb{C})$ analog of the boost velocity.

For the third remark we again comment on sign choices. In accord with our earlier finding, observe that had we replaced the matrices defining the $\hat{N}^j$ by their negatives in (3.239) through (3.241), the commutation rules (3.243) through (3.245) would be unchanged. However, as we will later see at the end of Exercise 8.2.14, the sign choice we have made is necessary for the construction of a natural map between the group $SL(2,\mathbb{C})$ and the Lorentz group.

In analogy to the study in Exercise 3.29 of the representations of the Lorentz group carried by various entities such as four-vectors, the last task of this exercise is to examine what representations are involved in the case of $SL(2,\mathbb{C})$. Following (3.221) and (3.222), form the matrices

$$\hat{A}^j = (\hat{L}^j + i\hat{N}^j)/2, \tag{7.3.256}$$

$$\hat{B}^j = (\hat{L}^j - i\hat{N}^j)/2, \tag{7.3.257}$$

using for the $\hat{L}^j$ and $\hat{N}^j$ the matrices (3.236) through (3.241). Verify that so doing yields the results

$$\hat{A}^j = 0, \tag{7.3.258}$$

$$\hat{B}^j = (-i/2)\sigma^j. \tag{7.3.259}$$

Continue on to show that

$$\sum_j (\hat{A}^j)^2 = 0, \tag{7.3.260}$$

$$\sum_j (\hat{B}^j)^2 = -(3/4)I. \tag{7.3.261}$$

Verify it follows from (3.260) and (3.261) that there are the results

$$j_1 = 0 \tag{7.3.262}$$

and

$$j_2 = 1/2. \tag{7.3.263}$$

You have shown that the use of $SL(2,\mathbb{C})$ produces the $\Gamma(0,1/2)$ representation of the Lorentz group.

We have seen that the representations of $s\ell(2,\mathbb{C})$ provided by the $\hat{L}^j$,$\hat{N}^j$ and the $\grave{\hat{L}}^j$,$\grave{\hat{N}}^j$ are not equivalent and that the representation provided by the $\hat{L}^j$, $\hat{N}^j$ is the $\Gamma(0,1/2)$ representation. What representation is provided by the $\grave{\hat{L}}^j$,$\grave{\hat{N}}^j$? Again following (3.221) and (3.222), form the matrices

$$\grave{\hat{A}}^j = (\grave{\hat{L}}^j + i\grave{\hat{N}}^j)/2, \tag{7.3.264}$$

$$\grave{\hat{B}}^j = (\grave{\hat{L}}^j - i\grave{\hat{N}}^j)/2, \tag{7.3.265}$$

using for the $\grave{\hat{L}}^j$ and $\grave{\hat{N}}^j$ the matrices (3.248) and (3.249). Verify that so doing yields the results

$$\grave{\hat{A}}^j = (\grave{\hat{L}}^j + i\grave{\hat{N}}^j)/2 = (\hat{L}^j - i\hat{N}^j)/2 = (-i/2)\sigma^j, \tag{7.3.266}$$

$$\grave{\hat{B}}^j = (\grave{\hat{L}}^j - i\grave{\hat{N}}^j)/2 = (\hat{L}^j + i\hat{N}^j)/2 = 0. \tag{7.3.267}$$

Continue on to show that

$$\sum_j (\grave{\hat{A}}^j)^2 = -(3/4)I, \tag{7.3.268}$$

$$\sum_j (\grave{\hat{B}}^j)^2 = 0. \tag{7.3.269}$$

Verify it follows from (3.268) and (3.269) that there are the results

$$j_1 = 1/2 \tag{7.3.270}$$

and

$$j_2 = 0. \tag{7.3.271}$$

You have shown that the use of the $\grave{\hat{L}}^j$,$\grave{\hat{N}}^j$ produces the $\Gamma(1/2, 0)$ representation of the Lorentz group. Correspondingly, we may write

$$\bar{\Gamma}(0, 1/2) = \Gamma(1/2, 0) \tag{7.3.272}$$

where the bar denotes the result of using the grave conjugation operation $\grave{\mathcal{C}}$. It can be shown that there is the general Lorentz group/Lie-algebra representation conjugacy relation

$$\bar{\Gamma}(j_1, j_2) = \Gamma(j_2, j_1) \tag{7.3.273}$$

of which (3.235) and (3.272) are particular cases.

7.3.31. Review Exercise 7.3.30. There you learned that $SL(2, \mathbb{C})$ and the Lorentz group have the same Lie algebra. The purpose of this exercise is to show that a *subgroup* of $SL(3, \mathbb{C})$ also has the Lorentz group Lie algebra, and to discover what representation of the Lorentz group is provided by this subgroup. In Exercise 3.7.26 you learned that $s\ell(3, \mathbb{C})$ consists of 3×3 complex matrices with trace 0. A 3×3 complex matrix requires 9 complex and hence 18 real numbers for its specification. Requiring that the trace vanish imposes one complex and hence two real conditions among these numbers with the result that $s\ell(3, \mathbb{C})$ has $18 - 2 = 16$ real dimensions. For comparison, we know that $s\ell(2, \mathbb{C})$ and the Lorentz group Lie algebra have 6 real dimensions. Therefore, if this exercise is to succeed, we must find a suitable six-dimensional *subalgebra* of $s\ell(3, \mathbb{C})$.

Recall the $so(3, \mathbb{C})$ matrices L^j defined by (3.7.178) through (3.7.180). Note that they are traceless. Use them to define $s\ell(3, \mathbb{C})$ matrices $\check{L}^j$ and $\check{N}^j$ by the rules

$$\check{L}^1 = L^1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \tag{7.3.274}$$

$$\check{L}^2 = L^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \tag{7.3.275}$$

$$\check{L}^3 = L^3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}; \tag{7.3.276}$$

$$\check{N}^1 = iL^1 = i\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \tag{7.3.277}$$

$$\check{N}^2 = iL^2 = i\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \tag{7.3.278}$$

$$\check{N}^3 = iL^3 = i\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{7.3.279}$$

Note that, in analogy with (3.242),

$$\check{N}^j = i\check{L}^j. \tag{7.3.280}$$

Verify that the $\check{L}^j$ and $\check{N}^j$ obey the commutation rules

$$\{\check{L}^j, \check{L}^k\} = \sum_\ell \epsilon_{jk\ell}\check{L}^\ell, \tag{7.3.281}$$

$$\{\check{L}^j, \check{N}^k\} = \sum_\ell \epsilon_{jk\ell}\check{N}^\ell, \tag{7.3.282}$$

$$\{\check{N}^j, \check{N}^k\} = -\sum_\ell \epsilon_{jk\ell}\check{L}^\ell. \tag{7.3.283}$$

Evidently they span a *subalgebra* of $s\ell(3,\mathbb{C})$. Moreover, the structure constants in (3.281) through (3.283) are the same as those in (3.186) through (3.188). Therefore this subalgebra is the *same* as the Lie algebra of the Lorentz group.

What representation of the Lorentz group Lie algebra is provided by the $\check{L}^j$ and $\check{N}^j$? Review Exercise 3.29. Employ in (3.221) and (3.222) the $\check{L}^j$ and $\check{N}^j$ given by (3.274) through (3.279) to find the results

$$\check{A}^j = (\check{L}^j + i\check{N}^j)/2 = (L^j - L^j)/2 = 0, \tag{7.3.284}$$

$$\check{B}^j = (\check{L}^j - i\check{N}^j)/2 = (L^j + L^j)/2 = L^j. \tag{7.3.285}$$

Continue on to show that

$$\sum_j (\check{A}^j)^2 = 0, \tag{7.3.286}$$

$$\sum_j (\check{B}^j)^2 = \sum_j (L^j)^2 = -(2)I. \tag{7.3.287}$$

See (3.7.215). Verify it follows from (3.286) and (3.287) that there are the results

$$j_1 = 0 \tag{7.3.288}$$

and

$$j_2 = 1. \tag{7.3.289}$$

You have shown that the use of the $s\ell(3,\mathbb{C})$ subalgebra produces the $\Gamma(0,1)$ representation of the Lorentz group Lie algebra.

Suppose the grave Lie algebraic conjugacy operator defined by $\grave{\mathcal{C}}$ is applied to the elements $\check{L}^j$ and $\check{N}^j$ that comprise a basis for the $s\ell(3,\mathbb{C})$ Lie subalgebra. Show from the definition (3.7.225) and (3.274) through (3.279) that there are the results

$$\grave{\mathcal{C}}(\check{L}^j) = \grave{\check{L}}^j = -(\check{L}^j)^\dagger = \check{L}^j, \tag{7.3.290}$$

$$\grave{\mathcal{C}}(\check{N}^j) = \grave{\check{N}}^j = -(\check{N}^j)^\dagger = -\check{N}^j. \tag{7.3.291}$$

Evidently the $\check{L}^j$ are left in peace and the $\check{N}^j$ change sign. Verify, as expected, that the $\grave{\check{L}}^j$, $\grave{\check{N}}^j$ also provide a representation of the Lorentz group Lie algebra.

Suppose instead the breve Lie algebraic conjugacy operator defined by $\breve{\mathcal{C}}$ is applied to the elements $\check{L}^j$ and $\check{N}^j$ that comprise a basis for the $s\ell(3,\mathbb{C})$ Lie subalgebra. Show from the definition (3.7.224) and (3.274) through (3.279) that there are the results

$$\breve{\mathcal{C}}(\check{L}^j) = \breve{\check{L}}^j = \bar{\check{L}}^j = \check{L}^j, \tag{7.3.292}$$

$$\breve{\mathcal{C}}(\check{N}^j) = \breve{\check{N}}^j = \bar{\check{N}}^j = -\check{N}^j. \tag{7.3.293}$$

Evidently, for the $s\ell(3,\mathbb{C})$ Lie subalgebra, the grave and breve operations have the *same* effect.

Read again the part of Exercise 3.30 that showed the representations of the Lorentz group Lie algebra provided by $\hat{L}^j$,$\hat{N}^j$ and $\grave{\hat{L}}^j$,$\grave{\hat{N}}^j$ are not equivalent. Construct a similar proof that the representations of the Lorentz group Lie algebra provided by $\check{L}^j$,$\check{N}^j$ and $\grave{\check{L}}^j$,$\grave{\check{N}}^j$ are not equivalent. Also, state and prove an analog of (3.255). Finally, show that the $\grave{\check{L}}^j$,$\grave{\check{N}}^j$ produce the $\Gamma(1,0)$ representation of the Lorentz group Lie algebra so that

$$\bar{\Gamma}(0,1) = \Gamma(1,0), \tag{7.3.294}$$

which is again a particular case of (3.273).

7.3.32. Recall that under the action of a Lorentz transformation Λ the electromagnetic field tensor $F^{\mu\nu}$ defined by

$$F^{\mu\nu} = \begin{pmatrix} 0 & -B_z & B_y & E_x/c \\ B_z & 0 & -B_x & E_y/c \\ -B_y & B_x & 0 & E_z/c \\ -E_x/c & -E_y/c & -E_z/c & 0 \end{pmatrix} \tag{7.3.295}$$

transforms according to the rule

$$\hat{F}^{\alpha\beta} = \sum_{\mu\nu} \Lambda^{\alpha\mu}\Lambda^{\beta\nu}F^{\mu\nu} \Leftrightarrow \hat{F} = \Lambda F \Lambda^T. \tag{7.3.296}$$

Review Exercise 1.6.17. (Here we use a hat ˆ rather than a bar ¯ as a distinguishing mark because later we will want to use a bar to indicate complex conjugation.) This exercise is the first of two exercises whose purpose is to relate the transformation rule (3.296) to the Lorentz Lie algebra/group representations $\Gamma(0,1)$ and $\Gamma(1,0)$ found in Exercise 3.31 above.

It will be limited to Lorentz transformations that are near the identity transformation. It will be followed by a subsequent exercise that extends the results found here to all Lorentz transformations.

Before beginning our exploration it is convenient to introduce some new notation. Define a tensor-valued *function* $\mathcal{F}^{\mu\nu}$ of $\boldsymbol{E}$ and $\boldsymbol{B}$ by the rule

$$\mathcal{F}^{\mu\nu}(\boldsymbol{E},\boldsymbol{B}) = \begin{pmatrix} 0 & -B_z & B_y & E_x/c \\ B_z & 0 & -B_x & E_y/c \\ -B_y & B_x & 0 & E_z/c \\ -E_x/c & -E_y/c & -E_z/c & 0 \end{pmatrix}. \tag{7.3.297}$$

With this definition we may rewrite (3.295) in the form

$$F^{\mu\nu} = \mathcal{F}^{\mu\nu}(\boldsymbol{E},\boldsymbol{B}), \tag{7.3.298}$$

and we may also write

$$\hat{F}^{\mu\nu} = \mathcal{F}^{\mu\nu}(\hat{\boldsymbol{E}},\hat{\boldsymbol{B}}) = \begin{pmatrix} 0 & -\hat{B}_z & \hat{B}_y & \hat{E}_x/c \\ \hat{B}_z & 0 & -\hat{B}_x & \hat{E}_y/c \\ -\hat{B}_y & \hat{B}_x & 0 & \hat{E}_z/c \\ -\hat{E}_x/c & -\hat{E}_y/c & -\hat{E}_z/c & 0 \end{pmatrix} \tag{7.3.299}$$

where $\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{B}}$ are the transformed fields associated with $\hat{F}$. Finally, for compactness of notation, we will sometimes omit the tensor indices to simply write

$$F = \mathcal{F}(\boldsymbol{E},\boldsymbol{B}) \tag{7.3.300}$$

and

$$\hat{F} = \mathcal{F}(\hat{\boldsymbol{E}},\hat{\boldsymbol{B}}). \tag{7.3.301}$$

Now let us begin our exploration by considering some particular cases. Suppose Λ is the Lorentz transformation for a *small* rotation θ about the z axis,

$$\Lambda = \exp(\theta L^3) = I + \theta L^3 + O(\theta)^2. \tag{7.3.302}$$

Show that in this case (3.296) becomes

$$\hat{F} = \Lambda F \Lambda^T = (I+\theta L^3)F(I-\theta L^3) + O(\theta^2) = F + \theta\{L^3, F\} + O(\theta^2). \tag{7.3.303}$$

(Recall that L^3 is antisymmetric.) Verify that

$$\begin{aligned} FL^3 &= \begin{pmatrix} 0 & -B_z & B_y & E_x/c \\ B_z & 0 & -B_x & E_y/c \\ -B_y & B_x & 0 & E_z/c \\ -E_x/c & -E_y/c & -E_z/c & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} -B_z & 0 & 0 & 0 \\ 0 & -B_z & 0 & 0 \\ B_x & B_y & 0 & 0 \\ -E_y/c & E_x/c & 0 & 0 \end{pmatrix}, \end{aligned} \tag{7.3.304}$$

$$L^3F = [F^T(L^3)^T]^T = (FL^3)^T = \begin{pmatrix} -B_z & 0 & B_x & -E_y/c \\ 0 & -B_z & B_y & E_x/c \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{7.3.305}$$

and therefore

$$\{L^3, F\} = \begin{pmatrix} 0 & 0 & B_x & -E_y/c \\ 0 & 0 & B_y & E_x/c \\ -B_x & -B_y & 0 & 0 \\ E_y/c & -E_x/c & 0 & 0 \end{pmatrix}. \tag{7.3.306}$$

Verify that use of (3.297) through (3.306) yields the relations

$$\hat{E}_x = E_x - \theta E_y + O(\theta^2), \tag{7.3.307}$$

$$\hat{E}_y = E_y + \theta E_x + O(\theta^2), \tag{7.3.308}$$

$$\hat{E}_z = E_z + O(\theta^2); \tag{7.3.309}$$

$$\hat{B}_x = B_x - \theta B_y + O(\theta^2), \tag{7.3.310}$$

$$\hat{B}_y = B_y + \theta B_x + O(\theta^2), \tag{7.3.311}$$

$$\hat{B}_z = B_z + O(\theta^2). \tag{7.3.312}$$

What are we to make of the relations (3.307) through (3.312)? Verify that (3.307) through (3.309) can be rewritten in the form

$$\begin{pmatrix} \hat{E}_x \\ \hat{E}_y \\ \hat{E}_z \end{pmatrix} = \begin{pmatrix} 1 & -\theta & 0 \\ \theta & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} + O(\theta^2) \tag{7.3.313}$$

or, in matrix/vector notation,

$$\hat{\boldsymbol{E}} = \exp(\theta \check{L}^3)\boldsymbol{E} + O(\theta^2). \tag{7.3.314}$$

See (3.276). Verify that (3.310) through (3.312) can be rewritten analogously.

At this point, in anticipation of further results, it is convenient to define two three-dimensional complex vectors $\boldsymbol{F}^{\pm}$ (sometimes called *Faraday* vectors) by the rules

$$\boldsymbol{F}^{\pm} = \boldsymbol{E} \pm ic\boldsymbol{B}. \tag{7.3.315}$$

Verify that the results (3.307) through (3.312) can be rewritten in the form

$$\begin{pmatrix} \hat{F}_x^+ \\ \hat{F}_y^+ \\ \hat{F}_z^+ \end{pmatrix} = \begin{pmatrix} 1 & -\theta & 0 \\ \theta & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} F_x^+ \\ F_y^+ \\ F_z^+ \end{pmatrix} + O(\theta^2) \tag{7.3.316}$$

or, in matrix/vector notation,

$$\hat{\boldsymbol{F}}^+ = \exp(\theta \check{L}^3)\boldsymbol{F}^+ + O(\theta^2). \tag{7.3.317}$$

Let us extend our notation a bit further. Evidently a knowledge of $\boldsymbol{F}^{+}$ is equivalent to a knowledge of $\boldsymbol{E}$ and $\boldsymbol{B}$, and vice versa. Indeed, from (3.315) we see that

$$\boldsymbol{E} = \Re(\boldsymbol{F}^{+}) \tag{7.3.318}$$

and

$$\boldsymbol{B} = (1/c)\Im(\boldsymbol{F}^{+}). \tag{7.3.319}$$

We may therefore view $\boldsymbol{F}^{+}$ as being an argument for $\mathcal{F}$ so that (3.300) and (3.301) can also be written as

$$F = \mathcal{F}(\boldsymbol{F}^{+}) \tag{7.3.320}$$

and

$$\hat{F} = \mathcal{F}(\hat{\boldsymbol{F}}^{+}). \tag{7.3.321}$$

Verify, using this extended notation, that the results (3.301) through (3.303) and (3.317) can be rewritten in the form

$$\exp(\theta L^3)\mathcal{F}(\boldsymbol{F}^{+})[\exp(\theta L^3)]^T = \mathcal{F}[\exp(\theta \check{L}^3)\boldsymbol{F}^{+}] + O(\theta^2). \tag{7.3.322}$$

We have studied the case of a small *rotation* about the z axis. As a second particular case, suppose Λ is the Lorentz transformation for a small *boost* λ along the z axis,

$$\Lambda = \exp(\lambda N^3) = I + \lambda N^3 + O(\lambda)^2. \tag{7.3.323}$$

Show that in this case (3.296) becomes

$$\hat{F} = \Lambda F \Lambda^T = (I + \lambda N^3)F(I + \lambda N^3) + O(\lambda^2) = F + \lambda\{N^3, F\}_+ + O(\lambda^2). \tag{7.3.324}$$

(Here $\{*, *\}_+$ denotes an anticommutator, and we have used the fact that N^3 is symmetric.) Verify that

$$\begin{aligned} FN^3 &= \begin{pmatrix} 0 & -B_z & B_y & E_x/c \\ B_z & 0 & -B_x & E_y/c \\ -B_y & B_x & 0 & E_z/c \\ -E_x/c & -E_y/c & -E_z/c & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & E_x/c & B_y \\ 0 & 0 & E_y/c & -B_x \\ 0 & 0 & E_z/c & 0 \\ 0 & 0 & 0 & -E_z/c \end{pmatrix}, \end{aligned} \tag{7.3.325}$$

$$N^3F = [F^T(N^3)^T]^T = -(FN^3)^T = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -E_x/c & -E_y/c & -E_z/c & 0 \\ -B_y & B_x & 0 & E_z/c \end{pmatrix}, \tag{7.3.326}$$

and therefore

$$\{N^3, F\}_+ = \begin{pmatrix} 0 & 0 & E_x/c & B_y \\ 0 & 0 & E_y/c & -B_x \\ -E_x/c & -E_y/c & 0 & 0 \\ -B_y & B_x & 0 & 0 \end{pmatrix}. \tag{7.3.327}$$

Verify that use of (3.297) through (3.299), and (3.324) through (3.327) yields the relations

$$\hat{E}_x = E_x + \lambda c B_y + O(\lambda^2), \tag{7.3.328}$$

$$\hat{E}_y = E_y - \lambda c B_x + O(\lambda^2), \tag{7.3.329}$$

$$\hat{E}_z = E_z + O(\lambda^2); \tag{7.3.330}$$

$$\hat{B}_x = B_x - \lambda E_y/c + O(\lambda^2), \tag{7.3.331}$$

$$\hat{B}_y = B_y + \lambda E_x/c + O(\lambda^2), \tag{7.3.332}$$

$$\hat{B}_z = B_z + O(\lambda^2). \tag{7.3.333}$$

What are we to make of the transformation results (3.328) through (3.333)? Verify, in terms of the Faraday vector $\boldsymbol{F}^+$, that

$$\begin{aligned} \hat{F}^+_x &= \hat{E}_x + ic\hat{B}_x = E_x + icB_x + \lambda c B_y - i\lambda E_y + O(\lambda^2) \\ &= E_x + icB_x - i\lambda(E_y + icB_y) + O(\lambda^2) \\ &= F^+_x - i\lambda F^+_y + O(\lambda^2), \end{aligned} \tag{7.3.334}$$

$$\begin{aligned} \hat{F}^+_y &= \hat{E}_y + ic\hat{B}_y = E_y + icB_y - \lambda c B_x + i\lambda E_x + O(\lambda^2) \\ &= E_y + icB_y + i\lambda(E_x + icB_x) + O(\lambda^2) \\ &= F^+_y + i\lambda F^+_x + O(\lambda^2), \end{aligned} \tag{7.3.335}$$

$$\hat{F}^+_z = \bar{E}_z + ic\hat{B}_z = E_z + icB_z + O(\lambda^2) = F^+_z + O(\lambda^2). \tag{7.3.336}$$

Consequently the results (3.328) through (3.336) can be rewritten in the form

$$\begin{pmatrix} \hat{F}^+_x \\ \hat{F}^+_y \\ \hat{F}^+_z \end{pmatrix} = \begin{pmatrix} 1 & -i\lambda & 0 \\ i\lambda & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} F^+_x \\ F^+_y \\ F^+_z \end{pmatrix} + O(\lambda^2) \tag{7.3.337}$$

or, in matrix/vector notation,

$$\hat{\boldsymbol{F}}^+ = \exp(\lambda \check{N}^3)\boldsymbol{F}^+ + O(\lambda^2). \tag{7.3.338}$$

See (3.279). Finally, using the extended notation (3.320) and (3.321), verify that the results (3.324) through (3.338) can be rewritten in the form

$$\exp(\lambda N^3)\mathcal{F}(\boldsymbol{F}^+)[\exp(\lambda N^3)]^T = \mathcal{F}[\exp(\lambda\check{N}^3)\boldsymbol{F}^+] + O(\lambda^2). \tag{7.3.339}$$

7.3.33. This exercise is a sequel to Exercise 3.32. It established, for *small* θ and *small* λ, the key results (3.322) and (3.339). Associated with these results are the correspondences

$$\Lambda = \exp(\theta L^3) \Leftrightarrow \hat{\boldsymbol{F}}^+ = \exp(\theta \check{L}^3)\boldsymbol{F}^+ + O(\theta^2), \tag{7.3.340}$$

$$\Lambda = \exp(\lambda N^3) \Leftrightarrow \hat{\boldsymbol{F}}^+ = \exp(\lambda \check{N}^3)\boldsymbol{F}^+ + O(\lambda^2). \tag{7.3.341}$$

The purpose of this exercise is to extend these results and to relate them to the Lorentz Lie algebra/group representations $\Gamma(0,1)$ and $\Gamma(1,0)$.

The first extension is to observe that the $O(\theta^2)$ error terms in (3.322) and (3.340) and the $O(\lambda^2)$ error terms in (3.339) and (3.341) are in fact *identically zero*! To see this, in the case of the $O(\theta^2)$ error terms, observe that there is the *group* relation

$$\exp[\theta L^3] = \exp[(1/2)\theta L^3]\exp[(1/2)\theta L^3]. \tag{7.3.342}$$

Verify that employing this group relation in (3.322) yields the result

$$\begin{aligned}
&\exp(\theta L^3)\mathcal{F}(\boldsymbol{F}^+)[\exp(\theta L^3)]^T = \\
&\exp[(1/2)\theta L^3]\exp[(1/2)\theta L^3]\mathcal{F}(\boldsymbol{F}^+)\{\exp[(1/2)\theta L^3)]\}^T\{\exp[(1/2)\theta L^3)]\}^T = \\
&\exp[(1/2)\theta L^3]\mathcal{F}\{\exp[(1/2)\theta \check{L}^3]\boldsymbol{F}^+\}\{\exp[(1/2)\theta L^3)]\}^T + O[(\theta/2)^2] = \\
&\mathcal{F}\{\exp[(1/2)\theta \check{L}^3]\exp[(1/2)\theta \check{L}^3]\boldsymbol{F}^+\} + 2O[(\theta/2)^2] = \\
&\mathcal{F}[\exp(\theta \check{L}^3)\boldsymbol{F}^+] + (1/2)O(\theta^2).
\end{aligned} \tag{7.3.343}$$

That is, the possible $O(\theta^2)$ error term in (3.322) has been replaced by the possible $(1/2)O(\theta^2)$ error term in (3.343). Similarly, using in (3.322) the group relation

$$\exp[\theta L^3] = \{\exp[(1/n)\theta L^3]\}^n \tag{7.3.344}$$

will yield the result

$$\exp(\theta L^3)\mathcal{F}(\boldsymbol{F}^+)[\exp(\theta L^3)]^T = \mathcal{F}[\exp(\theta \check{L}^3)\boldsymbol{F}^+] + (1/n)O(\theta^2). \tag{7.3.345}$$

Therefore, upon letting $n \to \infty$, we find the result

$$\exp(\theta L^3)\mathcal{F}(\boldsymbol{F}^+)[\exp(\theta L^3)]^T = \mathcal{F}[\exp(\theta \check{L}^3)\boldsymbol{F}^+] \tag{7.3.346}$$

as claimed.

In an analogous way, *mutatis mutandis* and again using group properties, (3.339) becomes the relation

$$\exp(\lambda N^3)\mathcal{F}(\boldsymbol{F}^+)[\exp(\lambda N^3)]^T = \mathcal{F}[\exp(\lambda \check{N}^3)\boldsymbol{F}^+]. \tag{7.3.347}$$

At this point let us make a sanity check. Verify that

$$\exp(\lambda \check{N}^3) = \begin{pmatrix} \cosh(\lambda) & -i\sinh(\lambda) & 0 \\ i\sinh(\lambda) & \cosh(\lambda) & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{7.3.348}$$

and therefore in this case it follows that

$$\hat{F}_x^+ = \cosh(\lambda)F_x^+ - i\sinh(\lambda)F_y^+, \tag{7.3.349}$$

$$\hat{F}_y^+ = i\sinh(\lambda)F_x^+ + \cosh(\lambda)F_y^+, \tag{7.3.350}$$

$$\hat{F}_z^+ = F_z^+, \tag{7.3.351}$$

so that

$$\hat{E}_x = \Re(\hat{F}_x^+) = \cosh(\lambda)E_x + \sinh(\lambda)cB_y = \gamma E_x + \beta\gamma cB_y, \tag{7.3.352}$$

$$\hat{E}_y = \Re(\hat{F}_y^+) = \cosh(\lambda)E_y - \sinh(\lambda)cB_x = \gamma E_y - \beta\gamma cB_x, \tag{7.3.353}$$

$$\hat{E}_z = \Re(\hat{F}_z^+) = E_z; \tag{7.3.354}$$

$$\hat{B}_x = (1/c)\Im(\hat{F}_x^+) = \cosh(\lambda)B_x - \sinh(\lambda)(1/c)E_y = \gamma B_x - \beta\gamma(1/c)E_y, \tag{7.3.355}$$

$$\hat{B}_y = (1/c)\Im(\hat{F}_y^+) = \cosh(\lambda)B_y + \sinh(\lambda)(1/c)E_x = \gamma B_y + \beta\gamma(1/c)E_x, \tag{7.3.356}$$

$$\hat{B}_z = (1/c)\Im(\hat{F}_z^+) = B_z. \tag{7.3.357}$$

[See (3.210) and (3.211).] The relations (3.352) through (3.357) are the expected ones for a boost along the z axis, and therefore the sanity check has been passed.

So far we have considered rotations about the z axis and boosts along the z axis. The second extension is to consider other axes. Verify that, since we have already allowed $\boldsymbol{E}$ and $\boldsymbol{B}$ to be completely arbitrary, there must be the obvious generalizations: Suppose Λ is the Lorentz transformation for a general rotation,

$$\Lambda = \exp(\theta\boldsymbol{n}\cdot\boldsymbol{L}). \tag{7.3.358}$$

In this case (3.346) has the generalization

$$\exp(\theta\boldsymbol{n}\cdot\boldsymbol{L})\mathcal{F}(\boldsymbol{F}^+)[\exp(\theta\boldsymbol{n}\cdot\boldsymbol{L})]^T = \mathcal{F}[\exp(\theta\boldsymbol{n}\cdot\check{\boldsymbol{L}})\boldsymbol{F}^+]. \tag{7.3.359}$$

Or suppose Λ is the Lorentz transformation for a general boost,

$$\Lambda = \exp(\lambda\boldsymbol{m}\cdot\boldsymbol{N}). \tag{7.3.360}$$

In this case (3.347) has the generalization

$$\exp(\lambda\boldsymbol{m}\cdot\boldsymbol{N})\mathcal{F}(\boldsymbol{F}^+)[\exp(\lambda\boldsymbol{m}\cdot\boldsymbol{N})]^T = \mathcal{F}[\exp(\lambda\boldsymbol{m}\cdot\check{\boldsymbol{N}})\boldsymbol{F}^+]. \tag{7.3.361}$$

The final extension is to consider general Lorentz transformations

$$\Lambda(\lambda,\boldsymbol{m};\theta,\boldsymbol{n}) = \exp(\lambda\boldsymbol{m}\cdot\boldsymbol{N})\exp(\theta\boldsymbol{n}\cdot\boldsymbol{L}). \tag{7.3.362}$$

Recall (3.189). Verify that in this case there is the result

$$\begin{aligned}&\Lambda\mathcal{F}(\boldsymbol{F}^+)\Lambda^T = \exp(\lambda\boldsymbol{m}\cdot\boldsymbol{N})\exp(\theta\boldsymbol{n}\cdot\boldsymbol{L})\mathcal{F}(\boldsymbol{F}^+)[\exp(\theta\boldsymbol{n}\cdot\boldsymbol{L})]^T[\exp(\lambda\boldsymbol{m}\cdot\boldsymbol{N})]^T =\\ &\exp(\lambda\boldsymbol{m}\cdot\boldsymbol{N})\mathcal{F}[\exp(\theta\boldsymbol{n}\cdot\check{\boldsymbol{L}})\boldsymbol{F}^+][\exp(\lambda\boldsymbol{m}\cdot\boldsymbol{N})]^T =\\ &\mathcal{F}[\exp(\lambda\boldsymbol{m}\cdot\check{\boldsymbol{N}})\exp(\theta\boldsymbol{n}\cdot\check{\boldsymbol{L}})\boldsymbol{F}^+].\end{aligned} \tag{7.3.363}$$

Verify, consequently, that there is the general correspondence

$$\Lambda = \exp(\lambda\boldsymbol{m}\cdot\boldsymbol{N})\exp(\theta\boldsymbol{n}\cdot\boldsymbol{L}) \Leftrightarrow \hat{\boldsymbol{F}}^+ = \exp(\lambda\boldsymbol{m}\cdot\check{\boldsymbol{N}})\exp(\theta\boldsymbol{n}\cdot\check{\boldsymbol{L}})\boldsymbol{F}^+. \tag{7.3.364}$$

Your last task is to explore how these results relate to the Lorentz Lie algebra/group representations $\Gamma(0,1)$ and $\Gamma(0,1)$. You already found in Exercise 3.31 that the matrices $\check{\boldsymbol{L}}$ and $\check{\boldsymbol{N}}$ constitute a basis for the $\Gamma(0,1)$ representation of the Lorentz group Lie algebra.

Consequently, according to the right side equation in (3.364), the vectors $\boldsymbol{F}^{+}$ carry the $\Gamma(0,1)$ representation of the Lorentz Lie algebra/group. Now form the complex conjugate of the right side equation in (3.364). Verify that so doing yields the relation

$$\bar{\hat{\boldsymbol{F}}}^{+} = \exp(\lambda \boldsymbol{m}\cdot\bar{\check{\boldsymbol{N}}})\exp(\theta\boldsymbol{n}\cdot\bar{\check{\boldsymbol{L}}})\bar{\boldsymbol{F}}^{+} \tag{7.3.365}$$

where a bar $^{-}$ denotes complex conjugation. [Verify from (3.7.1) that $\overline{\exp(B)} = \exp(\bar{B})$ for any matrix B.] Next verify that

$$\bar{\hat{\boldsymbol{F}}}^{+} = \hat{\boldsymbol{F}}^{-} \text{ and } \bar{\boldsymbol{F}}^{+} = \boldsymbol{F}^{-}. \tag{7.3.366}$$

Also show from (3.290) through (3.293) that for the matrices $\check{\boldsymbol{L}}$ and $\check{\boldsymbol{N}}$ there are the relations

$$\bar{\check{\boldsymbol{L}}} = \grave{\check{\boldsymbol{L}}} \text{ and } \bar{\check{\boldsymbol{N}}} = \grave{\check{\boldsymbol{N}}}. \tag{7.3.367}$$

Verify, therefore, that (3.365) can be rewritten in the form

$$\hat{\boldsymbol{F}}^{-} = \exp(\lambda \boldsymbol{m}\cdot\grave{\check{\boldsymbol{N}}})\exp(\theta\boldsymbol{n}\cdot\grave{\check{\boldsymbol{L}}})\boldsymbol{F}^{-}. \tag{7.3.368}$$

Verify, accordingly, that the correspondence (3.364) implies the correspondence

$$\Lambda = \exp(\lambda\boldsymbol{m}\cdot\boldsymbol{N})\exp(\theta\boldsymbol{n}\cdot\boldsymbol{L}) \Leftrightarrow \hat{\boldsymbol{F}}^{-} = \exp(\lambda \boldsymbol{m}\cdot\grave{\check{\boldsymbol{N}}})\exp(\theta\boldsymbol{n}\cdot\grave{\check{\boldsymbol{L}}})\boldsymbol{F}^{-}, \tag{7.3.369}$$

and vice versa. You already found in Exercise 3.31 that the matrices $\grave{\check{\boldsymbol{L}}}$ and $\grave{\check{\boldsymbol{N}}}$ constitute a basis for the $\Gamma(1,0)$ representation of the Lorentz group Lie algebra. Consequently, according to (3.368), the vectors $\boldsymbol{F}^{-}$ carry the $\Gamma(1,0)$ representation of the Lorentz Lie algebra/group. Put another way, in transforming according to the rule (3.296), the electromagnetic field tensor $F^{\mu\nu}$ carries *both* the Lie algebra/group representations $\Gamma(0,1)$ and $\Gamma(1,0)$. Note since (3.296) involves only real quantities when acting on F and hence on $\boldsymbol{E}$ and $\boldsymbol{B}$, this net representation is real. See, for example, (3.352) through (3.357). We would now like to see in more detail how this net representation carries both the $\Gamma(0,1)$ and $\Gamma(1,0)$ representations. For this, see Exercise 7.3.35.

7.3.34. The purpose of this exercise is to develop some general purpose matrix machinery for working with complex matrices that will be of subsequent use.[8] Suppose k_1 and k_2 are two $n\times n$ possibly complex matrices. Decompose each k_j into real and imaginary parts by writing

$$k_j = \Re k_j + i\Im k_j \tag{7.3.370}$$

so that

$$\begin{aligned} k_1k_2 = (\Re k_1 + i\Im k_1)(\Re k_2 + i\Im k_2) = \\ (\Re k_1\Re k_2 - \Im k_1\Im k_2) + i(\Re k_1\Im k2 + \Im k_1\Re k_2). \end{aligned} \tag{7.3.371}$$

Next suppose we define k_3 by the rule

$$k_3 = k_1k_2 \tag{7.3.372}$$

[8] This machinery is analogous to some of the machinery in Section 3.9.

and also make the decomposition (3.370) for k_3. Verify it follows from (3.370) through (3.372) that

$$\Re k_3 = \Re k_1 \Re k_2 - \Im k_1 \Im k_2 \tag{7.3.373}$$

and

$$\Im k_3 = \Re k_1 \Im k_2 + \Im k_1 \Re k_2. \tag{7.3.374}$$

Now comes an interesting construction: Given any $n \times n$ matrix k define, in terms of k, a $2n \times 2n$ *real* matrix $K(k)$ by the rule

$$K(k) = \begin{pmatrix} \Re k & -\Im k \\ \Im k & \Re k \end{pmatrix}. \tag{7.3.375}$$

Here each entry on the right side of (3.375) is an $n \times n$ block.[9] Verify that for *real* scalars λ there is the scalar multiplication result

$$K(\lambda k) = \lambda K(k). \tag{7.3.376}$$

Verify that there is the additive isomorphism

$$K(k_1 + k_2) = K(k_1) + K(k_2). \tag{7.3.377}$$

More remarkably, verify using (3.370) through (3.375) that there is the multiplicative isomorphism

$$K(k_1 k_2) = K(k_1)K(k_2). \tag{7.3.378}$$

That is, show that there is the matrix relation

$$\begin{pmatrix} \Re k_3 & -\Im k_3 \\ \Im k_3 & \Re k_3 \end{pmatrix} = \begin{pmatrix} \Re k_1 & -\Im k_1 \\ \Im k_1 & \Re k_1 \end{pmatrix} \begin{pmatrix} \Re k_2 & -\Im k_2 \\ \Im k_2 & \Re k_2 \end{pmatrix}. \tag{7.3.379}$$

Let $I^{[n]}$ and $I^{[2n]}$ be the $n \times n$ and $2n \times 2n$ identity matrices, respectively. Verify that

$$K(I^{[n]}) = I^{[2n]}. \tag{7.3.380}$$

Suppose that k is invertible. Verify that

$$K(k^{-1}) = [K(k)]^{-1}. \tag{7.3.381}$$

For another remarkable result, let W be the matrix defined by (3.9.12), which we write more precisely as

$$W = \frac{1}{\sqrt{2}} \begin{pmatrix} I^{[n]} & iI^{[n]} \\ iI^{[n]} & I^{[n]} \end{pmatrix}. \tag{7.3.382}$$

Verify the similarity transformation relation

$$\begin{aligned} WK(k)W^{-1} &= (1/2) \begin{pmatrix} I^{[n]} & iI^{[n]} \\ iI^{[n]} & I^{[n]} \end{pmatrix} \begin{pmatrix} \Re k & -\Im k \\ \Im k & \Re k \end{pmatrix} \begin{pmatrix} I^{[n]} & -iI^{[n]} \\ -iI^{[n]} & I^{[n]} \end{pmatrix} \\ &= (1/2) \begin{pmatrix} I^{[n]} & iI^{[n]} \\ iI^{[n]} & I^{[n]} \end{pmatrix} \begin{pmatrix} \Re k + i\Im k & -i\Re k - \Im k \\ \Im k - i\Re k & -i\Im k + \Re k \end{pmatrix} \\ &= \begin{pmatrix} \Re k + i\Im k & 0 \\ 0 & \Re k - i\Im k \end{pmatrix} = \begin{pmatrix} k & 0 \\ 0 & \bar{k} \end{pmatrix}. \end{aligned} \tag{7.3.383}$$

[9] Note that if K is known, then k can be found from a knowledge of the upper and lower left blocks of K. Thus the mapping between k and K is invertible.

Again, each block occurring in (3.383) is $n \times n$. For bonus points show from (3.383) that

$$\det[K(k)] = |\det(k)|^2. \tag{7.3.384}$$

(Recall Exercise 3.3.2.) Finally, associated with the additive and multiplicative isomorphisms (3.376) through (3.378), there are various a Lie algebraic/group properties. Suppose b_1 and b_2 are two possibly complex $n \times n$ matrices. Using (3.376) through (3.378), verify that

$$\begin{aligned} K(\{b_1, b_2\}) &= K(b_1 b_2 - b_2 b_1) = K(b_1 b_2) - K(b_2 b_1) \\ &= K(b_1)K(b_2) - K(b_2)K(b_1) \\ &= \{K(b_1), K(b_2)\}. \end{aligned} \tag{7.3.385}$$

Thus K is a Lie product (commutator) isomorphism. Suppose b is a possibly complex $n \times n$ matrix. Let us try to evaluate $K[\exp(b)]$. Verify that there is the result

$$K[\exp(b)] = K[\sum_{\ell}(1/\ell!)b^{\ell}] = \sum_{\ell}(1/\ell!)K(b^{\ell}) = \sum_{\ell}(1/\ell!)[K(b)]^{\ell} = \exp[K(b)], \tag{7.3.386}$$

which provides a relation between group elements $\exp(b)$ and group elements $\exp[K(b)]$.

7.3.35. Review Exercise 7.3.34. The purpose of this exercise is to apply the matrix machinery developed there to the Lorentz group. Make the Ansatz

$$k = \exp(\lambda \boldsymbol{m} \cdot \check{\boldsymbol{N}}) \exp(\theta \boldsymbol{n} \cdot \check{\boldsymbol{L}}) \tag{7.3.387}$$

so that the right side of (3.364) can be rewritten in the form

$$\hat{\boldsymbol{F}}^{+} = k \boldsymbol{F}^{+}. \tag{7.3.388}$$

Decompose k into real and imaginary parts by writing

$$k = \Re k + i \Im k. \tag{7.3.389}$$

Verify that by so doing (3.388) can be rewritten in the form

$$\hat{\boldsymbol{E}} + ic\hat{\boldsymbol{B}} = (\Re k + i\Im k)(\boldsymbol{E} + ic\boldsymbol{B}) = (\Re k\boldsymbol{E} - \Im kc\boldsymbol{B}) + i(\Im k\boldsymbol{E} + \Re kc\boldsymbol{B}). \tag{7.3.390}$$

Now equate real and imaginary parts of (3.390) to obtain, because $\boldsymbol{E}$ and $c\boldsymbol{B}$ are real, the relations

$$\hat{\boldsymbol{E}} = \Re k\boldsymbol{E} - \Im kc\boldsymbol{B}, \tag{7.3.391}$$

$$c\hat{\boldsymbol{B}} = \Im k\boldsymbol{E} + \Re kc\boldsymbol{B}. \tag{7.3.392}$$

Introduce a *real* six-component vector u by the rule

$$u = (E_x, E_y, E_z; cB_x, cB_y, cB_z). \tag{7.3.393}$$

Verify that the relations (3.391) and (3.392) can be summarized in the form

$$\hat{u} = K(k)u. \tag{7.3.394}$$

Note that all quantities appearing in (3.394) are real. Verify from (3.384) that if (3.387) holds, then $\det[K(k)] = 1$. Moreover, in view of (3.378), the real 6×6 matrices $K(k)$ provide a representation of the Lorentz group; and evidently this is the representation carried by the $F^{\mu\nu}$.

Finally, look at (3.383) whose far right side involves k and $\bar{k}$. Verify from (3.367) and (3.387) that

$$\bar{k} = \exp(\lambda\boldsymbol{m}\cdot\bar{\check{\boldsymbol{N}}})\exp(\theta\boldsymbol{n}\cdot\bar{\check{\boldsymbol{L}}}) = \exp(\lambda\boldsymbol{m}\cdot\grave{\check{\boldsymbol{N}}})\exp(\theta\boldsymbol{n}\cdot\grave{\check{\boldsymbol{L}}}). \tag{7.3.395}$$

We already know, from the work of Exercise 3.31 and the earlier discussion in this exercise, that $\check{\boldsymbol{L}}$ and $\check{\boldsymbol{N}}$ and hence k carry the $\Gamma(0,1)$ representation of the Lorentz Lie algebra/group, and $\grave{\check{\boldsymbol{L}}}$ and $\grave{\check{\boldsymbol{N}}}$ and hence $\bar{k}$ carry the $\Gamma(1,0)$ representation of the Lorentz Lie algebra/group. Note that the right side of (3.383) is *block diagonal* for all k and that W as given by (3.382) is *fixed* and thus *independent* of k. Therefore the representation carried by the $K(k)$ is *reducible*. It follows that the representation carried by the $K(k)$ and hence by the $F^{\mu\nu}$ is *equivalent* to the *direct sum* representation

$$\Gamma(0,1)\ \oplus\ \Gamma(1,0) = \Gamma(0,1)\ \oplus\ \bar{\Gamma}(0,1) = \bar{\Gamma}(1,0)\ \oplus\ \Gamma(1,0). \tag{7.3.396}$$

Here use has also been made of (3.273). Moreover, upon combining (3.383), (3.387), and (3.395), we find for Lorentz group elements the relation

$$WK[\exp(\lambda\boldsymbol{m}\cdot\check{\boldsymbol{N}})\exp(\theta\boldsymbol{n}\cdot)\check{\boldsymbol{L}})]W^{-1} \;=$$

$$\begin{pmatrix} \exp(\lambda\boldsymbol{m}\cdot\check{\boldsymbol{N}})\exp(\theta\boldsymbol{n}\cdot\check{\boldsymbol{L}}) & 0 \\ 0 & \exp(\lambda\boldsymbol{m}\cdot\bar{\check{\boldsymbol{N}}})\exp(\theta\boldsymbol{n}\cdot\bar{\check{\boldsymbol{L}}}) \end{pmatrix}. \tag{7.3.397}$$

So far we have been employing the construction/definition (3.375) to set up the *group* isomorphism (3.378) between the complex 3×3 matrices k given by (3.387) that provide the $\Gamma(0,1)$ representation of the Lorentz group and the real 6×6 matrices $K(k)$ that provide a representation that is equivalent to the representation (3.396). To complete the story, we would like to have a corresponding *Lie algebraic* isomorphism. That is what (3.385) does. Verify that from (3.281) through (3.283) and (3.385) it follows that

$$\{K(\check{L}^j), K(\check{L}^k)\} = \sum_\ell \epsilon_{jk\ell}K(\check{L}^\ell), \tag{7.3.398}$$

$$\{K(\check{L}^j), K(\check{N}^k)\} = \sum_\ell \epsilon_{jk\ell}K(\check{N}^\ell), \tag{7.3.399}$$

$$\{K(\check{N}^j), K(\check{N}^k)\} = -\sum_\ell \epsilon_{jk\ell}K(\check{L}^\ell). \tag{7.3.400}$$

Moreover, the matrices $K(\check{L}^j)$ and $K(\check{N}^j)$ are all real: Verify, since the $\check{L}^j$ are real, it follows that

$$K(\check{L}^j) = \begin{pmatrix} \check{L}^j & 0 \\ 0 & \check{L}^j \end{pmatrix}. \tag{7.3.401}$$

Since the $\check{N}^j$ are pure imaginary, see (3.280), verify that

$$K(\check{N}^j) = \begin{pmatrix} 0 & -\check{L}^j \\ \check{L}^j & 0 \end{pmatrix}. \tag{7.3.402}$$

Consequently, the $K(\check{L}^j)$ and $K(\check{N}^j)$ provide a representation of the Lorentz group Lie algebra by 6×6 real matrices. (Is this representation related to the adjoint representation?) Finally, for any 3×3 matrix b, there is the relation

$$WK(b)W^{-1} = \begin{pmatrix} b & 0 \\ 0 & \bar{b} \end{pmatrix}. \tag{7.3.403}$$

See (3.383). It follows that the $K(\check{L}^j)$ and $K(\check{N}^j)$ provide, for the Lorentz group Lie algebra, a representation that is equivalent to the representation (3.396).

7.3.36. Review Exercise 3.7.41. It examines the relation between the group commutator and the associated Lie-algebraic commutator. In this exercise you will evaluate the Lorentz group commutator for *boosts* along the 3 and 1 axes.[10] Recall from (3.185) that the generators for these boosts satisfy the commutation rule

$$\{N^3, N^1\} = -L^2, \tag{7.3.404}$$

and the fact that this commutator involves a rotation generator is the source of Wigner (1902-1995) rotations and Thomas (1903-1992) precession. According to (3.184) the remaining commutation rules among the generators N^3, N^1, L^2 are

$$\{L^2, N^3\} = N^1, \tag{7.3.405}$$

$$\{L^2, N^1\} = -N^3. \tag{7.3.406}$$

Your task is to evaluate the Lorentz *group* commutator

$$h(s) = \exp(-sN^1)\exp(-sN^3)\exp(sN^1)\exp(sN^3). \tag{7.3.407}$$

That is, your task is to find the net effect of a boost along the 3 axis with rapidity s followed by a boost along the 1 axis with the same rapidity followed by a boost along the 3 axis with rapidity $-s$ finally followed by a boost along the 1 axis with rapidity $-s$. Roughly speaking, we may refer to this operation as the *concatenation* of four boosts along the sides of a square. Note that the rapidities along the 3 and 1 axes add to zero, and therefore we might intuit that there is no net boost. However, in view of (3.251), we might expect some net rotation. Finally note that, in view of the work of Exercise 3.28, you will *equivalently* be exploring some properties of $Sp(2,\mathbb{R})$. That is, (3.254) is also an $Sp(2,\mathbb{R})$ group commutator.

As an initial exploratory step, your first sub-task is to find $h(s)$ through terms of $O(s^3)$ making use of the BCH formula (3.7.41). This may be done in steps: Verify that

$$\begin{aligned}
&\exp(sN^1)\exp(sN^3) = \exp[sN^1 + sN^3 + (s^2/2)\{N^1, N^3\} \\
&+(s^3/12)\{N^1,\{N^1,N^3\}\} + (s^3/12)\{N^3,\{N^3,N^1\}\} + O(s^4)] = \\
&\exp[sN^1 + sN^3 + (s^2/2)L^2 \\
&+(s^3/12)\{N^1, L^2\} - (s^3/12)\{N^3, L^2\} + O(s^4)] = \\
&\exp[sN^1 + sN^3 + (s^2/2)L^2 + (s^3/12)N^3 + (s^3/12)N^1 + O(s^4)].
\end{aligned} \tag{7.3.408}$$

[10] Why not choose the 1 and 2 axes which, from a notational perspective, would be more natural? Although physically there should be no difference, we will see that our choice is computationally more convenient.

Similarly, verify that

$$\begin{aligned}&\exp(-sN^1)\exp(-sN^3) = \exp[-sN^1 - sN^3 + (s^2/2)\{N^1, N^3\}\\&-(s^3/12)\{N^1, \{N^1, N^3\}\} - (s^3/12)\{N^3, \{N^3, N^1\}\} + O(s^4)] =\\&\exp[-sN^1 - sN^3 + (s^2/2)L^2\\&-(s^3/12)\{N^1, L^2\} + (s^3/12)\{N^3, L^2\} + O(s^4)] =\\&\exp[-sN^1 - sN^3 + (s^2/2)L^2 - (s^3/12)N^3 - (s^3/12)N^1 + O(s^4)].\end{aligned} \tag{7.3.409}$$

Next show that (3.255) and (3.256) may be combined to give the result

$$\begin{aligned}&h(s) = \exp(-sN^1)\exp(-sN^3)\exp(sN^1)\exp(sN^3) =\\&\exp[-sN^1 - sN^3 + (s^2/2)L^2 - (s^3/12)N^3 - (s^3/12)N^1 + O(s^4)] \times\\&\exp[sN^1 + sN^3 + (s^2/2)L^2 + (s^3/12)N^3 + (s^3/12)N^1 + O(s^4)] =\\&\exp[-sN^1 - sN^3 + (s^2/2)L^2 + O(s^4)]\exp[sN^1 + sN^3 + (s^2/2)L^2 + O(s^4)] =\\&\exp[s^2L^2 + (1/2)\{-sN^1 - sN^3 + (s^2/2)L^2, sN^1 + sN^3 + (s^2/2)L^2\} + O(s^4)] =\\&\exp[s^2L^2 + (s^3/2)\{L^2, N^1 + N^3\} + O(s^4)] =\\&\exp[-(s^3/2)N^3 + (s^3/2)N^1 + O(s^4)]\exp[s^2L^2 + O(s^4)].\end{aligned} \tag{7.3.410}$$

[Note that, in passing from lines 2 and 3 in (3.257) to subsequent lines, the terms proportional to $s^3/12$ cancel through $O(s^3)$.] Observe that the far right side of (3.257) is written in polar form. [See Subsection 4.2.2, Exercise 4.2.5, Exercise 7.3.30, (3.186), and (3.241).] Therefore, through terms of $O(s^3)$, the grand result of the four boosts appearing in the Lorentz group commutator is a net rotation about the 2 axis proportional to s^2 and a boost in the $-\boldsymbol{e}_3 + \boldsymbol{e}_1$ direction proportional to s^3. [Verify that the same result can be obtained using (3.7.241).] Our intuition about there being no net boost is wrong, but not entirely wrong since there is no boost term proportional to s. That there is a net boost at all is a relativistic, $O[(v/c)^3]$, effect.

The composition of non-colinear boosts appears to be a complicated matter. What can be said about higher order terms? As it stands, the evaluation of (3.254) requires working with 4×4 matrices. However, since what really matters in this situation are the commutation rules which we know are the same for the Lorentz group and $SL(2, \mathbb{C})$, we could equally well evaluate the associated $SL(2, \mathbb{C})$ group commutator

$$\hat{h}(s) = \exp(-s\hat{N}^1)\exp(-s\hat{N}^3)\exp(s\hat{N}^1)\exp(s\hat{N}^3). \tag{7.3.411}$$

This task is simpler because it involves only the exponentiation and multiplication of 2×2 matrices, which will make it easier to work to higher order in s. Recall (3.232) through (3.237) and the results of Exercise 5.7.7 that summaries various properties of the Pauli matrices. Moreover, in view of (3.233), (3.235), and (2.327), we will be able to work with *real* matrices, which simplifies numerical computation. This fact is the actual reason for choosing the boosts to be along the 3 and 1 axes. Our choice was dictated by the way in which the Pauli matrices have been defined.

Still, there is work to be done. Verify the following preliminary results:

$$\begin{aligned}\exp(s\hat{N}^3) &= \exp[(s/2)\sigma^3] = \cosh[(s/2)\sigma^3] + \sinh[(s/2)\sigma^3]\\&= I\cosh(s/2) + \sigma^3\sinh(s/2),\end{aligned} \tag{7.3.412}$$

$$\exp(s\hat{N}^1) = I\cosh(s/2) + \sigma^1\sinh(s/2), \tag{7.3.413}$$

$$\exp(-s\hat{N}^3) = I\cosh(s/2) - \sigma^3\sinh(s/2), \tag{7.3.414}$$

$$\exp(-s\hat{N}^1) = I\cosh(s/2) - \sigma^1\sinh(s/2); \tag{7.3.415}$$

$$\begin{aligned}&\exp(s\hat{N}^1)\exp(s\hat{N}^3) = [I\cosh(s/2) + \sigma^1\sinh(s/2)][I\cosh(s/2) + \sigma^3\sinh(s/2)]\\ &= I\cosh^2(s/2) + (\sigma^3+\sigma^1)\cosh(s/2)\sinh(s/2) + \sigma^1\sigma^3\sinh^2(s/2)\\ &= I\cosh^2(s/2) + (1/2)(\sigma^3+\sigma^1)\sinh(s) - i\sigma^2\sinh^2(s/2),\end{aligned} \tag{7.3.416}$$

$$\begin{aligned}&\exp(-s\hat{N}^1)\exp(-s\hat{N}^3) =\\ &I\cosh^2(s/2) - (1/2)(\sigma^3+\sigma^1)\sinh(s) - i\sigma^2\sinh^2(s/2).\end{aligned} \tag{7.3.417}$$

Next show that combining (3.263) and (3.264) yields the result

$$\begin{aligned}&\exp(-s\hat{N}^1)\exp(-s\hat{N}^3)\exp(s\hat{N}^1)\exp(s\hat{N}^3) =\\ &[I\cosh^2(s/2) - (1/2)(\sigma^3+\sigma^1)\sinh(s) - i\sigma^2\sinh^2(s/2)]\times\\ &[I\cosh^2(s/2) + (1/2)(\sigma^3+\sigma^1)\sinh(s) - i\sigma^2\sinh^2(s/2)] =\\ &\Big[[I\cosh^2(s/2) - i\sigma^2\sinh^2(s/2)] - [(1/2)(\sigma^3+\sigma^1)\sinh(s)]\Big]\times\\ &\Big[[I\cosh^2(s/2) - i\sigma^2\sinh^2(s/2)] + [(1/2)(\sigma^3+\sigma^1)\sinh(s)]\Big] =\\ &[I\cosh^2(s/2) - i\sigma^2\sinh^2(s/2)]^2 - [(1/2)(\sigma^3+\sigma^1)\sinh(s)]^2\\ &-(i/2)\{\sigma^2,(\sigma^3+\sigma^1)\}\sinh^2(s/2)\sinh(s).\end{aligned} \tag{7.3.418}$$

Look at the contents of the last line of (3.265). It has three pieces, each of which can be expanded/simplified. Verify that for the first piece there is the result

$$\begin{aligned}&[I\cosh^2(s/2) - i\sigma^2\sinh^2(s/2)]^2 =\\ &I\cosh^4(s/2) - I\sinh^4(s/2) - 2i\sigma^2\cosh^2(s/2)\sinh^2(s/2) =\\ &I[\cosh^2(s/2) - \sinh^2(s/2)][\cosh^2(s/2) + \sinh^2(s/2)]\\ &-2i\sigma^2[\cosh(s/2)\sinh(s/2)]^2 = I\cosh(s) - (i/2)\sigma^2\sinh^2(s).\end{aligned} \tag{7.3.419}$$

Verify that for the second and third pieces there are the results

$$-[(1/2)(\sigma^3+\sigma^1)\sinh(s)]^2 = -I(1/2)\sinh^2(s) \tag{7.3.420}$$

and

$$-(i/2)\{\sigma^2,(\sigma^3+\sigma^1)\}\sinh^2(s/2)\sinh(s) = -(\sigma^3-\sigma^1)\sinh^2(s/2)\sinh(s). \tag{7.3.421}$$

Show that adding the results for these three pieces yields the final result

$$\begin{aligned}&\hat{h}(s) = \exp(-s\hat{N}^1)\exp(-s\hat{N}^3)\exp(s\hat{N}^1)\exp(s\hat{N}^3) = \\ &I\cosh(s) - (i/2)\sigma^2\sinh^2(s) \\ &-I(1/2)\sinh^2(s) \\ &-(\sigma^3 - \sigma^1)\sinh^2(s/2)\sinh(s) = \\ &I[\cosh(s) - (1/2)\sinh^2(s)] - (i/2)\sigma^2\sinh^2(s) \\ &-(\sigma^3 - \sigma^1)\sinh^2(s/2)\sinh(s) = \\ &I[\cosh(s) - (1/2)\sinh^2(s)] + \hat{L}^2\sinh^2(s) \\ &-2(\hat{N}^3 - \hat{N}^1)\sinh^2(s/2)\sinh(s).\end{aligned} \tag{7.3.422}$$

Examination of the far right side of the result (3.269) shows signs of a rotation about the 2 axis and a boost in the $-\boldsymbol{e}_3 + \boldsymbol{e}_1$ direction. But the result as it stands is not particularly illuminating because it is not written in polar form. What is needed to further clarify the situation is a polar decomposition for $\hat{h}(s)$ of the form

$$\hat{h}(s) = \exp(\lambda_3\hat{N}^3 + \lambda_1\hat{N}^1)\exp(\theta\hat{L}^2). \tag{7.3.423}$$

In this decomposition, which we know by Section 4.2 is always possible, the quantities (λ_3, λ_1) specify the net boost, and θ specifies the net rotation. Only then can definitive statements be made.[11] You will have the privilege of making this decomposition in the next exercise. Also, a numerical calculation will be described that confirms the correctness of our results.

In the meantime, your next subtask is to show that (3.257) and (3.269) agree through terms of order s^3. Verify, through terms of order s^3, that (3.269) has the expansion

$$\begin{aligned}&\hat{h}(s) = \exp(-s\hat{N}^1)\exp(-s\hat{N}^3)\exp(s\hat{N}^1)\exp(s\hat{N}^3) = \\ &I[\cosh(s) - (1/2)\sinh^2(s)] + \hat{L}^2\sinh^2(s) \\ &-2(\hat{N}^3 - \hat{N}^1)\sinh^2(s/2)\sinh(s) = \\ &I + s^2\hat{L}^2 - (s^3/2)(\hat{N}^3 - \hat{N}^1) + O(s^4).\end{aligned} \tag{7.3.424}$$

Using the BCH formula show that (3.271) can be rewritten in the form

$$\begin{aligned}&\hat{h}(s) = \exp(-s\hat{N}^1)\exp(-s\hat{N}^3)\exp(s\hat{N}^1)\exp(s\hat{N}^3) = \\ &I + s^2\hat{L}^2 - (s^3/2)(\hat{N}^3 - \hat{N}^1) + O(s^4) = \\ &[I - (s^3/2)\hat{N}^3 + (s^3/2)\hat{N}^1 + O(s^4)][I + s^2\hat{L}^2 + O(s^4)] = \\ &\exp[-(s^3/2)\hat{N}^3 + (s^3/2)\hat{N}^1 + O(s^4)]\exp[s^2\hat{L}^2 + O(s^4)],\end{aligned} \tag{7.3.425}$$

in agreement with (3.257). Verity, therefore, that

$$\lambda_3 = -s^3/2 + O(s^4), \tag{7.3.426}$$

[11] Moreover, although the matrices (N^3, N^1, L^2) and $(\hat{N}^3, \hat{N}^1, \hat{L}^2)$ obey the same *Lie* algebra, they do not obey the same *"multiplicative"* algebra. For example, $(\hat{N}^1)^2 = I/4$ but $(N^1)^2 \neq I/4$. Therefore, we should only expect agreement/correspondence when comparing Lie algebraic quantities in some Lie algebraic form, such as comparing $h(s)$ and $\hat{h}(s)$ in polar form.

$$\lambda_1 = s^3/2 + O(s^4), \tag{7.3.427}$$

$$\theta = s^2 + O(s^4). \tag{7.3.428}$$

7.3.37. Review Exercise 3.31. As described at the end of that exercise, what would be definitive would be to have a polar decomposition for $\hat{h}(s)$ of the form (3.270). The purpose of this exercise is to find and check this desired polar decomposition.

In view of (3.232) through (3.237), let us make a few cosmetic changes in (3.270) that will simplify our computations. What we will seek are quantities $(\bar{\lambda}_3, \bar{\lambda}_1, \bar{\theta})$ such that

$$\hat{h}(s) = \exp(\bar{\lambda}_3\sigma^3 + \bar{\lambda}_1\sigma^1)\exp(\bar{\theta}i\sigma^2). \tag{7.3.429}$$

That is, in passing from (3.270) to (3.276) and in view of (3.232) through (3.237), we have made the (temporary) substitutions

$$\bar{\lambda}_3 = (1/2)\lambda_3 \;\Leftrightarrow\; \lambda_3 = 2\bar{\lambda}_3, \tag{7.3.430}$$

$$\bar{\lambda}_1 = (1/2)\lambda_1 \;\Leftrightarrow\; \lambda_1 = 2\bar{\lambda}_1, \tag{7.3.431}$$

$$\bar{\theta} = -(1/2)\theta \;\Leftrightarrow\; \theta = -2\bar{\theta}. \tag{7.3.432}$$

We will also use the notation

$$\lambda = (\lambda_3^2 + \lambda_1^2)^{1/2}, \tag{7.3.433}$$

$$\bar{\lambda} = (\bar{\lambda}_3^2 + \bar{\lambda}_1^2)^{1/2}, \tag{7.3.434}$$

so that

$$\bar{\lambda} = (1/2)\lambda \;\Leftrightarrow\; \lambda = 2\bar{\lambda}. \tag{7.3.435}$$

At this point we could invoke the machinery of Exercise 4.2.5 with the hope of working out the desired results. Instead, let us take a different tack. Observe that the penultimate equation in (3.269), which reads

$$\begin{aligned}
&\hat{h}(s) = \\
&I[\cosh(s) - (1/2)\sinh^2(s)] \\
&-(i/2)\sigma^2\sinh^2(s) \\
&-\sigma^3\sinh^2(s/2)\sinh(s) \\
&+\sigma^1\sinh^2(s/2)\sinh(s),
\end{aligned} \tag{7.3.436}$$

provides a Pauli matrix expansion for $\hat{h}(s)$. We will expand $\hat{h}(s)$ as given by (3.276) in terms of the Pauli matrices and then use the orthogonality properties (5.7.42) to match coefficients in the anticipated Pauli matrix expansion and that given by (3.283).

In order to expand (3.276) here are things for you to check: First verify that

$$\begin{aligned}
&\exp(\bar{\theta}i\sigma^2) = I + (\bar{\theta}i\sigma^2) + (\bar{\theta}i\sigma^2)^2/2! + (\bar{\theta}i\sigma^2)^3/3! + \cdots = \\
&I + \bar{\theta}(i\sigma^2) - I\bar{\theta}^2/2! - (i\sigma^2)(\bar{\theta}^3/3!) + \cdots = \\
&I(1 - \bar{\theta}^2/2! + \cdots) + i\sigma^2(\bar{\theta} - \bar{\theta}^3/3! + \cdots) = \\
&I\cos(\bar{\theta}) + i\sigma^2\sin(\bar{\theta}).
\end{aligned} \tag{7.3.437}$$

Next verify that

$$\begin{aligned}
&\exp(\bar\lambda_3\sigma^3+\bar\lambda_1\sigma^1)=I+(\bar\lambda_3\sigma^3+\bar\lambda_1\sigma^1)+(\bar\lambda_3\sigma^3+\bar\lambda_1\sigma^1)^2/2!+(\bar\lambda_3\sigma^3+\bar\lambda_1\sigma^1)^3/3!+\cdots=\\
&I+(\bar\lambda_3\sigma^3+\bar\lambda_1\sigma^1)+I(\bar\lambda_3^2+\bar\lambda_1^2)/2!+(\bar\lambda_3\sigma^1+\bar\lambda_1\sigma^1)(\bar\lambda_3^2+\bar\lambda_1^2)/3!+\cdots=\\
&I\cosh[(\bar\lambda_3^2+\bar\lambda_1^2)^{1/2}]+(\bar\lambda_3\sigma^3+\bar\lambda_1\sigma^1)(\bar\lambda_3^2+\bar\lambda_1^2)^{-1/2}\sinh[(\bar\lambda_3^2+\bar\lambda_1^2)^{1/2}]=\\
&I\cosh(\bar\lambda)+(\bar\lambda_3\sigma^3+\bar\lambda_1\sigma^1)(1/\bar\lambda)\sinh(\bar\lambda).
\end{aligned}\tag{7.3.438}$$

Finally, verify that

$$\begin{aligned}
&\hat h(s)=\exp(\bar\lambda_3\sigma^3+\bar\lambda_1\sigma^1)\exp(\bar\theta i\sigma^2)=\\
&[I\cosh(\bar\lambda)+(\bar\lambda_3\sigma^3+\bar\lambda_1\sigma^1)(1/\bar\lambda)\sinh(\bar\lambda)][I\cos(\bar\theta)+i\sigma^2\sin(\bar\theta)]=\\
&I\cosh(\bar\lambda)\cos(\bar\theta)\\
&+i\sigma^2\cosh(\bar\lambda)\sin(\bar\theta)\\
&+(\bar\lambda_3\sigma^3+\bar\lambda_1\sigma^1)(1/\bar\lambda)\sinh(\bar\lambda)\cos(\bar\theta)+\\
&(\bar\lambda_3\sigma^3+\bar\lambda_1\sigma^1)(1/\bar\lambda)\sinh(\bar\lambda)i\sigma^2\sin(\bar\theta)=\\
&I\cosh(\bar\lambda)\cos(\bar\theta)\\
&+i\sigma^2\cosh(\bar\lambda)\sin(\bar\theta)\\
&+\sigma^3[\bar\lambda_3\cos(\bar\theta)-\bar\lambda_1\sin(\bar\theta)](1/\bar\lambda)\sinh(\bar\lambda)\\
&+\sigma^1[\bar\lambda_3\sin(\bar\theta)+\bar\lambda_1\cos(\bar\theta)](1/\bar\lambda)\sinh(\bar\lambda).
\end{aligned}\tag{7.3.439}$$

Now equate terms in the Pauli matrix expansion (3.283) with like terms in the Pauli matrix expansion (3.286). Show that so doing yields the relations

$$\cosh(\bar\lambda)\cos(\bar\theta)=\cosh(s)-(1/2)\sinh^2(s),\tag{7.3.440}$$

$$\cosh(\bar\lambda)\sin(\bar\theta)=-(1/2)\sinh^2(s),\tag{7.3.441}$$

$$[\bar\lambda_3\cos(\bar\theta)-\bar\lambda_1\sin(\bar\theta)](1/\bar\lambda)\sinh(\bar\lambda)=-\sinh^2(s/2)\sinh(s),\tag{7.3.442}$$

$$[\bar\lambda_3\sin(\bar\theta)+\bar\lambda_1\cos(\bar\theta)](1/\bar\lambda)\sinh(\bar\lambda)=\sinh^2(s/2)\sinh(s).\tag{7.3.443}$$

The terms on the left sides of these relations, which are the unknown terms, come from (3.286). And the terms on the right sides, which are known, come from (3.283).

The last step is to solve (3.287) through (3.290) for $\bar\lambda_3,\bar\lambda_1$, and $\bar\theta$. Upon dividing (3.288) by (3.287), show that

$$\tan(\bar\theta)=-(1/2)\sinh^2(s)/[\cosh(s)-(1/2)\sinh^2(s)].\tag{7.3.444}$$

Upon squaring (3.287) and (3.288) and adding the results, show that

$$\cosh^2(\bar\lambda)=[\cosh(s)-(1/2)\sinh^2(s)]^2+(1/4)[\sinh(s)]^4.\tag{7.3.445}$$

The results (3.291) and (3.292) determine $\bar\theta$ and $\bar\lambda$ as functions of s.

We would also like formulas for $\bar\lambda_3$ and $\bar\lambda_1$ as functions of s. Upon multiplying (3.289) by $\cos(\bar\theta)$ and (3.290) by $\sin(\bar\theta)$ and adding the results, show that

$$\bar\lambda_3(1/\bar\lambda)\sinh(\bar\lambda)=[-\cos(\bar\theta)+\sin(\bar\theta)]\sinh^2(s/2)\sinh(s)\tag{7.3.446}$$

so that

$$\bar{\lambda}_3 = [-\cos(\bar{\theta}) + \sin(\bar{\theta})][\sinh^2(s/2)\sinh(s)]/[(1/\bar{\lambda})\sinh(\bar{\lambda})]. \tag{7.3.447}$$

Upon multiplying (3.289) by $[-\sin(\bar{\theta})]$ and (3.290) by $\cos(\bar{\theta})$ and adding the results, show that

$$\bar{\lambda}_1(1/\bar{\lambda})\sinh(\bar{\lambda}) = [\cos(\bar{\theta}) + \sin(\bar{\theta})]\sinh^2(s/2)\sinh(s) \tag{7.3.448}$$

so that

$$\bar{\lambda}_1 = [\cos(\bar{\theta}) + \sin(\bar{\theta})][\sinh^2(s/2)\sinh(s)]/[(1/\bar{\lambda})\sinh(\bar{\lambda})]. \tag{7.3.449}$$

Considerable algebra has gone by in passing from (3.270) to (3.296). How do we know there have been no mistakes in deriving the results (3.291), (3.292), (3.294), and (3.296)? Shortly, as a sanity check, we will make low-order expansions in s for these results to verify that these expansions agree with what we already know. Then we will also describe numerical checks.

But before doing so, and assuming the correctness of (3.294) and (3.296), it is instructive to examine the direction of the net boost associated with the Lorentz group commutator. In view of (3.277), (3.278), and (3.280) through (3.282), define an angle χ by the relations

$$\cos(\chi) = \bar{\lambda}_3/\bar{\lambda} = \lambda_3/\lambda, \tag{7.3.450}$$

$$\sin(\chi) = \bar{\lambda}_1/\bar{\lambda} = \lambda_1/\lambda, \tag{7.3.451}$$

$$\tan(\chi) = \bar{\lambda}_1/\bar{\lambda}_3 = \lambda_1/\lambda_3. \tag{7.3.452}$$

Show from (3.294), (3.296), and (3.299) that

$$\begin{aligned}\tan(\chi) &= -[\cos(\bar{\theta}) + \sin(\bar{\theta})]/[\cos(\bar{\theta}) - \sin(\bar{\theta})] = \\ &= -[\cos(\theta/2) - \sin(\theta/2)]/[\cos(\theta/2) + \sin(\theta/2)]. \end{aligned}\tag{7.3.453}$$

We have learned that the direction of the net boost depends simply on $\bar{\theta}$ or, equivalently, on θ.

As promised, we now turn to making low-order expansions in s. Begin by verifying the expansions

$$\tan^{-1}(\psi) = \psi - \psi^3/3 + \cdots, \tag{7.3.454}$$

$$-(1/2)\sinh^2(s)/[\cosh(s) - (1/2)\sinh^2(s)] = -(1/2)s^2 + O(s^4). \tag{7.3.455}$$

Show if follows from (3.291) that

$$\bar{\theta} = -(1/2)s^2 + O(s^4). \tag{7.3.456}$$

Verify that employing (3.279) in (3.299) yields the result

$$\theta = s^2 + O(s^4), \tag{7.3.457}$$

in agreement with (3.275).

Next consider the relation (3.292). Verify, say with the use of *Mathematica* or by hand calculation, that there is the expansion

$$[\cosh(s) - (1/2)\sinh^2(s)]^2 + (1/4)[\sinh(s)]^4 = 1 + s^6/8 + O(s^8). \tag{7.3.458}$$

Note the remarkable fact that the coefficients of s^2 and s^4 vanish! Verify also the expansions

$$\cosh(\bar{\lambda}) = 1 + \bar{\lambda}^2/2! + O(\bar{\lambda}^4), \tag{7.3.459}$$

$$\cosh^2(\bar{\lambda}) = 1 + \bar{\lambda}^2 + O(\bar{\lambda}^4). \tag{7.3.460}$$

Show it follows, using (3.292) and (3.301 through (3.303), that

$$\bar{\lambda}^2 = s^6/8 + O(s^8), \tag{7.3.461}$$

and consequently

$$\bar{\lambda}(s) = |s^3|(1/\sqrt{8}) + O(s^5). \tag{7.3.462}$$

Finally, with the use of (3.282), show that

$$\lambda(s) = |s^3|(1/\sqrt{2}) + O(s^5). \tag{7.3.463}$$

How does the result (3.061) compare with (3.273) and (3.274)? From (3.273), (3.274), and (3.280) show that

$$\lambda(s) = |s^3|(1/\sqrt{2}) + O(s^4). \tag{7.3.464}$$

Evidently (3.306) and (3.307) are consistent.

What remains is to examine the small s behavior of $\bar{\lambda}_3$ and $\bar{\lambda}_1$ as given by (3.294) and (3.296). Begin by verifying the expansions

$$\cos(\bar{\theta}) = 1 - \bar{\theta}^2/2 + \cdots = 1 - (1/8)s^4 + O(s^6), \tag{7.3.465}$$

$$\sin(\bar{\theta}) = \bar{\theta} - (\bar{\theta})^3/6 + \cdots = -(1/2)s^2 + O(s^4), \tag{7.3.466}$$

$$[\mp\cos(\bar{\theta}) + \sin(\bar{\theta})] = \mp 1 - (1/2)s^2 + O(s^4), \tag{7.3.467}$$

$$\sinh^2(s/2)\sinh(s) = s^3/4 + O(s^5), \tag{7.3.468}$$

$$(1/\bar{\lambda})\sinh(\bar{\lambda}) = 1 + O(s^6). \tag{7.3.469}$$

Show, therefore, that (3.299) and (3.301) have the expansions

$$\bar{\lambda}_3 = -s^3/4 - s^5/8 + \cdots, \tag{7.3.470}$$

$$\bar{\lambda}_1 = s^3/4 - s^5/8 + \cdots, \tag{7.3.471}$$

and consequently λ_3 and λ_1 have the expansions

$$\lambda_3 = -s^3/2 - s^5/4 + \cdots, \tag{7.3.472}$$

$$\lambda_1 = s^3/2 - s^5/4 + \cdots, \tag{7.3.473}$$

in agreement with (3.278) and (3.279).

The last item of interest with regard to low-order expansions is the behavior of $\chi(s)$. From (3.319) and (3.320) we see that

$$\begin{aligned}\tan(\chi) &= \lambda_1/\lambda_3 = (s^3/2 - s^5/4 + \cdots)/(-s^3/2 - s^5/4 + \cdots)\\ &= -(1 - s^2/2 + \cdots)/(1 + s^2/2 + \cdots) = -(1 - s^2 + \cdots)\\ &= -1 + s^2 + \cdots .\end{aligned} \tag{7.3.474}$$

As expected from (3.300), (3.303), and (3.304), χ is indeed s dependent.

To fulfill our last promise, we now describe how the results (3.291), (3.292), (3.294), and (3.296) can also be checked numerically for any s using the charged particle beam transport code MaryLie. Because it is based on Lie-algebraic methods, MaryLie is capable of performing various operations related to the symplectic group and, as we have seen, what we are seeking are various relations among elements of $Sp(2,\mathbb{R})$. In particular for our present purposes, MaryLie can perform the following operations numerically:

1. Given a quadratic polynomial f_2, it can compute the symplectic matrix M associated with the linear symplectic map $\mathcal{M}$ in the relation
$$\mathcal{M} = \exp(:f_2:). \tag{7.3.475}$$

2. It can multiply symplectic matrices thereby implementing group-element multiplication for the symplectic group.

3. Given a symplectic matrix M, let $\mathcal{M}$ be the associated linear symplectic map. With M as input, MaryLie can compute the quadratic polynomials f_2^a and f_2^c in the decomposition
$$\mathcal{M} = \exp(:f_2^c)\exp(:f_2^a:). \tag{7.3.476}$$
See Section 7.6. That is, MaryLie can carry out (orthogonal) polar decomposition of symplectic matrices.

How can these tools be employed in the present context? Examination of (5.6.6), (5.6.7), and (5.6.11) through (5.6.14) shows that there are the following correspondences between Pauli matrices and quadratic polynomials:

$$f = (1/2)(-q^2 + p^2) \leftrightarrow \sigma^1, \tag{7.3.477}$$

$$b_0 = (1/2)(q^2 + p^2) \leftrightarrow J = i\sigma^2, \tag{7.3.478}$$

$$g = qp \leftrightarrow \sigma^3. \tag{7.3.479}$$

(Recall that these correspondences were set up in Section 5.5.) Consequently, there is also the correspondence

$$\begin{aligned}&\hat{h}(s) = \exp(-s\hat{N}^1)\exp(-s\hat{N}^3)\exp(s\hat{N}^1)\exp(s\hat{N}^3) =\\ &\exp[(s/2)\sigma^1)]\exp[(s/2)\sigma^3)]\exp[(-s/2)\sigma^1)]\exp[(-s/2)\sigma^3)] \leftrightarrow\\ &\exp[(s/2):qp:]\exp[(s/4):-q^2+p^2:]\exp[(-s/2):qp:]\exp[(-s/4):-q^2+p^2:].\end{aligned} \tag{7.3.480}$$

[Note that the order of the Lie transformations appearing in the right side of the correspondence (3.327) is opposite to the order of the related matrices on the left side of (3.327). This reversal in order is to be expected. See the discussion in Section 8.3.] What we will use MaryLie to carry out numerically are the operations

$$\begin{aligned}&\exp[(s/2):qp:]\exp[(s/4):-q^2+p^2:]\exp[(-s/2):qp:]\exp[(-s/4):-q^2+p^2:]\\ &= \exp(:f_2^c)\exp(:f_2^a:).\end{aligned} \tag{7.3.481}$$

That is, it will compute and multiply the four maps on the left side of (3.328) and then express the result in the factored product form shown on the right side. Specifically, it will output the quadratic polynomials f_2^c and f_2^a. When this is done, in view of (3.276) and the correspondences (3.324) through (3.326), we expect f_2^c and f_2^a will be given by the relations

$$f_2^c = \bar{\theta}b_0 = \bar{\theta}(1/2)(q^2 + p^2) \tag{7.3.482}$$

and

$$f_2^a = \bar{\lambda}_3 g + \bar{\lambda}_1 f = \bar{\lambda}_3 qp + \bar{\lambda}_1(1/2)(-q^2 + p^2). \tag{7.3.483}$$

`Exhibit 7.3.1: Sample MaryLie Run`

7.3.38. Review Exercises 3.27, 3.31, and 3.32. Exercise 3.27 found, among other things, the concatenation rule for two collinear boosts: rapidities simply add. And Exercises 3.31 and 3.32 found the concatenation rule for four boosts along the sides of a square. The purpose of this exercise is to find the concatenation rule for two non-collinear boosts. Specifically, given two real three-component vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, we wish to study group element $k(\boldsymbol{\mu}, \boldsymbol{\nu})$ defined by the product

$$k(\boldsymbol{\mu}, \boldsymbol{\nu}) = \exp(\boldsymbol{\nu} \cdot \boldsymbol{N}) \exp(\boldsymbol{\mu} \cdot \boldsymbol{N}). \tag{7.3.484}$$

Observe that the vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ determine (or in the collinear case lie in) a plane which for convenience, and without loss of generality, may be taken to be the 3,1 plane. Therefore we may make the decompositions

$$\boldsymbol{\mu} = \mu_3 \boldsymbol{e}_3 + \mu_1 \boldsymbol{e}_1, \tag{7.3.485}$$

$$\boldsymbol{\nu} = \nu_3 \boldsymbol{e}_3 + \nu_1 \boldsymbol{e}_1, \tag{7.3.486}$$

so that

$$\boldsymbol{\mu} \cdot \boldsymbol{N} = \mu_3 N^3 + \mu_1 N^1, \tag{7.3.487}$$

$$\boldsymbol{\nu} \cdot \boldsymbol{N} = \nu_3 N^3 + \nu_1 N^1, \tag{7.3.488}$$

and

$$\boldsymbol{\mu} \times \boldsymbol{\nu} = (\mu_3\nu_1 - \mu_1\nu_3)\boldsymbol{e}_2. \tag{7.3.489}$$

What we are interested in is finding the vector

$$\boldsymbol{\lambda} = \lambda_3 \boldsymbol{e}_3 + \lambda_1 \boldsymbol{e}_1 \tag{7.3.490}$$

and the angle θ such that

$$\exp(\boldsymbol{\nu} \cdot \boldsymbol{N}) \exp(\boldsymbol{\mu} \cdot \boldsymbol{N}) = \exp(\boldsymbol{\lambda} \cdot \boldsymbol{N}) \exp(\theta L^2). \tag{7.3.491}$$

To get a feel for what to expect, we may use the BCH formula to combine exponents on the left side of (3.338). Verify, through terms quadratic in the components of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, that

$$\begin{aligned} &\exp(\boldsymbol{\nu} \cdot \boldsymbol{N}) \exp(\boldsymbol{\mu} \cdot \boldsymbol{N}) = \exp[(\boldsymbol{\mu} + \boldsymbol{\nu}) \cdot \boldsymbol{N} + \{\boldsymbol{\nu} \cdot \boldsymbol{N}, \boldsymbol{\mu} \cdot \boldsymbol{N}\}/2 + \cdots] = \\ &\exp[(\boldsymbol{\mu} + \boldsymbol{\nu}) \cdot \boldsymbol{N} + \cdots] \exp[\{\boldsymbol{\nu} \cdot \boldsymbol{N}, \boldsymbol{\mu} \cdot \boldsymbol{N}\}/2 + \cdots]. \end{aligned} \tag{7.3.492}$$

Next show that

$$\{\boldsymbol{\nu}\cdot\boldsymbol{N},\boldsymbol{\mu}\cdot\boldsymbol{N}\}=(\mu_3\nu_1-\mu_1\nu_3)L^2=(\boldsymbol{\mu}\times\boldsymbol{\nu})\cdot\boldsymbol{L}. \tag{7.3.493}$$

Conclude, upon comparing the right sides of (3.338) and (3.339), that there are the results

$$\boldsymbol{\lambda}=\boldsymbol{\mu}+\boldsymbol{\nu}+\cdots\,, \tag{7.3.494}$$

$$\theta=(1/2)(\boldsymbol{\mu}\times\boldsymbol{\nu})\cdot\boldsymbol{e_2}+\cdots=(1/2)(\mu_3\nu_1-\mu_1\nu_3)+\cdots\,. \tag{7.3.495}$$

What we are next interested in are the higher-order terms in (3.341) and (3.342).

Again, since what really matters in this situation are the commutation rules which we know are the same for the Lorentz group and $SL(2,\mathbb{C})$, we can equally well evaluate the simpler associated $SL(2,\mathbb{C})$ [and $Sp(2,\mathbb{R})$] function

$$\hat{k}(\boldsymbol{\mu},\boldsymbol{\nu})=\exp(\boldsymbol{\nu}\cdot\hat{\boldsymbol{N}})\exp(\boldsymbol{\mu}\cdot\hat{\boldsymbol{N}}). \tag{7.3.496}$$

In this case (3.338) becomes

$$\exp(\boldsymbol{\nu}\cdot\hat{\boldsymbol{N}})\exp(\boldsymbol{\mu}\cdot\hat{\boldsymbol{N}})=\exp(\boldsymbol{\lambda}\cdot\hat{\boldsymbol{N}})\exp(\theta\hat{L}^2) \tag{7.3.497}$$

or, in terms of Pauli matrices,

$$\exp(\boldsymbol{\nu}\cdot\boldsymbol{\sigma}/2)\exp(\boldsymbol{\mu}\cdot\boldsymbol{\sigma}/2)=\exp(\boldsymbol{\lambda}\cdot\boldsymbol{\sigma}/2)\exp(-\theta i\sigma^2/2). \tag{7.3.498}$$

To further simplify calculations, and in analogy to what was done in Exercise 3.32, we will also make the (temporary) substitutions

$$\bar{\boldsymbol{\mu}}=(1/2)\boldsymbol{\mu}\ \Leftrightarrow\ \boldsymbol{\mu}=2\bar{\boldsymbol{\mu}},\ \text{etc.}; \tag{7.3.499}$$

$$\bar{\theta}=-(1/2)\theta\ \Leftrightarrow\ \theta=-2\bar{\theta}, \tag{7.3.500}$$

so that

$$\exp(\boldsymbol{\nu}\cdot\boldsymbol{\sigma}/2)=\exp(\bar{\nu}_3\sigma^3+\bar{\nu}_1\sigma^1),\ \text{etc.} \tag{7.3.501}$$

and

$$\exp(-\theta i\sigma^2/2)=\exp(\bar{\theta}i\sigma^2). \tag{7.3.502}$$

Moreover, we will use the notation

$$\mu=(\mu_3^2+\mu_1^2)^{1/2},\ \text{etc.}, \tag{7.3.503}$$

$$\bar{\mu}=(\bar{\mu}_3^2+\bar{\mu}_1^2)^{1/2},\ \text{etc.}, \tag{7.3.504}$$

so that

$$\bar{\mu}=(1/2)\mu\ \Leftrightarrow\ \mu=2\bar{\mu},\ \text{etc.} \tag{7.3.505}$$

What we will seek to find are the quantities $(\bar{\lambda}_3,\bar{\lambda}_1,\bar{\theta})$ such that

$$\exp(\bar{\nu}_3\sigma^3+\bar{\nu}_1\sigma^1)\exp(\bar{\mu}_3\sigma^3+\bar{\mu}_1\sigma^1)=\exp(\bar{\lambda}_3\sigma^3+\bar{\lambda}_1\sigma^1)\exp(\bar{\theta}i\sigma^2). \tag{7.3.506}$$

Consult again Exercise 3.32 to observe that it contains almost all the results necessary to complete the current exercise. For example, in view of (3.285), there is the analogous result

$$\begin{aligned}
&\exp(\bar\nu_3\sigma^3 + \bar\nu_1\sigma^1) = \exp(\bar{\boldsymbol{\nu}}\cdot\boldsymbol{\sigma}) = \\
&I\cosh(\bar\nu) + (\bar\nu_3\sigma^3 + \bar\nu_1\sigma^1)(1/\bar\nu)\sinh(\bar\nu) = \\
&I\cosh(\bar\nu) + \bar{\boldsymbol{\nu}}\cdot\boldsymbol{\sigma}(1/\bar\nu)\sinh(\bar\nu),
\end{aligned} \tag{7.3.507}$$

and there is an analogous result involving $\bar{\boldsymbol{\mu}}$. Finally, the right side of (3.353) has already been treated in (3.286).

Carry out the next step by showing that

$$\begin{aligned}
&\exp(\bar\nu_3\sigma^3 + \bar\nu_1\sigma^1)\exp(\bar\mu_3\sigma^3 + \bar\mu_1\sigma^1) = \exp(\bar{\boldsymbol{\nu}}\cdot\boldsymbol{\sigma})\exp(\bar{\boldsymbol{\mu}}\cdot\boldsymbol{\sigma}) = \\
&[I\cosh(\bar\nu) + \bar{\boldsymbol{\nu}}\cdot\boldsymbol{\sigma}(1/\bar\nu)\sinh(\bar\nu)][I\cosh(\bar\mu) + \bar{\boldsymbol{\mu}}\cdot\boldsymbol{\sigma}(1/\bar\mu)\sinh(\bar\mu)] = \\
&I\cosh(\bar\nu)\cosh(\bar\mu) \\
&+\bar{\boldsymbol{\mu}}\cdot\boldsymbol{\sigma}\cosh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu) \\
&+\bar{\boldsymbol{\nu}}\cdot\boldsymbol{\sigma}\cosh(\bar\mu)(1/\bar\nu)\sinh(\bar\nu) \\
&+(\bar{\boldsymbol{\nu}}\cdot\boldsymbol{\sigma})\cdot(\bar{\boldsymbol{\mu}}\cdot\boldsymbol{\sigma})(1/\bar\nu)\sinh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu) = \\
&I\cosh(\bar\nu)\cosh(\bar\mu) \\
&+[\bar{\boldsymbol{\mu}}\cosh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu) + \bar{\boldsymbol{\nu}}\cosh(\bar\mu)(1/\bar\nu)\sinh(\bar\nu)]\cdot\boldsymbol{\sigma} \\
&+[I\bar{\boldsymbol{\nu}}\cdot\bar{\boldsymbol{\mu}} + i(\bar{\boldsymbol{\nu}}\times\bar{\boldsymbol{\mu}})\cdot\boldsymbol{\sigma}](1/\bar\nu)\sinh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu) = \\
&I[\cosh(\bar\nu)\cosh(\bar\mu) + \bar{\boldsymbol{\nu}}\cdot\bar{\boldsymbol{\mu}}(1/\bar\nu)\sinh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu)] \\
&+i\sigma^2(\bar\nu_3\bar\mu_1 - \bar\nu_1\bar\mu_3)(1/\bar\nu)\sinh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu) \\
&+[\bar\mu_3\cosh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu) + \bar\nu_3\cosh(\bar\mu)(1/\bar\nu)\sinh(\bar\nu)]\sigma^3 \\
&+[\bar\mu_1\cosh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu) + \bar\nu_1\cosh(\bar\mu)(1/\bar\nu)\sinh(\bar\nu)]\sigma^1.
\end{aligned} \tag{7.3.508}$$

Now equate terms in the Pauli matrix expansion (3.286) with like terms in the Pauli matrix expansion (3.355). Show that so doing yields the relations

$$\cosh(\bar\lambda)\cos(\bar\theta) = \cosh(\bar\nu)\cosh(\bar\mu) + \bar{\boldsymbol{\nu}}\cdot\bar{\boldsymbol{\mu}}(1/\bar\nu)\sinh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu), \tag{7.3.509}$$

$$\cosh(\bar\lambda)\sin(\bar\theta) = (\bar\nu_3\bar\mu_1 - \bar\nu_1\bar\mu_3)(1/\bar\nu)\sinh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu), \tag{7.3.510}$$

$$[\bar\lambda_3\cos(\bar\theta) - \bar\lambda_1\sin(\bar\theta)](1/\bar\lambda)\sinh(\bar\lambda) = \bar\mu_3\cosh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu) + \bar\nu_3\cosh(\bar\mu)(1/\bar\nu)\sinh(\bar\nu), \tag{7.3.511}$$

$$[\bar\lambda_3\sin(\bar\theta) + \bar\lambda_1\cos(\bar\theta)](1/\bar\lambda)\sinh(\bar\lambda) = \bar\mu_1\cosh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu) + \bar\nu_1\cosh(\bar\mu)(1/\bar\nu)\sinh(\bar\nu). \tag{7.3.512}$$

The terms on the left sides of these relations, which are the unknown terms, come from (3.286). And the terms on the right sides, which are known, come from (3.355).

The last step is to solve (3.356) through (3.359) for $\bar\lambda_3, \bar\lambda_1$, and $\bar\theta$. Upon dividing (3.357) by (3.356), show that

$$\begin{aligned}
\tan(\bar\theta) \;=\; &(\bar\nu_3\bar\mu_1 - \bar\nu_1\bar\mu_3)(1/\bar\nu)\sinh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu)\times \\
&[\cosh(\bar\nu)\cosh(\bar\mu) + \bar{\boldsymbol{\nu}}\cdot\bar{\boldsymbol{\mu}}(1/\bar\nu)\sinh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu)]^{-1}.
\end{aligned} \tag{7.3.513}$$

Upon squaring (3.356) and (3.357) and adding the results, show that

$$\begin{aligned}\cosh^2(\bar\lambda) &= [\cosh(\bar\nu)\cosh(\bar\mu) + \bar{\boldsymbol\nu}\cdot\bar{\boldsymbol\mu}(1/\bar\nu)\sinh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu)]^2 \\ &\quad +[(\bar\nu_3\bar\mu_1 - \bar\nu_1\bar\mu_3)(1/\bar\nu)\sinh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu)]^2. \end{aligned} \qquad (7.3.514)$$

The results (3.360) and (3.361) determine $\bar\theta$ and $\bar\lambda$ as functions of $\bar{\boldsymbol\mu}$ and $\bar{\boldsymbol\nu}$.

We would also like formulas for $\bar\lambda_3$ and $\bar\lambda_1$ as functions of $\bar{\boldsymbol\mu}$ and $\bar{\boldsymbol\nu}$. Upon multiplying (3.358) by $\cos(\bar\theta)$ and (3.359) by $\sin(\bar\theta)$ and adding the results, show that

$$\begin{aligned}\bar\lambda_3(1/\bar\lambda)\sinh(\bar\lambda) &= [\bar\mu_3\cos(\bar\theta) + \bar\mu_1\sin(\bar\theta)]\cosh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu) \\ &\quad +[\bar\nu_3\cos(\bar\theta) + \bar\nu_1\sin(\bar\theta)]\cosh(\bar\mu)(1/\bar\nu)\sinh(\bar\nu) \end{aligned} \qquad (7.3.515)$$

so that

$$\begin{aligned}\bar\lambda_3 &= \{[\bar\mu_3\cos(\bar\theta) + \bar\mu_1\sin(\bar\theta)]\cosh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu) \\ &\quad +[\bar\nu_3\cos(\bar\theta) + \bar\nu_1\sin(\bar\theta)]\cosh(\bar\mu)(1/\bar\nu)\sinh(\bar\nu)\} \times \\ &\quad [(1/\bar\lambda)\sinh(\bar\lambda)]^{-1}. \end{aligned} \qquad (7.3.516)$$

Upon multiplying (3.358) by $[-\sin(\bar\theta)]$ and (3.359) by $\cos(\bar\theta)$ and adding the results, show that

$$\begin{aligned}\bar\lambda_1(1/\bar\lambda)\sinh(\bar\lambda) &= [-\bar\mu_3\sin\bar\theta) + \bar\mu_1\cos(\bar\theta)]\cosh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu) \\ &\quad +[-\bar\nu_3\sin(\bar\theta) + \bar\nu_1\cos(\bar\theta)]\cosh(\bar\mu)(1/\bar\nu)\sinh(\bar\nu) \end{aligned} \qquad (7.3.517)$$

so that

$$\begin{aligned}\bar\lambda_1 &= \{[-\bar\mu_3\sin(\bar\theta) + \bar\mu_1\cos(\bar\theta)]\cosh(\bar\nu)(1/\bar\mu)\sinh(\bar\mu) \\ &\quad +[-\bar\nu_3\sin(\bar\theta) + \bar\nu_1\cos(\bar\theta)]\cosh(\bar\mu)(1/\bar\nu)\sinh(\bar\nu)\} \times \\ &\quad [(1/\bar\lambda)\sinh(\bar\lambda)]^{-1}. \end{aligned} \qquad (7.3.518)$$

The results (3.360), (3.361), (3.363), and (3.365) can also be checked numerically for any $\boldsymbol\mu$ and $\boldsymbol\nu$ using the charged particle beam transport code MaryLie. How can MaryLie tools be employed in the present context? Based on the correspondences (3.324) through (3.326) there is also the correspondence

$$\begin{aligned}&\exp(\boldsymbol\nu\cdot\hat{\boldsymbol N})\exp(\boldsymbol\mu\cdot\hat{\boldsymbol N}) = \exp(\boldsymbol\nu\cdot\boldsymbol\sigma/2)\exp(\boldsymbol\mu\cdot\boldsymbol\sigma/2) = \\ &\exp(\bar\nu_3\sigma^3 + \bar\nu_1\sigma^1)\exp(\bar\mu_3\sigma^3 + \bar\mu_1\sigma^1) \leftrightarrow \\ &\exp[:\bar\mu_3 qp + \bar\mu_1(1/2)(-q^2+p^2):]\exp[:\bar\nu_3 qp + \bar\nu_1(1/2)(-q^2+p^2):]. \end{aligned} \qquad (7.3.519)$$

What we will use MaryLie to carry out numerically are the operations

$$\begin{aligned}&\exp[:\bar\mu_3 qp + \bar\mu_1(1/2)(-q^2+p^2):]\exp[:\bar\nu_3 qp + \bar\nu_1(1/2)(-q^2+p^2):] \\ &= \exp(:f_2^c)\exp(:f_2^a:). \end{aligned} \qquad (7.3.520)$$

That is, it will compute and multiply the two maps on the left side of (3.367) and then express the result in the factored product form shown on the right side. Specifically, it will output the quadratic polynomials f_2^c and f_2^a. When this is done, as in (3.329) and (3.330), we expect f_2^c and f_2^a will be given by the relations

$$f_2^c = \bar{\theta} b_0 = \bar{\theta}(1/2)(q^2 + p^2) \tag{7.3.521}$$

and

$$f_2^a = \bar{\lambda}_3 g + \bar{\lambda}_1 f = \bar{\lambda}_3 qp + \bar{\lambda}_1 (1/2)(-q^2 + p^2). \tag{7.3.522}$$

`Exhibit 7.3.2: Sample MaryLie Run`

7.4 Symplectic Map for Flow of Time-Independent Hamiltonian

Suppose the Hamiltonian of Theorem 6.4.1 does not explicitly depend on the time. [In this case, we say that the differential equations (1.5.6) generated by the Hamiltonian H are *autonomous*.] Then the symplectic map (6.4.1) obtained by following the Hamiltonian flow specified by H can be written immediately in the form

$$\mathcal{M} = \exp\{-(t^f - t^i) : H :\}. \tag{7.4.1}$$

That is, we have the relation

$$z^f = \mathcal{M} z^i. \tag{7.4.2}$$

To verify (4.1) and (4.2), let $\mathcal{M}$ act on z^i to give the result

$$z^f = \mathcal{M} z^i = \sum_{m=0}^{\infty} (1/m!)(t^f - t^i)^m : -H :^m z^i. \tag{7.4.3}$$

However, Taylor's theorem gives the result

$$z^f = z(t^f) = z(t^i) + \sum_{m=1}^{\infty} (1/m!)(t^f - t^i)^m (d/dt)^m z(t)|_{t^i}. \tag{7.4.4}$$

Also, Hamilton's equations of motion for the z's can be written in the form

$$\dot{z} = [z, H] = [-H, z] =: -H : z,$$

$$\ddot{z} = [-H, \dot{z}] =: -H : \dot{z} =: -H :^2 z,$$

$$(d^3 z)/(dt)^3 =: -H :^3 z, \text{ etc.} \tag{7.4.5}$$

Upon inserting the results of (4.5) into (4.4), we obtain the desired result (4.3).

At this point several observations are possible and in order. Suppose we replace the final time t^f by a general time t. Then (4.1) and (4.2) can be written in the form

$$z(t) = \mathcal{M}(t) z^i, \tag{7.4.6}$$

with

$$\mathcal{M}(t) = \exp\{(t - t^i) : -H :\}. \tag{7.4.7}$$

Note that here the Hamiltonian H depends on the initial conditions z^i, and does not depend explicitly on the time,

$$H = H(z^i). \tag{7.4.8}$$

Suppose that (4.7) is differentiated with respect to the time. Doing so gives, in accord with Appendix C, the result

$$\begin{aligned} \dot{\mathcal{M}} &= \exp\{(t - t^i) : -H :\} : -H : \\ &= \mathcal{M} : -H : . \end{aligned} \tag{7.4.9}$$

The relation (4.9) provides an equation of motion for $\mathcal{M}$ that, although only derived so far for the case of a time independent Hamiltonian, will eventually be shown to hold in general. Also, $\mathcal{M}$ evidently satisfies the initial condition

$$\mathcal{M}(t^i) = \mathcal{I}. \tag{7.4.10}$$

It will eventually be shown that the equation of motion (4.9) and the initial condition (4.10) specify $\mathcal{M}(t)$ completely. Conversely, if $\mathcal{M}(t)$ is known, the Hamiltonian H can be found, up to an immaterial constant, from the relation

$$: -H := \mathcal{M}^{-1}\dot{\mathcal{M}}. \tag{7.4.11}$$

See (5.3.15) and (5.3.16).

Next, suppose we differentiate (4.6) with respect to the time. Doing so and making use of (4.9) gives the result

$$\begin{aligned} \dot{z}(t) &= \dot{\mathcal{M}}(t) z^i = \mathcal{M} : -H : z^i \\ &= \mathcal{M}[-H, z^i]_{z^i} = \mathcal{M}[z^i, H]_{z^i}. \end{aligned} \tag{7.4.12}$$

The right side of (4.12) can be manipulated further using the relations (5.4.15) and (5.4.11) to give the result

$$\begin{aligned} \mathcal{M}[z^i, H]_{z^i} &= [\mathcal{M}z^i, \mathcal{M}H]_{z^i} = [\mathcal{M}z^i, H(\mathcal{M}z^i)]_{z^i} \\ &= [z(t), H(z(t))]_{z^i}. \end{aligned} \tag{7.4.13}$$

Also, we should really write $z(t)$ in the more explicit form

$$z = z(z^i, t) \tag{7.4.14}$$

to indicate that $z(t)$ depends on the initial conditions z^i. Finally, because the mapping between z^i and z is symplectic, the Poisson bracket on the right side of (4.13) can be rewritten to give the result

$$\begin{aligned} [z(t), H(z(t))]_{z^i} &= [z(z^i, t), H(z(z^i, t))]_{z^i} \\ &= [z, H(z)]_z. \end{aligned} \tag{7.4.15}$$

See (6.3.3) and (6.3.10) and let z^i play the role of $\overline{z}$. Putting all these results together, we find the final expected relation

$$\dot{z} = [z, H(z)]_z. \tag{7.4.16}$$

As a generalization of (4.1), suppose that the Hamiltonian H is not necessarily time independent, but does have the property that the Lie operators $: H(z,t) :$ for various times all commute. That is, one has the relation

$$\{: H(z,t) :,\ : H(z,t') :\} = 0 \text{ for all } t, t'. \tag{7.4.17}$$

Alternatively, because of the homomorphism (5.3.14), we may require that $H(z,t)$ and $H(z,t')$ be in involution for all t, t'. It can be shown that in either case the symplectic map obtained by following the Hamiltonian flow specified by H can be written in the form

$$\mathcal{M} = \exp(-\int_{t^i}^{t^f} : H : dt). \tag{7.4.18}$$

See Section 10.3.

Exercises

7.4.1. Verify in detail the steps leading from (4.12) to (4.16).

7.4.2. Prove (4.18) given the assumption (4.17).

7.4.3. Consider the map $\mathcal{M}(t)$ given by (1.4.13). Find H using (4.11).

7.4.4. Consider the map

$$q(q^i, p^i, t) = q^i(1 - tp^i)^2, \tag{7.4.19}$$

$$p(q^i, p^i, t) = p^i/(1 - tp^i). \tag{7.4.20}$$

Verify that this map is symplectic. Find H using (4.11). Sketch the flow generated by H.

7.4.5. Use the result (1.4.13) to carry out Exercise 5.4.5 and to derive (2.23). Use the results (1.4.21), (1.4.22), (1.4.23), and (1.4.24) to carry out Exercises 5.4.1 through 5.4.4 and Exercise 5.4.6, and to derive (2.20).

7.4.6. Consider the Hamiltonian H given by

$$H = (p^2 + \sigma q^2)/2 \tag{7.4.21}$$

and the linear symplectic map $\mathcal{M}$ generated by H,

$$\mathcal{M}(\sigma, t) = \exp(-t : H :). \tag{7.4.22}$$

Let $M(\sigma, t)$ be the symplectic matrix associated with $\mathcal{M}$ as in (7.2.1). Find M explicitly and show that, in accord with Poincaré's Theorem 3.3 given in Section 1.3, M is analytic in the variables σ and t. Write M in the form

$$M = \exp(JS) \tag{7.4.23}$$

and show that the Hamiltonian matrix (JS) is analytic in σ and t. Find the eigenvalues of M and (JS) and plot them in the complex plane as a function of σ. See Section 3.4 and Exercise 3.7.12. Show that the eigenvalues of M and (JS) have square-root branch points (singularities) at $\sigma = 0$.

7.4.7. Let H be a quadratic and time independent Hamiltonian, and write it in the form

$$H = (1/2)(z, Sz) \tag{7.4.24}$$

where S is a time independent symmetric matrix. It generates the linear symplectic map

$$\mathcal{M} = \exp(-t : H :) \tag{7.4.25}$$

Show that the symplectic matrix M associated with $\mathcal{M}$ is given by the relation

$$M = \exp(tJS). \tag{7.4.26}$$

7.5 Taylor Maps and Jets

Let $\mathcal{N}$ be a symplectic map, and suppose $\mathcal{N}$ sends the particular point $\tilde{z}^i$ to the point $\tilde{z}^f$. Consider points z near $\tilde{z}^i$ by writing the relation

$$z = \tilde{z}^i + \zeta, \tag{7.5.1}$$

and define points $\overline{z}$ near $\tilde{z}^f$ by writing the relation

$$\overline{z} = \tilde{z}^f + \overline{\zeta}. \tag{7.5.2}$$

Then, by construction, we have the relation

$$\overline{\zeta} = 0 \text{ if } \zeta = 0. \tag{7.5.3}$$

Also, the mappings (5.1) and (5.2) are symplectic. See Exercise 6.2.2. It follows from the group property for symplectic maps that the mapping between ζ and $\overline{\zeta}$, call it $\mathcal{M}$, is also symplectic. We write the relation

$$\overline{\zeta} = \mathcal{M}\zeta, \tag{7.5.4}$$

and observe that according to (5.3), $\mathcal{M}$ sends the origin into itself.

Suppose the map $\mathcal{N}$ is *analytic* in z around the point $\tilde{z}^i$. Then $\mathcal{M}$ will be analytic in ζ around the origin. Correspondingly, we may write a *Taylor* expansion of the form

$$\overline{\zeta}_a = \sum_b R_{ab}\zeta_b + \sum_{bc} T_{abc}\zeta_b\zeta_c + \sum_{bcd} U_{abcd}\zeta_b\zeta_c\zeta_d + \cdots \, . \tag{7.5.5}$$

Note that the expansion has no constant terms due to (5.3). Expansions of the form (5.5) often are used both in magnetic particle optics and in light ray optics. In this context the coefficients R describe paraxial optics, and the remaining coefficients T, U, $\cdots$ describe aberration effects. For this reason we will refer to expansions of the form (5.5) either as *Taylor maps* or *aberration expansions*.

We have already seen in Section 6.4 that Hamiltonian flows produce symplectic maps. Also, according to Theorem 3.3.3, if the Hamiltonian has suitable analytic properties, which is often the case, then the symplectic map it produces will also be analytic. See Chapter 26 and Appendix F. Thus, *analytic* symplectic maps are of great interest. Without loss of generality, such maps may be taken to be of the form (5.5).

Finally, suppose $f(\zeta)$, a function of the phase-space variables ζ, is analytic at the origin. Suppose further that the Taylor expansion of f begins with quadratic terms. Then evidently the symplectic map given by the Lie transformation $\exp(:f:)$ is of the form (5.5).

By combining (5.1) through (5.5) we may write the relation

$$\bar{z} = \mathcal{N} z \tag{7.5.6}$$

in the form

$$\begin{aligned} \bar{z}_a = \tilde{z}^f_a \quad &+ \quad \sum_b R_{ab}(z-\tilde{z}^i)_b \\ &+ \quad \sum_{bc} T_{abc}(z-\tilde{z}^i)_b(z-\tilde{z}^i)_c \\ &+ \quad \sum_{bcd} U_{abcd}(z-\tilde{z}^i)_b(z-\tilde{z}^i)_c(z-\tilde{z}^i)_d + \cdots . \end{aligned} \tag{7.5.7}$$

Let $g_a(m;z)$ denote a homogeneous polynomial of degree m in the components of z. With this notation (5.7) can also be written in the form

$$\bar{z}_a = \sum_{m=0}^{\infty} g_a[m;(z-\tilde{z}^i)]. \tag{7.5.8}$$

Suppose $\mathcal{N}'$ is some other symplectic map that sends $\tilde{z}^i$ to $\tilde{z}^f$, and suppose $\mathcal{N}'$ has an expansion of the form

$$\bar{z}_a = \sum_{m=0}^{\infty} g'_a[m;(z-\tilde{z}^i)]. \tag{7.5.9}$$

Since both $\mathcal{N}$ and $\mathcal{N}'$ send $\tilde{z}^i$ to $\tilde{z}^f$, we have the relation

$$g_a(0;z) = g'_a(0;z). \tag{7.5.10}$$

Now suppose that in fact g_a and g'_a agree for $m \leq k$:

$$g_a(m;z) = g'_a(m;z) \text{ for } m \leq k. \tag{7.5.11}$$

We express this condition symbolically by writing

$$\mathcal{N}' \overset{k}{\sim} \mathcal{N} \tag{7.5.12}$$

and say that $\mathcal{N}'$ and $\mathcal{N}$ are *equivalent* through terms of degree k. It can be shown, as the notation and terminology are meant to suggest, that (5.12) defines an equivalence relation among maps that send $\tilde{z}^i$ to $\tilde{z}^f$. See Exercise 5.2. With the aid of this equivalence relation, we may define equivalence classes of maps. For any given k, an equivalence class is called a *k-jet*. Evidently, an equivalence class (a k-jet) is determined by specifying the polynomials $h_a(0;z)$, $h_a(1;z)$, $\cdots$ $h_a(k;z)$. Put another way, we may say that two maps $\mathcal{N}$ and $\mathcal{N}'$ represent the same k-jet if their derivatives agree at $\tilde{z}_i$ through order k. Or, what amounts to the same thing, we may view a k-jet as being representated by a point $\tilde{z}^i$ and a Taylor series map (about this point) truncated beyond terms of degree k.

Finally, suppose a Taylor map is a Taylor expansion of a symplectic map. We will refer to the jet obtained by truncating such an expansion as a *symplectic jet*. It is important to note that a symplectic k-jet is generally *not* a symplectic map, but rather is a k-jet that satisfies the symplectic condition through terms of degree $(k-1)$. See Exercise 5.3.

Exercises

7.5.1. Suppose that $f(\zeta)$ is analytic at the origin and begins with quadratic terms. Show that the symplectic map given by $\exp(:f:)$ is of the form (5.5).

7.5.2. Show that (5.11) and (5.12) produce an equivalence relation on the set of differentiable maps. See Exercise 5.12.7.

7.5.3. Exercise about symplectic jets.

7.6 Factorization Theorem

Note what has been accomplished so far. Section 3.7 showed that matrices of the form JS with S symmetric produce a Lie algebra. It also showed that any symplectic matrix sufficiently near the identity can be written in the form $\exp(JS)$. Section 3.8 showed that any symplectic matrix can be written as the product of two exponentials. Similarly, Section 5.3 showed that the set of Lie operators $:f:$ forms a Lie algebra. And Section 7.1 showed that Lie transformations $\exp(:f:)$ are symplectic maps. Finally, we have just seen that if f is analytic at the origin and begins with quadratic terms, then the Lie transformation $\exp(:f:)$ produces a map of the form (5.5). What remains to be studied is the question of whether any symplectic map $\mathcal{M}$ can be written in exponential form. The answer to this question is given by the *factorization* theorem.

Theorem 6.1 Let $\mathcal{M}$ be an *analytic* symplectic map that sends the origin into itself. That is, the relation between z and $\bar{z}$ is assumed to be expressible in a Taylor series of the form

$$\bar{z}_a = \sum_b R_{ab} z_b + \sum_{bc} T_{abc} z_b z_c + \sum_{bcd} U_{abcd} z_b z_c z_d + \cdots \, . \tag{7.6.1}$$

In the terminology of the last section, truncating this series beyond terms of degree k yields, for any k, a symplectic k-jet. (Here, to avoid proliferation of notation, we again use the symbols $\bar{z}$ and z to denote general phase-space variables.) Then, remarkably, there are *unique* functions $f_2^c(z), f_2^a(z), f_3(z), f_4(z), \cdots$ such that the relation (6.1) can be written in the form

$$\bar{z} = \mathcal{M} z, \tag{7.6.2}$$

with $\mathcal{M}$ expressed as a product of Lie transformations in the form

$$\mathcal{M} = \exp(: f_2^c :) \exp(: f_2^a :) \exp(: f_3 :) \exp(: f_4 :) \cdots . \tag{7.6.3}$$

Furthermore, each of the functions $f_m(z)$ is a *homogeneous polynomial* of degree m in the variables z.

The proof of this theorem is best achieved in stages by verifying a series of lemmas.
Lemma 6.1 The matrix R of (6.1) is symplectic. To see this, compute the Jacobian matrix $M(z)$ corresponding to the transformation (6.1) using (6.1.2). We find the result

$$M(0) = R. \tag{7.6.4}$$

Then, since $\mathcal{M}$ is assumed to be symplectic, $M(z)$ must be symplectic for all z, and hence R must be symplectic.

Lemma 6.2 Let $g_1(z) \cdots g_{2n}(z)$ be a set of $2n$ functions. (Here, as usual, $2n$ is the dimensionality of the phase space in question.) Suppose these functions satisfy the relations

$$[z_a, g_b] = [z_b, g_a]. \tag{7.6.5}$$

Then such functions exist if and only if there is a function h such that

$$g_a = [h, z_a] =: h : z_a. \tag{7.6.6}$$

The function h is unique up to an additive constant.

To verify this assertion, first suppose that each g_a is given by (6.6). Then we quickly demonstrate (6.5). We find the result

$$\begin{aligned} [z_a, g_b] - [z_b, g_a] &= [z_a, [h, z_b]] - [z_b, [h, z_a]] \\ &= [h, [z_a, z_b]] = [h, J_{ab}] = 0. \end{aligned} \tag{7.6.7}$$

Here we have used the Jacobi identity (5.1.7).

Next, suppose (6.5) is true. We are now in a situation analogous to that of Section 6.4. Compare (6.4.30) and (6.5). As before, define functions η_c using (6.4.39),

$$\eta_c = \sum_d J_{dc} g_d, \tag{7.6.8}$$

and define an associated function h by the integral

$$h = -\int^z \sum_a \eta_a dz'_a. \tag{7.6.9}$$

As we have seen, the integral is path independent, and has the property

$$\partial h / \partial z_b = -\eta_b. \tag{7.6.10}$$

Following (6.4.31), we find that the Poisson bracket $[h, z_a]$ has the value

$$\begin{aligned}
[h, z_a] &= -[z_a, h] = -\sum_b J_{ab}(\partial h/\partial z_b) \\
&= \sum_b J_{ab}\eta_b = \sum_{bc} J_{ab}J_{cb}g_c \\
&= \sum_{bc} J_{ab}(J^T)_{bc}g_c = \sum_c (JJ^T)_{ac}g_c = g_a,
\end{aligned} \tag{7.6.11}$$

as desired. Here we have also used (6.8).

Lemma 6.3 Let f_m be a homogeneous polynomial in z of degree m. Also, let $\mathcal{P}_r$ denote the set of all homogeneous polynomials of degree r. Then, for any two homogeneous polynomials f_m and f_n, we have the relation

$$[f_m, f_n] \in \mathcal{P}_{m+n-2}. \tag{7.6.12}$$

To put it another way, define a *degree* functional by the rule

$$\deg(f_m) = m. \tag{7.6.13}$$

Then we have the relation

$$\deg([f_m, f_n]) = m + n - 2. \tag{7.6.14}$$

This lemma is obviously true because the Poisson bracket operation simply involves two differentiations and multiplication.

We now have the necessary tools to prove Theorem 6.1. First consider the linear part of the transformation (6.1) that is described by the matrix R. Since R is symplectic, it can be written in the standard form

$$R = PO. \tag{7.6.15}$$

See (2.2). Let $f_2^a(z)$ and $f_2^c(z)$ be the polynomials associated with R using (2.3) and (2.8). Then, according to (2.11), (2.12), and (1.23), we have the result

$$\exp(-:f_2^a:)\exp(-:f_2^c:)z = R^{-1}z. \tag{7.6.16}$$

Suppose both sides of (6.1) are acted on by $\exp(-:f_2^a:)\exp(-:f_2^c:)$. Doing so, and using (6.16), gives the result

$$\exp(-:f_2^a:)\exp(-:f_2^c:)\overline{z}_b = z_b + r_b(>1). \tag{7.6.17}$$

Here the notation $r_b(>m)$ denotes *any* "remainder" series consisting of terms of degree higher than m.

To proceed further, suppose the remainder terms $r_b(>1)$ are decomposed into second degree terms $g_b(2;z)$ and higher degree terms by writing the relations

$$r_b(>1) = g_b(2;z) + r_b(>2). \tag{7.6.18}$$

With this notation, we may rewrite (6.17) in the form

$$\exp(- : f_2^a :) \exp(- : f_2^c :)\bar{z}_b = z_b + g_b(2; z) + r_b(> 2). \tag{7.6.19}$$

Take the Poisson bracket of both sides of (6.19) with themselves for different values of the index b. Doing so, and making use of (5.4.15), (5.4.16), and (6.12), gives the result

$$J_{bc} = J_{bc} + [z_b, g_c(2)] + [g_b(2), z_c] + r_{bc}(> 1). \tag{7.6.20}$$

Finally, upon equating terms of like degree in (6.20), we find the relation

$$[z_b, g_c(2)] + [g_b(2), z_c] = 0. \tag{7.6.21}$$

At this point the results of Lemma 6.2 come into play. According to this lemma, there is a function h such that g_a is given by (6.6). Indeed, we may use (6.9) to compute h explicitly. Inserting the definition (6.8) for the functions η_a into (6.9) gives the result

$$h = -\int^z \sum_{ab} g_a J_{ab} dz'_b. \tag{7.6.22}$$

Suppose we consider the case where each g_a is a homogeneous polynomial of degree m, call it $g_a(m, z)$, and suppose we denote by f_{m+1} the result of computing h in this case. Then the path integral is conveniently evaluated along the path

$$z'_b = \tau z_b, \tag{7.6.23}$$

where the parameter τ ranges from 0 to 1. Use of (6.22) and (6.23) gives the result

$$f_{m+1}(z) = -[1/(m+1)] \sum_{ab} g_a(m; z) J_{ab} z_b. \tag{7.6.24}$$

As the notation suggests, $f_{m+1}(z)$ is a homogeneous polynomial of degree $(m+1)$. In particular, use of (6.24) with the functions $g_b(2; z)$ produces the third-degree polynomial $f_3(z)$.

As a warm-up exercise for the next step, consider the effect of applying the Lie transformation $\exp(- : f_3 :)$ to z. We find the result

$$\exp(- : f_3 :)z_b = z_b + \underbrace{: -f_3 : z_b}_{\text{quadratic terms}} + (1/2!)\underbrace{: -f_3 :^2 z_b}_{\text{cubic terms}} + \cdots . \tag{7.6.25}$$

Here, in accord with (6.13), the degrees of the various terms have been indicated.

Now apply $\exp(: -f_3 :)$ to both sides of (6.19). Doing so, and again making use of (6.14), gives the result

$$\begin{aligned} &\exp(- : f_3 :) \exp(- : f_2^a :) \exp(- : f_2^c :)\bar{z}_b = \\ &\quad z_b + : -f_3 : z_b + g_b(2; z) + r_b(> 2). \end{aligned} \tag{7.6.26}$$

However, according to Lemma 6.2, f_3 has the property

$$: -f_3 : z_b + g_b(2; z) = 0. \tag{7.6.27}$$

Consequently, (6.26) can be rewritten in the form

$$\exp(- : f_3 :) \exp(- : f_2^a :) \exp(- : f_2^c :)\overline{z}_b = z_b + r_b(> 2). \tag{7.6.28}$$

Comparison of the right sides of (6.17) and (6.28) shows that the degree of the remainder term has been raised by 1. At this stage it should also be clear that the degree of the remainder term can be increased indefinitely by finding $f_4, f_5, \cdots$ and applying the Lie transformations $\exp(: -f_4 :), \exp(- : f_5 :) \cdots$. That is, for any s we have the general result

$$\exp(- : f_s :) \cdots \exp(: -f_3 :) \exp(- : f_2^a :) \exp(- : f_2^c :)\overline{z}_b = z_b + r_b[> (s - 1)]. \tag{7.6.29}$$

We are ready for the final step. Rewrite (6.29) in the form

$$\overline{z}_b = \exp(: f_2^c :) \exp(: f_2^a :) \exp(: f_3 :) \cdots \exp(: f_s :)z_b + r_b[> (s - 1)], \tag{7.6.30}$$

and let $s \to \infty$. Then, if the remainder term tends to zero, we obtain the advertised result (6.2) and (6.3). Otherwise the result is true only formally. In this latter case the infinite product (6.3) is also not convergent.

We have proved a key result. Recall that in Section 6.4 it was shown that Hamiltonian flows produce symplectic maps. Also, Theorem 1.3.3 shows that for many systems of physical interest such maps are analytic. Now, thanks to Theorem 6.1, it is possible to describe the most general analytic symplectic map (which sends the origin into itself) simply in terms of various homogeneous polynomials. Finally, it will be shown in the next section that the restriction of preserving the origin can be removed by including Lie transformations of the form $\exp(: f_1 :)$ where f_1 is a suitably chosen polynomial linear in the z's. Consequently, any analytic symplectic map can be represented uniquely as a product of Lie transformations generated by homogeneous polynomials. Conversely, any product of Lie transformations generated by homogeneous polynomials is a symplectic map.

At this point, two comments are appropriate. First, suppose the factored product representation (6.3) is truncated at any point. Then the resulting expression is still a symplectic map because each term in the product is a symplectic map. Also, if the truncation consists of dropping all terms in the product (6.3) beyond $\exp(: f_m :)$ for some m, then according to (6.31) the power-series expansion for the truncated map agrees with that of the original Taylor map (6.1) through terms of degree $(m - 1)$. Consequently, a truncated product map provides a symplectic approximation to the exact map. By contrast, as we know from Exercise 5.3, simply truncating a Taylor map generally *violates* the symplectic condition.

Second, suppose (6.3) is decomposed, as shown below, into those factors involving only quadratic polynomials, and the remaining factors involving cubic and higher degree polynomials,

$$\mathcal{M} = \overbrace{\exp(: f_2^c :) \exp(: f_2^a :)}^{\text{“Gaussian optics''}} \times \overbrace{\exp(: f_3 :) \exp(: f_4 :) \cdots}^{\text{Aberrations or nonlinear corrections}} . \tag{7.6.31}$$

It will be demonstrated in subsequent sections that dropping all terms beyond those involving the quadratic polynomials leads to a lowest-order approximation for $\mathcal{M}$ that is equivalent to the paraxial Gaussian optics approximation in the case of light optics, and the usual linear matrix approximation in the case of charged-particle beam optics. Moreover, the remaining factors $\exp(:f_3:)\exp(:f_4:)\cdots$ represent aberrations or nonlinear corrections to the lowest-order approximation. In particular, in the case of charged-particle beam optics, the factor $\exp(:f_3:)$ describes various chromatic effects and the effects due to sextupole magnets. Similarly, the factor $\exp(:f_4:)$ describes higher-order chromatic effects, the effects due to iterated sextupoles, and the effects due to octupoles. Finally in some cases f_3, f_4, etc. also describe what may be called "kinematic" nonlinearities in the equations of motion. They arise, for example, from the fact that the equations of motion generated by the Hamiltonians (1.6.16) and (1.6.17) are intrinsically nonlinear even in the absence of electric and magnetic fields. Let D be an integer with $D \geq 3$. In general, as will be shown later, retaining in the product (6.31) only those terms with f_m satisfying $m \leq D$ amounts to neglecting aberrations of degree D and higher.

We close this section with a cautionary note: the quantities f_2^c and f_2^a that occur in the factorization (2.10), or more generally in the factorization (6.3), can have what may seem to be surprising properties. Suppose, for example, that $\mathcal{M}$ is a *linear* symplectic map. Employ the notation $z = (q_1, q_2, \cdots; p_1, p_2, \cdots)$ and write $\bar{z} = \mathcal{M}z = (\bar{q}_1, \bar{q}_2, \cdots; \bar{p}_1, \bar{p}_2, \cdots)$. Suppose moreover that $\mathcal{M}$ is known to have the property

$$\bar{p}_1 = \mathcal{M}p_1 = p_1. \tag{7.6.32}$$

Such maps obviously form a group, and any map of the form

$$\mathcal{M}_g = \exp(:g_2:) \tag{7.6.33}$$

where g_2 does not depend on the variable q_1,

$$\partial g_2/\partial q_1 = 0, \tag{7.6.34}$$

will have this property. Suppose that h_2 is another function that does not depend on q_1. Then it is clear that all linear combinations of g_2 and h_2 and their single and multiple Poisson brackets will also be independent of q_1. Thus, the set of all such functions forms a Lie subalgebra. Now let $\mathcal{M}$ be any product of maps which individually are exponentials of Lie operators associated with q_1-independent quadratic polynomials, and let f_2^c and f_2^a be the quadratic polynomials associated with a factorization of $\mathcal{M}$ in the form (2.10) or (6.3),

$$\mathcal{M} = \exp(:g_2:)\exp(:h_2:)\cdots = \exp(:f_2^c:)\exp(:f_2^a). \tag{7.6.35}$$

Then, it is tempting to assume that f_2^c and f_2^a will also be independent of q_1. This would be the case if f_2^c and f_2^a were in the Lie subalgebra generated by $g_2, h_2\cdots$. However, a simple counter-example shows that this need not be true. See Exercise 6.14. Although f_2^c and f_2^a are in the Lie algebra of quadratic polynomials, they need not be in the subalgebra generated by $g_2, h_2\cdots$. This is because, by their construction, f_2^c and f_2^a are required to have specific properties with respect to J, and it may happen that these properties can only be achieved by including Lie elements outside the subalgebra. For example, in the case of $sp(2,\mathbb{R})$, f_2^c is proportional to q^2+p^2; and f_2^a is a linear combination of $-q^2+p^2$ and qp.

See Section 5.6. Note that both f_2^c and f_2^a, when considered individually, depend on *both* the variables q and p.

In the context of Accelerator Physics a particularly confusing/irritating example of this phenomena occurs in the case of *static/time-independent* maps where the differential transit time t plays the role of q_1 in the present discussion and its conjugate momentum (energy) p_t plays the role of p_1. With the exception of maps for radio-frequency cavities and maps for magnets with time-dependent fields, most maps for accelerator beam-line elements are static. Yet when the f_2^c and f_2^a are computed for a *static* map, they may turn out to be *time dependent*.[12] But, of course, when the effects of these f_2^c and f_2^a are combined to compute a net map, this net map will leave p_t unchanged even though this fact is not readily apparent simply by looking at f_2^c and f_2^a.

Exercises

7.6.1. Verify (6.7).

7.6.2. Verify (6.8) through (6.10).

7.6.3. Show that any h that satisfies (6.6) is unique up to an additive constant.

7.6.4. Verify (6.12) and (6.14).

7.6.5. Verify (6.20).

7.6.6. Verify (6.11) and (6.22). Carry out the path integral described to get (6.24).

7.6.7. Suppose that $f(m, z)$ is a homogeneous polynomial of degree m. Then it must satisfy *Euler's* relation

$$\sum_a z_a(\partial f/\partial z_a) = mf. \tag{7.6.36}$$

See Exercise 1.5.11. Given (6.5), verify by direct calculation that f_{m+1} as given by (6.24) satisfies the relation

$$: f_{m+1} : z_b = g_b(m; z). \tag{7.6.37}$$

7.6.8. Verify (6.26).

7.6.9. Justify the passage from (6.29) to (6.30).

7.6.10. Consider the two-variable map, a variant of the *Hénon* map, given by the relations

$$\begin{aligned}\overline{q} &= \lambda[q + (q-p)^2],\\ \overline{p} &= (1/\lambda)[p + (q-p)^2],\end{aligned} \tag{7.6.38}$$

where λ is a parameter (positive or negative). Show that this map is symplectic. Find the factorization (6.3). That is, determine the polynomials f_m.

[12] For this and other reasons, the charged-particle beam transport code MaryLie does not work directly with the polynomials f_2^c and f_2^a. It works with 6×6 matrices M to represent the linear part of the map $\mathcal{M}$, and only computes f_2^c and f_2^a from M when requested. The static/dynamic nature of M is readily apparent upon inspection of its matrix elements.

7.6.11. Consider the two-variable map, a variant of the *Hénon* map, given by the relations

$$\overline{q} = q\cos\alpha + p\sin\alpha + p^2\cos\alpha,$$

$$\overline{p} = -q\sin\alpha + p\cos\alpha - p^2\sin\alpha, \tag{7.6.39}$$

where α is a parameter. Show that this map is symplectic. Find the factorization (6.3). That is, determine the polynomials f_m.

7.6.12. Given the factorization (6.31), show how to compute the R, T, and U of (6.1).

7.6.13. Find the restrictions on the coefficients R, T, and U in (6.1) that are entailed by the symplectic condition.

7.6.14. This exercise is a sequel to Exercise 5.6.7, which you should review. Its purpose is to examine two linear two-dimensional symplectic maps, find their single exponential forms where applicable, find their polar decompositions and associated polynomials f_2^a and f_2^c, and examine the q and p content of these polynomials.

As the first case to be studied, let $\mathcal{M}$ be a linear symplectic map acting on the two-dimensional phase space q_1, p_1 and described by the matrix L given by the relation

$$L = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}. \tag{7.6.40}$$

Evidently L is symplectic and $\mathcal{M}$ satisfies (6.32). Verify that, in fact, $\mathcal{M}$ can be written in the form

$$\mathcal{M} = \exp(:g_2:) \tag{7.6.41}$$

with

$$g_2 = -p_1^2/2. \tag{7.6.42}$$

Observe that g_2 does not depend on q_1.

Let us continue on to compute the quadratic polynomials f_2^c and f_2^a appearing in (2.10) to examine their q_1 behavior and then find them explicitly. Begin by polar decomposing L as in (6.15). From Exercise 5.6.7 we know that P is given by (5.6.34) and O is given by (5.6.40). If we write O and P in the exponential forms (3.8.17) and (3.8.25), we see that neither S^c nor S^a can vanish since neither O nor P are equal to the identity. Use the parameterization of (5.9.10) to write

$$JS^c = \beta_0 B^0, \tag{7.6.43}$$

$$JS^a = \phi F + \gamma G. \tag{7.6.44}$$

See (5.6.7), (5.6.13), and (5.6.14). Show that f_2^c and f_2^a are given by the relations

$$f_2^c = -(\beta_0/2)(p_1^2 + q_1^2), \tag{7.6.45}$$

$$f_2^a = -(\phi/2)(p_1^2 - q_1^2) - \gamma q_1 p_1. \tag{7.6.46}$$

Because neither S^c nor S^a can vanish, verify that consequently both f_2^c and f_2^a, unlike g_2, must depend on q_1.

Let us now compute f_2^c and f_2^a explicitly. With regard to O, use (5.9.12) to deduce the relation

$$O = \exp(\beta_0 B^0) = I \cos\beta_0 + B^0 \sin\beta_0, \tag{7.6.47}$$

and thereby obtain the results

$$\cos\beta_0 = 2/\sqrt{5}, \tag{7.6.48}$$

$$\sin\beta_0 = 1/\sqrt{5}, \tag{7.6.49}$$

$$\tan(\beta_0) = 1/2, \tag{7.6.50}$$

$$\beta_0 = \tan^{-1}(1/2) = .463\cdots . \tag{7.6.51}$$

Here, in view of the 2π periodicity of the right side of (6.47), we have restricted β_0 to the interval $\beta_0 \in [-\pi, \pi]$.

With regard to P, deduce from (5.9.11) the relation

$$\begin{aligned} P = \exp(\phi F + \gamma G) &= I \cosh[(\phi^2+\gamma^2)^{1/2}] \\ &+ [(\phi F + \gamma G)]/(\phi^2+\gamma^2)^{1/2}] \sinh[(\phi^2+\gamma^2)^{1/2}]. \end{aligned} \tag{7.6.52}$$

Using the explicit form (5.6.27) for P, take the trace of both sides of (6.52) to find the result

$$\cosh[(\phi^2+\gamma^2)^{1/2}] = (1/2)\sqrt{5}. \tag{7.6.53}$$

Next multiply both sides of (6.52) by F and again take traces to find the result

$$[\phi/(\phi^2+\gamma^2)^{1/2}] \sinh[(\phi^2+\gamma^2)^{1/2}] = -1/\sqrt{5}. \tag{7.6.54}$$

Finally, multiply both sides of (6.52) by G and take traces to find the result

$$[\gamma/(\phi^2+\gamma^2)^{1/2}] \sinh[(\phi^2+\gamma^2)^{1/2}] = 1/(2\sqrt{5}). \tag{7.6.55}$$

Show that (6.54) and (6.55) are consistent with (6.53), and from them deduce the relation

$$\phi = -2\gamma. \tag{7.6.56}$$

Solve (6.53) through (6.55) to obtain the results

$$\phi = -.430\cdots , \tag{7.6.57}$$

$$\gamma = .215\cdots . \tag{7.6.58}$$

Taken together, the relations (6.45) and (6.46), with β_0 and ϕ and γ given by (6.51) and (6.57) and (6.58), specify f_2^c and f_2^a explicitly. Note that both f_2^c and f_2^a depend on both q_1 and p_1 even though (6.32) holds and g_2 is independent of q_1.

As the second case to be studied, let $\mathcal{M}$ be the linear symplectic map described by the matrix M given by the relation

$$M = -L = \begin{pmatrix} -1 & -1 \\ 0 & -1 \end{pmatrix}. \tag{7.6.59}$$

Evidently $\mathcal{M}$ is symplectic. Moreover, we know from Exercise 3.7.12 that this $\mathcal{M}$ cannot be written in single exponential form.

Let us continue on to compute and find explicitly the quadratic polynomials f_2^c and f_2^a for this $\mathcal{M}$. From Exercise 5.6.7 we know that M has the polar decomposition (5.6.41). It follows from (5.6.42) that f_2^a is the same as the f_2^a found in the first part of this exercise. What remains is to find f_2^c. Evidently the Ansatz (6.45) continues to hold, but with a different value of β_0. Show that in this case there is the relation

$$O' = \exp(\beta_0 B^0) = I\cos\beta_0 + B^0\sin\beta_0 \tag{7.6.60}$$

with O' given by (6.43). Show that now there are the results

$$\cos\beta_0 = -2/\sqrt{5}, \tag{7.6.61}$$

$$\sin\beta_0 = -1/\sqrt{5}, \tag{7.6.62}$$

$$\tan(\beta_0) = 1/2, \tag{7.6.63}$$

$$\beta_0 = -\pi + \tan^{-1}(1/2) = -\pi + .463\cdots = -2.677\cdots . \tag{7.6.64}$$

7.7 Inclusion of Translations

Consider transformations of the form

$$\bar{z}_b = z_b + \delta_b, \tag{7.7.1}$$

where the quantities $\delta_1, \cdots \delta_{2n}$ are parameters. It is easy to verify that (7.1) is a symplectic map. See Exercise 6.2.2. Define a related set of parameters δ_a^* by the rule

$$\delta_a^* = \sum_b J_{ab}\delta_b, \text{ or } \delta^* = J\delta. \tag{7.7.2}$$

Also, define a first-degree polynomial $g_1(z)$ by the rule

$$\begin{aligned} g_1(z) &= \sum_{ab} J_{ab} z_a \delta_b = (z, \delta^*) = (z, J\delta) \\ &= (J^T z, \delta) = (\delta, J^T z) = -(\delta, Jz) \\ &= -(\delta, z^*). \end{aligned} \tag{7.7.3}$$

Then, use of (6.10) shows that g_1 obeys the relations

$$\begin{aligned} :g_1: z_b &= [g_1, z_b] = -[z_b, g_1] \\ &= -\partial g_1/\partial z_b^* = \delta_b, \end{aligned} \tag{7.7.4}$$

$$:g_1:^m z_b = 0 \text{ for } m > 1. \tag{7.7.5}$$

Consequently, we have the relation

$$\exp(:g_1:) z_b = z_b + \delta_b. \tag{7.7.6}$$

That is, Lie transformations of the form $\exp(: g_1 :)$ produce translations in phase space, and any translation can be written in this form.

Suppose the Taylor map (6.1) is generalized by the addition of constant terms to give a transformation of the form

$$\overline{z}_a = \delta_a + \sum_b R_{ab} z_b + \sum_{bc} T_{abc} z_b z_c + \sum_{bcd} U_{abcd} z_b z_c z_d + \cdots . \tag{7.7.7}$$

Then, a slight modification of Theorem (6.1) shows that (6.31) is generalized to become the relation

$$\begin{aligned} &\exp(- : f_s :) \cdots \exp(- : f_3 :) \exp(- : f_2^a :) \exp(- : f_2^c :) \overline{z}_b \\ &\qquad = z_b + \delta_b + r_b[> (s-1)]. \end{aligned} \tag{7.7.8}$$

Here the homogeneous polynomials f_m are the same as before. Next use (7.6) in (7.8) to get the relation

$$\begin{aligned} &\exp(- : f_s :) \cdots \exp(- : f_3 :) \exp(- : f_2^a :) \exp(- : f_2^c :) \overline{z}_b \\ &\qquad = \exp(: g_1 :) z_b + r_b[> (s-1)]. \end{aligned} \tag{7.7.9}$$

Finally, rewrite (7.9) in the form

$$\overline{z}_b = \exp(: f_2^c :) \exp(: f_2^a :) \exp(: f_3 :) \cdots \exp(: f_s :) \exp(: g_1 :) z_b + r_b[> (s-1)], \tag{7.7.10}$$

and let $s \to \infty$. We see that the generalized transformation (7.7) can be written in the form

$$\overline{z} = \mathcal{M} z \tag{7.7.11}$$

where $\mathcal{M}$ has the factorization

$$\mathcal{M} = [\exp(: f_2^c :) \exp(: f_2^a :) \exp(: f_3 :) \exp(: f_4 :) \cdots] \exp(: g_1 :). \tag{7.7.12}$$

As in Section 7.5, let $\mathcal{N}$ be a symplectic map, and suppose $\mathcal{N}$ sends the particular point $\tilde{z}^i$ to the point $\tilde{z}^f$. Also, again suppose $\mathcal{N}$ is analytic in z around the point $\tilde{z}^i$. Then, the relations (5.1), (5.2), and (5.5) can also be written in the form

$$\begin{aligned} \overline{z}_a \;=\; & \tilde{z}_a^f + \sum_b R_{ab} (z - \tilde{z}^i)_b + \sum_{bc} T_{abc} (z - \tilde{z}^i)_b (z - \tilde{z}^i)_c \\ + \; & \sum_{bcd} U_{abcd} (z - \tilde{z}^i)_b (z - \tilde{z}^i)_c (z - \tilde{z}^i)_d + \cdots . \end{aligned} \tag{7.7.13}$$

Let $h_1(z)$ be a first-degree polynomial defined by the relation

$$h_1(z) = (z, J\tilde{z}^i). \tag{7.7.14}$$

Then, by construction and the discussion at the beginning of this section, $h_1(z)$ has the property

$$\exp(: h_1 :) z = z + \tilde{z}^i, \tag{7.7.15}$$

or

$$\exp(: h_1 :)(z - \tilde{z}^i) = z. \tag{7.7.16}$$

Apply $\exp(:h_1:)$ to both sides of (7.13). Doing so, and making use of (7.16) and (5.4.11), gives the result

$$\exp(:h_1:)\overline{z}_a = \tilde{z}^f_a + \sum_b R_{ab}z_b + \sum_{bc} T_{abc}z_b z_c + \sum_{bcd} U_{abcd}z_b z_c z_d + \cdots . \tag{7.7.17}$$

We are now again in the situation described by (7.7) with $\tilde{z}^f$ playing the role of k. Consequently, we may write the relation

$$\begin{aligned} \exp(-:f_s:) \quad \cdots \quad & \exp(-:f_3:)\exp(-:f_2^a:)\exp(-:f_2^c:)\exp(:h_1:)\overline{z}_b \\ = \quad & \exp(:g_1:)z_b + r_b[>(s-1)]. \end{aligned} \tag{7.7.18}$$

Here the homogeneous polynomials f_m are again the same as before, and g_1 is given by the relation

$$g_1(z) = (z, J\tilde{z}^f). \tag{7.7.19}$$

Finally, rewrite (7.18) in the form

$$\begin{aligned} \overline{z}_b \quad = \quad & \exp(-:h_1:)\exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_3:)\cdots\exp(:f_s:)\exp(:g_1:)z_b \\ + \quad & \exp(-:h_1:)r_b[>(s-1)]. \end{aligned} \tag{7.7.20}$$

Again let $s \to \infty$. Then, providing the remainder term tends to zero,

$$\lim_{s\to\infty} \exp(-:h_1:)r_b[>(s-1)] = 0, \tag{7.7.21}$$

we have the result

$$\overline{z} = \mathcal{N}z \tag{7.7.22}$$

where $\mathcal{N}$ has the factorization

$$\mathcal{N} = \exp(-:h_1:)[\exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_3:)\exp(:f_4:)\cdots]\exp(:g_1:). \tag{7.7.23}$$

We conclude that the general analytic symplectic map $\mathcal{N}$ given by (7.13) can be written in the factored product form (7.23).

Note that Sections 5.1 and 5.3 showed that the set of Lie operators forms an infinite-dimensional Lie algebra, and Section 6.2 showed that symplectic maps form a group. Theorem 6.1 and (7.23) [see also (8.1) and (8.2)] show that Lie operators form the Lie algebra of the group of symplectic maps. *Thus, the group of symplectic maps is an infinite-dimensional Lie group, and its Lie algebra is the Lie algebra of Lie operators.* From (2.10) we see that the subgroup of all symplectic maps that preserve the origin and are linear, namely $Sp(2n,\mathbb{R})$, has as its Lie algebra $sp(2n,\mathbb{R})$ all Lie operators of the form $:f_2:$. Next consider $SpM(2n,\mathbb{R})$, the group of all symplectic maps that preserve the origin. From (6.3) we see that its Lie algebra, $spm(2n,\mathbb{R})$, consists of all Lie operators of the form $:f_m:$ with $m = 2,3,\cdots$. Finally, consider $ISpM(2n,\mathbb{R})$, the group of all symplectic maps. We see from (7.23) [see also (8.1) and (8.2)] that its Lie algebra, $ispm(2n,\mathbb{R})$, consists of all Lie operators of the form $:f_m:$ with $m = 1,2,3\cdots$.

We close this section with the remark that if one considers the set of all invertible analytic maps, and not just the subset of analytic symplectic maps, then this set of all invertible analytic maps also forms a group. This group, sometimes called the group of *analytic diffeomorphisms*, is also an infinite-dimensional Lie group, and has as its Lie algebra the set of all general Lie operators of the form (5.3.17) with the g_b being analytic functions. This group is sometimes called $\mathit{Diff}(R^m)$ where m is the dimension of the space under consideration. If one is more careful, which we generally are not because we assume analyticity, one should make distinctions between maps that are merely continuous (C^0), or have some specified number of derivatives (C^k), or have an infinite number of derivatives (C^∞), or are analytic (C^ω).[13] With more careful notation, the group of *analytic* diffeomorphisms should be called $\mathit{Diff}^{\,\omega}(R^m)$. We also remark that often C^∞ functions are called *smooth*.

Exercises

7.7.1. Verify (7.8).

7.7.2. Find the factorization of the form (7.12) for the map (6.2.10). Show that Lie operators of the form $: f_1 :$ and $: f_2 :$ generate a Lie algebra under commutation. This is the Lie algebra $isp(2n, \mathbb{R})$, the Lie algebra of the inhomogeneous symplectic group $ISp(2n, \mathbb{R})$. What is the dimension of this Lie algebra?

Show that the polynomials f_0 (where f_0 = any constant) and f_1 generate a Lie algebra under the Poisson bracket operation. This is the Lie algebra of the *Heisenberg* group. What is its dimension?

Show that the polynomials f_0, f_1, and f_2 generate a Lie algebra under the Poisson bracket operation. This is the Lie algebra of the *Jacobi* group. For lack of a standard notation, we will denote the Jacobi group by the symbols $J(2n, \mathbb{R})$, and its Lie algebra by $j(2n, \mathbb{R})$. What is the dimension of this Lie algebra? Show that this algebra is homomorphic to that of the inhomogeneous symplectic group. See (5.3.14), (5.3.15), and (5.3.16).

7.7.3. Consider the matrices Q, P, and E defined by the rules

$$Q = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{7.7.24}$$

$$P = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \tag{7.7.25}$$

$$E = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{7.7.26}$$

Show that these matrices form a Lie algebra with the commutator as a Lie product. Consider a two-dimensional phase space with coordinates q and p. Show that the functions q, p, and 1 form a Lie algebra with the Poisson bracket as a Lie product. According to Exercise 7.2

[13] In the context of Accelerator Physics where one is often concerned with charged-particle motion in electromagnetic fields in vacuum, this analyticity can be proved. See Appendix F.

above, this Lie algebra is the Heisenberg Lie algebra in the case of a two-dimensional phase space. Show that the commutator Lie algebra associated with the matrices Q, P, and E has the same structure constants as the Heisenberg Lie algebra, and therefore provides a matrix representation of the Heisenberg Lie algebra. Show that this representation is *not* the adjoint representation.

7.7.4. As shown in Section 7.2, the symplectic group $Sp(2n,\mathbb{R})$ acts transitively on punctured phase space. Consequently, according to the discussion in Section 5.12, it must be possible to view punctured phase space as a coset space of $Sp(2n,\mathbb{R})$ with respect to one of its subgroups. The purpose of this exercise is to find this subgroup. Following the general procedure of Section 5.12, we must look for all $Sp(2n,\mathbb{R})$ transformations of punctured phase space that leave some point fixed. Without loss of generality, this point can be taken to be the point z^1 given by (2.30).

Let M be a symplectic matrix that preserves the vector (phase-space point) z^1,

$$Mz^1 = z^1. \tag{7.7.27}$$

Show from (2.29) and (7.27) that the M_{a1} matrix elements obey the relations

$$M_{a1} = \delta_{a1}. \tag{7.7.28}$$

Let $\mathcal{M}(M)$ be a symplectic map associated with M by the relation

$$\mathcal{M}(M)z_a = \sum_b (M^T)_{ab} z_b. \tag{7.7.29}$$

Show from (7.29) that $\mathcal{M}$ is symplectic, and from (7.28) and (7.29) that $\mathcal{M}$ satifies the relation

$$\mathcal{M}q_1 = q_1. \tag{7.7.30}$$

Conversely, show that if $\mathcal{M}$ is a symplectic map of the form (7.29) that also satisfies (7.30), then (7.28) and (7.27) are satisfied. Consequently, we can concentrate on finding the general solution to (7.30).

We begin by working near the identity, and write $\mathcal{M}$ in the form

$$\mathcal{M} = \exp(:f_2:). \tag{7.7.31}$$

Show that requiring (7.30) for $\mathcal{M}$ near the identity is equivalent to requiring that f_2 satisfy the relation

$$:f_2:q_1 = 0. \tag{7.7.32}$$

Show that any f_2 that satisfies (7.32) cannot depend on p_1, and is therefore a linear combination of the polynomials q_1^2, $q_1\tilde{f}_1$, and $\tilde{f}_2$. Here $\tilde{f}_1$ and $\tilde{f}_2$ denote homogeneous polynomials in the remaining variables $q_2 \cdots q_n$ and $p_2 \cdots p_n$ of degrees 1 and 2, respectively. Show that the polynomials $\tilde{f}_2$ give a representation of the Lie algebra $sp[2(n-1), R]$. Review Exercise 7.2, and consider the Poisson bracket Lie algebra generated by f_0, f_1, and f_2, the Jacobi Lie algebra $j(2n, \mathbb{R}R)$. You should have found that it has dimension $n(2n+3)+1$. Show that the polynomials q_1^2, $q_1\tilde{f}_1$, and $\tilde{f}_2$ have the Lie algebra $j[2(n-1), R]$ under the Poisson bracket operation, and that the Lie operators $: q_1^2 :$, $: q_1\tilde{f}_1 :$, and $: \tilde{f}_2 :$ have that same Lie algebra under commutation. Show that the dimensions of $sp(2n,\mathbb{R})$ and $j[2(n-1),\mathbb{R}]$ are related by the equation

$$\dim\{sp(2n,\mathbb{R})\} - \dim\{j[2(n-1),\mathbb{R}]\} = 2n. \tag{7.7.33}$$

Let $J[2(n-1),\mathbb{R}]$ be the Lie group generated by the Lie operators $: q_1^2 :$, $: q_1\tilde{f}_1 :$, and $: \tilde{f}_2 :$. Show that the general $\mathcal{M}$ in $J[2(n-1),\mathbb{R}]$ can be written in the form

$$\mathcal{M} = \exp(\alpha : q_1^2 :) \exp(: \tilde{f}_2^c :) \exp(: \tilde{f}_2^a :) \exp(: q_1\tilde{f}_1 :), \tag{7.7.34}$$

where α is an arbitrary parameter. Show that (7.34) is the most general $Sp(2n,\mathbb{R})$ transfer map that satisfies the relation

$$\mathcal{M}q_1 = q_1. \tag{7.7.35}$$

Hint: If you are having difficulty, see the beginning of Section 9.4.

Let H be the subgroup of $Sp(2n,\mathbb{R})$ consisting of all matrices M that satisfy (7.27). Suppose M^1 and M^2 are in H. Then we have the relation

$$\begin{aligned}
\mathcal{M}(M^1)\mathcal{M}(M^2)z_a &= \mathcal{M}(M^1)\sum_b [(M^2)^T]_{ab} z_b \\
&= \sum_{b,c} [(M^2)^T]_{ab}[(M^1)^T]_{bc} z_c \\
&= \sum_c [(M^2)^T(M^1)^T]_{ac} z_c = \sum_c [(M^1M^2)^T]_{ac} z_c \\
&= \mathcal{M}(M^1M^2)z_a,
\end{aligned} \tag{7.7.36}$$

or, more compactly put,

$$\mathcal{M}(M^1)\mathcal{M}(M^2) = \mathcal{M}(M^1M^2). \tag{7.7.37}$$

From (7.37) we see that the subgroup H is a matrix realization of $J[2(n-1),\mathbb{R}]$. Let G be the group $Sp(2n,\mathbb{R})$. Show that the coset space G/H has dimension $2n$, the expected dimension for the phase space under consideration.

7.7.5. Section 6.1 defined a symplectic map to be a map (of a $2n$-dimensional space into itself) whose Jacobian matrix M is symplectic. Suppose that M is instead required to be *orthogonal*. Show that such maps also form a group, which might be called the group of orthogonal maps. Consider the Taylor expansion of such a map in the 2-dimensional case. Show that, unlike the expansion for symplectic maps, *only* constant and linear terms can occur in the Taylor expansion! You have verified a special case of the fact that, unlike symplectic maps, orthogonal maps are trivial in the sense that they consist only of translations and *linear* orthogonal transformations.

7.8 Other Factorizations

In addition to the factorization (7.23), there are other factorizations that are often useful. First, as will be shown in Chapter 9, it is possible to bring all first degree polynomials over to the left. In this case, $\mathcal{N}$ has the factorization

$$\mathcal{N} = \exp(:f_1:)\exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_3:)\exp(:f_4:)\cdots . \tag{7.8.1}$$

Here the polynomials f_m are generally different from those in (7.23).[14] The factorization (8.1) will be called *forward* or *ascending* factorization. Second, it is often useful to have a factorization of the form

$$\mathcal{N} = \cdots\exp(:f_4:)\exp(:f_3:)\exp(:f_2^a:)\exp(:f_2^c:)\exp(:f_1:). \tag{7.8.2}$$

Here again the polynomials f_m are generally different from those in (8.1) or (7.23). The factorization (8.2) will be called *reverse* or *descending* factorization. Finally, it is often useful to have *mixed* factorizations where the f_m terms with $m > 1$ are ascending or descending, and the $\exp(:f_1:)$ term is at the beginning, or at the end.

Exercises

7.8.1. Find other factorizations for the inhomogeneous symplectic group. See Exercises (3.9.2) and (7.2).

7.9 Coordinates and Connectivity

Suppose G is a finite-dimensional Lie group, and let $B_1,\ B_2,\ \cdots\ B_n$ be a basis for the associated Lie algebra L. To be more specific, suppose G is realized as a group of matrices, and suppose that some element g in G is sufficiently near the identity so that it can be written in the form

$$g = \exp(\sum_{j=1}^{n} \xi_j B_j). \tag{7.9.1}$$

The parameters ξ_j are called *canonical coordinates* (for g) of the *first kind.* Another possibility is to write g in the form

$$g = \exp(\eta_1 B_1)\exp(\eta_2 B_2)\cdots\exp(\eta_n B_n). \tag{7.9.2}$$

The parameters η_j are called *canonical coordinates* of the *second kind.*[15] For the case of a finite-dimensional Lie group, at least in some neighborhood of the identity, one may pass in principle from one kind of coordinates to the other with the aid of the BCH formula. (See Section 3.7. See also Exercise 10.4.2.) This may not be possible in the infinite-dimensional case because then the BCH series may not have any domain of convergence. See Section 33.7.

[14] In Section 6.6 we saw that the $2n$ functions required to specify a map in $2n$ variables can be related to a single function in the symplectic case. This fact is also evident from (8.1) if we formally regard the f_j as the homogeneous parts of a single function f.

[15] Note that there are in principle as many as $n!$ different canonical coordinates of the second kind because, since the B_j may not commute, the order of the factors in (9.2) may be important.

Reference to (7.23) shows that if the factorization process succeeds, the general analytic symplectic map $\mathcal{N}$ has been given coordinates that are a hybrid of canonical coordinates of the first and second kinds. The map is written as a product of exponentials as in (9.2), and each exponential is a sum of terms as in (9.1).

Suppose $\mathcal{N}$ can be factored as in (7.23). Then it is easy to see that $\mathcal{N}$ is *connected* to the identity map $\mathcal{I}$ by a continuous family of symplectic maps. Indeed, let λ be a parameter and let $\mathcal{N}(\lambda)$ be the map

$$\mathcal{N}(\lambda) = \exp(-\lambda : h_1 :)[\exp(\lambda : f_2^c :)\exp(\lambda : f_2^a :)\exp(\lambda : f_3 :)\cdots]\exp(\lambda : g_1 :). \tag{7.9.3}$$

It is evident that $\mathcal{N}(\lambda)$ is a symplectic map for all λ with the properties

$$\mathcal{N}(0) = \mathcal{I}\ ,\ \mathcal{N}(1) = \mathcal{N}. \tag{7.9.4}$$

The argument just given lacks generality because we had to assume that $\mathcal{N}$ is analytic and that the factorization process converges. However, we can do better. Suppose we assume only that $\mathcal{N}$ has at least a first few derivatives so that (7.13) can be written in the form

$$\bar{z}_a = \tilde{z}_a^f + \sum_b R_{ab}(z-\tilde{z}^i)_b + O[(z-\tilde{z}^i)^2]. \tag{7.9.5}$$

Then, by arguments similar to those of Section 7.7, the map $\mathcal{N}$ can be rewritten in the form

$$\mathcal{N} = \exp(- : h_1 :)\exp(: f_2^c :)\exp(: f_2^a :)\mathcal{P}\exp(: g_1 :) \tag{7.9.6}$$

where $\mathcal{P}$ is a symplectic map that sends the origin into itself and has an expansion of the form

$$\bar{z}_a = \mathcal{P}z_a = z_a + W_a(z) \tag{7.9.7}$$

with

$$W_a(z) = O[(z)^2]. \tag{7.9.8}$$

Define a map $\mathcal{P}(\lambda)$ by the relation

$$\bar{z}_a = \mathcal{P}(\lambda)z_a = (1/\lambda)[\lambda z_a + W_a(\lambda z)] = z_a + (1/\lambda)W_a(\lambda z). \tag{7.9.9}$$

Here λz denotes the collection of quantities λz_b. Evidently, in view of (9.8), we have the relations

$$\mathcal{P}(0) = \mathcal{I}\ ,\ \mathcal{P}(1) = \mathcal{P}. \tag{7.9.10}$$

Also, $\mathcal{P}(\lambda)$ is a symplectic map for all values of λ. To see this, observe that $\mathcal{P}(\lambda)$ can be written as a product of three maps in the form

$$\mathcal{P}(\lambda) = [(1/\lambda)\mathcal{I}][\mathcal{P}][\lambda\mathcal{I}]. \tag{7.9.11}$$

Although the maps $[(1/\lambda)\mathcal{I}]$ and $[\lambda\mathcal{I}]$ are not symplectic (if $\lambda \neq 1$), the product (9.11) is. Indeed, denoting by $P(\lambda, z)$ and $P(z)$ the Jacobian matrices of $\mathcal{P}(\lambda)$ and $\mathcal{P}$, respectively, use of the chain rule gives the relation

$$P(\lambda, z) = [(1/\lambda)I][P(\lambda z)][\lambda I] = P(\lambda z). \tag{7.9.12}$$

Since we know that $\mathcal{P}$ is a symplectic map, $P(z)$ will be a symplectic matrix. According to (9.12) $P(\lambda, z)$, being equal to $P(\lambda z)$, is also a symplectic matrix because $P(z)$ is a symplectic matrix for *all* values of z. Finally, because $P(\lambda, z)$ is a symplectic matrix, $\mathcal{P}(\lambda)$ is a symplectic map.[16] In analogy with (9.6), we now define $\mathcal{N}(\lambda)$ by writing

$$\mathcal{N}(\lambda) = \exp(-\lambda : h_1 :) \exp(\lambda : f_2^c :) \exp(\lambda : f_2^a :) \mathcal{P}(\lambda) \exp(\lambda : g_1 :). \tag{7.9.13}$$

Evidently $\mathcal{N}(\lambda)$ is a symplectic map for all λ and has the desired properties (9.4). We have shown that if a symplectic map $\mathcal{N}$ has at least a first few derivatives, then it is connected to the identity map by a continuous family of symplectic maps.

Of course, the family just constructed will generally differ from that given by (9.3). There are many families of symplectic maps that connect a given symplectic map to the identity map.[17] We note that, according to Section 6.4, for each family there is a corresponding Hamiltonian that generates it.

7.10 Storage Requirements

How much computer memory is required to store a symplectic map in the Taylor form (7.7), and how much memory is required to store the corresponding Lie form (8.1)? Suppose the Taylor map (7.7) is truncated by discarding all terms having degree $(D+1)$ and higher. We denote this truncated Taylor map by $\mathcal{T}_{D+1}$. Then (7.7) has the truncated form

$$\tilde{z}_a = \mathcal{T}_{D+1} z_a = \sum_{m=0}^{D} g_a(m; z) \tag{7.10.1}$$

where, as in Section 7.6, each $g_a(m; z)$ is a homogeneous polynomial of degree m. According to (8.1) and our discussion of the relation between Taylor and Lie maps, there is a map $\mathcal{M}$ in Lie form that corresponds to $\mathcal{T}_{D+1}$ (they both have the same Taylor expansions through terms of degree D), and this map has a truncated factored product representation of the form

$$\mathcal{M} = \exp(: f_1 :) \exp(: f_2^c :) \exp(: f_2^a :) \exp(: f_3 :) \exp(: f_4 :) \cdots \exp(: f_{D+1} :). \tag{7.10.2}$$

Let $S(m, d)$ be the total number of monomials in d variables having degrees 1 through m. We know from (3.40) that the total number of monomials in d variables having degree m is $N(m, d)$. Consequently, $S(m, d)$ is given by the relation

$$S(m, d) = \sum_{k=1}^{m} N(k, d). \tag{7.10.3}$$

The sum (10.3) can be evaluated to give the result

$$S(m, d) = \begin{pmatrix} m+d \\ m \end{pmatrix} - 1 = \frac{(m+d)!}{m!d!} - 1. \tag{7.10.4}$$

[16] This scaling by some factor λ that we have used is sometimes caled *Alexander's Trick*.

[17] For example, in (9.13) one might replace λ as the coefficient of f_2^c by λ^2. The reader should be able to construct other examples.

See Exercises 10.1 and 10.2. Table 10.1 below lists values of $S(m,d)$ for various values of m and d.

Now let $S_L(D,d)$ be the number of storage locations required to specify the Lie map (10.2). We know that the specification of an f_m requires $N(m,d)$ numbers (with $d=6$ for the case of a 6-dimensional phase space). Consequently, comparison of (10.2) and (10.3) gives the result

$$S_L(D,d) = S(D+1,d) = \left(\begin{array}{c} D+d+1 \\ D+1 \end{array}\right) - 1 = \frac{(D+d+1)!}{(D+1)!d!} - 1. \tag{7.10.5}$$

Correspondingly, let S_T be the number of locations required to specify the truncated Taylor map (10.1). We know that the specification of a particular $g_a(m,z)$ requires $N(m,d)$ numbers. Consequently, $S_T(D,d)$ must be given by the relation

$$S_T(D,d) = d\sum_{k=0}^{D} N(k,d) = d[S(D,d)+1] = \frac{d(D+d)!}{D!d!}. \tag{7.10.6}$$

Note that d must be even in both (10.5) and (10.6) because phase space is even dimensional.

Finally, let us compare $S_T(D,d)$ and $S_L(D,d)$ for various values of d and D. Table 10.2 below lists values of S_T, S_L, and the ratio S_T/S_L, for $d=4$ and $d=6$ and various values of D. We conclude that (for modest values of D) storing a 6-dimensional phase-space map in Taylor form requires about 3 times more storage locations than the equivalent Lie form. For large D values this ratio approaches 6. This difference in storage requirements for the Taylor and Lie forms of a symplectic map arises from the fact that the Taylor form makes no use of the symplectic condition. Indeed, the Lie form contains exactly the minimal information required to specify a symplectic map while the Taylor form has all the coefficients required to specify the most general (analytic diffeomorphic) map.[18]

[18] Observe that $S_L(3,6)$, the number of storage locations required to store a 6-variable symplectic map through third order in Lie form, has the value $S_L(3,6) = 209$. Curiously, according to the Los Angles Times, in 1987 (when MaryLie 3.0 was being written) the Chinese Communist Party Central Committee, China's highest governing body, had 209 members.

Table 7.10.1: Number of monomials of degree 1 through m in various numbers of variables.

m	$S(m,4)$	$S(m,5)$	$S(m,6)$	$S(m,7)$	$S(m,8)$	$S(m,9)$	$S(m,10)$	$S(m,11)$
1	4	5	6	7	8	9	10	11
2	14	20	27	35	44	54	65	77
3	34	55	83	119	164	219	285	363
4	69	125	209	329	494	714	1000	1364
5	125	251	461	791	1286	2001	3002	4367
6	209	461	923	1715	3002	5004	8007	12375
7	329	791	1715	3431	6434	11439	19447	31823
8	494	1286	3002	6434	12869	24309	43757	75581
9	714	2001	5004	11439	24309	48619	92377	167959
10	1000	3002	8007	19447	43757	92377	184755	352715
11	1364	4367	12375	31823	75581	167959	352715	705431
12	1819	6187	18563	50387	125969	293929	646645	1352077

Table 7.10.2: Storage Requirements for Taylor and Lie Maps.

D	$S_T(D,4)$	$S_L(D,4)$	$S_T(D,4)/S_L(D,4)$	$S_T(D,6)$	$S_L(D,6)$	$S_T(D,6)/S_L(D,6)$
2	60	34	1.8	168	83	2.0
3	140	69	2.0	504	209	2.4
4	280	125	2.2	1260	461	2.7
5	504	209	2.4	2772	923	3.0
6	840	329	2.6	5544	1715	3.2
7	1320	494	2.7	10,296	3002	3.4
8	1980	714	2.8	18,018	5004	3.6
9	2860	1000	2.9	30,030	8007	3.8
10	4004	1364	2.9	48,048	12,375	3.9
11	5460	1819	3.0	74,256	18,563	4.0
12	7280	2379	3.1	111,384	27,131	4.1

Exercises

7.10.1. Verify (10.3) through (10.6). [Hint: Use the relations (3.52) through (3.54).] Show that S can be generated using the recursion relation

$$S(m,d) = S(m,d-1) + S(m-1,d) + 1 \tag{7.10.7}$$

with the initial conditions

$$S(m,1) = m, \tag{7.10.8}$$

$$S(1,d) = d. \tag{7.10.9}$$

Show that S also satisfies the relation

$$S(m,d) = S(m-1,d) + N(m,d). \tag{7.10.10}$$

7.10.2. The relation (10.4) can be derived directly from the definition of $S(m,d)$ as the total number of monomials in d variables having degrees 1 through m. Let $S_0(m,d)$ be the total number of monomials in d variables having degrees 0 through m. Then we evidently have the relation

$$S_0(m,d) = S(m,d) + 1. \tag{7.10.11}$$

Show that from its definition $S_0(m,d)$ obeys the relations

$$S_0(1,1) = 2, \tag{7.10.12}$$

$$S_0(m,1) = m + 1, \tag{7.10.13}$$

$$S_0(2,2) = 6. \tag{7.10.14}$$

Next show that S_0 obeys the recursion relation

$$S_0(m,d) = S_0(m-1,d) + S_0(m,d-1). \tag{7.10.15}$$

Hint: The number of monomials in d variables having degrees 0 through m is the number of monomials having degree 0 through $m-1$, which is $S_0(m-1,d)$, plus the number of monomials having degree m. We have already agreed to let $N(m,d)$ be the number of monomials having degree m. Homogeneous monomials of degree m in the d variables $z_1 \cdots z_d$ can be viewed as monomials in the variables $z_1 \cdots z_{d-1}$ of degree 0 through m multiplied by the appropriate power of z_d to make the total degree exactly m. Thus, we have the relation

$$N(m,d) = S_0(m,d-1). \tag{7.10.16}$$

Finally, use the recursion relation (10.15) with the inital conditions (10.12) through (10.14) to show that S_0 is given by the relation

$$S_0(m,d) = \left(\begin{array}{c} m+d \\ m \end{array} \right) = \frac{(m+d)!}{m!d!}. \tag{7.10.17}$$

Hint: Recall that the binomial coefficients satisfy the recursion relation (3.52).

7.10.3. Evaluate the ratio S_T/S_L for various values of d (say $d=4$ and $d=6$) and various values of D. Show that this ratio approaches d in the limit of large D.

7.10.4. Compute the quantity $[S_L(D+1,d) - S_L(D,d)]/S_L(D,d)$ for large D. This quantity is the limiting fractional incremental increase in storage required to include one order higher aberration effects.

Bibliography

Combinatorics

[1] A. Nijenhuis and H.S. Wilf, *Combinatorial Algorithms for Computers and Calculators*, Academic Press (1978).

[2] J. Riordan, *An Introduction to Combinatorial Analysis*, John Wiley (1958).

[3] The method of Exercise 3.12 is due to M. Venturini.

Taylor Maps and Jets

[4] E. Hille and R. Phillips, *Functional Analysis and Semi-groups*, American Mathematical Society Colloquium Publications, Volume 31 (1957).

[5] P.W. Michor, *Manifolds of Differentiable Mappings*, Shiva Publishing Limited (1980).

Factorization

[6] A. Dragt and J. Finn, "Lie Series and Invariant Functions for Analytic Symplectic Maps", *J. Math. Phys.* **17**, 2215 (1976).

Wigner Rotation

[7] See the Web site https://en.wikipedia.org/wiki/Wigner_rotation.

Jacobi Group

[8] R. Berndt and R. Schmidt, *Elements of the Representation Theory of the Jacobi Group*, (Progress in Mathematics; Vol. 163), Birkhäuser Verlag (1998).

Connectivity

[9] P.J. Chanell, "Hamiltonian suspensions of symplectomorphisms: an alternative approach to design problems", *Physica D* **127**, pp. 117-130 (1999).

Group Theory and Analyticity

[10] B. Beers and A. Dragt, "New Theorems about Spherical Harmonic Expansions and $SU(2)$", *Journal of Mathematical Physics*, Volume 11, pp. 2313-2328 (1970).

Computation of Charged-Particle Beam Transport

[11] A. Dragt et al., *MaryLie 3.0 Users' Manual* (2003). See www.physics.umd.edu/dsat/.

Chapter 8

A Calculus for Lie Transformations and Noncommuting Operators

Section 6.4 showed that Hamiltonian flows produce symplectic maps, and Sections 7.2 and 7.7 showed that the general analytic symplectic map (7.7.13) can be written in the factored product form (7.7.23). In addition, (7.4.1) gives an explicit representation for the symplectic map in the case of a time-independent Hamiltonian. See also (7.4.18). In subsequent sections these results will be applied to charged particle beam transport, light optics, and orbits in circular machines. The purpose of this chapter is to provide a collection of formulas for the manipulation of Lie transformations and noncommuting operators in general. Some formulas will be used to compute the product of two symplectic maps when each is written in factored product form. Others will be used to combine various exponents in a factored product decomposition into a single exponent. Still others will be used to produce factored product decompositions. Where necessary, discussion will be restricted to symplectic maps that send the origin into itself. (See Section 7.6.) This restriction will subsequently be removed in Chapter 9.

8.1 Adjoint Lie Operators and the Adjoint Lie Algebra

Work with noncommuting quantities is often facilitated by the concept of an *adjoint* Lie operator. Let $:f:$ be some Lie operator, and let $:g:$ be any other Lie operator. The *adjoint* of the Lie operator $:f:$, which will be denoted by the symbol $\#:f:\#$, is a kind of super operator that acts on other Lie operators according to the rule

$$\#:f:\#:g:=\{:f:,:g:\}. \tag{8.1.1}$$

Here, the right side of (1.1) denotes the commutator as in (5.3.10). Thanks to (5.3.14), the relation (1.1) can also be written in the form

$$\#:f:\#:g:=\{:f:,:g:\}=:[f,g]:. \tag{8.1.2}$$

We see that adjoint Lie operators act on Lie operators, and send them to other Lie operators. Furthermore, this action is linear. That is, we have the relation

$$\# : f : \#(a : g : + \, b : h :) = a\# : f : \# : g : + b\# : f : \# : h : . \tag{8.1.3}$$

We remark that the word *adjoint* is much overused in mathematics, and is not to be confused in this context with the Hermitian conjugate (also sometimes referred to as a Hermitian adjoint) defined in (7.3.15). To simplify notation in some cases where no confusion can arise, the set of colons in the symbol $\# : f : \#$ for the adjoint of the Lie operator $: f :$ will often be omitted. That is, the abbreviated symbol $\#f\#$ will often be used to serve for the complete symbol $\# : f : \#$.

Powers of $\# : f : \#$ or $\#f\#$ can be defined by repeated application of (1.1). For example, $\#f\#^2$ is defined by the relation

$$\#f\#^2 : g := \{: f :, \{: f :, : g :\}\}. \tag{8.1.4}$$

Also, $\#f\#$ to the zero power is defined to be the identity operator,

$$\#f\#^0 : g :=: g : . \tag{8.1.5}$$

The set of adjoint Lie operators $\# : f : \#$ can also be made into a Lie algebra in its own right. This Lie algebra is called the *adjoint Lie algebra*. First, there is obviously the relation

$$a\#f\# + b\#g\# = \#(af + bg)\#. \tag{8.1.6}$$

That is, the set of adjoint Lie operators forms a linear vector space. Next, we define the Lie product of any two adjoint Lie operators to be their commutator,

$$\{\#f\#, \#g\#\} = \#f\#\#g\# - \#g\#\#f\#. \tag{8.1.7}$$

We note that this definition of a Lie product obviously satisfies the requirements 1 through 4 listed in Section (3.7) for a Lie algebra. It also satisfies requirement 5 since commutators satisfy the Jacobi condition. Of course, we must also show that the Lie product (commutator) of two adjoint Lie operators is again an adjoint Lie operator. Let $: h :$ be an arbitrary Lie operator. Then we have the results

$$\#f\#\#g\# : h := \#f\#\{: g :, : h :\} = \{: f :, \{: g :, : h :\}\},$$

$$\#g\#\#f\# : h := \#g\#\{: f :, : h :\} = \{: g :, \{: f :, : h :\}\},$$

$$\begin{aligned} \{\#f\#, \#g\#\} : h : &= \{: f :, \{: g :, : h :\}\} - \{: g :, \{: f :, : h :\}\} \\ &= \{: f :, \{: g :, : h :\}\} + \{: g :, \{: h :, : f :\}\}. \end{aligned} \tag{8.1.8}$$

Now use the Jacobi condition for the Lie algebra of Lie operators to find the relation

$$\{: f :, \{: g :, : h :\}\} + \{: g :, \{: h :, : f :\}\} = -\{: h :, \{: f :, : g :\}\} = \{\{: f :, : g :\}, : h :\}. \tag{8.1.9}$$

We see that (1.8) can be rewritten in the form

$$\{\#f\#, \#g\#\} : h :=\{\{: f :, : g :\}, : h :\} = \#\{: f :, : g :\}\# : h : . \tag{8.1.10}$$

Since the Lie operator $: h :$ is arbitrary, it follows that we have the result

$$\{\#f\#, \#g\#\} = \#\{: f :, : g :\}\# = \# : [f, g] : \#. \tag{8.1.11}$$

Thus the Lie product of two adjoint Lie operators is indeed an adjoint Lie operator.

Our discussion should have a familiar ring. It parallels, in fact, the material at the end of Section 3.7 and the treatment of Lie operators given in Section 5.3. Reviewing these sections, we see that the commutator Lie algebra of Lie operators is actually the adjoint Lie algebra of the underlying Poisson bracket Lie algebra. And, consequently, the “adjoint” we have been discussing is really the “adjoint-adjoint” of the basic Poisson bracket Lie algebra.

Exercises

8.1.1. Prove (1.3).

8.1.2. Prove (1.6). Verify that requirements 1 through 5 listed in Section 3.7 for a Lie algebra are satisfied by the adjoint Lie algebra.

8.1.3. Describe in detail how the adjoint Lie algebra is the adjoint-adjoint of the basic Poisson bracket Lie algebra. What would the adjoint-adjoint-adjoint Lie algebra be?

8.2 Formulas Involving Adjoint Lie Operators

There are several useful formulas involving adjoint Lie operators. First, we have the relations

$$\#f\#^0 : g :=: g :=: (: f :^0 g) :,$$

$$\#f\# : g := \{: f :, : g :\} =: [f, g] :=: (: f : g) : . \tag{8.2.1}$$

Here use has been made of (5.3.14). From these relations we have by induction the general result

$$\#f\#^n : g :=: (: f :^n g) : . \tag{8.2.2}$$

Second, the definition of $\#f\#$ can be extended to let $\#f\#$ act on any sum or product, or sum of products, or even power series, of Lie operators. Suppose $F(: g :, : h :, \cdots)$ is any function of a collection of Lie operators $: g :, : h :, \cdots$. Then we define the action of $\#f\#$ on F in analogy to (1.1) by the rule

$$\#f\#F = \{: f :, F\}. \tag{8.2.3}$$

As a special case of (2.3) we have the relation

$$\#f\#(: g :: h :) = \{: f :, : g :: h :\}$$

$$\begin{aligned} &= \{:f:,:g:\}:h: + :g:\{:f:,:h:\} \\ &= (\#f\#:g:):h: + :g:(\#f\#:h:). \end{aligned} \tag{8.2.4}$$

We see that the adjoint Lie operator $\#f\#$ is a *derivation* with respect to the multiplication of Lie operators.

Now suppose that $:f:$ and $:g:$ are any two Lie operators. We then find that

$$\exp(:f:):g:\exp(-:f:) = \exp(\#f\#):g:. \tag{8.2.5}$$

Here, as the notation suggests,

$$\exp(\#f\#) = \sum_{m=0}^{\infty} \#f\#^m/m!. \tag{8.2.6}$$

This result is sometimes called *Hadamard's lemma.*[1] To see that (2.5) is correct, consider the operator function $O(\tau)$ defined by the equation

$$O(\tau) = \exp(\tau:f:):g:\exp(-\tau:f:), \tag{8.2.7}$$

where τ is a parameter. Then we have the relation

$$O(0) =:g:. \tag{8.2.8}$$

Further, we find by differentiation of (2.7) the relation

$$dO/d\tau =:f:O - O:f:= \{:f:,O\} = \#f\#O. \tag{8.2.9}$$

The solution to this differential equation with the initial condition (2.8) is given by the relation

$$O(\tau) = \exp(\tau\#f\#):g:. \tag{8.2.10}$$

Now set $\tau = 1$ in (2.10) to obtain the desired result.

From (2.2) it follows that we also have the relation

$$\exp(\#f\#):g:=:\exp(:f:)g:. \tag{8.2.11}$$

Consequently, (2.5) can also be written in the form

$$\exp(:f:):g:\exp(-:f:) =:\exp(:f:)g:. \tag{8.2.12}$$

Because $\#f\#$ is a derivation, see (2.4), there is an even more general result. Let $F(:g:,:h:,\cdots)$ be a function of a collection of Lie operators of the type described above. Then we have the relations

$$\exp(:f:)F(:g:,:h:,\cdots)\exp(-:f:) = \exp(\#f\#)F(:g:,:h:,\cdots), \tag{8.2.13}$$

$$\begin{aligned} \exp(\#f\#)F(:g:,:h:,\cdots) &= F(\exp(\#f\#):g:,\exp(\#f\#):h:,\cdots) \\ &= F(:\exp(:f:)g:,:\exp(:f:)h:,\cdots). \end{aligned} \tag{8.2.14}$$

[1] A Web search reveals that there are also other Hadamard lemmas.

As a special case of (2.13) and (2.14) we find the results

$$\exp(:f:) :g:^m \exp(-:f:) = [\exp(\#f\#) :g:]^m, \tag{8.2.15}$$

$$\exp(:f:)\exp(:g:)\exp(-:f:) = \exp[\exp(\#f\#) :g:]. \tag{8.2.16}$$

This discussion should also have a familiar ring. See Section 5.4. Here we are exploiting the fact that exp(#f#) is an *isomorphism* with respect to Lie operator multiplications.

The relations (2.13) and (2.14) can also be derived directly. Consider, for example, the simple case

$$\begin{aligned} &\exp(:f:) :g::h: \exp(-:f:) = \exp(\#f\#) :g::h: \\ &= (\exp(\#f\#) :g:)(\exp(\#f\#) :h:) =: \exp(:f:)g :: \exp(:f:)h : . \end{aligned} \tag{8.2.17}$$

The relation (2.17) can also be found by using the fact that the expression $\exp(-:f:)\exp(:f:)$ is the identity operator and employing (2.11) and its analog for $:h:$,

$$\begin{aligned} \exp(:f:) :g::h: \exp(-:f:) &= \exp(:f:) :g: \exp(-:f:)\exp(:f:) :h: \exp(-:f:) \\ &= :\exp(:f:)g :: \exp(:f:)h : . \end{aligned} \tag{8.2.18}$$

We can carry (2.15) a step further using (2.11) to find the relation

$$\exp(:f:) :g:^m \exp(-:f:) =: \exp(:f:)g :^m, \tag{8.2.19}$$

which is a generalization of (2.12). Moreover, (2.19) in turn, or direct use of (2.16) and (2.11), yields the relation

$$\begin{aligned} \exp(:f:)\exp(:g:)\exp(-:f:) &= \exp[:\exp(:f:)g:] \\ &= \exp[:g(\exp:f:z):]. \end{aligned} \tag{8.2.20}$$

This relation gives a result for the multiplication of a particular combination of Lie transformations.

The relations (2.2), (2.13), and (2.14) have obvious generalizations to the case of several Lie operators. Consider, for example, the case of two Lie operators $:e:$ and $:f:$. Then, (2.2) has the generalization

$$\#e\#^m\#f\#^n :g:=: (:e:^m:f:^n g): . \tag{8.2.21}$$

Indeed, suppose E is any function consisting of sums, products, sums of products, or even power series in two arguments. Then (2.2) has the generalization

$$E(\#e\#,\#f\#) :g:=: E(:e:,:f:)g : . \tag{8.2.22}$$

Analogous results hold for any number of Lie operators and functions of any number of arguments.

As for the relations (2.13) and (2.14), they can be generalized to any number of factors. For example, for the case of two factors, we have the results

$$\begin{aligned} &\exp(:e:)\exp(:f:)F(:g:,:h:,\cdots)\exp(-:f:)\exp(-:e:) \\ &\quad = \exp(\#e\#)(\exp\#f\#)F(:g:,:h:,\cdots), \end{aligned} \tag{8.2.23}$$

$$\begin{aligned} &\exp(\#e\#)\exp(\#f\#)F(:g:,:h:,\cdots) \\ &\quad = F(:\exp(:e:)\exp(:f:)g:,\ :\exp(:e:)\exp(:f:)h:,\cdots). \end{aligned} \tag{8.2.24}$$

Analogous results hold for any number of factors. Consider, for example, the factors that compose (7.6.3). In this case we have the result

$$\mathcal{M}F(:g:,:h:,\cdots)\mathcal{M}^{-1}$$

$$= \exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_3:)\exp(:f_4:)\cdots\times$$

$$F(:g:,:h:,\cdots)\cdots\exp(-:f_4:)\exp(-:f_3:)\exp(-:f_2^a:)\exp(-:f_2^c:)$$

$$= \exp(\#f_2^c\#)\exp(\#f_2^a\#)\exp(\#f_3\#)\exp(\#f_4\#)\cdots F(:g:,:h:,\cdots)$$

$$= F(\exp(\#f_2^c\#)\exp(\#f_2^a\#)\exp(\#f_3\#)\exp(\#f_4\#)\cdots:g:,\cdots)$$

$$= F(:\exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_3:)\exp(:f_4:)\cdots g:,\cdots)$$

$$= F(:\mathcal{M}g:,:\mathcal{M}h:,\cdots)$$

$$= F(:g(\mathcal{M}z):,:h(\mathcal{M}z):,\cdots). \tag{8.2.25}$$

As special cases of (2.25) we have the result

$$\mathcal{M}:g(z):\mathcal{M}^{-1} =: \mathcal{M}g(z) :=: g(\mathcal{M}z):, \tag{8.2.26}$$

which is an extension of (2.12), and the result

$$\mathcal{M}[\exp:g(z):]\mathcal{M}^{-1} = \exp:\mathcal{M}g(z) := \exp:g(\mathcal{M}z):, \tag{8.2.27}$$

which is an extension of (2.20).

We close this section with another useful result for the multiplication of Lie transformations. It is the analog of formulas (3.7.33) and (3.7.34) for Lie operators. Suppose $:f:$ and $:g:$ are any two Lie operators. Then one has the BCH formula

$$\exp(s:f:)\exp(t:g:) = \exp(s:f: + t:g:$$

$$+(st/2)\{:f:,:g:\} + (s^2t/12)\{:f:,\{:f:,:g:\}\}$$

$$+ (st^2/12)\{:g:,\{:g:,:f:\}\} + \cdots). \tag{8.2.28}$$

Moreover, using (5.3.14) and (2.2), (2.28) can also be written in the form

$$\exp(s:f:)\exp(t:g:) = \exp(:h:) \tag{8.2.29}$$

with

$$h = sf + tg + (st/2)[f,g]$$

$$+ (s^2t/12):f:^2 g + (st^2/12):g:^2 f + \cdots. \tag{8.2.30}$$

Exercises

8.2.1. Prove (2.2).

8.2.2. Prove (2.4).

8.2.3. Carry out the steps that lead to (2.5), (2.11), and (2.12). Also verify (2.5) term by term for at least the first few terms by comparing power series expansions.

8.2.4. Prove (2.13) and (2.14). Hint: Imitate the proof of (2.5) and (2.12). Also verify (2.13) term by term for at least the first few terms by comparing power series expansions.

8.2.5. Construct a general proof of (2.13) and (2.14) by employing the method used to prove (2.17).

8.2.6. Prove (2.20).

8.2.7. Prove (2.21) and (2.22).

8.2.8. Prove (2.23), (2.24), and (2.25). Show that (2.25) also holds for $\mathcal{M}$ of the form (7.7.23).

8.2.9. Prove (2.30).

8.2.10. Review Exercise 3.7.31. Since the Lie algebras $su(2)$ and $so(3,\mathbb{R})$ are the same, we may expect a close relation between the groups $SU(2)$ and $SO(3,\mathbb{R})$. The purpose of this exercise is to show that there is a two-to-one homomorphism between $SU(2)$ and $SO(3,\mathbb{R})$. We will also find several formulas, involving $SU(2)$ and $SO(3,\mathbb{R})$ and their Lie algebras, that will be useful for later work.

Suppose $v \in SU(2)$. Consider matrices $\bar{K}^\alpha(v)$ defined by the relation

$$\bar{K}^\alpha(v) = v^\dagger K^\alpha v. \tag{8.2.31}$$

Verify that the $\bar{K}^\alpha(v)$ are anti-Hermitian and traceless. It follows, since the K^β form a basis for the set of 2×2 traceless anti-Hermitian matrices, that there must be a relation of the form

$$\bar{K}^\alpha(v) = \sum_\beta M_{\alpha\beta}(v) K^\beta \tag{8.2.32}$$

where $M(v)$ is a 3×3 matrix to be determined. Show, in view of the definitions (3.7.169) through (3.7.171), that (2.32) is equivalent to the relations

$$v^\dagger \sigma^\alpha v = \sum_\beta M(v)_{\alpha\beta}\sigma^\beta, \tag{8.2.33}$$

which may be viewed as defining $M(v)$. Indeed, from this result deduce, with the aid of (3.7.168), the relation

$$M_{\alpha\beta}(v) = (1/2)\mathrm{tr}(v^\dagger\sigma^\alpha v\sigma^\beta). \tag{8.2.34}$$

Let us find some of the properties of $M(v)$. Verify that

$$M(I) = I, \tag{8.2.35}$$

$$M(-I) = I, \tag{8.2.36}$$

and, more generally,

$$M(-v) = M(v). \tag{8.2.37}$$

Verify also that

$$M_{\beta\alpha}(v) = (1/2)\mathrm{tr}(v^\dagger\sigma^\beta v\sigma^\alpha) = (1/2)\mathrm{tr}(v\sigma^\alpha v^\dagger\sigma^\beta) = M_{\alpha\beta}(v^\dagger) \tag{8.2.38}$$

so that there is the relation

$$M^T(v) = M(v^\dagger). \tag{8.2.39}$$

[Note that $-v \in SU(2)$ and $v^\dagger \in SU(2)$ if $v \in SU(2)$.]

Evidently (2.34) is a rule that sends any matrix $v \in SU(2)$ to a corresponding matrix $M(v)$. Moreover, for any two $SU(2)$ elements v_1 and v_2, this rule has the property

$$M(v_1v_2) = M(v_1)M(v_2), \tag{8.2.40}$$

and therefore is in fact a group homomorphism. To verify this assertion, show that

$$\begin{aligned}M_{\alpha\beta}(v_1v_2) &= (1/2)\mathrm{tr}[(v_1v_2)^\dagger\sigma^\alpha v_1v_2\sigma^\beta)] = (1/2)\mathrm{tr}[(v_2)^\dagger(v_1)^\dagger\sigma^\alpha v_1v_2\sigma^\beta)]\\ &= (1/2)\sum_\gamma M(v_1)_{\alpha\gamma}\mathrm{tr}(v_2^\dagger\sigma^\gamma v_2\sigma^\beta) = \sum_\gamma M(v_1)_{\alpha\gamma}M(v_2)_{\gamma\beta} = [M(v_1)M(v_2)]_{\alpha\beta}.\end{aligned} \tag{8.2.41}$$

Check the chain of deductions

$$I = M(I) = M(v^\dagger v) = M(v^\dagger)M(v) = M^T(v)M(v) \tag{8.2.42}$$

to conclude that M is orthogonal and that

$$M^{-1}(v) = M(v^{-1}). \tag{8.2.43}$$

Argue, based on (2.35) and the topology (connectedness) of $SU(2)$, that, by continuity, M must have determinant $+1$, and therefore $M \in SO(3,\mathbb{R})$. [We will see below that $M(v)$ is a real matrix.] Thus, (2.34) provides a map from $SU(2)$ to $SO(3,\mathbb{R})$ and, in view of (2.37) and (2.40), this map is a two-to-one homomorphism.

Even more explicit results are possible: Suppose that v is parameterized as in (3.7.187). Then (2.31) and (2.32) can be rewritten in the form

$$\begin{aligned}\bar{K}^\alpha(\theta,\boldsymbol{n}) &= v(\theta,\boldsymbol{n})^\dagger K^\alpha v(\theta,\boldsymbol{n})\\ &= \exp(-\theta\boldsymbol{n}\cdot\boldsymbol{K})K^\alpha\exp(\theta\boldsymbol{n}\cdot\boldsymbol{K})\\ &= \sum_\beta M_{\alpha\beta}(\theta,\boldsymbol{n})K^\beta,\end{aligned} \tag{8.2.44}$$

and (2.35) takes the form

$$M(0,\boldsymbol{n}) = I. \tag{8.2.45}$$

Next we will find a differential equation for M. Show that differentiating the first and third terms in (2.44) yields the result

$$\begin{aligned}&\partial_\theta \bar{K}^\alpha(\theta, \boldsymbol{n}) = \partial_\theta[\exp(-\theta \boldsymbol{n}\cdot \boldsymbol{K})K^\alpha \exp(\theta \boldsymbol{n}\cdot \boldsymbol{K})]\\ &= \exp(-\theta \boldsymbol{n}\cdot \boldsymbol{K})\{K^\alpha, \boldsymbol{n}\cdot \boldsymbol{K}\}\exp(\theta \boldsymbol{n}\cdot \boldsymbol{K}).\end{aligned} \tag{8.2.46}$$

But, according to (3.7.183) and (3.7.201), there is the relation

$$\begin{aligned}&\{K^\alpha, \boldsymbol{n}\cdot \boldsymbol{K}\} = \{\boldsymbol{e_\alpha}\cdot \boldsymbol{K}, \boldsymbol{n}\cdot \boldsymbol{K}\} = (\boldsymbol{e}_\alpha \times \boldsymbol{n})\cdot \boldsymbol{K}\\ &= -(\boldsymbol{n}\times \boldsymbol{e}_\alpha)\cdot \boldsymbol{K} = -[(\boldsymbol{n}\cdot \boldsymbol{L})\boldsymbol{e}_\alpha]\cdot \boldsymbol{K}\end{aligned} \tag{8.2.47}$$

so that (2.46) can also be written in the form

$$\begin{aligned}&\partial_\theta \bar{K}^\alpha(\theta, \boldsymbol{n}) = -\exp(-\theta \boldsymbol{n}\cdot \boldsymbol{K})[(\boldsymbol{n}\cdot \boldsymbol{L})\boldsymbol{e}_\alpha]\cdot \boldsymbol{K}\exp(\theta \boldsymbol{n}\cdot \boldsymbol{K})\\ &= -\sum_\beta \exp(-\theta \boldsymbol{n}\cdot \boldsymbol{K})[(\boldsymbol{n}\cdot \boldsymbol{L})\boldsymbol{e}_\alpha]_\beta \boldsymbol{K}^\beta \exp(\theta \boldsymbol{n}\cdot \boldsymbol{K})\\ &= \sum_\beta (\boldsymbol{n}\cdot \boldsymbol{L})_{\alpha\beta}\exp(-\theta \boldsymbol{n}\cdot \boldsymbol{K})\boldsymbol{K}^\beta \exp(\theta \boldsymbol{n}\cdot \boldsymbol{K})\\ &= \sum_{\beta\gamma}(\boldsymbol{n}\cdot \boldsymbol{L})_{\alpha\beta}M_{\beta\gamma}(\theta, \boldsymbol{n})K^\gamma = \sum_\gamma [(\boldsymbol{n}\cdot \boldsymbol{L})M(\theta, \boldsymbol{n})]_{\alpha\gamma}K^\gamma.\end{aligned} \tag{8.2.48}$$

Verify, by working with components, that here we have used the antisymmetry of the L^α to correctly make the calculation

$$\begin{aligned}&-[(\boldsymbol{n}\cdot \boldsymbol{L})\boldsymbol{e}_\alpha]_\beta = -(\boldsymbol{e}_\beta, [\boldsymbol{n}\cdot \boldsymbol{L}]\boldsymbol{e}_\alpha)\\ &= -[\boldsymbol{n}\cdot \boldsymbol{L}]_{\beta\alpha} = [\boldsymbol{n}\cdot \boldsymbol{L}]_{\alpha\beta}.\end{aligned} \tag{8.2.49}$$

On the other hand, differentiating the first and last terms in (2.44) and changing summation indices yields the result

$$\partial_\theta \bar{K}^\alpha(\theta, \boldsymbol{n}) = \sum_\gamma [\partial_\theta M(\theta, \boldsymbol{n})]_{\alpha\gamma}K^\gamma. \tag{8.2.50}$$

By comparing (2.48) and (2.50) conclude that M satisfies the differential equation

$$\partial_\theta M(\theta, \boldsymbol{n}) = (\boldsymbol{n}\cdot \boldsymbol{L})M(\theta, \boldsymbol{n}). \tag{8.2.51}$$

Show that (2.51) with the initial condition (2.45) has the unique solution

$$M(\theta, \boldsymbol{n}) = \exp(\theta \boldsymbol{n}\cdot \boldsymbol{L}) = R(\theta, \boldsymbol{n}). \tag{8.2.52}$$

You have demonstrated that

$$\exp(-\theta \boldsymbol{n}\cdot \boldsymbol{K})K^\alpha \exp(\theta \boldsymbol{n}\cdot \boldsymbol{K}) = \sum_\beta R(\theta, \boldsymbol{n})_{\alpha\beta}K^\beta, \tag{8.2.53}$$

and (2.34) becomes

$$R_{\alpha\beta}(v) = (1/2)\mathrm{tr}(v^\dagger \sigma^\alpha v \sigma^\beta) \ \Leftrightarrow\ R[\exp(\theta \boldsymbol{n}\cdot \boldsymbol{K})] = \exp(\theta \boldsymbol{n}\cdot \boldsymbol{L}). \tag{8.2.54}$$

(Here the symbol $\Leftrightarrow$ is used to indicate logical implication in both directions.) Should a relation of the form (2.54) be surprising? Not from a group-theoretic perspective. The matrices v and $v^\dagger$ on the right side of (2.54) each carry a spin 1/2 representation of $SU(2)$. The matrix R on the left carries a spin 1 representation. We know that two spin 1/2 representations can be combined to produce a spin 1 representation, and evidently the Pauli matrices σ^α and σ^β on the right act as Clebsch-Gordan coefficients to pick out this representation.

To find further consequences of (2.53), multiply both sides by a_α and sum over α to get the result

$$\exp(-\theta \boldsymbol{n}\cdot\boldsymbol{K})(\sum_\alpha a_\alpha K^\alpha)\exp(\theta\boldsymbol{n}\cdot\boldsymbol{K}) = \sum_{\alpha\beta} R(\theta,\boldsymbol{n})_{\alpha\beta} a_\alpha K^\beta$$
$$= \sum_{\alpha\beta} R^T(\theta,\boldsymbol{n})_{\beta\alpha} a_\alpha K^\beta = \sum_{\alpha\beta} R^{-1}(\theta,\boldsymbol{n})_{\beta\alpha} a_\alpha K^\beta. \tag{8.2.55}$$

Verify that (2.55) can be written in the more compact form

$$\exp(-\theta \boldsymbol{n}\cdot\boldsymbol{K})(\boldsymbol{a}\cdot\boldsymbol{K})\exp(\theta\boldsymbol{n}\cdot\boldsymbol{K}) = \sum_{\alpha\beta} R^{-1}(\theta,\boldsymbol{n})_{\beta\alpha} a_\alpha K^\beta$$
$$= \sum_\beta [R^{-1}(\theta,\boldsymbol{n})\boldsymbol{a}]_\beta K^\beta = [R^{-1}(\theta,\boldsymbol{n})\boldsymbol{a}]\cdot\boldsymbol{K}. \tag{8.2.56}$$

Show, by making the replacement $\theta \to -\theta$, that there is also the general result

$$\exp(\theta \boldsymbol{n}\cdot\boldsymbol{K})(\boldsymbol{a}\cdot\boldsymbol{K})\exp(-\theta\boldsymbol{n}\cdot\boldsymbol{K}) = [R(\theta,\boldsymbol{n})\boldsymbol{a}]\cdot\boldsymbol{K}. \tag{8.2.57}$$

Note that, according to (3.7.200), the matrix R is produced/generated by exponentiating elements in the adjoint representation of $SU(2)$. Recall from Section 3.7.7 that the adjoint representation is defined completely in terms of the structure constants. It can be shown that the relations (2.56) and (2.57) hold for *any* set of matrices K^α that satisfy the commutation relations (3.7.173). See Section 8.1. Show, for example, that

$$\exp(-\psi L^j)L^k\exp(\psi L^j) = L^k\cos\psi - \{L^j,L^k\}\sin\psi \;\text{ for }\; j\neq k. \tag{8.2.58}$$

Verify the general result

$$R(\boldsymbol{a}\cdot\boldsymbol{L})R^{-1} = (R\boldsymbol{a})\cdot\boldsymbol{L}. \tag{8.2.59}$$

Suppose that $\boldsymbol{a}$ and $\boldsymbol{b}$ are any three-component vectors. As an application of (2.59), verify that

$$R(\boldsymbol{a}\times\boldsymbol{b}) = R(\boldsymbol{a}\cdot\boldsymbol{L})\boldsymbol{b} = R(\boldsymbol{a}\cdot\boldsymbol{L})R^{-1}R\boldsymbol{b} = [(R\boldsymbol{a})\cdot\boldsymbol{L}]R\boldsymbol{b} = (R\boldsymbol{a})\times(R\boldsymbol{b}), \tag{8.2.60}$$

as expected from the geometric definition of the cross product. Here we have also used (3.7.201). Suppose O is a 3×3 orthogonal matrix. Show that

$$\begin{aligned} O(\boldsymbol{a}\times\boldsymbol{b}) &= O(\boldsymbol{a}\cdot\boldsymbol{L})\boldsymbol{b} = O(\boldsymbol{a}\cdot\boldsymbol{L})O^{-1}O\boldsymbol{b} \\ &= \det(O)[(O\boldsymbol{a})\cdot\boldsymbol{L}]O\boldsymbol{b} = \det(O)[(O\boldsymbol{a})\times(O\boldsymbol{b})], \end{aligned} \tag{8.2.61}$$

on account of which $\boldsymbol{a} \times \boldsymbol{b}$ is called a *psuedo* vector if $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors.

Observe also that, according to (3.7.189) through (3.7.191), for fixed $\boldsymbol{n}$ we must have $\theta \in [0, 4\pi)$ to achieve a closed path in $SU(2)$; and, by (3.7.200), (3.7.203), and (3.7.204), in doing so an associated closed path in $SO(3,\mathbb{R})$ gets covered *twice*. Therefore, as already asserted, the homomorphism between $SU(2)$ and $SO(3,\mathbb{R})$ is two to one. For a further discussion of the topologies of $SU(2)$ and $SO(3,\mathbb{R})$, see Exercise 8.2.11.

Let us use the relation (2.54) to explore once again the relation between $su(2)$ and $so(3,\mathbb{R})$. This can be done by studying (2.54) for elements v near the identity. Suppose v is written in the form

$$v = \exp(\epsilon K) = I + \epsilon K + O(\epsilon^2) \tag{8.2.62}$$

where ϵ is small and $K \in su(2)$ is therefore any 2×2 traceless anti-Hermitian matrix. That is, following the terminology of Exercise 3.7.31, K can be written in the form

$$K = \boldsymbol{a} \cdot \boldsymbol{K} = \sum_{\gamma=1}^{3} a_\gamma K^\gamma = \sum_{\gamma=1}^{3} a_\gamma (-i/2)\sigma^\gamma \tag{8.2.63}$$

where $\boldsymbol{a}$ is a real vector. Then we have

$$v^\dagger = \exp(\epsilon K^\dagger) = I + \epsilon K^\dagger + O(\epsilon^2), \tag{8.2.64}$$

and (2.54) yields

$$\begin{aligned} R_{\alpha\beta}(v) &= (1/2)\mathrm{tr}[(I + \epsilon K^\dagger)\sigma^\alpha (I + \epsilon K)\sigma^\beta] + O(\epsilon^2) \\ &= (1/2)\mathrm{tr}(\sigma^\alpha \sigma^\beta) + (\epsilon/2)\mathrm{tr}(\sigma^\alpha K \sigma^\beta + K^\dagger \sigma^\alpha \sigma^\beta) + O(\epsilon^2) \\ &= \delta_{\alpha\beta} + \epsilon L_{\alpha\beta} + O(\epsilon^2) = [\exp(\epsilon L)]_{\alpha\beta} + O(\epsilon^2) \end{aligned} \tag{8.2.65}$$

where

$$L_{\alpha\beta}(K) = (1/2)\mathrm{tr}(\sigma^\alpha K \sigma^\beta + K^\dagger \sigma^\alpha \sigma^\beta). \tag{8.2.66}$$

Show, using the fact that K is anti-Hermitian and the properties of the trace operation and of the Pauli matrices (see Exercise 5.7.7), that (2.66) can also be written in the form

$$\begin{aligned} L_{\alpha\beta}(\boldsymbol{a} \cdot \boldsymbol{K}) &= L_{\alpha\beta}(K) = (1/2)\mathrm{tr}(\sigma^\alpha K \sigma^\beta + K^\dagger \sigma^\alpha \sigma^\beta) = (1/2)\mathrm{tr}(K \sigma^\beta \sigma^\alpha - K \sigma^\alpha \sigma^\beta) \\ &= (-1/2)(-i/2)\sum_{\gamma=1}^{3} a_\gamma \, \mathrm{tr}(\sigma^\gamma \{\sigma^\alpha, \sigma^\beta\}) = (-4i)(-1/2)(-i/2)\sum_{\gamma=1}^{3} a_\gamma (L^\gamma)_{\alpha\beta} \\ &= (\boldsymbol{a} \cdot \boldsymbol{L})_{\alpha\beta}. \end{aligned} \tag{8.2.67}$$

Thus, the relations (2.66) and (2.67) provide an explicit isomorphism between $su(2)$ and $so(3,\mathbb{R})$. Indeed, we have the relations

$$\{L(K), L(K'\} = L(\{K, K'\}) \text{ and, specifically, } \{L(K^\alpha), L(K^\beta)\} = L(\{K^\alpha, K^\beta\}). \tag{8.2.68}$$

8.2.11. The purpose of this exercise is to study the topology of $SU(2)$ and $SO(3,\mathbb{R})$. Consider all points in three-dimensional space of the form $\theta\boldsymbol{n}$ where $\boldsymbol{n}$ is an arbitrary unit vector and $0 \le \theta \le \theta_{\max}$. They evidently comprise the interior and surface of a ball in

3-dimensional space with radius $\theta_{\max}$. By looking at (3.7.203), show that the topology of $SO(3,\mathbb{R})$ is the same as that of the interior and surface of a ball of radius π with opposite pairs of points on the surface identified. By looking at (3.7.189), show that the topology of $SU(2)$ is the same as that of the interior and surface of a ball of radius 2π with all points on the surface identified.

The topology of $SU(2)$ can also be examined without use of the exponential function. Suppose v is any 2×2 matrix with complex entries written in the form

$$v = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}. \tag{8.2.69}$$

Now require that v be unitary. Show that the conditions

$$v^\dagger v = vv^\dagger = I \tag{8.2.70}$$

yield, among others, the relations

$$\delta = \bar{\alpha} \tag{8.2.71}$$

and

$$\gamma = -\bar{\beta} \tag{8.2.72}$$

so that v takes the form

$$v = \begin{pmatrix} \alpha & \beta \\ -\bar{\beta} & \bar{\alpha} \end{pmatrix}. \tag{8.2.73}$$

Next introduce the general parameterizations

$$\alpha = w_0 + iw_3 \tag{8.2.74}$$

and

$$\beta = w_2 + iw_1 \tag{8.2.75}$$

where all the w_j are real. Show that, in terms of these parameters, v takes the form

$$v = \begin{pmatrix} w_0 + iw_3 & w_2 + iw_1 \\ -w_2 + iw_1 & w_0 - iw_3 \end{pmatrix}. \tag{8.2.76}$$

Verify that v can also be written in the form

$$v = w_0\sigma^0 + i\sum_{j=1}^{3} w_j\sigma^j. \tag{8.2.77}$$

Lastly, show that requiring v to have determinant 1 yields the relation

$$\det(v) = \sum_{j=0}^{3} w_j^2 = 1, \tag{8.2.78}$$

and that this relation also guarantees that v as given by (2.76) or (2.77) is unitary.

Equation (2.78) is that for S^3, the 3-dimensional surface of a sphere in 4-dimensional space. Thus, $SU(2)$ has the topology of S^3.[2] This manifold is known to be *simply connected*, and therefore $SU(2)$ is simply connected. Simply connected means that any closed curve can be shrunk to a point. By contrast, $SO(3,\mathbb{R})$ is not simply connected. Consider the ball with radius π that parameterizes $SO(3,\mathbb{R})$. Let P be any path in the ball that stretches between antipodal points on the surface of the ball. Since antipodal points on the surface are to be identified, this path is a closed curve. Show that it cannot be shrunk to a point while remaining closed. A more detailed study shows that $SO(3,\mathbb{R})$ is doubly connected. Given a multiply connected manifold, there is a standard procedure in topology for constructing an associated singly-connected manifold. This singly-connected manifold is called the *covering* manifold of the original manifold. In the same spirit, $SU(2)$ is said to be the covering group of $SO(3,\mathbb{R})$.

It can be shown that all the $SO(n,\mathbb{R})$ for $n \geq 3$ are doubly connected. [We have already learned that $SO(2,\mathbb{R})$ is infinitely connected. See Section 5.9.1.] Consequently each has a two-fold covering group. These groups are called $Spin(n,\mathbb{R})$. For small n there are the redundancies $Spin(3,\mathbb{R}) = SU(2)$, $Spin(4,\mathbb{R}) = SU(2) \times SU(2)$, $Spin(5,\mathbb{R}) = USp(4)$, and $Spin(6,\mathbb{R}) = SU(4)$.

Let return to the case of $SO(3,\mathbb{R})$. Comparison of (5.10.22) and (2.78), and reference to Exercises 5.10.13 and 5.10.14, show that v is a unit quaternion matrix. For this reason, the quantities $w_0 \cdots w_3$ are sometimes called quaternion parameters. The quaternion parameterization of $SU(2)$ can be extended to a quaternion parameterization of $SO(3,\mathbb{R})$ with the aid of (2.54) and (2.77).[3] Show that doing so gives the result

$$R_{\alpha\beta}(w) = \delta_{\alpha\beta}(w_0^2 - \sum_{\gamma=1}^{3} w_\gamma^2) + 2w_\alpha w_\beta + 2w_0 \sum_{\gamma=1}^{3} \epsilon_{\alpha\beta\gamma} w_\gamma. \tag{8.2.79}$$

Students of dynamics or quantum mechanics may be familiar with the use of Euler angles to parameterize elements in both $SU(2)$ and $SO(3,\mathbb{R})$. For example, we may write

$$R(\phi,\theta,\psi) = \exp(\phi L^3)\exp(\theta L^2)\exp(\psi L^3). \tag{8.2.80}$$

See (3.7.195) and (3.7.208). However, when studying rigid-body dynamics, this is not always a good idea because the Euler angles ϕ and ψ are not uniquely defined when $\theta = 0$ and $\theta = \pi$. [Only the quantity $(\phi + \psi)$ plays a role when $\theta = 0$, and only the quantity $(\phi - \psi)$ plays a role when $\theta = \pi$.] That is, the quantities ϕ, θ, ψ do not provide good coordinate patches in the neighborhoods $\theta \simeq 0$ and $\theta \simeq \pi$. Correspondingly, the equations of motion for rigid-body motion in terms of Euler angles have singularities at these values of θ, and are therefore not well suited for numerical integration.[4] By contrast, the equations of motion are regular everywhere when quaternion parameters are employed. The only penalty to be paid for this advantage is that equations of motion must be integrated for four parameters

[2] Put another way, the manifold S^3 can be given a group structure, namely that of $SU(2)$. It can be shown that S^1 and S^3 are the only spheres that can be given a group structure.

[3] The so called *Cayley-Klein* parameters for specifying rotations are closely related to quaternion parameters. Also, sometimes quaternion parameters are called *Euler-Rodrigues* parameters.

[4] The associated problems encountered in numerical integration are sometimes referred to as *gimbal lock*. Google the words *Euler angle evil*.

instead of three. Moreover, the equations of motion preserve the relation (2.78), and the extent to which numerical integration preserves this relation can be used as a check on the accuracy of the procedure. For further discussion, see Section 11.1.

8.2.12. In Exercise 3.7.35 you verified that the Lie algebras $su(4)$ and $so(6,\mathbb{R})$ have the same dimension, namely dimension 15. The purpose of this exercise is to verify that $su(4)$ and $so(6,\mathbb{R})$ are in fact the same (equivalent) over the real field. Moreover, we will learn that there is a corresponding two-to-one homomorphism between the groups $SU(4)$ and $SO(6,\mathbb{R})$ just as there is a two-to-one homomorphism between the groups $SU(2)$ and $SO(3,\mathbb{R})$. See Exercises 3.7.31, 8.2.10, and 8.2.11.

We begin by exploiting a mathematical fact familiar from the relativistic treatment of electromagnetism. Let A be a general antisymmetric 4×4 matrix. It can be written in the form

$$A = \begin{pmatrix} 0 & -B_z & B_y & E_x/c \\ B_z & 0 & -B_x & E_y/c \\ -B_y & B_x & 0 & E_z/c \\ -E_x/c & -E_y/c & -E_z/c & 0 \end{pmatrix} \tag{8.2.81}$$

where the quantities E_α/c and B_α are arbitrary. See (1.6.56). Then we know from its use in relativistic electromagnetic theory that there is the *mathematical* identity

$$\det(A) = [(1/c)\boldsymbol{E}\cdot\boldsymbol{B}]^2. \tag{8.2.82}$$

See Exercise 1.6.17. At this point introduce variables $z_1, z_2, \cdots, z_6$ by the rules

$$E_z/c = iz_1 + z_2, \tag{8.2.83}$$

$$B_z = -iz_1 + z_2, \tag{8.2.84}$$

$$E_x/c = iz_3 + z_4, \tag{8.2.85}$$

$$B_x = -iz_3 + z_4, \tag{8.2.86}$$

$$E_y/c = iz_5 + z_6, \tag{8.2.87}$$

$$B_y = -iz_5 + z_6. \tag{8.2.88}$$

(While workable, this ordering may seem a little strange. It will be of use in Exercise 27.5.4.) Verify that in terms of these variables there is the relation

$$(1/c)\boldsymbol{E}\cdot\boldsymbol{B} = \sum_{\alpha=1}^{6} z_\alpha^2. \tag{8.2.89}$$

Let us write

$$A = A(z) = A(z_1 \cdots z_6) = \sum_{\alpha=1}^{6} z_\alpha A^\alpha \tag{8.2.90}$$

where the A^α are matrices to be determined. Show, using (2.82) and (2.89), that

$$\det(A) = \Big[\sum_{\alpha=1}^{6} z_\alpha^2\Big]^2. \tag{8.2.91}$$

Verify that

$$A(z) = \begin{pmatrix} 0 & iz_1 - z_2 & -iz_5 + z_6 & iz_3 + z_4 \\ -iz_1 + z_2 & 0 & iz_3 - z_4 & iz_5 + z_6 \\ iz_5 - z_6 & -iz_3 + z_4 & 0 & iz_1 + z_2 \\ -iz_3 - z_4 & -iz_5 - z_6 & -iz_1 - z_2 & 0 \end{pmatrix} \tag{8.2.92}$$

so that the A^α are given by the relations

$$A^1 = \begin{pmatrix} 0 & i & 0 & 0 \\ -i & 0 & -0 & 0 \\ 0 & 0 & 0 & i \\ 0 & 0 & -i & 0 \end{pmatrix} = -\begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}, \tag{8.2.93}$$

$$A^2 = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix} = i\begin{pmatrix} -\sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}, \tag{8.2.94}$$

$$A^3 = \begin{pmatrix} 0 & 0 & 0 & i \\ 0 & 0 & i & 0 \\ 0 & -i & 0 & 0 \\ -i & 0 & 0 & 0 \end{pmatrix} = i\begin{pmatrix} 0 & \sigma^1 \\ -\sigma^1 & 0 \end{pmatrix}, \tag{8.2.95}$$

$$A^4 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix} = i\begin{pmatrix} 0 & \sigma^2 \\ \sigma^2 & 0 \end{pmatrix}, \tag{8.2.96}$$

$$A^5 = \begin{pmatrix} 0 & 0 & -i & 0 \\ 0 & 0 & 0 & i \\ i & 0 & 0 & 0 \\ 0 & -i & 0 & 0 \end{pmatrix} = i\begin{pmatrix} 0 & -\sigma^3 \\ \sigma^3 & 0 \end{pmatrix}, \tag{8.2.97}$$

$$A^6 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \sigma^0 \\ -\sigma^0 & 0 \end{pmatrix}. \tag{8.2.98}$$

Here use has been made of the Pauli matrices given by (5.7.3). Evidently, the A^α span the space of 4×4 antisymmetric matrices when working over the complex field.

Verify that the A^α have the properties

$$(A^\alpha)^T = -A^\alpha, \tag{8.2.99}$$

$$(A^\alpha)^\dagger = -(-1)^\alpha A^\alpha, \tag{8.2.100}$$

and that they obey the multiplication rules

$$A^\alpha (A^\alpha)^\dagger = (A^\alpha)^\dagger A^\alpha = I, \tag{8.2.101}$$

$$A^1A^2 = A^2A^1 = i\begin{pmatrix} \sigma^0 & 0 \\ 0 & -\sigma^0 \end{pmatrix}, \tag{8.2.102}$$

$$A^1A^3 = -A^3A^1 = -iA^5, \tag{8.2.103}$$

$$A^1A^4 = A^4A^1 = -i\begin{pmatrix} 0 & \sigma^0 \\ \sigma^0 & 0 \end{pmatrix}, \tag{8.2.104}$$

$$A^1A^5 = -A^5A^1 = iA^3, \tag{8.2.105}$$

$$A^1A^6 = A^6A^1 = \begin{pmatrix} 0 & -\sigma^2 \\ \sigma^2 & 0 \end{pmatrix}, \tag{8.2.106}$$

$$A^2A^3 = A^3A^2 = -i\begin{pmatrix} 0 & \sigma^3 \\ \sigma^3 & 0 \end{pmatrix}, \tag{8.2.107}$$

$$A^2A^4 = -A^4A^2 = A^6, \tag{8.2.108}$$

$$A^2A^5 = A^5A^2 = -i\begin{pmatrix} 0 & \sigma^1 \\ \sigma^1 & 0 \end{pmatrix}, \tag{8.2.109}$$

$$A^2A^6 = -A^6A^2 = -A^4, \tag{8.2.110}$$

$$A^3A^4 = A^4A^3 = i\begin{pmatrix} -\sigma^3 & 0 \\ 0 & \sigma^3 \end{pmatrix}, \tag{8.2.111}$$

$$A^3A^5 = -A^5A^3 = -iA^1, \tag{8.2.112}$$

$$A^3A^6 = A^6A^3 = -i\begin{pmatrix} \sigma^1 & 0 \\ 0 & \sigma^1 \end{pmatrix}, \tag{8.2.113}$$

$$A^4A^5 = A^5A^4 = i\begin{pmatrix} -\sigma^1 & 0 \\ 0 & \sigma^1 \end{pmatrix}, \tag{8.2.114}$$

$$A^4A^6 = -A^6A^4 = A^2, \tag{8.2.115}$$

$$A^5A^6 = A^6A^5 = i\begin{pmatrix} \sigma^3 & 0 \\ 0 & \sigma^3 \end{pmatrix}. \tag{8.2.116}$$

Verify, by looking at (2.102) through (2.116), that there is the rule

$$A^\alpha A^\beta = (-1)(-1)^{\alpha+\beta}A^\beta A^\alpha \text{ for } \alpha \neq \beta. \tag{8.2.117}$$

Show that combining (2.100) and (2.117) gives the relations

$$A^\alpha(A^\beta)^\dagger = -A^\beta(A^\alpha)^\dagger \text{ for } \alpha \neq \beta, \tag{8.2.118}$$

$$(A^\alpha)^\dagger A^\beta = -(A^\beta)^\dagger A^\alpha \text{ for } \alpha \neq \beta. \tag{8.2.119}$$

Show that combining (2.101), (2.118), and (2.119) gives the relations

$$A^\alpha(A^\beta)^\dagger + A^\beta(A^\alpha)^\dagger = (A^\alpha)^\dagger A^\beta + (A^\beta)^\dagger A^\alpha = 2\delta_{\alpha\beta}I. \tag{8.2.120}$$

Finally, verify that the right sides of (2.102) through (2.116) are traceless. Combine this fact with (2.101) to derive the result

$$\mathrm{tr}[A^\alpha(A^\beta)^\dagger] = 4\delta_{\alpha\beta}. \tag{8.2.121}$$

For the moment, let v be any 4×4 matrix. Use it to define quantities $\bar{A}^{\alpha}(v)$ by the rule

$$\bar{A}^{\alpha}(v) = v^T A^{\alpha} v. \tag{8.2.122}$$

Note the similarity of (2.122) to (2.31) except that † has been replaced by T. Verify that the $\bar{A}^{\alpha}(v)$ are antisymmetric for any choice of v. It follows, since the A^{α} form a basis for the set of 4×4 antisymmetric matrices, that there must be a relation of the form

$$\bar{A}^{\alpha}(v) = v^T A^{\alpha} v = \sum_{\beta} R_{\alpha\beta}(v) A^{\beta} \tag{8.2.123}$$

where $R(v)$ is a 6×6 matrix to be determined. Verify, in view of (2.121), that there is the explicit formula

$$R_{\alpha\beta}(v) = (1/4)\mathrm{tr}[v^T A^{\alpha} v (A^{\beta})^{\dagger}]. \tag{8.2.124}$$

Let us find some of the properties of $R(v)$. Verify that

$$R(I) = I, \tag{8.2.125}$$

$$R(-I) = I, \tag{8.2.126}$$

and, more generally,

$$R(-v) = R(v). \tag{8.2.127}$$

The rule (2.124) is also a homomorphism,

$$R(v_1 v_2) = R(v_1) R(v_2). \tag{8.2.128}$$

Check this assertion by verifying the computation

$$\begin{aligned} R_{\alpha\beta}(v_1 v_2) &= (1/4)\mathrm{tr}[(v_1 v_2)^T A^{\alpha} v_1 v_2 (A^{\beta})^{\dagger}] = (1/4)\mathrm{tr}[v_2^T v_1^T A^{\alpha} v_1 v_2 (A^{\beta})^{\dagger}] \\ &= (1/4)\sum_{\gamma} R(v_1)_{\alpha\gamma} \mathrm{tr}[v_2^T A^{\gamma} v_2 (A^{\beta})^{\dagger}] = \sum_{\gamma} R(v_1)_{\alpha\gamma} R(v_2)_{\gamma\beta} = [R(v_1)R(v_2)]_{\alpha\beta}. \end{aligned} \tag{8.2.129}$$

From (2.123) deduce the relation

$$v^T \Big[\sum_{\alpha} z_{\alpha} A^{\alpha}\Big] v = \sum_{\alpha} z_{\alpha} v^T A^{\alpha} v = \sum_{\beta} \Big[\sum_{\alpha} z_{\alpha} R_{\alpha\beta}(v)\Big] A^{\beta}. \tag{8.2.130}$$

Define variables $\hat{z}_{\beta}$ by writing

$$\hat{z}_{\beta} = \sum_{\alpha} z_{\alpha} R_{\alpha\beta}(v). \tag{8.2.131}$$

Show that (2.130) can be written more compactly in the form

$$v^T A(z) v = A(\hat{z}). \tag{8.2.132}$$

Take the determinant of both sides of (2.132). Show that doing so yields, in view of (2.91), the result

$$[\det(v)]^2 \Big[\sum_{\alpha} z_{\alpha}^2\Big]^2 = \Big[\sum_{\beta} \hat{z}_{\beta}^2\Big]^2. \tag{8.2.133}$$

Take the square roots of both sides of (2.133) to find the result

$$[\det(v)]\Big[\sum_{\alpha} z_{\alpha}^2\Big] = \Big[\sum_{\beta} \hat{z}_{\beta}^2\Big]. \tag{8.2.134}$$

In taking this square root there is, of course, a sign ambiguity. This ambiguity is overcome by employing (2.125) and imposing continuity. For a Pfaffian-based approach that avoids this ambiguity, see Exercise 8.2.13.

Next manipulate the ingredients on the right side of (2.134) to show that

$$\begin{aligned}\sum_{\beta} \hat{z}_{\beta}^2 &= \sum_{\beta}\Big[\sum_{\alpha} z_{\alpha}R_{\alpha\beta}\Big]\Big[\sum_{\gamma} z_{\gamma}R_{\gamma\beta}\Big] \\ &= \sum_{\alpha\beta\gamma} z_{\alpha}R_{\alpha\beta}z_{\gamma}R_{\gamma\beta} = \sum_{\alpha\beta\gamma} z_{\alpha}R_{\alpha\beta}(R^T)_{\beta\gamma}z_{\gamma} = \sum_{\alpha\gamma} z_{\alpha}(RR^T)_{\alpha\gamma}z_{\gamma}.\end{aligned} \tag{8.2.135}$$

It follows that there is the relation

$$\det(v)\sum_{\alpha} z_{\alpha}^2 = \sum_{\alpha\gamma} z_{\alpha}(RR^T)_{\alpha\gamma}z_{\gamma}. \tag{8.2.136}$$

Prove from (2.136) that

$$R(v)R^T(v) = \det(v)I. \tag{8.2.137}$$

Finally, assume that v has unit determinant. Then we find that

$$R(v)R^T(v) = I, \tag{8.2.138}$$

the matrix R is *orthogonal*. At this stage we have found that (2.124) provides a homomorphism of $SL(4,C)$ into $SO(6,C)$. Verify, as a sanity check, that both $SL(4,C)$ and $SO(6,C)$ have dimension 30.

Next assume that v is also unitary so that $v \in SU(4)$. Then deduce the chain of relations

$$I = R(I) = R(v^{\dagger}v) = R(v^{\dagger})R(v) \tag{8.2.139}$$

to conclude that

$$R^T(v) = R(v^{\dagger}). \tag{8.2.140}$$

Also, we claim that $R(v)$ is *real* if $v \in SU(4)$. Verify that taking the complex conjugate of both sides of (2.124) and employing the invariance of the trace under transposing and cyclic permutation gives (with a * denoting complex conjugation) the result

$$\begin{aligned}R_{\alpha\beta}^*(v) &= (1/4)\mathrm{tr}[v^{\dagger}(A^{\alpha})^* v^*(A^{\beta})^T] = (1/4)\mathrm{tr}[A^{\beta}v^{\dagger}(A^{\alpha})^{\dagger}v^*] \\ &= (1/4)\mathrm{tr}[v^*A^{\beta}v^{\dagger}(A^{\alpha})^{\dagger}] = (1/4)\mathrm{tr}[(v^{\dagger})^T A^{\beta}v^{\dagger}(A^{\alpha})^{\dagger}] \\ &= R_{\beta\alpha}(v^{\dagger}) = [R(v^{\dagger})]_{\beta\alpha} = [R^T(v)]_{\beta\alpha} = R_{\alpha\beta}(v).\end{aligned} \tag{8.2.141}$$

Thus the mapping (2.124) has the property that $R(v) \in SO(6,\mathbb{R})$ if $v \in SU(4)$. In view of (2.127) and (2.128), this mapping is a two-to-one homomorphism.

Finally, we remark that $SU(4)$ is known to be simply connected. It follows that $SO(6,\mathbb{R})$ cannot be simply connected. Indeed, $SO(6,\mathbb{R})$ is known to be doubly connected, and its covering group is $SU(4)$.

At this point one might wonder again about employing the T operation in (2.123) and (2.124) rather than † operation as was done in (2.34). The reason for this choice lies in the Clebsch-Gordan series for $SU(4)$. The lowest dimensional representations for $SU(4)$, those of dimension 4, are *not* self conjugate. Rather, there are two distinct representations which we may call 4 and $\bar{4}$. The Clebsch-Gordan series for the direct products of these two representations, and it is something like a direct product that is going on in (2.124), are

$$4 \times 4 = 6 + 10, \tag{8.2.142}$$

$$4 \times \bar{4} = 1 + 15. \tag{8.2.143}$$

Thus, only by avoiding complex conjugation do we have any hope of obtaining something of dimension 6, the dimension that is required for the lowest dimensional representation of $SO(6,\mathbb{R})$.

We still have to address the relation between the two Lie algebras $su(4)$ and $so(6,\mathbb{R})$. This can be done by studying (2.124) for elements v near the identity. Suppose v is written in the form

$$v = \exp(\epsilon K) = I + \epsilon K + O(\epsilon^2) \tag{8.2.144}$$

where ϵ is small and $K \in su(4)$ is therefore any 4×4 traceless anti-Hermitian matrix. Then we have

$$v^T = \exp(\epsilon K^T) = I + \epsilon K^T + O(\epsilon^2), \tag{8.2.145}$$

and (2.124) yields

$$\begin{aligned} R_{\alpha\beta}(v) &= (1/4)\mathrm{tr}[(I+\epsilon K^T)A^\alpha(I+\epsilon K)(A^\beta)^\dagger] + O(\epsilon^2) \\ &= (1/4)\mathrm{tr}[A^\alpha(A^\beta)^\dagger] + (\epsilon/4)\mathrm{tr}[A^\alpha K(A^\beta)^\dagger + K^T A^\alpha(A^\beta)^\dagger] + O(\epsilon^2) \\ &= \delta_{\alpha\beta} + \epsilon L_{\alpha\beta} + O(\epsilon^2) = [\exp(\epsilon L)]_{\alpha\beta} + O(\epsilon^2) \\ &\Leftrightarrow \quad R[\exp(\epsilon K)] = \exp(\epsilon L) + O(\epsilon^2) \end{aligned} \tag{8.2.146}$$

where

$$L_{\alpha\beta}(K) = (1/4)\mathrm{tr}[A^\alpha K(A^\beta)^\dagger + K^T A^\alpha (A^\beta)^\dagger] = (1/4)\mathrm{tr}[K(A^\beta)^\dagger A^\alpha + K^T A^\alpha (A^\beta)^\dagger]. \tag{8.2.147}$$

Here we have made use of the trace property (3.6.130).

We note that, from the homomorphism property (2.128) and the infinitesimal relation (2.146), it follows that there is also the global result

$$R[\exp(K)] = \exp(L). \tag{8.2.148}$$

To verify this claim, show that

$$\begin{aligned} R[\exp(K)] &= R\{[\exp(K/\ell)]^\ell\} = \{R[\exp(K/\ell)]\}^\ell \\ &= \{\exp(L/\ell) + O[(1/\ell)^2]\}^\ell \\ &= \{\exp(L/\ell)\}^\ell + \ell O[(1/\ell)^2] \\ &= \exp(L) + \ell O[(1/\ell)^2]. \end{aligned} \tag{8.2.149}$$

Now let $\ell \to \infty$ in (2.149) to obtain the global result (2.148).

Let us examine the properties of L. We already know that R is orthogonal and real, and therefore L must be antisymmetric and real. It is valuable to check that these results can also be verified directly from (2.147). Show from (2.101) and (2.147) that

$$L_{\alpha\alpha}(K) = (1/4)\mathrm{tr}[KI + K^T I] = (1/2)\,\mathrm{tr}(K) = 0 \tag{8.2.150}$$

because K must be traceless to be in $su(4)$. Next show using (2.18) and (2.19) that, for $\alpha \neq \beta$,

$$\begin{aligned} L_{\alpha\beta}(K) &= (1/4)\mathrm{tr}[K(A^\beta)^\dagger A^\alpha + K^T A^\alpha (A^\beta)^\dagger] \\ &= -(1/4)\mathrm{tr}[K(A^\alpha)^\dagger A^\beta + K^T A^\beta (A^\alpha)^\dagger] \\ &= -L_{\beta\alpha}(K). \end{aligned} \tag{8.2.151}$$

Taken together, (2.150) and (2.151) show that L is antisymmetric. Next work on showing that L is real. Let * denote the operation of complex conjugation. To show that L is real, verify the chain of deductions

$$\begin{aligned} [L_{\alpha\beta}(K)]^* &= (1/4)\{\mathrm{tr}[K(A^\beta)^\dagger A^\alpha + K^T A^\alpha (A^\beta)^\dagger]\}^* \\ &= (1/4)\,\mathrm{tr}[K^*(A^\beta)^T (A^\alpha)^* + K^\dagger (A^\alpha)^* (A^\beta)^T] \\ &= (1/4)\,\mathrm{tr}\{(K^\dagger)^T (A^\beta)^T (A^\alpha)^* + K^\dagger [(A^\alpha)^T]^\dagger (A^\beta)^T\} \\ &= -(1/4)\,\mathrm{tr}\{K^T A^\beta (A^\alpha)^\dagger + K(A^\alpha)^\dagger A^\beta\} \\ &= -L_{\beta\alpha}(K) = L_{\alpha\beta}(K). \end{aligned} \tag{8.2.152}$$

Here we have used (2.99) and the fact that K must be anti-Hermitian to be in $su(4)$,

$$K^\dagger = -K, \tag{8.2.153}$$

and the antisymmetry conditions (2.150) and (2.151).

It remains to be verified that $L(K)$ is a homomorphism (and potentially an isomorphism). Let $(K^1, K^2, \cdots, K^{15})$ be a set of basis elements for $su(4)$. Form *group commutator* elements v by the rule

$$v = \exp(\epsilon K^\alpha)\exp(\epsilon K^\beta)\exp(-\epsilon K^\alpha)\exp(-\epsilon K^\beta). \tag{8.2.154}$$

Recall Exercise 3.7.41. Show, using the BCH series (3.7.41), that

$$v = \exp(\epsilon^2\{K^\alpha, K^\beta\}) + O(\epsilon^3), \tag{8.2.155}$$

from which it follows, using (2.148), that

$$R(v) = \exp[\epsilon^2 L(\{K^\alpha, K^\beta\})] + O(\epsilon^3). \tag{8.2.156}$$

Show from (2.128), (2.148), and BCH that

$$\begin{aligned} R(v) &= R[\exp(\epsilon K^\alpha)]R[\exp(\epsilon K^\beta)]R[\exp(-\epsilon K^\alpha)]R[\exp(-\epsilon K^\beta)] \\ &= \exp[\epsilon L(K^\alpha)]\exp[\epsilon L(K^\beta)]\exp[-\epsilon L(K^\alpha)]\exp[-\epsilon L(K^\beta)] \\ &= \exp[\epsilon^2\{L(K^\alpha), L(K^\beta)\}] + O(\epsilon^3). \end{aligned} \tag{8.2.157}$$

By equating powers of ϵ in (2.156) and (2.157), show that

$$\{L(K^{\alpha}), L(K^{\beta})\} = L(\{K^{\alpha}, K^{\beta}\}). \tag{8.2.158}$$

Suppose the quantities $c^{\gamma}_{\alpha\beta}$ are structure constants for $su(4)$ so that

$$\{K^{\alpha}, K^{\beta}\} = \sum_{\gamma} c^{\gamma}_{\alpha\beta} K^{\gamma}. \tag{8.2.159}$$

Show from (2.158) and (2.159) that there is the relation

$$\{L(K^{\alpha}), L(K^{\beta})\} = \sum_{\gamma} c^{\gamma}_{\alpha\beta} L(K^{\gamma}), \tag{8.2.160}$$

thereby verifying that, for suitable basis choices, $su(4)$ and $so(6,\mathbb{R})$ have the same structure constants.

We still want to know what particular $K \in su(4)$ produces what $L \in so(6,\mathbb{R})$, and we want to verify that every $L \in so(6,\mathbb{R})$ arises from some $K \in su(4)$ so that (2.150) is, in fact, an isomorphism. We begin this task by looking at specific cases. Listed below are three typical elements in $su(4)$:

$$K^1 = \begin{pmatrix} 0 & i & 0 & 0 \\ i & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = i \begin{pmatrix} \sigma^1 & 0 \\ 0 & 0 \end{pmatrix}, \tag{8.2.161}$$

$$K^2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = i \begin{pmatrix} \sigma^2 & 0 \\ 0 & 0 \end{pmatrix}, \tag{8.2.162}$$

$$K^3 = \begin{pmatrix} i & 0 & 0 & 0 \\ 0 & -i & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = i \begin{pmatrix} \sigma^3 & 0 \\ 0 & 0 \end{pmatrix}. \tag{8.2.163}$$

The matrix K^1 is symmetric, pure imaginary, and has zeroes on the diagonal, which makes it anti-Hermitian and traceless. There are 6 linearly independent matrices of this kind in $su(4)$. The matrix K^2 is antisymmetric and real, which makes it anti-Hermitian and traceless. There are also 6 linearly independent matrices of this kind in $su(4)$. The element K^3 is diagonal and pure imaginary, which makes it anti-Hermitian, and it is traceless. There are 3 linearly independent matrices of this kind in $su(4)$ for a total count of $6 + 6 + 3 = 15$, the dimension of $su(4)$. Show that the associated L matrices are given by the relations

$$L^1 = L(K^1) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}, \tag{8.2.164}$$

$$L^2 = L(K^2) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \end{pmatrix}, \tag{8.2.165}$$

$$L^3 = L(K^3) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix}. \tag{8.2.166}$$

We see that the associated L matrices are real and antisymmetric, as expected. In particular, L^1 is a linear combination of generators for rotations in the 3,6 and 4,5 planes; L^2 is a linear combination of generators for rotations in the 3,5 and 4,6 planes; and L^3 is a linear combination of generators for rotations in the 3,4 and 5,6 planes. Verify that the two generators in each linear combination commute. Verify that

$$\{K^1, K^2\} = -2K^3, \text{ etc.} \tag{8.2.167}$$

so that the elements K^1 through K^3 form some kind of $su(2)$ [or $so(3, R)$] within $su(4)$. Verify, in accord with (2.158), the relations

$$\{L(K^1), L(K^2)\} = -2L(K^3) = L(\{K^1, K^2\}), \text{ etc.} \tag{8.2.168}$$

Suppose the remaining elements of $su(4)$ are also considered so that we are working with a complete set of basis elements $(K^1, K^2, \cdots, K^{15})$ for $su(4)$. Then presumably their associated matrices $L^\alpha = L(K^\alpha)$ form a basis for $so(6, \mathbb{R})$. In fact we know from general principles that this must be the case. These principles are described in Section 8.9, and applied to the problem at hand in Exercise 8.9.19.

8.2.13. Review Section 3.13.3 and Exercise 8.2.12. The purpose of this exercise is to derive various results of Execise 8.2.12 with the aid of Pfaffians. Let A be a general 4×4 antisymmetric matrix written in the form

$$A = \begin{pmatrix} 0 & a & b & c \\ -a & 0 & d & e \\ -b & -d & 0 & f \\ -c & -e & -f & 0 \end{pmatrix}. \tag{8.2.169}$$

Then the Pfaffian of A is given by the relation

$$\mathrm{Pf}(A) = af - be + dc. \tag{8.2.170}$$

Suppose A is given in the form (2.81). Show, for this parameterization, that

$$\mathrm{Pf}(A) = -(1/c)\boldsymbol{E} \cdot \boldsymbol{B}. \tag{8.2.171}$$

Use the Pfaffian property (3.13.65) to derive (2.82). Use the Pfaffian property (3.13.66) to derive the relation

$$[\det(v)]\Big[\sum_{\alpha} z_{\alpha}^{2}\Big] = \Big[\sum_{\beta} \hat{z}_{\beta}^{2}\Big], \tag{8.2.172}$$

which is a specific square root of the relation (2.133), and yields (2.134) directly without any ambiguity in sign.

8.2.14. Review Exercise 2.10 that studied the relation between $SU(2)$ and $SO(3,\mathbb{R})$. In particular, it derived the formula (2.54) that maps $SU(2)$ to $SO(3,\mathbb{R})$ thereby demonstrating that $SU(2)$ is the covering group for $SO(3,\mathbb{R})$. The purpose of this exercise is to find analogous results for the relation between $SL(2,\mathbb{C})$ and the Lorentz group.

Begin by setting up some notation and definitions for later use. Let σ^{α} for $\alpha = 1,2,3$ be the usual Pauli matrices and let σ^4 be the 2×2 identity matrix,

$$\sigma^4 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{8.2.173}$$

Verify, in accord with Exercise 5.7.7, that there are the relations

$$\operatorname{tr} \sigma^{\alpha}\sigma^{\beta} = 2\delta^{\alpha\beta} \text{ for } \alpha,\beta = 1,2,3,4. \tag{8.2.174}$$

Let x^{μ} and y^{μ} be any two four-vectors. Make the definition

$$x \star y = \sum_{\mu=1}^{4} x^{\mu} y^{\mu}. \tag{8.2.175}$$

Note that (2.175) is just the ordinary Euclidean scalar product (x, y). By extension of notation, make the definition

$$x \star \sigma = \sum_{\mu=1}^{4} x^{\mu}\sigma^{\mu}. \tag{8.2.176}$$

Verify that $x \star \sigma$ is Hermitian if x is real, and anti-Hermitian if x is pure imaginary. Verify that

$$\operatorname{tr}(x \star \sigma) = 2x^4. \tag{8.2.177}$$

Verify that

$$\operatorname{tr}[(x \star \sigma)(y \star \sigma)] = 2(x \star y) \tag{8.2.178}$$

and, as a special case,

$$\operatorname{tr}[(x \star \sigma)^2] = 2(x \star x). \tag{8.2.179}$$

Show that $x \star \sigma$ has the explicit matrix form

$$x \star \sigma = \begin{pmatrix} x^4 + x^3 & x^1 - ix^2 \\ x^1 + ix^2 & x^4 - x^3 \end{pmatrix}. \tag{8.2.180}$$

Given a 2×2 matrix M, is there a four-vector x such that

$$M = x \star \sigma? \tag{8.2.181}$$

Any 2 × 2 matrix M can be written in the form

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{8.2.182}$$

where the quantities a through d are arbitrary. Verify, upon examination of (2.180) and (2.182), that for (2.181) to hold it must be possible to satisfy the relations

$$x^4 + x^3 = a, \tag{8.2.183}$$

$$x^1 - ix^2 = b, \tag{8.2.184}$$

$$x^1 + ix^2 = c, \tag{8.2.185}$$

$$x^4 - x^3 = d. \tag{8.2.186}$$

Show that the equation set (2.183) through (2.186) has the unique solution

$$x^1 = (1/2)(b + c), \tag{8.2.187}$$

$$x^2 = (i/2)(b - c), \tag{8.2.188}$$

$$x^3 = (1/2)(a - d), \tag{8.2.189}$$

$$x^4 = (1/2)(a + d). \tag{8.2.190}$$

Consequently, for any M, there exists a unique x such that (2.181) holds. That is, the σ^α form a basis for the set of all 2 × 2 matrices. Show that x is given in terms of M by the relations

$$x^\mu = (1/2)\,\mathrm{tr}(\sigma^\mu M). \tag{8.2.191}$$

Show that x is real if M is Hermitian, and is pure imaginary if M is anti-Hermitian.

Finally, make the definition

$$x \cdot y = x^4 y^4 - \sum_{\mu=1}^{3} x^\mu y^\mu. \tag{8.2.192}$$

Note that (2.192) is the Lorentz inner/scalar product.

We are now ready to make yet another remarkable statement about the Pauli matrices: Verify from (2.180) by explicit calculation that there is the relation

$$\det(x \star \sigma) = x \cdot x. \tag{8.2.193}$$

How could we have guessed that at least something like (2.193) should hold? Review Exercise 3.7.17. According to (3.7.150), for any 2 × 2 matrix A, there is the relation

$$\det(A) = \{[\mathrm{tr}(A)]^2 - \mathrm{tr}(A^2)\}/2. \tag{8.2.194}$$

Make the substitution

$$A = x \star \sigma. \tag{8.2.195}$$

Verify from (2.177) and (2.179) that

$$[\mathrm{tr}(A)]^2 = 4(x^4)^2 \tag{8.2.196}$$

and

$$\mathrm{tr}(A^2) = 2(x \star x). \tag{8.2.197}$$

Show, therefore, that use of (2.194) yields the result

$$\det(x \star \sigma) = [4(x^4)^2 - 2(x \star x)]/2 = [2(x^4)^2 - 2\sum_{\mu=1}^{3}(x^u)^2]/2 = x \cdot x, \tag{8.2.198}$$

in agreement with (2.193).

With the definitions and results just developed now at hand, we are ready to further explore the connection between the Lorentz group and $SL(2,\mathbb{C})$. Let v be any element of $SL(2,\mathbb{C})$ so that

$$\det v = 1. \tag{8.2.199}$$

Verify that then $v^\dagger$ is also in $SL(2,\mathbb{C})$ so that

$$\det v^\dagger = 1. \tag{8.2.200}$$

Next let x be any real four-vector. Consider the matrix $v(x \star \sigma)v^\dagger$. It is evidently 2×2. Show that it is also Hermitian,

$$[v(x \star \sigma)v^\dagger]^\dagger = v(x \star \sigma)v^\dagger. \tag{8.2.201}$$

Since $v(x \star \sigma)v^\dagger$ is a 2×2 Hermitian matrix, there must be a real and unique four-vector $\hat{x}$ such that

$$\hat{x} \star \sigma = v(x \star \sigma)v^\dagger. \tag{8.2.202}$$

Show, by taking determinants of both sides of (2.202), that there is the relation

$$\hat{x} \cdot \hat{x} = x \cdot x. \tag{8.2.203}$$

It follows that $\hat{x}$ and x are related by a Lorentz transformation! Thus, for each element $v \in SL(2,\mathbb{C})$ there is a Lorentz transformation $\Lambda(v)$.

How might one have guessed that this should be the case? We know from Exercise 7.3.30 that the $v \in SL(2,\mathbb{C})$ carry the representation $\Gamma(0,1/2)$ and we might expect, as can be proved, that the $v^\dagger$ would carry the representation $\Gamma(1/2,0)$. Since the right side of (2.202) involves both v and $v^\dagger$ in a "multiplicative" way, we might expect that what we are doing in (2.202) would involve the representation $\Gamma(0,1/2) \times \Gamma(1/2,0)$. But for the Lorentz group it is easy to see that there is the Clebsch-Gordan result

$$\Gamma(0,1/2) \times \Gamma(1/2,0) = \Gamma(1/2,1/2). \tag{8.2.204}$$

And, according to Exercise 7.3.29, $\Gamma(1/2,1/2)$ is the representation carried by Lorentz transformation matrices Λ acting on four-vectors.

Your next task is to find the matrix elements of Λ as functions of v. Let us examine and manipulate the "contents" of $v(x \star \sigma)v^\dagger$, the right side of (2.202). Verify that

$$v(x \star \sigma)v^\dagger = v(\sum_\nu x^\nu \sigma^\nu)v^\dagger = \sum_\nu (v\sigma^\nu v^\dagger)x^\nu. \tag{8.2.205}$$

Show that the matrices $v\sigma^\nu v^\dagger$ are Hermitian,

$$(v\sigma^\nu v^\dagger)^\dagger = v\sigma^\nu v^\dagger. \tag{8.2.206}$$

Verify it follows that there are *real* coefficients, call them $\Lambda^{\xi\nu}(v)$, such that

$$v\sigma^\nu v^\dagger = \sum_{\xi=1}^{4} \Lambda^{\xi\nu}(v)\sigma^\xi. \tag{8.2.207}$$

Show that, correspondingly, there is the relation

$$\Lambda^{\mu\nu}(v) = (1/2)\,\mathrm{tr}(\sigma^\mu v\sigma^\nu v^\dagger). \tag{8.2.208}$$

We already know that Λ is a real matrix. Still, it would be good to reverify directly that Λ as given by (2.208) is real even though some of the matrices appearing on the right side of (2.208) may be complex. Show, using (3.6.129) and (3.6.130), that

$$\begin{aligned}[\Lambda^{\mu\nu}(v)]^* &= (1/2)\,\mathrm{tr}[(\sigma^\mu v\sigma^\nu v^\dagger)^\dagger] = (1/2)\,\mathrm{tr}[v\sigma^\nu v^\dagger\sigma^\mu] \\ &= (1/2)\,\mathrm{tr}[\sigma^\mu v\sigma^\nu v^\dagger] = \Lambda^{\mu\nu}(v). \end{aligned} \tag{8.2.209}$$

Now what can be said about the vector $\hat{x}$ and its relation to x? Verify, in view of (2.202), that the components of $\hat{x}$ are given by the relations

$$\hat{x}^\mu = (1/2)\,\mathrm{tr}[\sigma^\mu v(x \star \sigma)v^\dagger]. \tag{8.2.210}$$

Next use (2.205) to find that

$$(1/2)\,\mathrm{tr}[\sigma^\mu v(x \star \sigma)v^\dagger] = (1/2)\sum_\nu [\mathrm{tr}(\sigma^\mu v\sigma^\nu v^\dagger)]x^\nu. \tag{8.2.211}$$

Finally, upon combining (2.208), (2.210), and (2.211), show that

$$\hat{x}^\mu = \sum_\nu \Lambda^{\mu\nu}(v)x^\nu \text{ or, in matrix/vector form, } \hat{x} = \Lambda x. \tag{8.2.212}$$

We have learned, as anticipated by our notation, that the quantities $\Lambda^{\mu\nu}(v)$ defined by (2.208) are the entries of a Lorentz transformation matrix.

What can be said about group properties? Suppose, to begin, that v is the 2×2 identity matrix σ^4. Verify that in this case use of (2.208) gives the result

$$\Lambda^{\mu\nu}(\sigma^4) = (1/2)\,\mathrm{tr}[\sigma^\mu\sigma^4\sigma^\nu(\sigma^4)^\dagger] = (1/2)\,\mathrm{tr}[\sigma^\mu\sigma^\nu] = \delta^{\mu\nu}. \tag{8.2.213}$$

That is, the image of the identity element in $SL(2,\mathbb{C})$ is the identity matrix in the Lorentz group. Next suppose that v is of the product form

$$v = uw \tag{8.2.214}$$

where u and w are both elements of $SL(2,\mathbb{C})$. In this case verify that use of (2.208) gives the result

$$\Lambda^{\mu\nu}(uw) = (1/2)\,\mathrm{tr}[\sigma^\mu uw\sigma^\nu(uw)^\dagger] = (1/2)\,\mathrm{tr}[\sigma^\mu u(w\sigma^\nu w^\dagger)u^\dagger]. \tag{8.2.215}$$

But, in analogy to (2.207), verify that

$$w\sigma^\nu w^\dagger = \sum_{\xi=1}^{4}\Lambda^{\xi\nu}(w)\sigma^\xi. \tag{8.2.216}$$

Show that combining (2.215) and (2.216) gives the intermediate results

$$\Lambda^{\mu\nu}(uw) = (1/2)\,\mathrm{tr}[\sigma^\mu u(\sum_{\xi=1}^{4}\Lambda^{\xi\nu}(w)\sigma^\xi)u^\dagger] = \sum_{\xi=1}^{4}\Lambda^{\xi\nu}(w)(1/2)\,\mathrm{tr}[\sigma^\mu(u\sigma^\xi u^\dagger)]. \tag{8.2.217}$$

But, again in analogy to (2.207), verify that

$$u\sigma^\xi u^\dagger = \sum_{\rho=1}^{4}\Lambda^{\rho\xi}(u)\sigma^\rho. \tag{8.2.218}$$

Show that combining (2.217) and (2.218) gives the final result

$$\begin{aligned}\Lambda^{\mu\nu}(uw) &= \sum_{\xi=1}^{4}\Lambda^{\xi\nu}(w)(1/2)\,\mathrm{tr}[\sigma^\mu(u\sigma^\xi u^\dagger)]\\ &= \sum_{\xi=1}^{4}\Lambda^{\xi\nu}(w)(1/2)\,\mathrm{tr}[\sigma^\mu\sum_{\rho=1}^{4}\Lambda^{\rho\xi}(u)\sigma^\rho]\\ &= \sum_{\xi=1}^{4}\sum_{\rho=1}^{4}\Lambda^{\rho\xi}(u)\Lambda^{\xi\nu}(w)(1/2)\,\mathrm{tr}[\sigma^\mu\sigma^\rho]\\ &= \sum_{\xi=1}^{4}\sum_{\rho=1}^{4}\Lambda^{\rho\xi}(u)\Lambda^{\xi\nu}(w)\delta^{\mu\rho}\\ &= \sum_{\xi=1}^{4}\Lambda^{\mu\xi}(u)\Lambda^{\xi\nu}(w) \qquad (8.2.219)\end{aligned}$$

or, in index-free notation,

$$\Lambda(uw) = \Lambda(u)\Lambda(w). \tag{8.2.220}$$

Complete our group property study with two calculations: First set

$$w = \sigma^4 \tag{8.2.221}$$

in (2.220) and deduce that

$$\Lambda(\sigma^4) = I \tag{8.2.222}$$

in agreement with (2.213). Second, set

$$w = v^{-1} \tag{8.2.223}$$

in (2.220) and deduce that

$$\Lambda(v^{-1}) = [\Lambda(v)]^{-1}. \tag{8.2.224}$$

The relation (2.220) and those that follow from it show that Lorentz transformation matrices Λ provide a representation of $SL(2,\mathbb{C})$. That is, the map (2.208) that sends elements of $SL(2,\mathbb{C})$ into elements of the Lorentz group is a homomorphism. At this point it is important to observe, by inspection, that the map (2.208) has the two-to-one property

$$\Lambda(-v) = \Lambda(v). \tag{8.2.225}$$

Therefore (2.208) is not an isomorphism. As we will see, $SL(2,\mathbb{C})$ is the covering group of the Lorentz group.

In Exercises 7.7.27 and 7.7.30 it was shown that the Lorentz group and $SL(2,\mathbb{C})$ have identical Lie algebras and analogous polar decompositions. And in this exercise we have seen that (2.208) provides a two-to-one homomorphic relation between the Lorentz group and $SL(2,\mathbb{C})$. The remainder of this exercise explores how these results fit together.

Suppose, employing the polar decomposition (7.3.241), that v is written in the form

$$v = \exp(\lambda \boldsymbol{m}\cdot\hat{\boldsymbol{N}})\exp(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}}). \tag{8.2.226}$$

Show that employing this factorization in (2.220) yields the result

$$\Lambda(v) = \Lambda[\exp(\lambda\boldsymbol{m}\cdot\hat{\boldsymbol{N}})]\Lambda[\exp(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}})]. \tag{8.2.227}$$

Your next tasks will be to work out results for each factor on the right side of (2.227).

Begin Evaluation of Second Factor

Begin with the second factor. For it you will need to work out

$$\Lambda^{\mu\nu}(w) = (1/2)\,\mathrm{tr}(\sigma^\mu w\sigma^\nu w^\dagger) \tag{8.2.228}$$

with

$$w = \exp(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}}) = \exp(\theta\boldsymbol{n}\cdot\boldsymbol{K}) = \exp[\theta(-i/2)\boldsymbol{n}\cdot\boldsymbol{\sigma}]. \tag{8.2.229}$$

See (7.3.232) through (7.3.234). Verify that in this case w is unitary.

The simplest matrix element to work out is Λ^{44}. Verify using (2.208) that

$$\Lambda^{44}(w) = (1/2)\,\mathrm{tr}(\sigma^4 w\sigma^4 w^\dagger) = (1/2)\,\mathrm{tr}(ww^\dagger) = (1/2)\,\mathrm{tr}(\sigma^4) = 1. \tag{8.2.230}$$

The next simplest cases are the $\Lambda^{\alpha 4}$ and $\Lambda^{4\alpha}$ with $\alpha = 1,2,3$. Verify that for these α

$$\Lambda^{\alpha 4}(w) = (1/2)\,\mathrm{tr}(\sigma^\alpha w\sigma^4 w^\dagger) = (1/2)\,\mathrm{tr}(\sigma^\alpha) = 0 \tag{8.2.231}$$

and

$$\Lambda^{4\alpha}(w) = (1/2)\,\mathrm{tr}(\sigma^4 w \sigma^\alpha w^\dagger) = (1/2)\,\mathrm{tr}(w^\dagger w \sigma^\alpha) = (1/2)\,\mathrm{tr}(\sigma^\alpha) = 0. \tag{8.2.232}$$

[Here, in writing (2.232), use has been made of the trace relation (3.6.130).] So it has now been established that $\Lambda(w)$ is of the form

$$\Lambda(w) = \begin{pmatrix} * & * & * & 0 \\ * & * & * & 0 \\ * & * & * & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{8.2.233}$$

To fill in the missing entries in (2.233) the simplest way at this point is to use what has already been accomplished in Exercises 3.7.31 and 2.10, which you should review. Note, again using (3.6.130), that (2.208) can be rewritten in the form

$$\Lambda^{\mu\nu}(v) = (1/2)\,\mathrm{tr}(v^\dagger \sigma^\mu v \sigma^\nu), \tag{8.2.234}$$

which has the same form as (2.54). Consequently, the missing entries are the elements of the 3×3 matrix given by (2.54). It follows, as expected, that for the w given by (2.229) $\Lambda(w)$ is given by the relation

$$\Lambda(w) = \exp(\theta \boldsymbol{n} \cdot \boldsymbol{L}) \tag{8.2.235}$$

or

$$\Lambda[\exp(\theta \boldsymbol{n} \cdot \hat{\boldsymbol{L}})] = \exp(\theta \boldsymbol{n} \cdot \boldsymbol{L}) \tag{8.2.236}$$

where the $\boldsymbol{L}$ are the matrices given by (7.3.177) through (7.3.179) and whose upper left 3×3 submatrices are the matrices given by (3.7.178) through (3.7.180).

As a sanity check of (2.236), consider the simple case where

$$\boldsymbol{n} = \boldsymbol{e}_3 \tag{8.2.237}$$

so that

$$w = \exp(\theta \boldsymbol{n} \cdot \hat{\boldsymbol{L}}) = \exp(\theta \hat{L}^3) = \exp[\theta(-i/2)\sigma^3]. \tag{8.2.238}$$

Verify that

$$w = \exp[\theta(-i/2)\sigma^3] = \exp(\theta K^3) = \begin{pmatrix} \exp(-i\theta/2) & 0 \\ 0 & \exp(i\theta/2) \end{pmatrix}. \tag{8.2.239}$$

See (3.7.171) and (3.7.194). Also, compare (2.236) with the analogous result (2.54). Show that for the w given by (2.235) and (2.236) there is the result

$$\Lambda(w) = \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 & 0 \\ \sin(\theta) & \cos(\theta) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \exp(\theta L^3) \tag{8.2.240}$$

with L^3 given by (7.3.179. See (3.7.207). Note that, according to (2.240), $\Lambda(w)$ is periodic in θ with period 2π,

$$\Lambda(w)|_{\theta+2\pi} = \Lambda(w)|_\theta. \tag{8.2.241}$$

But, according to (2.239), w is not; it has period 4π,

$$w|_{\theta+4\pi} = w|_{\theta}, \tag{8.2.242}$$

and

$$w|_{\theta+2\pi} = -w|_{\theta}. \tag{8.2.243}$$

Show that this result is consistent with (2.225). Show also that results completely analogous to (2.251) through (2.243) hold for any choice of the unit vector $\boldsymbol{n}$ and are consistent with (2.225).

The work of Exercise 2.10, which is what we have been using, involved the formation and solution of a differential equation. Another way to proceed, which is the same in essence, is to evaluate (2.229) and then (2.228) for small θ, and then build up to large values of θ by repeatedly using the group property (2.220). Let us explore this approach, which will turn out to be simpler.

Suppose, in view of (2.229), we work with a w and a $w^\dagger$ of the forms

$$w = \exp(\epsilon \boldsymbol{n} \cdot \hat{\boldsymbol{L}}) = \exp[\epsilon(-i/2)\boldsymbol{n} \cdot \boldsymbol{\sigma}] = I + \epsilon(-i/2)\boldsymbol{n} \cdot \boldsymbol{\sigma} + O(\epsilon^2), \tag{8.2.244}$$

$$w^\dagger = I + \epsilon(+i/2)\boldsymbol{n} \cdot \boldsymbol{\sigma} + O(\epsilon^2). \tag{8.2.245}$$

Show that employing this w and $w^\dagger$ pair in (2.228) gives the result

$$\begin{aligned}
\Lambda^{\mu\nu}(w) &= (1/2)\,\mathrm{tr}(\sigma^\mu w \sigma^\nu w^\dagger) \\
&= (1/2)\,\mathrm{tr}[\sigma^\mu(I + \epsilon(-i/2)\boldsymbol{n} \cdot \boldsymbol{\sigma})\sigma^\nu(I + \epsilon(+i/2)\boldsymbol{n} \cdot \boldsymbol{\sigma})] + O(\epsilon^2) \\
&= (1/2)\,\mathrm{tr}\{\sigma^\mu\sigma^\nu + \epsilon(-i/2)\sigma^\mu[(\boldsymbol{n} \cdot \boldsymbol{\sigma})\sigma^\nu - \sigma^\nu(\boldsymbol{n} \cdot \boldsymbol{\sigma})]\} + O(\epsilon^2) \\
&= \delta^{\mu\nu} + \epsilon(-i/4)\,\mathrm{tr}\{\sigma^\mu[(\boldsymbol{n} \cdot \boldsymbol{\sigma})\sigma^\nu - \sigma^\nu(\boldsymbol{n} \cdot \boldsymbol{\sigma})]\} + O(\epsilon^2).
\end{aligned} \tag{8.2.246}$$

Verify that

$$(\boldsymbol{n} \cdot \boldsymbol{\sigma})\sigma^\nu - \sigma^\nu(\boldsymbol{n} \cdot \boldsymbol{\sigma}) = \sum_{\xi=1}^{3} n_\xi\{\sigma^\xi, \sigma^\nu\}. \tag{8.2.247}$$

Make the definition

$$n_4 = 0 \tag{8.2.248}$$

so that (2.247) can also be written in the form

$$(\boldsymbol{n} \cdot \boldsymbol{\sigma})\sigma^\nu - \sigma^\nu(\boldsymbol{n} \cdot \boldsymbol{\sigma}) = \sum_{\xi=1}^{4} n_\xi\{\sigma^\xi, \sigma^\nu\}. \tag{8.2.249}$$

Pause to Develop Needed Mathematical Results

At this point we pause in our present calculations to define and develop the properties of a remarkable tensor that will be of subsequent use. Let $U^{\alpha\beta\gamma}$ be the tensor defined in terms of the Pauli matrices σ^1 through σ^4 by the rule

$$U^{\alpha\beta\gamma} = \mathrm{tr}[\sigma^\alpha(\sigma^\beta\sigma^\gamma - \sigma^\gamma\sigma^\beta)]. \tag{8.2.250}$$

Evidently U has the symmetry property

$$U^{\alpha\gamma\beta} = -U^{\alpha\beta\gamma}. \tag{8.2.251}$$

That is, U is antisymmetric under the interchange of its last two indices. Next verify, because of the trace relation (3.6.130), that

$$\begin{aligned} U^{\alpha\beta\gamma} &= \mathrm{tr}[\sigma^\alpha(\sigma^\beta\sigma^\gamma - \sigma^\gamma\sigma^\beta)] = \mathrm{tr}[\sigma^\alpha\sigma^\beta\sigma^\gamma - \sigma^\alpha\sigma^\gamma\sigma^\beta] \\ &= \mathrm{tr}[\sigma^\beta\sigma^\gamma\sigma^\alpha - \sigma^\gamma\sigma^\beta\sigma^\alpha] = \mathrm{tr}[\sigma^\beta\sigma^\gamma\sigma^\alpha - \sigma^\beta\sigma^\alpha\sigma^\gamma] \\ &= \mathrm{tr}[\sigma^\beta(\sigma^\gamma\sigma^\alpha - \sigma^\alpha\sigma^\gamma)] = -U^{\beta\alpha\gamma}. \end{aligned} \tag{8.2.252}$$

Thus, U is also antisymmetric under the interchange of its first two indices. Show that it follows from (2.251) and (2.252) that U is *completely* antisymmetric: That is, U changes sign under the interchange of any pair of indices.

To continue, define an antisymmetric tensor A with Pauli matrix entries by the rule

$$A^{\mu\nu} = \{\sigma^\mu, \sigma^\nu\}. \tag{8.2.253}$$

Show that

$$A = 2i \begin{pmatrix} 0 & \sigma^3 & -\sigma^2 & 0 \\ -\sigma^3 & 0 & \sigma^1 & 0 \\ \sigma^2 & -\sigma^1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{8.2.254}$$

Look at the matrices L^α for $\alpha = 1,2,3$ defined by (3.180) through (3.182). Also define a matrix L^4 by the rule

$$L^4 = \mathbf{0} \tag{8.2.255}$$

where $\mathbf{0}$ is the 4×4 matrix with all entries having value zero. Show that

$$A = -2i \sum_{\alpha=1}^{4} L^\alpha \sigma^\alpha. \tag{8.2.256}$$

Consequently there is the associated component relation

$$\{\sigma^\mu, \sigma^\nu\} = A^{\mu\nu} = -2i \sum_{\alpha=1}^{4} (L^\alpha)^{\mu\nu} \sigma^\alpha \tag{8.2.257}$$

from which it follows that

$$\mathrm{tr}[\sigma^\beta(\sigma^\mu\sigma^\nu - \sigma^\nu\sigma^\mu)] = \mathrm{tr}[\sigma^\beta\{\sigma^\mu, \sigma^\nu\}] = -2i \sum_{\alpha=1}^{4} (L^\alpha)^{\mu\nu}\, \mathrm{tr}(\sigma^\beta\sigma^\alpha) = -4i(L^\beta)^{\mu\nu}. \tag{8.2.258}$$

Finally, show that comparison of (2.249) and (2.257) gives the result

$$A^{\beta\mu\nu} = -4i(L^\beta)^{\mu\nu}. \tag{8.2.259}$$

With these definitions before us, show that

$$(\boldsymbol{n}\cdot\boldsymbol{\sigma})\sigma^\nu - \sigma^\nu(\boldsymbol{n}\cdot\boldsymbol{\sigma}) = \sum_{\xi=1}^{4} n_\xi\{\sigma^\xi,\sigma^\nu\} = \sum_{\xi=1}^{4} n_\xi A^{\xi\nu} = -2i\sum_{\xi=1}^{4}\sum_{\alpha=1}^{4} n_\xi (L^\alpha)^{\xi\nu}\sigma^\alpha. \quad (8.2.260)$$

Continue on to show that

$$\mathrm{tr}\{\sigma^\mu[(\boldsymbol{n}\cdot\boldsymbol{\sigma})\sigma^\nu - \sigma^\nu(\boldsymbol{n}\cdot\boldsymbol{\sigma})]\} = -2i\sum_{\xi=1}^{4}\sum_{\alpha=1}^{4} n_\xi (L^\alpha)^{\xi\nu}\,\mathrm{tr}(\sigma^\mu\sigma^\alpha) = -4i\sum_{\xi=1}^{4} n_\xi (L^\mu)^{\xi\nu}. \quad (8.2.261)$$

But, from the relation (2.259) and the symmetry properties of A, it follows that

$$(L^\mu)^{\xi\nu} = -(L^\xi)^{\mu\nu}. \quad (8.2.262)$$

Show, therefore, that (2.261) can be rewritten in the form

$$\mathrm{tr}\{\sigma^\mu[(\boldsymbol{n}\cdot\boldsymbol{\sigma})\sigma^\nu - \sigma^\nu(\boldsymbol{n}\cdot\boldsymbol{\sigma})]\} = -4i\sum_{\xi=1}^{4} n_\xi (L^\mu)^{\xi\nu} = 4i\sum_{\xi=1}^{4} n_\xi (L^\xi)^{\mu\nu} = 4i(\boldsymbol{n}\cdot\boldsymbol{L})^{\mu\nu}. \quad (8.2.263)$$

Resume and Complete Evaluation of Second Factor

Verify, as a result of these intervening calculations, that combining (2.246) and (2.263) yields the pleasingly simple result

$$\Lambda^{\mu\nu}(w) = \delta^{\mu\nu} + \epsilon(\boldsymbol{n}\cdot\boldsymbol{L})^{\mu\nu} + O(\epsilon^2) \quad (8.2.264)$$

or, in matrix form,

$$\Lambda(w) = I + \epsilon\boldsymbol{n}\cdot\boldsymbol{L} + O(\epsilon^2). \quad (8.2.265)$$

Show, using (2.244) and (2.265), that there is the relation

$$\Lambda[\exp(\epsilon\boldsymbol{n}\cdot\hat{\boldsymbol{L}})] = \exp(\epsilon\boldsymbol{n}\cdot\boldsymbol{L})] + O(\epsilon^2). \quad (8.2.266)$$

As promised, we now intend to use (2.266) and the group property (2.219) to work out results for finite values of θ. Let ℓ be a positive integer and set

$$\epsilon = \theta/\ell. \quad (8.2.267)$$

Show that

$$\begin{aligned}\Lambda[\exp(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}})] &= \Lambda[\exp(\ell\epsilon\boldsymbol{n}\cdot\hat{\boldsymbol{L}})] = \{\Lambda[\exp(\epsilon\boldsymbol{n}\cdot\hat{\boldsymbol{L}})]\}^\ell \\ &= [\exp(\epsilon\boldsymbol{n}\cdot\boldsymbol{L}) + O(\epsilon^2)]^\ell = \{\exp(\epsilon\boldsymbol{n}\cdot\boldsymbol{L}) + O[(\theta/\ell)^2]\}^\ell \\ &= \exp(\ell\epsilon\boldsymbol{n}\cdot\boldsymbol{L}) + \ell O[(\theta/\ell)^2] = \exp(\theta\boldsymbol{n}\cdot\boldsymbol{L}) + O(\theta^2/\ell). \quad (8.2.268)\end{aligned}$$

Verify that taking the limit $\ell \to \infty$ in (2.268) yields the expected result

$$\Lambda[\exp(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}})] = \exp(\theta\boldsymbol{n}\cdot\boldsymbol{L}). \quad (8.2.269)$$

Begin Evaluation of First Factor

What remains is to work out results for the first factor on the right side of (2.226). For it you will need to work out $\Lambda(u)$ with

$$u = \exp(\lambda \boldsymbol{m} \cdot \hat{\boldsymbol{N}}) = \exp[\lambda(1/2)\boldsymbol{m} \cdot \boldsymbol{\sigma}]. \tag{8.2.270}$$

See (7.3.235) through (7.3.237). Verify that u is the exponential of a Hermitian matrix and is therefore Hermitian and positive definite.[5] See Exercise 3.7.44. According to (2.208) what we now need to compute are the entries

$$\Lambda^{\mu\nu}(u) = (1/2)\,\mathrm{tr}(\sigma^{\mu} u \sigma^{\nu} u^{\dagger}) = (1/2)\,\mathrm{tr}(\sigma^{\mu} u \sigma^{\nu} u). \tag{8.2.271}$$

To proceed, we will adopt a strategy analogous to what we employed for the second factor. We will evaluate (2.266) for small λ and then build up to large values of λ by repeatedly using the group property (2.219).

Suppose, in view of (2.242), we work with a u of the form

$$u = \exp(\epsilon \boldsymbol{m} \cdot \hat{\boldsymbol{N}}) = \exp[\epsilon(1/2)\boldsymbol{m} \cdot \boldsymbol{\sigma}] = I + \epsilon(1/2)\boldsymbol{m} \cdot \boldsymbol{\sigma} + O(\epsilon^2). \tag{8.2.272}$$

Show that employing this u in (2.243) gives the result

$$\begin{aligned}\Lambda^{\mu\nu}(u) &= (1/2)\,\mathrm{tr}(\sigma^{\mu} u \sigma^{\nu} u)\\ &= (1/2)\,\mathrm{tr}[\sigma^{\mu}(I + \epsilon(1/2)\boldsymbol{m} \cdot \boldsymbol{\sigma})\sigma^{\nu}(I + \epsilon(1/2)\boldsymbol{m} \cdot \boldsymbol{\sigma})] + O(\epsilon^2)\\ &= (1/2)\,\mathrm{tr}\{\sigma^{\mu}\sigma^{\nu} + \epsilon(1/2)\sigma^{\mu}[(\boldsymbol{m} \cdot \boldsymbol{\sigma})\sigma^{\nu} + \sigma^{\nu}(\boldsymbol{m} \cdot \boldsymbol{\sigma})]\} + O(\epsilon^2)\\ &= \delta^{\mu\nu} + \epsilon(1/4)\,\mathrm{tr}\{\sigma^{\mu}[(\boldsymbol{m} \cdot \boldsymbol{\sigma})\sigma^{\nu} + \sigma^{\nu}(\boldsymbol{m} \cdot \boldsymbol{\sigma})]\} + O(\epsilon^2).\end{aligned} \tag{8.2.273}$$

Verify that

$$(\boldsymbol{m} \cdot \boldsymbol{\sigma})\sigma^{\nu} + \sigma^{\nu}(\boldsymbol{m} \cdot \boldsymbol{\sigma}) = \sum_{\xi=1}^{3} m_{\xi}\{\sigma^{\xi}, \sigma^{\nu}\}_{+}. \tag{8.2.274}$$

Make the definition

$$m_4 = 0 \tag{8.2.275}$$

so that (2.246) can also be written in the form

$$(\boldsymbol{m} \cdot \boldsymbol{\sigma})\sigma^{\nu} + \sigma^{\nu}(\boldsymbol{m} \cdot \boldsymbol{\sigma}) = \sum_{\xi=1}^{4} m_{\xi}\{\sigma^{\xi}, \sigma^{\nu}\}_{+}. \tag{8.2.276}$$

[5] Note that, unlike in the context of the second factor where we encounter both w and $-w$, in the context of the first factor we will not encounter both u and $-u$ because if u is positive definite, then $-u$ is not, and in the context of the first factor we only encounter the positive definite case.

Again Pause to Develop Needed Mathematical Results

At this point we pause in our present calculations to define and develop the properties of another remarkable tensor that will be of subsequent use. Let $V^{\alpha\beta\gamma}$ be the tensor defined in terms of the Pauli matrices σ^1 through σ^4 by the rule

$$V^{\alpha\beta\gamma} = \mathrm{tr}[\sigma^\alpha(\sigma^\beta\sigma^\gamma + \sigma^\gamma\sigma^\beta)]. \tag{8.2.277}$$

Evidently V has the symmetry property

$$V^{\alpha\gamma\beta} = V^{\alpha\beta\gamma}. \tag{8.2.278}$$

That is, V is symmetric under the interchange of its last two indices. Next verify, because of the trace relation (3.6.130), that

$$\begin{aligned} V^{\alpha\beta\gamma} &= \mathrm{tr}[\sigma^\alpha(\sigma^\beta\sigma^\gamma + \sigma^\gamma\sigma^\beta)] = \mathrm{tr}[\sigma^\alpha\sigma^\beta\sigma^\gamma + \sigma^\alpha\sigma^\gamma\sigma^\beta] \\ &= \mathrm{tr}[\sigma^\beta\sigma^\gamma\sigma^\alpha + \sigma^\gamma\sigma^\beta\sigma^\alpha] = \mathrm{tr}[\sigma^\beta\sigma^\gamma\sigma^\alpha + \sigma^\beta\sigma^\alpha\sigma^\gamma] \\ &= \mathrm{tr}[\sigma^\beta(\sigma^\gamma\sigma^\alpha + \sigma^\alpha\sigma^\gamma)] = V^{\beta\alpha\gamma}. \end{aligned} \tag{8.2.279}$$

Thus, V is also symmetric under the interchange of its first two indices. Show that it follows from (2.250) and (2.251) that V is *completely* symmetric: That is, V is unchanged under any permutation of its indices.

To continue, define a symmetric tensor S with Pauli matrix entries by the rule

$$S^{\mu\nu} = \{\sigma^\mu, \sigma^\nu\}_+. \tag{8.2.280}$$

Show that

$$S = 2\begin{pmatrix} \sigma^4 & 0 & 0 & \sigma^1 \\ 0 & \sigma^4 & 0 & \sigma^2 \\ 0 & 0 & \sigma^4 & \sigma^3 \\ \sigma^1 & \sigma^2 & \sigma^3 & \sigma^4 \end{pmatrix}. \tag{8.2.281}$$

Look at the matrices N^α for $\alpha = 1, 2, 3$ defined by (3.180) through (3.182). Also define a matrix N^4 by the rule

$$N^4 = I \tag{8.2.282}$$

where I is the 4×4 identity matrix. Show that

$$S = 2\sum_{\alpha=1}^{4} N^\alpha\sigma^\alpha. \tag{8.2.283}$$

Consequently there is the associated component relation

$$\{\sigma^\mu, \sigma^\nu\}_+ = S^{\mu\nu} = 2\sum_{\alpha=1}^{4}(N^\alpha)^{\mu\nu}\sigma^\alpha \tag{8.2.284}$$

from which it follows that

$$\mathrm{tr}[\sigma^\beta(\sigma^\mu\sigma^\nu + \sigma^\nu\sigma^\mu)] = \mathrm{tr}[\sigma^\beta\{\sigma^\mu, \sigma^\nu\}_+] = 2\sum_{\alpha=1}^{4}(N^\alpha)^{\mu\nu}\,\mathrm{tr}(\sigma^\beta\sigma^\alpha) = 4(N^\beta)^{\mu\nu}. \tag{8.2.285}$$

Finally, show that comparison of (2.249) and (2.257) gives the result

$$V^{\beta\mu\nu} = 4(N^{\beta})^{\mu\nu}. \tag{8.2.286}$$

With these definitions before us, show that

$$(\boldsymbol{m}\cdot\boldsymbol{\sigma})\sigma^{\nu} + \sigma^{\nu}(\boldsymbol{m}\cdot\boldsymbol{\sigma}) = \sum_{\xi=1}^{4} m_{\xi}\{\sigma^{\xi},\sigma^{\nu}\}_{+} = \sum_{\xi=1}^{4} m_{\xi}S^{\xi\nu} = 2\sum_{\xi=1}^{4}\sum_{\alpha=1}^{4} m_{\xi}(N^{\alpha})^{\xi\nu}\sigma^{\alpha}. \tag{8.2.287}$$

Continue on to show that

$$\mathrm{tr}\{\sigma^{\mu}[(\boldsymbol{m}\cdot\boldsymbol{\sigma})\sigma^{\nu} + \sigma^{\nu}(\boldsymbol{m}\cdot\boldsymbol{\sigma})]\} = 2\sum_{\xi=1}^{4}\sum_{\alpha=1}^{4} m_{\xi}(N^{\alpha})^{\xi\nu}\,\mathrm{tr}[\sigma^{\mu}\sigma^{\alpha}] = 4\sum_{\xi=1}^{4} m_{\xi}(N^{\mu})^{\xi\nu}. \tag{8.2.288}$$

But, from the relation (2.258) and the symmetry properties of V, it follows that

$$(N^{\mu})^{\xi\nu} = (N^{\xi})^{\mu\nu}. \tag{8.2.289}$$

Show, therefore, that (2.260) can be rewritten in the form

$$\mathrm{tr}\{\sigma^{\mu}[(\boldsymbol{m}\cdot\boldsymbol{\sigma})\sigma^{\nu}+\sigma^{\nu}(\boldsymbol{m}\cdot\boldsymbol{\sigma})]\} = 4\sum_{\xi=1}^{4} m_{\xi}(N^{\mu})^{\xi\nu} = 4\sum_{\xi=1}^{4} m_{\xi}(N^{\xi})^{\mu\nu} = 4(\boldsymbol{m}\cdot\boldsymbol{N})^{\mu\nu}. \tag{8.2.290}$$

Resume and Complete Evaluation of First Factor

Show, as a result of these intervening calculations, that combining (2.269) and (2.286) yields the pleasingly simple result

$$\Lambda^{\mu\nu}(u) = \delta^{\mu\nu} + \epsilon(\boldsymbol{m}\cdot\boldsymbol{N})^{\mu\nu} + O(\epsilon^2) \tag{8.2.291}$$

or, in matrix form,

$$\Lambda(u) = I + \epsilon\boldsymbol{m}\cdot\boldsymbol{N} + O(\epsilon^2). \tag{8.2.292}$$

[We remark that had we replaced the matrices in (7.3.235) through (7.3.327) by their negatives, which would not have affected the commutation rules (7.328) through (7.3.240), the + signs in (2.291) and (2.292) would have been replaced by − signs. This change would lead to unpleasant consequences in what follows.]

As promised, we now intend to use (2.292) and the group property (2.219) to work out results for finite values of λ. Review the steps (2.266) through (2.269) that led from (2.265) to (2.269). Demonstrate that analogous steps lead from (2.292) to the expected and desired relation

$$\Lambda[\exp(\lambda\boldsymbol{m}\cdot\hat{\boldsymbol{N}})] = \exp(\lambda\boldsymbol{m}\cdot\boldsymbol{N}). \tag{8.2.293}$$

What would have been the resulting relation had the + sign in (2.292) been a − sign?

Summary of Results

We began this exercise with the knowledge that the $SL(2,\mathbb{C})$ and Lorentz groups have the same Lie algebras and analogous polar decompositions. We therefore expect they are closely related. In this exercise we found that (2.208) provides a map that sends elements of $SL(2,\mathbb{C})$ into elements of the Lorentz group, and (2.219) showed that this map is a homomorphism. Subsequently application of this homomorphism to elements of $SL(2,\mathbb{C})$ written in polar form produced the factorization (2.226). The results for each factor were then found to be given by (2.269) and (2.293). Verify that combining (2.226), (2.269), and (2.293) produces the relation

$$\begin{aligned}\Lambda[\exp(\lambda\boldsymbol{m}\cdot\hat{\boldsymbol{N}})\exp(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}})] &= \Lambda[\exp(\lambda\boldsymbol{m}\cdot\hat{\boldsymbol{N}})]\Lambda[\exp(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}})] \\ &= \exp(\lambda\boldsymbol{m}\cdot\boldsymbol{N})\exp(\theta\boldsymbol{n}\cdot\boldsymbol{L}). \qquad (8.2.294)\end{aligned}$$

This relation shows that the map (2.208) has the property that every element in the Lorentz group has a preimage in $SL(2,\mathbb{C})$. From (2.224) it follows that every element of the Lorentz group has (at least) two distinct preimages in $SL(2,\mathbb{C})$. Therefore the map (2.208) is not an isomorphism. To be more precise, examination of (2.293) shows that the mapping between the first factors of $SL(2,\mathbb{C})$ and of the Lorentz group is one to one. And examination of the discussion associated with (2.232) shows that the relation between the second factors of $SL(2,\mathbb{C})$ and of the Lorentz group is two to one in complete analogy to the relation between $SU(2)$ and $SO(3,\mathbb{R})$. We conclude that the map between $SL(2,\mathbb{C})$. and the Lorentz group provided by (2.208) and (2.294) is two to one.

What can be said about the topologies of the two groups? Since the topology of $SL(2,\mathbb{C})$ is $E^3\times S^3$, and both factors are simply connected, $SL(2,\mathbb{C})$ can be shown to be simply connected. From (7.3.186) we see that the topology of the Lorentz group is $E^3\times SO(3,\mathbb{R})$. Since E^3 is simply connected and $SO(3,\mathbb{R})$ is doubly connected, the Lorentz group can be shown to be doubly connected. All these topological statements are consistent with the nature of the map between $SL(2,\mathbb{C})$ and the Lorentz group provided by (2.208) and (2.294). It follows that $SL(2,\mathbb{C})$ is the covering group of the Lorentz group.

8.2.15. Let $\boldsymbol{\phi}$ and $\boldsymbol{k}$ be two real three-component vectors. Use them to parameterize the $s\ell(2,\mathbb{C})$ element $\hat{S}$ defined by

$$\hat{S} = \boldsymbol{\phi}\cdot\hat{\boldsymbol{L}} + \boldsymbol{k}\cdot\hat{\boldsymbol{N}}. \qquad (8.2.295)$$

In turn, let v be the $SL(2,\mathbb{C})$ element defined by

$$v = \exp(\hat{S}). \qquad (8.2.296)$$

Prove that for the map (2.208) there is the relation

$$\Lambda(v) = \Lambda[\exp(\hat{S})] = \exp(\boldsymbol{\phi}\cdot\boldsymbol{L} + \boldsymbol{k}\cdot\boldsymbol{N}). \qquad (8.2.297)$$

Suggestion: First show that for small ϵ there is the result

$$\Lambda[\exp(\epsilon\hat{S})] = \exp[\epsilon(\boldsymbol{\phi}\cdot\boldsymbol{L} + \boldsymbol{k}\cdot\boldsymbol{N})] + O(\epsilon^2). \qquad (8.2.298)$$

This small ϵ result can be obtained by direct evaluation of (2.208) or, more easily, by use of (2.269), (2.293), and the BCH formula. Then use (2.298) and the homomorphism relation (2.219) to work out results for any finite value of ϵ including $\epsilon = 1$.

8.2.16. When can and when cannot an element in $SL(2,\mathbb{C})$ be written in single exponential form? And what can be said about the related question for the Lorentz group? Your task for this exercise is to answer these questions.

Begin with the case of $SL(2,\mathbb{C})$. Recall that any 2×2 matrix v can be diagonalized if its eigenvalues are distinct. That is, if the eigenvalues of v are distinct, it can be written in the form

$$v = ada^{-1} \tag{8.2.299}$$

where a is a nonsingular matrix, d is the diagonal matrix

$$d = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \tag{8.2.300}$$

and λ_1, λ_2 are its eigenvalues. Verify that in this case

$$\det(v) = \det(d) = \lambda_1\lambda_2. \tag{8.2.301}$$

Moreover,

$$\lambda_1\lambda_2 = 1 \tag{8.2.302}$$

if v is an element of $SL(2,\mathbb{C})$. Verify therefore that, in the $SL(2,\mathbb{C})$ and distinct eigenvalue case, d can be written in the form

$$d = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix} = \exp(\alpha\sigma^3) \tag{8.2.303}$$

where

$$\alpha = \log(\lambda). \tag{8.2.304}$$

Consequently, verify that any element of $SL(2,\mathbb{C})$ with distinct eigenvalues can be written in the forms

$$v = ada^{-1} = a[\exp(\alpha\sigma^3)]a^{-1} = \exp(\alpha a\sigma^3 a^{-1}). \tag{8.2.305}$$

Verify that the matrix $a\sigma^3a^{-1}$ is traceless, and therefore there is a three-component (possibly complex) vector $\boldsymbol{\beta}$ such that

$$\alpha a\sigma^3 a^{-1} = \boldsymbol{\beta}\cdot\boldsymbol{\sigma}. \tag{8.2.306}$$

Show that the net result of these deliberations is the relation

$$v = \exp(\boldsymbol{\beta}\cdot\boldsymbol{\sigma}), \tag{8.2.307}$$

which demonstrates that any element of $SL(2,\mathbb{C})$ with distinct eigenvalues can be written in single exponential form.

There remain the cases in which the eigenvalues of v are not distinct, in which cases because of (2.302) there are the eigenvalue possibilities $1, 1$ and $-1, -1$. Suppose that in these non distinct cases that v can nevertheless be diagonalized. Show that then there are the two possibilities

$$v = \sigma^4 = \exp(\mathbf{0}) \tag{8.2.308}$$

and

$$v = -\sigma^4 = \exp(i\pi\sigma^3). \tag{8.2.309}$$

Here, as before, $\mathbf{0}$ is the matrix with all zero entries. Evidently in both these cases v has again been written in single exponential form.

The last possibility for these non distinct cases is that v cannot be diagonalized. Verify that for this possibility there are only the cases

$$v = aj_{\pm}a^{-1} \tag{8.2.310}$$

where $j_{\pm}$ are the two *Jordan* normal form elements

$$j_{+} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \tag{8.2.311}$$

and

$$j_{-} = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}. \tag{8.2.312}$$

Consider the case with eigenvalues $1, 1$ and the Jordan normal form j_{+} given by (2.311). Show that in this case

$$j_{+} = \exp[(1/2)(\sigma^1 + i\sigma^2)] \tag{8.2.313}$$

from which it follows that

$$v = \exp[(1/2)a(\sigma^1 + i\sigma^2)a^{-1}]. \tag{8.2.314}$$

Verify that the exponent appearing in (2.314) is in $s\ell(2, \mathbb{C})$. You have demonstrated that in the $1, 1$ eigenvalue case every such v in $SL(2, \mathbb{C})$ can be written in single exponential form with the exponent being an element in $s\ell(2, \mathbb{C})$.

The final remaining case is that with eigenvalues $-1, -1$ and the Jordan normal form j_{-} given by (2.312). From the work of Exercise 3.7.12, which you should review, we know that j_{-} cannot be written in single exponential form. Show it follows that all $SL(2, \mathbb{C})$ elements that have j_{-} as their Jordan normal form cannot be written in single exponential form.

The case of $SL(2, \mathbb{C})$ has been dispatched: All elements in $SL(2, \mathbb{C})$ can be written in single exponential form except those having j_{-} as their Jordan normal form.

We now turn to the case of the Lorentz group. According to Exercise 2.15 above, all the Lorentz group elements associated with the $SL(2, \mathbb{C})$ elements that can be written in single exponential form can also be written in single exponential form. What about the Lorentz group elements associated with the $SL(2, \mathbb{C})$ elements whose Jordan normal form is j_{-}? Let v_{-} be any such $SL(2, \mathbb{C})$ element. What we wish to determine is the nature of $\Lambda(v_{-})$. Show that if a v in $SL(2, \mathbb{C})$ cannot be written in single exponential form, then $-v$ can. In particular, show that $-j_{-}$ has Jordan normal form j_{+}. (Again see the work of Exercise 3.7.12.) Verify it follows that $-v_{-}$ can be written in single exponential form, and therefore $\Lambda(-v_{-})$ can also be written in single exponential form. But from (2.225) we know that

$$\Lambda(v_{-}) = \Lambda(-v_{-}). \tag{8.2.315}$$

Therefore $\Lambda(v_{-})$, which is the same as $\Lambda(-v_{-})$, can be written in single exponential form. Conclude that *all* Lorentz group elements can be written in single exponential form!

8.2.17. Review Exercise 7.3.30. It showed that the use of $s\ell(2,\mathbb{C})$ provides the $\Gamma(1/2,0)$ and $\Gamma(0,1/2)$ representations of the Lorentz group Lie algebra. Review Exercise 7.3.34 that constructed an isomorphism between $n \times n$ possibly complex matrices and $2n \times 2n$ real matrices. The purpose of this exercise and the next is to describe how the results of Exercises 7.3.30 and 7.3.34 may be used to characterize/determine the effect of Lorentz transformations on what we will call Dirac 4-*spinors*.

Recall the $SL(2,\mathbb{C})$ group elements $\hat{\Lambda}$ given by (7.3.246). They are 2×2 possibly complex matrices. According to Exercise 7.3.30 they carry the representation $\Gamma(0,1/2)$ of the Lorentz group. Make the Ansatz

$$k = \hat{\Lambda} \tag{8.2.316}$$

and use (7.3.375) to define associated *real* 4×4 matrices. For example, suppose

$$\hat{\Lambda} = \exp(\theta \hat{L}^3). \tag{8.2.317}$$

Verify that in this case

$$\hat{\Lambda} = \exp(\theta \hat{L}^3) = \begin{pmatrix} \exp(-i\theta/2) & 0 \\ 0 & \exp(i\theta/2) \end{pmatrix} = \tag{8.2.318}$$

$$\begin{pmatrix} \cos(\theta/2) & 0 \\ 0 & \cos(\theta/2) \end{pmatrix} + i \begin{pmatrix} -\sin(\theta/2) & 0 \\ 0 & \sin(\theta/2) \end{pmatrix}, \tag{8.2.319}$$

and

$$K(k) = K(\hat{\Lambda}) = K[\exp(\theta \hat{L}^3)] = \begin{pmatrix} \cos(\theta/2) & 0 & \sin(\theta/2) & 0 \\ 0 & \cos(\theta/2) & 0 & -\sin(\theta/2) \\ -\sin(\theta/2) & 0 & \cos(\theta/2) & 0 \\ 0 & \sin(\theta/2) & 0 & \cos(\theta/2) \end{pmatrix}. \tag{8.2.320}$$

Verify that when $\theta = 2\pi$ there are the results

$$\hat{\Lambda} = \exp(2\pi \hat{L}^3) = -I^{[2]} \tag{8.2.321}$$

and

$$K(\hat{\Lambda}) = K[\exp(2\pi \hat{L}^3)] = -I^{[4]}. \tag{8.2.322}$$

Since the $\hat{\Lambda}$ are 2×2 matrices, they act on two-dimensional objects/arrays. Since the $K(\hat{\Lambda})$ are 4×4 matrices, they act on four-dimensional objects/arrays. But the 2×2 and 4×4 matrix/group elements given by (2.321) and (2.322) and corresponding to the Lorentz transformation consisting of a $\theta = 2\pi$ rotation about the z axis are *not* identity matrices. It follows that the objects/arrays on which they act are *not* vectors.[6] Rather, they are 2-*spinors* and 4-*spinors*, respectively.[7] Also note that, by the constructions (7.3.246) and (7.3.375), all elements $K(\hat{\Lambda})$ are continuously connected to the identity matrix $I^{[4]}$. As further evidence

[6] By their transformation properties ye shall know them.

[7] Sometimes what we have called 4-spinors are called *bispinors*.

that the four-component arrays that we have called 4-spinors are not vectors, observe that, according to (7.3.167), $-I^{[4]}$ is not a Lorentz transformation in the identity component of the Lorentz group.

Let b be any of the possibly complex 2×2 $s\ell(2, \mathbb{C})$ basis elements (7.3.236) through (7.3.241) and consider the associated *real* 4×4 matrices $K(b)$. What representation of the Lorentz group Lie algebra is provided by the matrices $K(\hat{L}^j)$ and $K(\hat{N}^j)$? Verify, using (7.3.383), that

$$WK(b)W^{-1} = \begin{pmatrix} b & 0 \\ 0 & \bar{b} \end{pmatrix}. \tag{8.2.323}$$

We know from the work of Exercise 7.3.30 that the matrices b provide the $\Gamma(0, 1/2)$ representation of the Lorentz group Lie algebra. Momentarily we will demonstrate that the matrices $\bar{b}$ provide the $\Gamma(1/2, 0)$ representation of the Lorentz group Lie algebra. That is, for the case of $s\ell(2, \mathbb{C})$, complex conjugation is equivalent to the grave operation. It follows from (2.323) that the matrices $K(b)$ provide the direct sum

$$\Gamma(0, 1/2) \,\oplus\, \Gamma(1/2, 0) = \Gamma(0, 1/2) \,\oplus\, \bar{\Gamma}(0, 1/2) = \bar{\Gamma}(1/2, 0) \,\oplus\, \Gamma(1/2, 0) \tag{8.2.324}$$

representation of the Lorentz group Lie algebra.

Now work on the advertised demonstration: We already know that complex conjugation is the result of the breve operation and that this operation is a conjugacy operation. See (3.7.223). For present purposes what we need to show is that, in the case of $s\ell(2, \mathbb{C})$, the breve/bar operation and the grave operation are equivalent in the sense of (3.7.218). Verify that for the bar and grave operations there are the relations

$$(\bar{\hat{L}}^1, \bar{\hat{L}}^2, \bar{\hat{L}}^3, \bar{\hat{N}}^1, \bar{\hat{N}}^2, \bar{\hat{N}}^3) = (-\hat{L}^1, \hat{L}^2, -\hat{L}^3, \hat{N}^1, -\hat{N}^2, \hat{N}^3), \tag{8.2.325}$$

$$(\grave{\hat{L}}^1, \grave{\hat{L}}^2, \grave{\hat{L}}^3, \grave{\hat{N}}^1, \grave{\hat{N}}^2, \grave{\hat{N}}^3) = (\hat{L}^1, \hat{L}^2, \hat{L}^3, -\hat{N}^1, -\hat{N}^2, -\hat{N}^3). \tag{8.2.326}$$

See (7.3.248) and (7.3.249). We will next show is that the contents of the right sides of (2.325) and (2.326) are related by a similarity transformation. Recall the relation (3.7.234) which involved the matrix

$$J_2 = i\sigma^2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{8.2.327}$$

Using the anticommutation relations (5.7.41) verify that

$$J_2 \sigma^1 J_2^{-1} = -\sigma^1, \tag{8.2.328}$$

$$J_2 \sigma^2 J_2^{-1} = \sigma^2, \tag{8.2.329}$$

$$J_2 \sigma^3 J_2^{-1} = -\sigma^3. \tag{8.2.330}$$

Show it follows from the definitions (7.3.236) through (7.3.241) and the results (2.328) through (3.230) that there are the similarity relations

$$J_2(-\hat{L}^1, \hat{L}^2, -\hat{L}^3, \hat{N}^1, -\hat{N}^2, \hat{N}^3)J_2^{-1} = (\hat{L}^1, \hat{L}^2, \hat{L}^3, -\hat{N}^1, -\hat{N}^2, -\hat{N}^3). \tag{8.2.331}$$

Thus the right sides of (2.325) and (2.326) are related by a similarity transformation, and therefore the left sides of (2.325) and (2.326) are related by a similarity transformation.

There is a second way of verifying that the matrices $K(\hat{L}^j)$ and $K(\hat{N}^j)$ provide a representation of $s\ell(2,\mathbb{C})$ that is equivalent to the direct sum (2.324). It essentially amounts to the previous argument, but appears to be more direct. Verify, using (7.3.236) through (7.3.241), that there are the *six* explicit results

$$K(\hat{L}^1) = (1/2)K(-i\sigma^1) = (1/2)\begin{pmatrix} \mathbf{0} & \sigma^1 \\ -\sigma^1 & \mathbf{0} \end{pmatrix}, \tag{8.2.332}$$

$$K(\hat{L}^2) = (1/2)K(-i\sigma^2) = (1/2)\begin{pmatrix} -i\sigma^2 & \mathbf{0} \\ \mathbf{0} & -i\sigma^2 \end{pmatrix}, \tag{8.2.333}$$

$$K(\hat{L}^3) = (1/2)K(-i\sigma^3) = (1/2)\begin{pmatrix} \mathbf{0} & \sigma^3 \\ -\sigma^3 & \mathbf{0} \end{pmatrix}, \tag{8.2.334}$$

$$K(\hat{N}^1) = (1/2)K(\sigma^1) = (1/2)\begin{pmatrix} \sigma^1 & \mathbf{0} \\ \mathbf{0} & \sigma^1 \end{pmatrix}, \tag{8.2.335}$$

$$K(\hat{N}^2) = (1/2)K(\sigma^2) = (1/2)\begin{pmatrix} \mathbf{0} & i\sigma^2 \\ -i\sigma^2 & \mathbf{0} \end{pmatrix}, \tag{8.2.336}$$

$$K(\hat{N}^3) = (1/2)K(\sigma^3) = (1/2)\begin{pmatrix} \sigma^3 & \mathbf{0} \\ \mathbf{0} & \sigma^3 \end{pmatrix}. \tag{8.2.337}$$

Next let V be the matrix defined by the rule

$$V = (1/\sqrt{2})\begin{pmatrix} I^{[2]} & iI^{[2]} \\ i\sigma^2 & \sigma^2 \end{pmatrix}. \tag{8.2.338}$$

Note that V involves $i\sigma^2$, which is J_2 in disguise, just as (2.331) involves J_2. Show that V is unitary so that

$$V^{-1} = V^\dagger = (1/\sqrt{2})\begin{pmatrix} I^{[2]} & -i\sigma^2 \\ -iI^{[2]} & \sigma^2 \end{pmatrix}. \tag{8.2.339}$$

Finally, by executing the indicated matrix multiplications, verify the six similarity relation results

$$VK(\hat{L}^1)V^{-1} = (-i/2)\begin{pmatrix} \sigma^1 & \mathbf{0} \\ \mathbf{0} & \sigma^1 \end{pmatrix}, \tag{8.2.340}$$

$$VK(\hat{L}^2)V^{-1} = (-i/2)\begin{pmatrix} \sigma^2 & \mathbf{0} \\ \mathbf{0} & \sigma^2 \end{pmatrix}, \tag{8.2.341}$$

$$VK(\hat{L}^3)V^{-1} = (-i/2)\begin{pmatrix} \sigma^3 & \mathbf{0} \\ \mathbf{0} & \sigma^3 \end{pmatrix}, \tag{8.2.342}$$

$$VK(\hat{N}^1)V^{-1} = (1/2)\begin{pmatrix} \sigma^1 & \mathbf{0} \\ \mathbf{0} & -\sigma^1 \end{pmatrix}, \tag{8.2.343}$$

$$VK(\hat{N}^2)V^{-1} = (1/2)\begin{pmatrix} \sigma^2 & \mathbf{0} \\ \mathbf{0} & -\sigma^2 \end{pmatrix}, \tag{8.2.344}$$

$$VK(\hat{N}^3)V^{-1} = (1/2)\begin{pmatrix} \sigma^3 & \mathbf{0} \\ \mathbf{0} & -\sigma^3 \end{pmatrix}. \tag{8.2.345}$$

In summary, there are the results

$$VK(\hat{L}^j)V^{-1} = \begin{pmatrix} \hat{L}^j & \mathbf{0} \\ \mathbf{0} & \grave{\hat{L}}^j \end{pmatrix} = \begin{pmatrix} \hat{L}^j & \mathbf{0} \\ \mathbf{0} & \hat{L}^j \end{pmatrix}, \tag{8.2.346}$$

$$VK(\hat{N}^j)V^{-1} = \begin{pmatrix} \hat{N}^j & \mathbf{0} \\ \mathbf{0} & \grave{\hat{N}}^j \end{pmatrix} = \begin{pmatrix} \hat{N}^j & \mathbf{0} \\ \mathbf{0} & -\hat{N}^j \end{pmatrix}. \tag{8.2.347}$$

Evidently the matrices on the right sides of (2.340) through (2.347) are block diagonal. Moreover, the upper blocks carry the representation $\Gamma(0, 1/2)$ of $s\ell(2, \mathbb{C})$ and the lower blocks carry the representation $\Gamma(1/2, 0)$. See Exercise 7.3.30. Thus the full matrices carry the representation

$$\Gamma(0, 1/2) \oplus \Gamma(1/2, 0). \tag{8.2.348}$$

.

8.2.18. This exercise is a continuation of the previous exercise and is devoted to sketching the relation of the $s\ell(2, \mathbb{C})$ representation provided by the real 4×4 matrices $K(\hat{L}^j)$ and $K(\hat{N}^j)$, see (7.3.385), to some of the mathematical machinery associated with some forms of the *Dirac* (1902-1984) equation.

Background

Discussion of the Dirac equation often begins with with the introduction of four 4×4 *gamma* matrices, denoted as γ^μ with $\mu = 1 \cdot \cdot 4$, which are required to satisfy the *anticommutation* rules

$$\{\gamma^\mu, \gamma^\nu\}_+ = \gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu = 2g^{\mu\nu}I^{[4]}. \tag{8.2.349}$$

The relations (2.349) are sometimes called the *Dirac algebra*.[8] For our purposes it is convenient to take the γ^μ to be the matrices

$$\gamma^1 = \begin{pmatrix} \mathbf{0} & i\sigma^3 \\ i\sigma^3 & \mathbf{0} \end{pmatrix} = i \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & -0 \end{pmatrix}, \tag{8.2.350}$$

$$\gamma^2 = \begin{pmatrix} iI^{[2]} & \mathbf{0} \\ \mathbf{0} & -iI^{[2]} \end{pmatrix} = i \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \tag{8.2.351}$$

$$\gamma^3 = \begin{pmatrix} \mathbf{0} & -i\sigma^1 \\ -i\sigma^1 & \mathbf{0} \end{pmatrix} = i \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}, \tag{8.2.352}$$

[8] Dirac algebra is a particular case of *Clifford* (1845-1879) algebra.

$$\gamma^4 = \begin{pmatrix} \mathbf{0} & -\sigma^2 \\ -\sigma^2 & \mathbf{0} \end{pmatrix} = i \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}. \tag{8.2.353}$$

Verify that the γ^μ do indeed satisfy the Dirac algebra (2.349). Note that they are all *purely imaginary.* Representations of the Dirac algebra that are purely imaginary are called *Majorana* (1906-1938) representations. Thus the Ansätze (2.350) through (2.353) provide a Majorana representation.[9]

Verify that for this Majorana representation there are the results

$$(\gamma^\mu)^\dagger = -\gamma^\mu \text{ for } \mu = 1, 2, 3 \tag{8.2.354}$$

and

$$(\gamma^4)^\dagger = \gamma^4. \tag{8.2.355}$$

Verify that, in view of (2.349), the results (2.354) and (2.355) are equivalent to the relations

$$(\gamma^\mu)^\dagger = \gamma^4\gamma^\mu\gamma^4 \text{ for } \mu = 1 \cdot\cdot 4. \tag{8.2.356}$$

Finally, verify that γ^4 can be written in the form

$$\gamma^4 = i \begin{pmatrix} \mathbf{0} & J_2 \\ J_2 & \mathbf{0} \end{pmatrix}. \tag{8.2.357}$$

Definition and some Properties of γ^5

It is useful to define a fifth 4×4 gamma matrix γ^5 in terms of the γ^μ by the rule

$$\gamma^5 = i\gamma^1\gamma^2\gamma^3\gamma^4. \tag{8.2.358}$$

Verify that it too is purely imaginary in a Majorana representation.[10] Show that, for the above Majorana representation given by (2.350) through (2.353), there is the result

$$\gamma^5 = i \begin{pmatrix} \mathbf{0} & -I^{[2]} \\ I^{[2]} & \mathbf{0} \end{pmatrix} = -iJ \tag{8.2.359}$$

where here J denotes the 4×4 version of (3.1.1),

$$J = \begin{pmatrix} \mathbf{0} & I^{[2]} \\ -I^{[2]} & \mathbf{0} \end{pmatrix}. \tag{8.2.360}$$

It follows that γ^5 is antisymmetric in our Majorana representation,

$$(\gamma^5)^T = -\gamma^5. \tag{8.2.361}$$

[9] There are several possible Majorana representations. As trivial examples of how one Majorana representation may be converted into another, the γ^μ or any subset of the γ^μ may be multiplied by -1. Also, the γ^μ for $\mu = 1 \cdot 3$ may be permuted among each other. See, in addition, Exercise 2.21.

[10] The notation γ^5 can be misleading. Whereas we will see that in some contexts the contravariant index ν in γ^ν for $\nu = 1 \cdot\cdot 4$ can be lowered using $g_{\mu\nu}$, there is in the Dirac machinery no analogous operation for γ^5. Thus, in the Dirac machinery, the 5 in γ^5 is simply a label and not a contravariant index.

The matrix γ^5 defined by (2.358) has some other properties that follow from the Dirac algebra relations (2.349) and therefore hold in in any representation of the γ^μ. Verify from (2.349) and (2.358) that

$$\{\gamma^\mu, \gamma^5\}_+ = 0, \tag{8.2.362}$$

$$\{\gamma^\mu\gamma^\nu, \gamma^5\} = 0, \tag{8.2.363}$$

and

$$(\gamma^5)^2 = I^{[4]}. \tag{8.2.364}$$

Definition of the $\hat{\sigma}^{\mu\nu}$ and their Relation to the $K(\hat{L}^j)$ and $K(\hat{N}^j)$

Also, for $\mu, \nu = 1 \cdot\cdot 4$, define matrices $\hat{\sigma}^{\mu\nu}$ by the *commutators*

$$\hat{\sigma}^{\mu\nu} = (1/2)\{\gamma^\mu, \gamma^\nu\} \Rightarrow \hat{\sigma}^{\mu\nu} = \gamma^\mu\gamma^\nu \text{ for } \mu \neq \nu. \tag{8.2.365}$$

[Here, in writing the second relation in (2.365), we have used the Dirac algebra.] Note that, because the γ^μ are purely imaginary in a Majorana representation, the $\hat{\sigma}^{\mu\nu}$ are purely real.[11] We will soon see that the $\hat{\sigma}^{\mu\nu}$ are related to the Lorentz group Lie generator matrices $K(\hat{L}^j)$ and $K(\hat{N}^j)$ and, since these K matrices are real, we would like the $\hat{\sigma}^{\mu\nu}$ to be real. That is the reason for our use of a Majorana representation for the γ^μ; and indeed it will transpire that we have chosen a particular Majorana representation to make things work out neatly.

Verify that there are the *six* results

$$\hat{\sigma}^{12} = \gamma^1\gamma^2 = \begin{pmatrix} \mathbf{0} & \sigma^3 \\ -\sigma^3 & \mathbf{0} \end{pmatrix}, \tag{8.2.366}$$

$$\hat{\sigma}^{23} = \gamma^2\gamma^3 = \begin{pmatrix} \mathbf{0} & \sigma^1 \\ -\sigma^1 & \mathbf{0} \end{pmatrix}, \tag{8.2.367}$$

$$\hat{\sigma}^{31} = \gamma^3\gamma^1 = \begin{pmatrix} -i\sigma^2 & \mathbf{0} \\ \mathbf{0} & -i\sigma^2 \end{pmatrix}, \tag{8.2.368}$$

$$\hat{\sigma}^{41} = \gamma^4\gamma^1 = \begin{pmatrix} \sigma^1 & \mathbf{0} \\ \mathbf{0} & \sigma^1 \end{pmatrix}, \tag{8.2.369}$$

$$\hat{\sigma}^{42} = \gamma^4\gamma^2 = \begin{pmatrix} \mathbf{0} & i\sigma^2 \\ -i\sigma^2 & \mathbf{0} \end{pmatrix}, \tag{8.2.370}$$

$$\hat{\sigma}^{43} = \gamma^4\gamma^3 = \begin{pmatrix} \sigma^3 & \mathbf{0} \\ \mathbf{0} & \sigma^3 \end{pmatrix}, \tag{8.2.371}$$

and that all other $\hat{\sigma}^{\mu\nu}$ results can be obtained from the antisymmetry relation

$$\hat{\sigma}^{\mu\nu} = -\hat{\sigma}^{\nu\mu}. \tag{8.2.372}$$

[11] Observe also that we depart from the usual Dirac literature definition/notation $\sigma^{\mu\nu} = (i/2)\{\gamma^\mu, \gamma^\nu\}$ by omitting a factor of i and placing a hat $\hat{}$ on σ to indicate that a change of notation has occurred. As described in the end of Exercise 3.7.43, we do this to avoid introducing mathematically unnecessary factors of i.

For comparison, recall the *six* results (2.332) through (2.337). Verify from (7.3.236) through (7.3.241), (2.332) through (2.337), and (2.366) through (2.371) that there are the relations

$$(1/2)\hat{\sigma}^{12} = (1/2)K(-i\sigma^3) = K(\hat{L}^3), \tag{8.2.373}$$

$$(1/2)\hat{\sigma}^{23} = (1/2)K(-i\sigma^1) = K(\hat{L}^1), \tag{8.2.374}$$

$$(1/2)\hat{\sigma}^{31} = (1/2)K(-i\sigma^2) = K(\hat{L}^2), \tag{8.2.375}$$

$$(1/2)\hat{\sigma}^{41} = (1/2)K(\sigma^1) = K(\hat{N}^1), \tag{8.2.376}$$

$$(1/2)\hat{\sigma}^{42} = (1/2)K(\sigma^2) = K(\hat{N}^2), \tag{8.2.377}$$

$$(1/2)\hat{\sigma}^{43} = (1/2)K(\sigma^3) = K(\hat{N}^3). \tag{8.2.378}$$

We see that, when the Majorana representation (2.350) through (2.353) is employed for the γ^μ, the matrices $(1/2)\hat{\sigma}^{\mu\nu}$ are simply a relabeling of the matrices $K(\hat{L}^j)$ and $K(\hat{N}^j)$.

It follows that, at least in this case, the matrices $(1/2)\hat{\sigma}^{\mu\nu}$ satisfy the Lorentz group Lie algebra commutation rules. For example, verify using (2.373) through (2.375), (7.3.385), and (7.3.243) that there is the commutation rule

$$\{(1/2)\hat{\sigma}^{12}, (1/2)\hat{\sigma}^{23}\} = \{K(\hat{L}^3), K(\hat{L}^1)\} = K(\{\hat{L}^3, \hat{L}^1\}) = K(\hat{L}^2) = (1/2)\hat{\sigma}^{31}. \tag{8.2.379}$$

Actually, the same commutation rules for the matrices $(1/2)\hat{\sigma}^{\mu\nu}$ hold for any representation choice for the γ^μ, and are in fact a consequence of the Dirac algebra (2.349). Consider again, for example, the commutator $\{(1/2)\hat{\sigma}^{12}, (1/2)\hat{\sigma}^{23}\}$. Verify, using (2.349) and (2.365), that

$$\begin{gathered}\{(1/2)\hat{\sigma}^{12}, (1/2)\hat{\sigma}^{23}\} = (1/4)\{\gamma^1\gamma^2, \gamma^2\gamma^3\} = (1/4)(\gamma^1\gamma^2\gamma^2\gamma^3 - \gamma^2\gamma^3\gamma^1\gamma^2) = \\ (1/4)(-\gamma^1\gamma^3 - \gamma^3\gamma^1\gamma^2\gamma^2) = (1/4)(\gamma^3\gamma^1 + \gamma^3\gamma^1) = (1/2)\gamma^3\gamma^1 = (1/2)\hat{\sigma}^{31},\end{gathered} \tag{8.2.380}$$

as expected. The matrices $(1/2)\hat{\sigma}^{\mu\nu}$ depend on the choice of representation for the γ^μ, but their commutation rules do not.

Additional Notation

At his point it is useful to introduce additional notation. Since the γ^μ (for $\mu = 1 \cdots 5$) are purely imaginary in a Majorana representation, we may write them in the form

$$\gamma^\mu = i\gamma_r^\mu \tag{8.2.381}$$

where the γ_r^μ are purely *real*. For example, from (2.353) we see that

$$\gamma_r^4 = \begin{pmatrix} \mathbf{0} & J_2 \\ J_2 & \mathbf{0} \end{pmatrix}, \tag{8.2.382}$$

and from (2.359) we see that

$$\gamma_r^5 = -J. \tag{8.2.383}$$

Verify that both γ_r^4 and γ_r^5 are antisymmetric in our Majorana representation,

$$(\gamma_r^4)^T = -\gamma_r^4 \text{ and } (\gamma_r^5)^T = -\gamma_r^5. \tag{8.2.384}$$

Lorentz Invariance of γ^5

We are now prepared to further explore some properties of Dirac matrices and 4-spinors. The first step has to do with γ^5 and consists of verifying that it is invariant under the action the Lorentz group. That is, there is the result

$$\{K(k), \gamma^5\} = 0 \Leftrightarrow K^{-1}\gamma^5 K = \gamma^5 \tag{8.2.385}$$

where k is any element in $SL(2, \mathbb{C})$. We will arrive at this result by several steps.

Begin by verifying from (2.363) and (2.365) that

$$\{\gamma^5, \hat{\sigma}^{\mu\nu}\} = 0. \tag{8.2.386}$$

At this point review Exercise 7.3.30. Use (2.316) and (7.3.246) to write

$$k(\lambda, \boldsymbol{m}; \theta, \boldsymbol{n}) = \exp(\lambda \boldsymbol{m} \cdot \hat{\boldsymbol{N}}) \exp(\theta \boldsymbol{n} \cdot \hat{\boldsymbol{L}}). \tag{8.2.387}$$

Verify it follows from (7.3.378) and (7.3.386) that

$$\begin{aligned} K(k) &= K[\exp(\lambda \boldsymbol{m} \cdot \hat{\boldsymbol{N}}) \exp(\theta \boldsymbol{n} \cdot \hat{\boldsymbol{L}})] \\ &= K[\exp(\lambda \boldsymbol{m} \cdot \hat{\boldsymbol{N}})] K[\exp(\theta \boldsymbol{n} \cdot \hat{\boldsymbol{L}})] \\ &= \exp[K(\lambda \boldsymbol{m} \cdot \hat{\boldsymbol{N}})] \exp[K(\theta \boldsymbol{n} \cdot \hat{\boldsymbol{L}})]. \end{aligned} \tag{8.2.388}$$

As a further step, verify it follows from (2.373) through (2.378) and (2.386) that

$$\{K(\hat{L}^j), \gamma^5\} = 0 \tag{8.2.389}$$

and

$$\{K(\hat{N}^j), \gamma^5\} = 0. \tag{8.2.390}$$

Use these results to show that

$$\{\exp[K(\theta \boldsymbol{n} \cdot \hat{\boldsymbol{L}})], \gamma^5\} = 0 \tag{8.2.391}$$

and

$$\{\exp[K(\lambda \boldsymbol{m} \cdot \hat{\boldsymbol{N}})], \gamma^5\} = 0; \tag{8.2.392}$$

and therefore it follows from (2.388) that (2.385) holds. Note that, in view of (2.381), there is the equivalent *real* matrix relation

$$\{K(k), \gamma_r^5\} = 0. \tag{8.2.393}$$

Properties of γ_r^4

Next, in preparation for future use, let us study various properties of γ_r^4. You have already verified that it is antisymmetric. Recall (2.384). Verify that it also satisfies the relation

$$(\gamma_r^4)^2 = -I^{[4]}. \tag{8.2.394}$$

Verify from (2.381) that

$$(\gamma^\mu)^\dagger = -i(\gamma_r^\mu)^T. \tag{8.2.395}$$

Insert this result into (2.356) to obtain the result

$$-i(\gamma_r^\mu)^T = (i)^3\gamma_r^4\gamma_r^\mu\gamma_r^4 \Leftrightarrow (\gamma_r^\mu)^T = \gamma_r^4\gamma_r^\mu\gamma_r^4. \tag{8.2.396}$$

Finally employ (2.394) in (2.396) to obtain (for $\mu = 1 \cdot\cdot 4$) the result

$$(\gamma_r^\mu)^T = -(\gamma_r^4)^{-1}\gamma_r^\mu\gamma_r^4 = -\gamma_r^4\gamma_r^\mu(\gamma_r^4)^{-1}. \tag{8.2.397}$$

The matrix γ_r^4 has one further, and remarkable, property with regard to $SL(2,\mathbb{C})$/Lorentz transformations. Namely, suppose k is any element in $SL(2,\mathbb{C})$. Then there is the result

$$K^T(k)\gamma_r^4K(k) = \gamma_r^4. \tag{8.2.398}$$

We/you will also prove this result in steps.

Begin by verifying the following relations:

$$\hat{\sigma}^{\mu\nu} = \gamma^\mu\gamma^\nu = (i)^2\gamma_r^\mu\gamma_r^\nu = -\gamma_r^\mu\gamma_r^\nu, \tag{8.2.399}$$

$$\begin{aligned}&\gamma_r^4\hat{\sigma}^{\mu\nu}(\gamma_r^4)^{-1} = -\gamma_r^4\gamma_r^\mu\gamma_r^\nu(\gamma_r^4)^{-1} = -\gamma_r^4\gamma_r^\mu(\gamma_r^4)^{-1}\gamma_r^4\gamma\nu_r(\gamma_r^4)^{-1} = \\ &-(\gamma_r^\mu)^T(\gamma_r^\nu)^T = -(\gamma_r^\nu\gamma_r^\mu)^T = (\gamma_r^\mu\gamma_r^\nu)^T = -(\hat{\sigma}^{\mu\nu})^T,\end{aligned} \tag{8.2.400}$$

$$(\hat{\sigma}^{\mu\nu})^T = -\gamma_r^4\hat{\sigma}^{\mu\nu}(\gamma_r^4)^{-1}. \tag{8.2.401}$$

Recall that the $\hat{\sigma}^{\mu\nu}$ are proportional to the $K(\hat{N}^j)$ and the $K(\hat{L}^j)$. See (2.373) through (2.378). Show from (2.401) that

$$[K(\hat{N}^j)]^T = -\gamma_r^4K(\hat{N}^j)(\gamma_r^4)^{-1} \tag{8.2.402}$$

and

$$[K(\hat{L}^j)]^T = -\gamma_r^4K(\hat{L}^j)(\gamma_r^4)^{-1}. \tag{8.2.403}$$

Let us pause for a moment. Note that (2.402) and (2.403) can be rewritten in the form

$$-[K(\hat{N}^j)]^T = \gamma_r^4K(\hat{N}^j)(\gamma_r^4)^{-1} \tag{8.2.404}$$

and

$$-[K(\hat{L}^j)]^T = \gamma_r^4K(\hat{L}^j)(\gamma_r^4)^{-1}. \tag{8.2.405}$$

Observe that the left side of (2.404) is the result of applying the tilde conjugacy operation to $K(\hat{N}^j)$, and the left side of (2.405) is the result of applying the tilde conjugacy operation to $K(\hat{L}^j)$. Recall (3.7.129). Consequently, the relations (2.404) and (2.405) show that the representation of $s\ell(2,\mathbb{C})$ provided by the $K(\hat{N}^j)$ and the $K(\hat{L}^j)$ is *self conjugate* (for the tilde conjugacy relation) under the similarity transformation provided by γ_r^4.

Recall (7.3.377). Continue on to verify it follows from (2.402) and (2.403) that

$$[K(\lambda\boldsymbol{m}\cdot\hat{\boldsymbol{N}})]^T = -\gamma_r^4K(\lambda\boldsymbol{m}\cdot\hat{\boldsymbol{N}})(\gamma_r^4)^{-1} \tag{8.2.406}$$

and

$$[K(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}})]^T = -\gamma_r^4K(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}})(\gamma_r^4)^{-1}. \tag{8.2.407}$$

Using (2.387), (2.406), and (2.407) verify that

$$\begin{aligned}
&K^T(k) = \{\exp[K(\lambda\boldsymbol{m}\cdot\hat{\boldsymbol{N}})]\exp[K(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}})]\}^T = \\
&\{\exp[K(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}})]\}^T\{\exp[K(\lambda\boldsymbol{m}\cdot\hat{\boldsymbol{N}})]\}^T = \\
&\exp\{[K(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}})]^T\}\exp\{[K(\lambda\boldsymbol{m}\cdot\hat{\boldsymbol{N}})]^T\} = \\
&\exp[-\gamma_r^4K(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}})(\gamma_r^4)^{-1}]\exp[-\gamma_r^4K(\lambda\boldsymbol{m}\cdot\hat{\boldsymbol{N}})(\gamma_r^4)^{-1}] = \\
&\gamma_r^4\exp[-K(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}})](\gamma_r^4)^{-1}\gamma_r^4\exp[-K(\lambda\boldsymbol{m}\cdot\hat{\boldsymbol{N}})](\gamma_r^4)^{-1} = \\
&\gamma_r^4\{\exp[K(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}})]\}^{-1}\{\exp[K(\lambda\boldsymbol{m}\cdot\hat{\boldsymbol{N}})]\}^{-1}(\gamma_r^4)^{-1} = \\
&\gamma_r^4\{\exp[K(\lambda\boldsymbol{m}\cdot\hat{\boldsymbol{N}})]\exp[K(\theta\boldsymbol{n}\cdot\hat{\boldsymbol{L}})]\}^{-1}(\gamma_r^4)^{-1} = \\
&\gamma_r^4[K(k)]^{-1}(\gamma_r^4)^{-1},
\end{aligned} \tag{8.2.408}$$

from which it follows that

$$K^T(k)\gamma_r^4 = \gamma_r^4K^{-1}(k) \tag{8.2.409}$$

and therefore

$$K^T(k)\gamma_r^4K(k) = \gamma_r^4. \tag{8.2.410}$$

Show, using (2.384), that there is also the relation

$$K^T(k)(\gamma_r^4)^TK(k) = (\gamma_r^4)^T. \tag{8.2.411}$$

Note that (2.408), upon comparing its beginning and end, can be written as

$$K^T(k) = \gamma_r^4[K(k)]^{-1}(\gamma_r^4)^{-1}. \tag{8.2.412}$$

This relation shows that, when the Majorana matrices (2.350) through (2.353) are employed, the matrices $K^T(k)$ and $K^{-1}(k)$ are related by the similarity transformation provided by γ_r^4. Persuade yourself, upon reflection, that this group element relation is a consequence of the Lie algebraic self conjugacy relations (2.404) and (2.405).

The relation (2.411) has an important consequence. Let M be *any* 4×4 matrix. Verify that

$$[\gamma_r^4K]^TMK = K^T(\gamma_r^4)^TMK = K^T(\gamma_r^4)^TKK^{-1}MK = (\gamma_r^4)^TK^{-1}MK. \tag{8.2.413}$$

This will prove to be a key result.

Conjugate 4-Spinors and Bilinear Forms

With γ_r^4 in mind, we are ready for further definitions/constructions. Suppose u is some real 4-spinor. Define a related *conjugate* 4-spinor $\bar{u}$ by the rule

$$\bar{u} = \gamma_r^4u. \tag{8.2.414}$$

Note, in view of (2.394), that this barring operation is what we might call an *anti-involution*,

$$\bar{\bar{u}} = (\gamma_r^4)^2u = -I^{[4]}u = -u. \tag{8.2.415}$$

We are going to be working with quantities of the form $(\bar{u}, v)$ where v is any other real 4-spinor and the usual real (no complex conjugate) scalar product is employed. Upon employing (2.414) and (2.384), verify that we may write

$$(\bar{u}, v) = (\gamma_r^4 u, v) = (u, [\gamma_r^4]^T v) = -(u, \gamma_r^4 v) = -(u, v)_{\gamma_r^4}. \tag{8.2.416}$$

Here we have introduced, for any matrix G, the definition

$$(u, v)_G = (u, Gv). \tag{8.2.417}$$

Quantities of the form (u, Gv) are called *bilinear forms*. Evidently, according to (2.416), the introduction of conjugate 4-spinors is equivalent (apart from a minus sign) to the use of the bilinear form associated with γ_r^4. Finally, suppose G is a *symmetric* matrix S or an *antisymmetric* matrix A. Verify that in these cases there are the relations

$$(u, v)_S = (v, u)_S \text{ and } (u, v)_A = -(v, u)_A. \tag{8.2.418}$$

Because of these symmetries under the interchange of the arguments u and v, the bilinear forms $(u, v)_S$ and $(u, v)_A$ are said to be *symmetric* and *antisymmetric*, respectively. Note that, by (2.384) and this definition, that the bilinear form $(u, v)_{\gamma_r^4}$ is antisymmetric.

Concept of Transformation Properties

With these definitions behind us, suppose $K(k)$ with $k \in SL(2, \mathbb{C})$ acts on u to produce a transformed spinor that we denote as $\breve{u}$,

$$\breve{u} = K(k)u. \tag{8.2.419}$$

We will also use the notation $\bar{\breve{u}}$ to denote the conjugate of $\breve{u}$. Therefore we may write

$$\bar{\breve{u}} = \gamma_r^4 K(k)u. \tag{8.2.420}$$

Finally, let v be any other real 4-spinor, and act on it to obtain the *transformed* 4-spinor

$$\breve{v} = K(k)v. \tag{8.2.421}$$

Here we again beg the reader's forgiveness for awkward notation. In the present context a bar $\bar{}$ simply denotes multiplication by γ_r^4 as in (2.414); and a breve $\breve{}$ is simply a distinguishing mark as in (2.419). Also, we observe that the term *conjugate* has many meanings/applications. Here we speak of conjugate spinors. In Exercise 3.7.36 we spoke of conjugate matrices and representations.

What we/you will soon explore are the *transformation properties* of $(\bar{\breve{u}}, M\breve{v})$ under the action of $K(k)$ where, as in (2.413), M is any 4×4 matrix. Verify, using the notation just introduced, the properties of K and γ_r^4, and (2.413) that

$$\begin{aligned}(\bar{\breve{u}}, M\breve{v}) = (\gamma_r^4 K u, MKv) = (u, [\gamma_r^4 K]^T MKv) = (u, (\gamma_r^4)^T K^{-1} MKv) = \\ (\gamma_r^4 u, K^{-1} MKv) = (\bar{u}, K^{-1} MKv).\end{aligned} \tag{8.2.422}$$

By looking at the beginning and end of (2.419) we see that it can be rewritten in the form

$$(\bar{\breve{u}}, M\breve{v}) = (\bar{u}, K^{-1} MKv). \tag{8.2.423}$$

This, like (2.413), is also a key result.

Transformation Properties of $I^{[4]}$ and γ^5

Let us consider various possibilities for the matrix M. The simplest possibility is $M = I^{[4]}$. In this case (2.423) becomes

$$(\bar{\breve{u}}, I^{[4]}\breve{v}) = (\bar{u}, K^{-1}I^{[4]}Kv) = (\bar{u}, I^{[4]}v) \Leftrightarrow (\bar{\breve{u}}, I^{[4]}\breve{v}) = (\bar{u}, I^{[4]}v) \text{ or } (\bar{\breve{u}}, \breve{v}) = (\bar{u}, v). \quad (8.2.424)$$

We may say that $I^{[4]}$ is *invariant* (behaves like a *scalar*) under the action of K. Suppose we also describe the result (2.424) in terms of the associated bilinear form $(u, v)_{\gamma_r^4}$. Verify that there is the relation

$$\begin{aligned}&(\breve{u}, I^{[4]}\breve{v})_{\gamma_r^4} = (Ku, Kv)_{\gamma_r^4} = (Ku, \gamma_r^4 Kv) = (u, K^T\gamma_r^4 Kv) = (u, \gamma_r^4 v) = (u, v)_{\gamma_r^4}\\ &\Leftrightarrow (\bar{\breve{u}}, I^{[4]}\breve{v}) = (\bar{u}, I^{[4]}v) \text{ or } (\bar{\breve{u}}, \breve{v}) = (\bar{u}, v) \text{ or } (Ku, Kv)_{\gamma_r^4} = (u, v)_{\gamma_r^4}. \qquad (8.2.425)\end{aligned}$$

Verify that

$$(\bar{u}, v) = (\gamma_r^4 u, v) = (u, [\gamma_r^4]^\dagger, v) = -(u, \gamma_r^4 v) = -(u, v)_{\gamma_r^4}. \qquad (8.2.426)$$

Out of curiosity, verify that

$$(u, v)_{\gamma_r^4} = -u_4v_1 + u_3v_2 - u_2v_3 + u_1v_4. \qquad (8.2.427)$$

Note that the associated bilinear form is manifestly antisymmetric, as expected from (2.384).

The next more complicated possibility is $M = \gamma_r^5$. In this case use of (2.423) and (2.385) gives the result

$$(\bar{\breve{u}}, \gamma_r^5\breve{v}) = (\bar{u}, K^{-1}\gamma_r^5 Kv) = (\bar{u}, \gamma_r^5 v) \qquad (8.2.428)$$

so we may say that γ_r^5 is also a scalar/invariant under the action of K. Verify that, in terms of bilinear forms, there is the associated result

$$\begin{aligned}&(\breve{u}, \breve{v})_{\gamma_r^4\gamma_r^5} = (\breve{u}, \gamma_r^4\gamma_r^5\breve{v}) = (Ku, \gamma_r^4\gamma_r^5 Kv) = (Ku, \gamma_r^4 K\gamma_r^5 v) =\\ &(u, K^T\gamma_r^4 K\gamma_r^5 v) = (u, \gamma_r^4\gamma_5^5 v) = (u, v)_{\gamma_r^4\gamma_r^5}. \qquad (8.2.429)\end{aligned}$$

Verify that

$$(\bar{u}, \gamma_r^5 v) = (\gamma_r^4 u, \gamma_r^5 v) = -(u, \gamma_r^4\gamma_r^5 v) = -(u, v)_{\gamma_r^4\gamma_r^5}. \qquad (8.2.430)$$

Verify using (2.362) and (2.384) that the associated bilinear form is antisymmetric. Out of curiosity verify that

$$(u, v)_{\gamma_r^4\gamma_r^5} = +u_1v_2 - u_2v_1 - u_3v_4 + u_4v_3, \qquad (8.2.431)$$

which is manifestly antisymmetric. We have learned that there are two invariant bilinear forms, namely (2.427) and (2.431), and both are antisymmetric.

Transformation Properties of the γ^μ

With these instructive but relatively simple observations behind us, let us consider other possibilities for the matrix M. Suppose we select the possibilities $M = \gamma^\mu$ for $\mu = 1 \cdot \cdot 4$. Your task in this case is to show that

$$(\bar{\breve{u}}, \gamma^\mu\breve{v}) = \sum_\nu \Lambda^{\mu\nu}(k)(\bar{u}, \gamma^\nu v). \qquad (8.2.432)$$

Note that, because of (2.381), we may also write the real matrix relations

$$(\breve{\bar{u}}, \gamma_r^\mu \breve{v}) = \sum_\nu \Lambda^{\mu\nu}(k)(\bar{u}, \gamma_r^\nu v). \tag{8.2.433}$$

In view of (2.432), it is commonly (although somewhat imprecisely) said that the γ^μ (equivalently, the γ_r^μ) behave like a four-vector under the action of $K(k)$ for $k \in SL(2,\mathbb{C})$ in the same spirit that $I^{[4]}$ and γ_r^5 are said to behave like scalars.[12] According to (2.423) there is the relation

$$(\breve{\bar{u}}, \gamma^\mu \breve{v}) = (\bar{u}, K^{-1}\gamma^\mu K v). \tag{8.2.434}$$

Show that (2.432) is established if there is the relation

$$K^{-1}(k)\gamma^\mu K(k) = \sum_\nu \Lambda^{\mu\nu}(k)\gamma^\nu. \tag{8.2.435}$$

Let us work on verifying (2.435). It helps to separate the problem into two parts. Make the decomposition

$$k = k_b k_r \tag{8.2.436}$$

where

$$k_b = \exp(\lambda \boldsymbol{m} \cdot \hat{\boldsymbol{N}}) \tag{8.2.437}$$

is the $SL(2,\mathbb{C})$ element for a *boost* and

$$k_r = \exp(\theta \boldsymbol{n} \cdot \hat{\boldsymbol{L}}) \tag{8.2.438}$$

is the $SL(2,\mathbb{C})$ element for a *rotation*. Then, by (7.3.378), verify that

$$K(k) = K(k_b k_r) = K(k_b)K(k_r). \tag{8.2.439}$$

Correspondingly, verify that

$$K^{-1}(k)\gamma^\alpha K(k) = K^{-1}(k_r)K^{-1}(k_b)\gamma^\alpha K(k_b)K(k_r). \tag{8.2.440}$$

Conjecture that (2.435) holds for *pure* boosts and for *pure* rotations,

$$K^{-1}(k_b)\gamma^\alpha K(k_b) \overset{?}{=} \sum_\beta \Lambda^{\alpha\beta}(k_b)\gamma^\beta \ \text{ and } \ K^{-1}(k_r)\gamma^\beta K(k_r) \overset{?}{=} \sum_\delta \Lambda^{\beta\delta}(k_r)\gamma^\delta. \tag{8.2.441}$$

Verify the desired result (2.435) then follows from (2.440), (2.441), and (2.219):

$$\begin{aligned} &K^{-1}(k)\gamma^\alpha K(k) = K^{-1}(k_r)K^{-1}(k_b)\gamma^\alpha K(k_b)K(k_r) = \\ &K^{-1}(k_r)\sum_\beta \Lambda^{\alpha\beta}(k_b)\gamma^\beta K(k_r) = \sum_\beta \Lambda^{\alpha\beta}(k_b)K^{-1}(k_r)\gamma^\beta K(k_r) = \\ &\sum_{\beta\delta} \Lambda^{\alpha\beta}(k_b)\Lambda^{\beta\delta}(k_r)\gamma^\delta = \sum_\delta \Lambda^{\alpha\delta}(k_b k_r)\gamma^\delta = \sum_\delta \Lambda^{\alpha\delta}(k)\gamma^\delta. \end{aligned} \tag{8.2.442}$$

[12] Since the γ^μ for $\mu = 1 \cdot \cdot 4$ behave like four-vectors, it makes sense to raise and lower their indices (if desired) using the metric tensor g. See, for example, (2.478).

What remains is to verify the conjectures (2.441).

For simplicity, first work on verifying the second conjecture in (2.441). Verify that

$$K(k_r) = K[\exp(\theta \boldsymbol{n} \cdot \hat{\boldsymbol{L}})) = \exp[\theta K(\boldsymbol{n} \cdot \hat{\boldsymbol{L}})]. \tag{8.2.443}$$

See (7.3.386). Next verify that combining (2.443) and the second conjecture in (2.441) yields the conjecture

$$\exp[-\theta K(\boldsymbol{n} \cdot \hat{\boldsymbol{L}})]\gamma^\beta \exp[\theta K(\boldsymbol{n} \cdot \hat{\boldsymbol{L}})] \overset{?}{=} \sum_\delta \Lambda^{\beta\delta}(k_r)\gamma^\delta. \tag{8.2.444}$$

At this point review Sections 8.1 and 8.2 that described the concept of an adjoint Lie operator, using the symbol #, in the context of Lie operators $: f :$. The same concept can be applied to matrices in terms of commutators. Let A and B be any two $n \times n$ matrices. In this context, define a matrix adjoint operator $\#A\#$ that acts on matrices B by the rule

$$\#A\#B = \{A, B\}. \tag{8.2.445}$$

Then, in complete analogy to the discussion in Sections 8.1 and 8.2, there is the relation

$$\exp(-A)B\exp(A) = \exp(-\#A\#)B, \tag{8.2.446}$$

which may be thought of as the matrix version of Hadamard's lemma. See also (27.11.20) through (27.11.24) in Section 27.11 where this result is again used.

Verify that applying the matrix adjoint operator machinery just described produces, in the present context, the relation

$$\exp[-\theta K(\boldsymbol{n} \cdot \hat{\boldsymbol{L}})]\gamma^\beta \exp[\theta K(\boldsymbol{n} \cdot \hat{\boldsymbol{L}})] = \exp[-\theta \#K(\boldsymbol{n} \cdot \hat{\boldsymbol{L}})\#]\gamma^\beta. \tag{8.2.447}$$

Consequently, employing (2.447) in (2.444) produces the conjecture

$$\exp[-\theta \#K(\boldsymbol{n} \cdot \hat{\boldsymbol{L}})\#]\gamma^\beta \overset{?}{=} \sum_\delta \Lambda^{\beta\delta}[\exp(\theta \boldsymbol{n} \cdot \hat{\boldsymbol{L}})]\gamma^\delta. \tag{8.2.448}$$

Verify that employing (2.269) in (2.448) produces the logically equivalent conjecture

$$\exp[-\theta \#K(\boldsymbol{n} \cdot \hat{\boldsymbol{L}})\#]\gamma^\beta \overset{?}{=} \sum_\delta [\exp(\theta \boldsymbol{n} \cdot \boldsymbol{L})]^{\beta\delta}\gamma^\delta, \tag{8.2.449}$$

which, in turn, produces the logically equivalent conjectures

$$-\theta\{K(\boldsymbol{n} \cdot \hat{\boldsymbol{L}}), \gamma^\beta\} \overset{?}{=} \theta \sum_\delta [\boldsymbol{n} \cdot \boldsymbol{L})]^{\beta\delta}\gamma^\delta, \tag{8.2.450}$$

$$-\{K(\boldsymbol{n} \cdot \hat{\boldsymbol{L}}), \gamma^\beta\} \overset{?}{=} \sum_\delta [\boldsymbol{n} \cdot \boldsymbol{L})]^{\beta\delta}\gamma^\delta. \tag{8.2.451}$$

Let us seek to verify (2.451) for the three specific cases $\boldsymbol{n} = \boldsymbol{e}_1$, $\boldsymbol{n} = \boldsymbol{e}_2$, and $\boldsymbol{n} = \boldsymbol{e}_3$. Verify that if we can do so, then (2.451) follows by linearity. Consider, for example, the case $\boldsymbol{n} = \boldsymbol{e}_3$ so that the conjecture (2.451) becomes the conjecture

$$-\{K(\hat{L}^3), \gamma^\beta\} \overset{?}{=} \sum_\delta [L^3]^{\beta\delta}\gamma^\delta. \tag{8.2.452}$$

Verify that using (2.373) and (2.366) in (2.452) produces the logically equivalent conjectures

$$-(1/2)\{\hat{\sigma}^{12}), \gamma^\beta\} \overset{?}{=} \sum_\delta [L^3]^{\beta\delta}\gamma^\delta, \tag{8.2.453}$$

$$-(1/2)\{\gamma^1\gamma^2, \gamma^\beta\} \overset{?}{=} \sum_\delta [L^3]^{\beta\delta}\gamma^\delta. \tag{8.2.454}$$

Using the Dirac algebra, verify the following commutation relations:

$$-(1/2)\{\gamma^1\gamma^2, \gamma^1\} = -\gamma^2, \tag{8.2.455}$$

$$-(1/2)\{\gamma^1\gamma^2, \gamma^2\} = \gamma^1, \tag{8.2.456}$$

$$-(1/2)\{\gamma^1\gamma^2, \gamma^3\} = 0, \tag{8.2.457}$$

$$-(1/2)\{\gamma^1\gamma^2, \gamma^4\} = 0. \tag{8.2.458}$$

Recall from (7.3.182) that there is the result

$$L^3 = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{8.2.459}$$

Verify, upon comparing (2.455) through (2.459), that the conjecture (2.454) is correct, and therefore the conjecture (2.452) is correct. Similarly, verify that (2.451) also holds for $\boldsymbol{n} = \boldsymbol{e}_1$ and $\boldsymbol{n} = \boldsymbol{e}_2$, and therefore holds for all $\boldsymbol{n}$. Finally, proceed backwards through the chain of logically equivalent conjectures to verify that the initial conjecture, the second conjecture in (2.441), is indeed correct.

Now work on verifying the first conjecture in (2.441). The procedure for doing so is analogous to that just used to verify the second conjecture. Carry out the verification using the relations and conjectures below:

$$K(k_b) = K[\exp(\lambda \boldsymbol{m} \cdot \hat{\boldsymbol{N}})) = \exp[\lambda K(\boldsymbol{m} \cdot \hat{\boldsymbol{N}})], \tag{8.2.460}$$

$$\exp[-\lambda K(\boldsymbol{m} \cdot \hat{\boldsymbol{N}})]\gamma^\beta \exp[\lambda K(\boldsymbol{m} \cdot \hat{\boldsymbol{N}})] \overset{?}{=} \sum_\delta \Lambda^{\beta\delta}(k_b)\gamma^\delta, \tag{8.2.461}$$

$$\exp[-\lambda K(\boldsymbol{m} \cdot \hat{\boldsymbol{N}})]\gamma^\beta \exp[\lambda K(\boldsymbol{m} \cdot \hat{\boldsymbol{N}})] = \exp[-\lambda \#K(\boldsymbol{m} \cdot \hat{\boldsymbol{N}})\#]\gamma^\beta, \tag{8.2.462}$$

$$\exp[-\lambda \#K(\boldsymbol{m} \cdot \hat{\boldsymbol{N}})\#]\gamma^\beta \overset{?}{=} \sum_\delta \Lambda^{\beta\delta}[\exp(\lambda \boldsymbol{m} \cdot \hat{\boldsymbol{N}})]\gamma^\delta, \tag{8.2.463}$$

$$\exp[-\lambda \#K(\boldsymbol{m} \cdot \hat{\boldsymbol{N}})\#]\gamma^\beta \overset{?}{=} \sum_\delta [\exp(\lambda \boldsymbol{m} \cdot \boldsymbol{N})]^{\beta\delta}\gamma^\delta, \tag{8.2.464}$$

$$-\lambda\{K(\boldsymbol{m} \cdot \hat{\boldsymbol{N}}), \gamma^\beta\} \overset{?}{=} \lambda \sum_\delta [\boldsymbol{m} \cdot \boldsymbol{N})]^{\beta\delta}\gamma^\delta, \tag{8.2.465}$$

$$-\{K(\boldsymbol{m} \cdot \hat{\boldsymbol{N}}), \gamma^\beta\} \overset{?}{=} \sum_\delta [\boldsymbol{m} \cdot \boldsymbol{N})]^{\beta\delta}\gamma^\delta, \tag{8.2.466}$$

$$-\{K(\hat{N}^3),\gamma^\beta\} \stackrel{?}{=} \sum_\delta [N^3]^{\beta\delta}\gamma^\delta, \tag{8.2.467}$$

$$-(1/2)\{\hat{\sigma}^{43}),\gamma^\beta\} \stackrel{?}{=} \sum_\delta [N^3]^{\beta\delta}\gamma^\delta, \tag{8.2.468}$$

$$-(1/2)\{\gamma^4\gamma^3,\gamma^\beta\} \stackrel{?}{=} \sum_\delta [N^3]^{\beta\delta}\gamma^\delta, \tag{8.2.469}$$

$$-(1/2)\{\gamma^4\gamma^3,\gamma^1\} = 0, \tag{8.2.470}$$

$$-(1/2)\{\gamma^4\gamma^3,\gamma^2\} = 0, \tag{8.2.471}$$

$$-(1/2)\{\gamma^4\gamma^3,\gamma^3\} = \gamma^4, \tag{8.2.472}$$

$$-(1/2)\{\gamma^4\gamma^3,\gamma^4\} = \gamma^3, \tag{8.2.473}$$

$$N^3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{8.2.474}$$

Let us pause for a moment to reflect on what has been discovered/accomplished here. We were able to contemplate/determine the action of the $K(k)$ on the γ^μ by the fact that both $K(k)$ and the γ^μ were known 4×4 matrices so that the quantities $K^{-1}\gamma^\mu K$ could be evaluated. Equivalently, because the γ^μ obeyed the Dirac algebra, we were able to use this algebra to determine the action of the $K(\hat{L}^j)$ and the $K(\hat{N}^j)$ on the γ^β as in (2.452) through (2.458) and (2.467) through (2.473).

Transformation Properties of Products of gamma matrices

Consider the case of a product of two *distinct* gamma matrices, $M = \gamma^\mu\gamma^\nu$ with $\mu \neq \mu$. Verify that

$$K^{-1}\gamma^\mu\gamma^\nu K = K^{-1}\gamma^\mu K K^{-1}\gamma^\nu K. \tag{8.2.475}$$

Now use (2.435) in (2.475) to obtain the result

$$K^{-1}\gamma^\mu\gamma^\nu K = \sum_{\alpha\beta} \Lambda^{\mu\alpha}(k)\Lambda^{\nu\beta}(k)\gamma^\alpha\gamma^\beta. \tag{8.2.476}$$

Evidently $\gamma^\mu\gamma^\nu$ transforms like a second rank tensor. Note, according to (2.365), that $\hat{\sigma}^{\mu\nu}$ involves the commutator $\{\gamma^\mu,\gamma^\nu\}$. Show it follows that

$$K^{-1}\hat{\sigma}^{\mu\nu} K = \sum_{\alpha\beta} \Lambda^{\mu\alpha}(k)\Lambda^{\nu\beta}(k)\hat{\sigma}^{\alpha\beta}, \tag{8.2.477}$$

the $\hat{\sigma}^{\mu\nu}$ transform like a second rank antisymmetric tensor.

You have determined the transformation properties of the commutator $\{\gamma^\mu,\gamma^\nu\}$. What can be said about the transformation properties of the anticommutator $\{\gamma^\mu,\gamma^\nu\}_+$? Evaluate $K^{-1}\{\gamma^\mu,\gamma^\nu\}_+K$ using (2.349). Also evaluate it using (2.476), and then simplify your result using (2.349). Verify, with the aid of (6.2.51), that (2.349) and (2.476) are compatible.

Using the Dirac algebra, verify that

$$\sum_{\nu} \gamma_\nu \gamma^\nu = \sum_{\mu\nu} g_{\mu\nu} \gamma^\mu \gamma^\nu = 4I^{[4]}, \tag{8.2.478}$$

and observe from the far right side of (2.478) that $\sum_\nu \gamma_\nu \gamma^\nu$ is invariant, as the notation suggests.

Consider the case of three gamma matrices. Verify, using the Dirac algebra, that the product of any three distinct gamma matrices can be written in the form $\gamma^5\gamma^\mu$. Verify that

$$K^{-1}\gamma^5\gamma^\mu K = \gamma^5 K^{-1}\gamma^\mu K = \sum_\nu \Lambda^{\mu\nu}(k)\gamma^5\gamma^\nu. \tag{8.2.479}$$

Evidently the quantities $\gamma^5\gamma^\mu$, like the quantities γ^μ, also behave as four-vectors.

Consider the case of four gamma matrices. Verify that the product of any four distinct gamma matrices must be proportional to γ^5, whose transformation properties are given by (2.385). Even a bit more can be said. From (2.435), and an obvious extension of (2.476) to the case of four gamma matrices, verify that the gamma matrix products $\gamma^\mu\gamma^\nu\gamma^\sigma\gamma^\tau$ transform like a rank four tensor. Verify, moreover, that there is the relation

$$\gamma^5 = (i/4!) \sum_{\mu\nu\sigma\tau} \epsilon_{\mu\nu\sigma\tau} \gamma^\mu\gamma^\nu\gamma^\sigma\gamma^\tau. \tag{8.2.480}$$

Observe that the invariance of γ^5, already established, is consistent with the invariant appearance of this relation.

Consider the case of five or more gamma matrices. Verify that any product of five or more gamma matrices may be reduced to the product of four or less gamma matrices using the Dirac algebra, and the transformation properties of these products have already been determined.

Summary of Results and Relation to the Clebsch-Gordan Series

In summary, we have found the following: two "scalars", $I^{[4]}$ and γ^5; two "four-vectors", γ^μ and $\gamma^5\gamma^\mu$; and one "antisymmetric tensor", $\hat{\sigma}^{\mu\nu}$. Why should this be? Look again at the bilinear form $(\bar{\breve{u}}, M\breve{v})$ that appears on the left side of (2.423). It contains the two 4-spinors $\breve{u}$ and $\breve{v}$, each of which carries the representation $\Gamma(0,1/2) \oplus \Gamma(1/2,0)$, and a *fixed* matrix M. Evidently the quantities $(\bar{\breve{u}}, M\breve{v})$ are essentially tensor products and may be expected to contain whatever representations of $s\ell(2,\mathbb{C})$ occur in the tensor product of $\Gamma(0,1/2) \oplus \Gamma(1/2,0)$ with itself. Let us see what representations can be found in this tensor product. Verify the tensor product result

$$\begin{aligned}&[\Gamma(0,1/2) \oplus \Gamma(1/2,0)] \times [\Gamma(0,1/2) \oplus \Gamma(1/2,0)] = \\ &[\Gamma(0,1/2) \times \Gamma(0,1/2)] \oplus [\Gamma(0,1/2) \times \Gamma(1/2,0)] \oplus \\ &[\Gamma(1/2,0) \times \Gamma(0,1/2)] \oplus [\Gamma(1/2,0) \times \Gamma(1/2,0)].\end{aligned} \tag{8.2.481}$$

Next verify the Clebsch-Gordan results

$$\Gamma(0,1/2) \times \Gamma(0,1/2) = \Gamma(0,1) \oplus \Gamma(0,0), \tag{8.2.482}$$

$$\Gamma(0,1/2) \times \Gamma(1/2,0) = \Gamma(1/2,1/2), \tag{8.2.483}$$

$$\Gamma(1/2,0) \times \Gamma(0,1/2) = \Gamma(1/2,1/2), \tag{8.2.484}$$

$$\Gamma(1/2,0) \times \Gamma(1/2,0) = \Gamma(1,0) \oplus \Gamma(0,0). \tag{8.2.485}$$

Show it follows that there is the grand Clebsch-Gordan result

$$\begin{aligned}&[\Gamma(0,1/2) \oplus \Gamma(1/2,0)] \times [\Gamma(0,1/2) \oplus \Gamma(1/2,0)] = \\ &2\Gamma(0,0) \oplus 2\Gamma(1/2,1/2) \oplus [\Gamma(0,1) \oplus \Gamma(1,0)],\end{aligned} \tag{8.2.486}$$

which corresponds to the two scalars, two four-vectors, and one antisymmetric tensor, as described above.

Group-Theoretical Nature of Dirac Gamma Matrices

You have verified that the four matrices γ^μ given by (2.350) through (2.353) satisfy the Dirac algebra (2.349). Moreover, in view of (2.365) and (2.373) through (2.378), the γ^μ are related to $s\ell(2,\mathbb{C})$ in that they have the further remarkable property that they *factorize* the $K(\hat{L}^j)$ and $K(\hat{N}^j)$. That is, each $K(\hat{L}^j)$ and $K(\hat{N}^j)$ can be written as the product of two gamma matrices.

But do the gamma matrices have additional group-theoretical significance? Let p be a four-vector with covariant components p_β and suppose a Lorentz group element associated with the $SL(2,\mathbb{C})$ element k acts on p to produce the four-vector $\breve{p}$. Verify that, according to (1.6.286), (1.6.287), and (1.6.289), the four-vector $\breve{p}$ will have covariant components $\breve{p}_\alpha$ given by the relation

$$\breve{p}_\alpha = \sum_\beta \{[\Lambda^T(k)]^{-1}\}^{\alpha\beta} p_\beta = \sum_\beta \{[\Lambda^{-1}(k)]^T\}^{\alpha\beta} p_\beta = \sum_\beta [\Lambda^{-1}(k)]^{\beta\alpha} p_\beta. \tag{8.2.487}$$

[Here we must again apologize for confusing notation: The matrix K defined by (1.6.287) is *not* the matrix $K(k)$ that appears in (7.3.375).] Now consider the matrix, call it $C(k)$, given by the rule

$$C(k) = \sum_\alpha \breve{p}_\alpha K^{-1}(k)\gamma^\alpha K(k) = K^{-1}(k)[\sum_\alpha \breve{p}_\alpha \gamma^\alpha]K(k). \tag{8.2.488}$$

According to (2.435) there is the relation

$$K^{-1}(k)\gamma^\alpha K(k) = \sum_\delta \Lambda^{\alpha\delta}(k)\gamma^\delta. \tag{8.2.489}$$

Now employ (2.487) and then (2.489) in (2.488) to show that

$$\begin{aligned} C(k) &= K^{-1}(k)[\sum_\alpha \breve{p}_\alpha \gamma^\alpha]K(k) = K^{-1}\{\sum_{\alpha\beta} p_\beta \gamma^\alpha [\Lambda^{-1}(k)]^{\beta\alpha}\}K(k) \\ &= \sum_{\alpha\beta\delta} p_\beta \gamma^\delta [\Lambda^{-1}(k)]^{\beta\alpha}\Lambda(k)^{\alpha\delta} = \sum_{\beta\delta} p_\beta \gamma^\delta [\Lambda^{-1}(k)\Lambda(k)]^{\beta\delta} \\ &= \sum_{\beta\delta} p_\beta \gamma^\delta \{I^{[4]}\}^{\beta\delta} = \sum_\delta p_\delta \gamma^\delta. \end{aligned} \tag{8.2.490}$$

You have shown that the matrix $C(k)$ is, in fact, *independent* of k.

Now suppose that p is the momentum four-vector for a particle that has mass m and is at rest,

$$p_\beta = (0, 0, 0, mc). \tag{8.2.491}$$

In this case there is the relation

$$\sum_\delta p_\delta \gamma^\delta = mc\gamma^4 \tag{8.2.492}$$

so that

$$C(k) = mc\gamma^4. \tag{8.2.493}$$

Next verify from (2.355) that γ^4 must be diagonalizable and have real eigenvalues. And verify from the Dirac algebra condition $(\gamma^4)^2 = I^{[4]}$ that its eigenvalues, call them τ, must satisfy $\tau^2 = 1$. Let v be a 4-spinor which is an eigenvector of γ^4 with eigenvalue τ,

$$\gamma^4 v = \tau v. \tag{8.2.494}$$

We will determine such eigenvectors shortly. They will turn out to be complex, but that need not concern us. Meanwhile define, as before, $\breve{v}$ by the rule

$$\breve{v} = K(k)v. \tag{8.2.495}$$

With these definitions in mind, verify the relations

$$mc\gamma^4 v = mc\tau v \tag{8.2.496}$$

and

$$C(k)v = K^{-1}(k)[\sum_\alpha \breve{p}_\alpha \gamma^\alpha]\breve{v} = mc\gamma^4 v = \tau mcv. \tag{8.2.497}$$

Show it follows from (2.495) and (2.497) that

$$[\sum_\alpha \breve{p}_\alpha \gamma^\alpha]\breve{v} = \tau m K(k)v = \tau mc\breve{v}. \tag{8.2.498}$$

Let us rewrite (2.498) in the form

$$[\sum_\alpha \gamma^\alpha \breve{p}_\alpha]\breve{v} = \tau mc\breve{v}. \tag{8.2.499}$$

Observe that on the left side of (2.499) there are the k dependent quantities $\breve{p}_\alpha$ and $\breve{v}$ that carry the $\Gamma(1/2, 1/2)$ and $[\Gamma(0, 1/2) \oplus \Gamma(1/2, 0)]$ representations of $s\ell(2, \mathbb{C})$, respectively. And on the right side of (2.499) we find just $\breve{v}$ which carries the $[\Gamma(0, 1/2) \oplus \Gamma(1/2, 0)]$ representation of $s\ell(2, \mathbb{C})$. Also observe that matrices γ^α are *three*-index quantities because there is the superscript α and the two matrix indices for each gamma matrix. Thus, from a group-theoretical perspective, the three-index quantities γ^α are *Clebsch-Gordan coefficients* that couple the representations $\Gamma(1/2, 1/2)$ and $[\Gamma(0, 1/2) \oplus \Gamma(1/2, 0)]$ to produce the representation $[\Gamma(0, 1/2) \oplus \Gamma(1/2, 0)]$.[13]

[13] Also, from the same group-theoretical perspective, the matrices $M = I^{[4]}$, $M = \gamma^5$, $M = \gamma^\mu$, $M = \gamma^5\gamma^\mu$, and $M = \hat{\sigma}^{\mu\nu}$ are Clebsch-Gordan coefficients that project out from the tensor product appearing on the left side of (2.481) the representations $\Gamma(0, 0)$, $\Gamma(1/2, 1/2)$, and $[\Gamma(0, 1) \oplus \Gamma(1, 0)]$.

Let us check that this possibility/conclusion makes group-theoretical sense. Verify the tensor product result

$$\begin{aligned}&\Gamma(1/2,1/2)\times[\Gamma(0,1/2)\oplus\Gamma(1/2,0)] = \\ &\Gamma(1/2,1/2)\times\Gamma(0,1/2)\oplus\Gamma(1/2,1/2)\times\Gamma(1/2,0).\end{aligned} \tag{8.2.500}$$

Verify the Clebsch-Gordan results

$$\Gamma(1/2,1/2)\times\Gamma(0,1/2)=\Gamma(1/2,0)\oplus\Gamma(1/2,1), \tag{8.2.501}$$

$$\Gamma(1/2,1/2)\times\Gamma(1/2,0)=\Gamma(0,1/2)\oplus\Gamma(1,1/2). \tag{8.2.502}$$

Verify, therefore, that there is the grand Clebsch-Gordan result

$$\begin{aligned}&\Gamma(1/2,1/2)\times[\Gamma(0,1/2)\oplus\Gamma(1/2,0)] = \\ &\Gamma(1/2,0)\oplus\Gamma(1/2,1)\oplus\Gamma(0,1/2)\oplus\Gamma(1,1/2) = \\ &[\Gamma(0,1/2)\oplus\Gamma(1/2,0)]\oplus\Gamma(1/2,1)\oplus\Gamma(1,1/2).\end{aligned} \tag{8.2.503}$$

Evidently the role of the γ^α is to project out, from the tensor product with which (2.503) begins, the representation that appears in square brackets in the last line of (2.503).

Application to the Dirac Equation

The Dirac equation for a *free* particle of mass m reads

$$i\hbar\sum_\mu \gamma^\mu\partial_\mu\psi(x) = mc\psi(x) \tag{8.2.504}$$

where $\psi(x)$ is a 4-spinor *field* that depends on the space-time coordinates x. Dirac's equation is often written in natural units (for which $\hbar = c = 1$) and the summation convention is employed so that it takes the elegant form

$$i\gamma^\mu\partial_\mu\psi = m\psi. \tag{8.2.505}$$

One may also employ the *Feynman* (1918-1988) *slash* notation

$$\partial\!\!\!/ = \sum_\mu \gamma^\mu\partial_\mu \tag{8.2.506}$$

to achieve the even more elegant form

$$i\partial\!\!\!/\psi = m\psi. \tag{8.2.507}$$

Dirac's memorial stone, near Newton's monument in Westminster Abbey, displays his equation in the form

$$i\gamma\cdot\partial\psi = m\psi. \tag{8.2.508}$$

To see this stone, Google the two words Dirac Westminster.

For pedagogical reasons we will use Dirac's equation in the form (2.504) in order to keep track of dimensions. Note that here again it is evident that the gamma matrices play the role of Clebsch-Gordan coefficients. On the left side of (2.504) we have ∂_μ and ψ which carry the representations $\Gamma(1/2,1/2)$ and $\Gamma(0,1/2)\oplus\Gamma(1/2,0)$, respectively. The γ^μ couple them to produce the representation $\Gamma(0,1/2)\oplus\Gamma(1/2,0)$, which is also the representation carried by the ψ appearing on the right side of (2.504). Thus, Dirac's equation is group-theoretically consistent.

Let us seek a plane-wave solution to (2.504) of the form

$$\psi(x) = w\exp[-i(\sum_\mu p_\mu x^\mu)/\hbar]. \tag{8.2.509}$$

As done before in (2.491), we will take p_μ to be specified by the relation

$$p_\mu = (0,0,0,mc), \tag{8.2.510}$$

and we will assume w is a 4-spinor that is independent of x. This ψ has no *spatial* dependence (ψ is translationally invariant) which implies that this proposed Ansatz is intended to describe a free-particle at rest.[14] With regard to temporal dependence, show that

$$\partial_4\psi = (\partial/\partial x^4)\psi = -i(mc/\hbar)w\exp[-i(\sum_\mu p_\mu x^\mu)/\hbar]. \tag{8.2.511}$$

Therefore, in this case, verify that the left side of (2.504) becomes

$$i\hbar\sum_\mu \gamma^\mu\partial_\mu\psi(x) = mc\gamma^4\psi(x) \tag{8.2.512}$$

so that (2.504) becomes

$$mc\gamma^4\psi(x) = mc\psi(x). \tag{8.2.513}$$

Verify that canceling out common factors from both sides of (2.513) yields for w the relation

$$\gamma^4 w = w. \tag{8.2.514}$$

That is, w must be an eigenvector of γ^4 with eigenvalue +1. We will soon see that γ^4 has eigenvectors with eigenvalues ±1 so that (2.494) has solutions for $\tau = \pm 1$. Note that (2.514) is consistent with (2.494).

Let us pause/digress briefly to discuss some commonly employed terminology for solutions of the Dirac equation. Verify, using (2.510) and (1.6.42), that the the Ansatz (2.509) can be rewritten in the form

$$\psi = w\exp[-i\mathcal{E}_0 t/\hbar] \tag{8.2.515}$$

where $\mathcal{E}_0$ is the (manifestly positive) rest energy,

$$\mathcal{E}_0 = mc^2. \tag{8.2.516}$$

[14] We remark that once "at rest" solutions have been found, all other free-particle solutions can be found by acting on the at rest solutions with suitable Lorentz transformations.

Despite the appearance of a *minus* sign in the exponent appearing in (2.515), and also in (2.509), this proposed solution is called a *positive energy* solution. The reason for this terminology has to do with the analogous case of the nonrelativistic *Schrödinger* (1887-1961) equation. The Schrödinger equation for the Schrödinger wave function, call it $\chi(\boldsymbol{r},t)$, reads

$$i\hbar(\partial/\partial t)\chi(\boldsymbol{r},t) = H(q,p)\chi(\boldsymbol{r},t) \tag{8.2.517}$$

where H is the (assumed time independent) Hamiltonian. If we make the separation of variables Ansatz

$$\chi(\boldsymbol{r},t) = u(\boldsymbol{r})f(t) \tag{8.2.518}$$

and specify that u is an eigenfunction of H with eigenvalue E so that

$$Hu = Eu, \tag{8.2.519}$$

then (2.517) has the solution

$$\chi(\boldsymbol{r},t) = u(\boldsymbol{r})\exp[-iEt/\hbar]. \tag{8.2.520}$$

Evidently the argument of the exponential function appearing in (2.520) is *negative* imaginary as the time t becomes evermore positive providing the energy E is *positive*, and vice versa. Observe that the argument of the exponential function appearing in (2.509) eventually becomes negative imaginary as t becomes evermore positive. Correspondingly, (2.509) is called a positive energy solution of the Dirac equation.

Now return to the main discussion. As promised, let us find the eigenvectors of γ^4. Begin by writing w in the form

$$w = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \tag{8.2.521}$$

where the quantities a through d are to be determined from (2.514). Verify using (2.353) that

$$\gamma^4 w = \begin{pmatrix} id \\ -ic \\ ib \\ -ia \end{pmatrix}, \tag{8.2.522}$$

and consequently the condition (2.514) yields the relations

$$d = -ia \tag{8.2.523}$$

$$c = ib. \tag{8.2.524}$$

Therefore any eigenvector of γ^4 having eigenvalue $+1$ must be of the form

$$w = \begin{pmatrix} a \\ b \\ ib \\ -ia \end{pmatrix}. \tag{8.2.525}$$

Let us pause at this point to seek eigenvectors of γ^4 with eigenvalue -1. Take the complex conjugate of both sides of (2.514) to find the relation

$$(\gamma^4)^* w^* = (\gamma^4 w)^* = w^*. \tag{8.2.526}$$

Since γ^4 is purely imaginary (in a Majorana representation), there is the relation

$$(\gamma^4)^* = -\gamma^4, \tag{8.2.527}$$

from which it follows that

$$\gamma^4 w^* = -w^*. \tag{8.2.528}$$

Therefore the eigenvectors of γ^4 with eigenvalue -1 are of the form

$$w^* = \begin{pmatrix} a^* \\ b^* \\ -ib^* \\ ia^* \end{pmatrix}. \tag{8.2.529}$$

Let us continue on. Evidently there is a two-fold degeneracy for both the eigenvalue $+1$ and -1 eigenvectors. To break this degeneracy it is convenient to employ the matrix M^3 defined by the rule

$$M^3 = i\hat{\sigma}^{12}. \tag{8.2.530}$$

Verify, using the Dirac algebra and (2.349), that there are the relations

$$(M^3)^\dagger = M^3 \tag{8.2.531}$$

and

$$(M^3)^2 = I^{[4]}. \tag{8.2.532}$$

It follows that M^3 can be diagonalized, has real eigenvalues, and the square of these eigenvalues is $+1$. Verify, again using the Dirac algebra, that

$$\{M^3, \gamma^4\} = 0, \tag{8.2.533}$$

which shows that γ^4 and M^3 can be diagonalized simultaneously.

Let us find the simultaneous eigenvectors of γ^4 and M^3. For w of the form (2.521) show, using (2.366), that

$$M^3 w = \begin{pmatrix} ic \\ -id \\ -ia \\ ib \end{pmatrix}. \tag{8.2.534}$$

Therefore, if we demand that

$$M^3 w = w, \tag{8.2.535}$$

verify that there are the relations

$$c = -ia \tag{8.2.536}$$

and

$$d = ib. \tag{8.2.537}$$

Suppose we now seek to satisfy all the demands (2.523), (2.524), (2.536), and (2.537). Verify that combining (2.523) and (2.537) yields the relation

$$b = -a. \tag{8.2.538}$$

Show that inserting (2.538) into (2.525) yields the vector, which we will call $w^{+\uparrow}$, given by the relation

$$w^{+\uparrow} = \begin{pmatrix} a \\ -a \\ -ia \\ -ia \end{pmatrix}. \tag{8.2.539}$$

(Here, in anticipation of subsequent results, we introduce the symbols ↑ and ↓ to indicate what will soon be identified as having spin up and spin down.) Verify that $w^{+\uparrow}$ is a simultaneous eigenvector of both γ^4 and M^3 with eigenvalues $+1$ for each.

Alternatively, if we demand that

$$M^3 w = -w, \tag{8.2.540}$$

verify that there are the relations

$$c = ia \tag{8.2.541}$$

and

$$d = -ib. \tag{8.2.542}$$

Verify that combining (2.523) and (2.542) in this case now yields the relation

$$b = a. \tag{8.2.543}$$

Show that inserting (2.543) into (2.525) yields the vector, which we will call $w^{+\downarrow}$, given by the relation

$$w^{+\downarrow} = \begin{pmatrix} a \\ a \\ ia \\ -ia \end{pmatrix}. \tag{8.2.544}$$

Verify that $w^{+\downarrow}$ is a simultaneous eigenvector of both γ^4 and M^3 with eigenvalue $+1$ for γ^4 and eigenvalue -1 for M^3. Note that the quantity a appearing in $w^{+\uparrow}$ and $w^{+\downarrow}$ is arbitrary and may, for example, be set to 1.

Finally, let us make the definition

$$S^3 = iK(\hat{L}^3). \tag{8.2.545}$$

Verify using (2.373) and (2.530) that

$$S^3 = (1/2)M^3. \tag{8.2.546}$$

Consequently, $w^{+\uparrow}$ is a simultaneous eigenvector of both γ^4 and S^3 with eigenvalue $+1$ for γ^4 and eigenvalue $+1/2$ for S^3. And $w^{+\downarrow}$ is a simultaneous eigenvector of both γ^4 and S^3 with eigenvalue $+1$ for γ^4 and eigenvalue $-1/2$ for S^3. We may say that $w^{+\uparrow}$ is the spinor part of the ψ for a spin 1/2 particle (at rest) having spin up, and $w^{+\downarrow}$ is the spinor part of the ψ for a spin 1/2 particle (at rest) having spin down.

So far we have been discussing positive energy solutions of the Dirac equation. We close this exercise with a brief discussion of negative energy solutions, which also exist. Suppose, instead of (2.509), we make the Ansatz

$$\psi(x) = w \exp[+i(\sum_{\mu} p_\mu x^\mu)/\hbar]. \tag{8.2.547}$$

Observe that the argument of the exponential function appearing in (2.547) eventually becomes *positive* imaginary as t becomes evermore positive. Correspondingly, (2.547) is called a *negative* energy solution.[15] Show that the Ansatz (2.547) satisfies the Dirac equation (2.504) provided w satisfies the relation

$$\gamma^4 w = -w. \tag{8.2.548}$$

That is, w must be an eigenvector of γ^4 with eigenvalue -1.

Consider the vectors (4-spinors) $w^{-\uparrow}$ and $w^{-\downarrow}$ defined by

$$w^{-\uparrow} = (w^{+\downarrow})^* \tag{8.2.549}$$

and

$$w^{-\downarrow} = (w^{+\uparrow})^*. \tag{8.2.550}$$

Verify that

$$\gamma^4 w^{-\uparrow} = [(\gamma^4)^* w^{+\downarrow}]^* = [-\gamma^4 w^{+\downarrow}]^* = -(w^{+\downarrow})^* = -w^{-\uparrow}. \tag{8.2.551}$$

Similarly, verify that

$$\gamma^4 w^{-\downarrow} = -w^{-\downarrow}. \tag{8.2.552}$$

Next, observe from (2.545) that

$$(S^3)^* = -S^3. \tag{8.2.553}$$

Consequently, verify that

$$S^3 w^{-\uparrow} = [(S^3)^* w^{+\downarrow}]^* = [-S^3 w^{+\downarrow}]^* = (1/2)[w^{+\downarrow}]^* = (1/2)w^{-\uparrow}. \tag{8.2.554}$$

Similarly, verify that

$$S^3 w^{-\downarrow} = -(1/2)w^{-\downarrow}. \tag{8.2.555}$$

Thus, $w^{-\uparrow}$ and $w^{-\downarrow}$ behave under the action of γ^4 and S^3 as their notation suggests. Finally, verify from (2.539), (2.544), (2.549), and (2.550) that there are the explicit results

$$w^{-\uparrow} = \begin{pmatrix} a^* \\ a^* \\ -ia^* \\ ia^* \end{pmatrix}, \tag{8.2.556}$$

[15] Motivated by the relation $E = h\nu$, positive energy solutions are also sometimes called *positive frequency* solutions, and negative energy solutions are sometimes called *negative frequency* solutions.

$$w^{-\downarrow} = \begin{pmatrix} a^* \\ -a^* \\ ia^* \\ ia^* \end{pmatrix}. \tag{8.2.557}$$

Here again the quantity a is arbitrary and may, for example, be set to 1.

In summary, when both positive and negative energy solutions (for a particle at rest) are considered, we have seen that there are four possibilities for w, namely $w^{+\uparrow}$, $w^{+\downarrow}$, $w^{-\uparrow}$, and $w^{-\downarrow}$.

What is the use of the negative energy solutions? Very much further discussion would bring us too far afield. But remarkably it can be shown that, taken together, the positive and negative energy solutions can be used to construct a four-component *quantum* field involving creation and destruction operators for particles and their antiparticles. In this construction destruction operators are associated with positive energy solutions and creation operators are associated with negative energy solutions. The result of this construction is a theory that describes simultaneously both spin 1/2 particles and their antimatter counterparts (for example electrons and positrons), and these particles obey Fermi-Dirac statistics. (Thus the four-fold nature of the Dirac equation associated with the four possibilities for w is revealed to be related to the four possibilities of spin up and spin down and matter and antimatter.) Moreover, in the full quantum field version of Dirac theory, there is a ground state called the *vacuum* which corresponds to a (unique) state for which there are no particles. And in the quantum field theory version both particle and antiparticle states have positive energies, and there is complete symmetry between matter and antimatter.

8.2.19. The two preceding Exercises 2.17 and 2.18 in this chapter have, among other things, explored the relations between the 2×2 complex matrices $\hat{L}^j$ and $\hat{N}^j$, that carry the $\Gamma(0,1/2)$ of the Lorentz group Lie algebra, and the 4×4 real matrices $K(\hat{L}^j)$ and $K(\hat{N}^j)$. According to Exercise 7.3.30, the 2×2 complex matrices $\grave{\hat{L}}^j$ and $\grave{\hat{N}}^j$ carry the $\Gamma(1/2,0)$ of the Lorentz group Lie algebra. The purpose of this exercise is to consider the complementary question: what are the relations between the 2×2 complex matrices $\grave{\hat{L}}^j$ and $\grave{\hat{N}}^j$ and the 4×4 real matrices $K(\grave{\hat{L}}^j)$ and $K(\grave{\hat{N}}^j)$? We know from the work of Exercise 2.17 that the matrix set $K(\hat{L}^j)$,$K(\hat{N}^j)$ carries the Lorentz group Lie algebra representation $\Gamma(0,1/2)\oplus\Gamma(1/2,0)$. Could it be that the matrix set $K(\grave{\hat{L}}^j)$,$K(\grave{\hat{N}}^j)$ also carries this representation? If so, the two sets must be related by a similarity transformation (and vice verse).

Begin our/your investigation by showing from (7.3.248), (7.3.249), and (7.3.376) that there are the relations

$$K(\grave{\hat{L}}^j) = K(\hat{L}^j) \tag{8.2.558}$$

and

$$K(\grave{\hat{N}}^j) = -K(\hat{N}^j). \tag{8.2.559}$$

Next verify, using (2.366) through (2.368), (2.373) through (2.375), and the Dirac algebra, that

$$(\gamma^4)^{-1}K(\hat{L}^j)\gamma^4 = K(\hat{L}^j) \tag{8.2.560}$$

from which, by (2.558), it follows that

$$(\gamma^4)^{-1}K(\hat{L}^j)\gamma^4 = K(\grave{\hat{L}}^j). \tag{8.2.561}$$

Also verify, using (2.369) through (2.371), (2.376) through (2.378), and the Dirac algebra, that

$$(\gamma^4)^{-1}K(\hat{N}^j)\gamma^4 = -K(\hat{N}^j) \tag{8.2.562}$$

from which, by (2.559), it follows that

$$(\gamma^4)^{-1}K(\hat{N}^j)\gamma^4 = K(\grave{\hat{N}}^j). \tag{8.2.563}$$

Taken together, (2.561) and (2.563) show that the two sets $K(\hat{L}^j)$,$K(\hat{N}^j)$ and $K(\grave{\hat{L}}^j)$,$K(\grave{\hat{N}}^j)$ are indeed related by a similarity transformation, the similarity transformation provided by γ^4.[16] Finally, verify that the two sets are also related by the similarity transformation provided by $\gamma^5\gamma^4$.

Let us summarize our findings. From Exercise 7.3.30 we learned that the $\Gamma(0, 1/2)$ representation provided by the matrices $\hat{L}^j$,$\hat{N}^j$ is different from (not equivalent to) the $\Gamma(1/2, 0)$ representation provided by the matrices $\grave{\hat{L}}^j$,$\grave{\hat{N}}^j$. Now we have learned that the representations provided by the two sets of matrices $K(\hat{L}^j)$,$K(\hat{N}^j)$ and $K(\grave{\hat{L}}^j)$,$K(\grave{\hat{N}}^j)$ are the same/equivalent, namely the representation $\Gamma(0, 1/2) \oplus \Gamma(1/2, 0)$.

8.2.20. (Under construction) Review Exercise 2.19. The purpose of this exercise is to do something analogous for the representations $\Gamma(0, 1)$, $\Gamma(1, 0)$, and $\Gamma(0, 1) \oplus \Gamma(1, 0)$ using K.

8.2.21. (Under construction) Review Exercise 7.3.34. The purpose of this exercise is to explore, at least to some extent, what special properties matrices $K(k)$ might have if the matrices k have special properties. In particular, we will study what can be said about the case

$$k \in SL(2, \mathbb{C}) = Sp(2, \mathbb{C}) \Leftrightarrow k^T J_2 k = J_2 \tag{8.2.564}$$

without immediately invoking the Dirac machinery.

Apply (7.3.378) to the symplectic condition in (2.557) to show that

$$K(k^T)K(J_2)K(k) = K(J_2). \tag{8.2.565}$$

Using the definition (7.3.375), verify the following relations:

$$K(J_2) = \begin{pmatrix} J_2 & \mathbf{0} \\ \mathbf{0} & J_2 \end{pmatrix} = J', \tag{8.2.566}$$

$$K(k^T) = \begin{pmatrix} \Re k^T & -\Im k^T \\ \Im k^T & \Re k^T \end{pmatrix} = \begin{pmatrix} (\Re k)^T & -(\Im k)^T \\ (\Im k)^T & (\Re k)^T \end{pmatrix}, \tag{8.2.567}$$

and

$$[K(k)]^T = \begin{pmatrix} (\Re k)^T & (\Im k)^T \\ -(\Im k)^T & (\Re k)^T \end{pmatrix}. \tag{8.2.568}$$

[16] Note that the relations (2.561) and (2.563) can also be written in the form $(\gamma^4)^{-1}K(\grave{\hat{L}}^j)\gamma^4 = K(\hat{L}^j)$, etc. Thus, there is complete symmetry between the use of the $\Gamma(0, 1/2)$ and the $\Gamma(1/2, 0)$ representations.

Let J be the matrix

$$J = \begin{pmatrix} \mathbf{0} & I^{[2]} \\ -I^{[2]} & \mathbf{0} \end{pmatrix}. \tag{8.2.569}$$

Verify that

$$J^{-1}[K(k)]^T J = K(k^T). \tag{8.2.570}$$

Show, using (2.559) and (2.563), that (2.558) can be rewritten in the form

$$J^{-1}[K(k)]^T J J' K(k) = J', \tag{8.2.571}$$

from which it follows that

$$[K(k)]^T [JJ'] K(k) = [JJ']. \tag{8.2.572}$$

Define a matrix S by the rule

$$S = JJ' \tag{8.2.573}$$

so that (2.565) can be written as

$$[K(k)]^T S K(k) = S. \tag{8.2.574}$$

[Note that, despite our notation, this matrix S has nothing to do with the matrix S^3 defined by (2.538) and (2.539).] Verify that

$$S = \begin{pmatrix} \mathbf{0} & J_2 \\ -J_2 & \mathbf{0} \end{pmatrix}. \tag{8.2.575}$$

Verify that J and J' commute,

$$\{J, J'\} = 0. \tag{8.2.576}$$

Show that

$$S^T = S, \tag{8.2.577}$$

$$\det(S) = 1, \tag{8.2.578}$$

and

$$S^2 = I^{[4]}. \tag{8.2.579}$$

Evidently S is *real*. Verify from (2.570) and (2.572) that S is also orthogonal,

$$S^T S = I^{[4]}. \tag{8.2.580}$$

A brute force verification using (2.568), while tedious, is possible. But note that (2.570) through (2.572) follow directly from (2.566) and (2.569) and the relations

$$J^T = -J,\ (J')^T = -J',\ \det(J) = \det(J') = 1. \tag{8.2.581}$$

Upon comparing (2.567) with the symplectic condition in (2.557) we see that they have a similar form but J_2 is antisymmetric while S is symmetric. Since S is symmetric, it can be used to define an inner product, and the relation (2.567) can be viewed as a *preservation* condition for this inner product. Suppose u and v are any two four-component real arrays and we define their inner product $(u, v)_S$ by the rule

$$(u, v)_S = (u, Sv) \tag{8.2.582}$$

where the inner product on the right side of (2.575) is the usual real inner product. Note that, since S is real, the quantity $(u, v)_S$ is real if u and v are real. Then, using (2.567), verify that

$$(Ku, Kv)_S = (Ku, SKv) = (u, K^T SKv) = (u, Sv) = (u, v)_S. \tag{8.2.583}$$

That is, the inner product $(u, v)_S$ is invariant under the action of $K(k)$. Moreover, make the definition

$$(u, v)_{S\gamma_r^5} = (u, S\gamma_r^5 v). \tag{8.2.584}$$

Note that $(u, v)_{S\gamma_r^5}$ is also *real*. Then, using (2.567) and (2.384), verify that

$$\begin{aligned}&(Ku, Kv)_{S\gamma_r^5} = (Ku, S\gamma_r^5 Kv) = (Ku, SK\gamma_r^5 v) = \\ &(u, K^T SK\gamma_r^5 v) = (u, S\gamma_r^5 v) = (u, v)_{S\gamma_r^5}.\end{aligned} \tag{8.2.585}$$

Therefore the inner product $(u, v)_{S\gamma_r^5}$ is also invariant under the action of $K(k)$.

But now we are confronted with an embarrassment of riches!

Out of curiosity, verify that

$$(u, v)_S = (u, Sv) = u_1v_4 - u_2v_3 - u_3v_2 + u_4v_1. \tag{8.2.586}$$

8.2.22. (Under Construction) Exercise on use of S rather than γ_r^4 and the existence of more bilinear forms.

8.2.23. (Under Construction) Other Majorana representations.

What is the nature of the inner product $(*, *)_S$? Evidently S is real. Verify from (2.329) and (2.331) that S is also orthogonal,

$$S^T S = I^{[4]}. \tag{8.2.587}$$

Therefore there must be a real orthogonal matrix O such that

$$O^T SO = D \Leftrightarrow S = ODO^T \tag{8.2.588}$$

where D is diagonal. Verify using (2.335) and (2.331) that there is the relation

$$D^2 = O^T SOO^T SO = O^T S^2 O = O^T O = I^{[4]}, \tag{8.2.589}$$

and therefore the diagonal entries of D must have absolute value 1.

Let us try to find O and D. To continue, verify that

$$S = K(-iJ_2) = K(\sigma^2). \tag{8.2.590}$$

Next, using the relation between $\boldsymbol{\sigma}$ and $\boldsymbol{K}$ given by (3.7.169) through (3.7.171), verify that (8.2.57) can also be written in the form

$$\exp[(-i/2)\theta\boldsymbol{n}\cdot\boldsymbol{\sigma}](\boldsymbol{a}\cdot\boldsymbol{\sigma})\exp[(i/2)\theta\boldsymbol{n}\cdot\boldsymbol{\sigma}] = [R(\theta,\boldsymbol{n})\boldsymbol{a}]\cdot\boldsymbol{\sigma}. \tag{8.2.591}$$

(Sorry yet again about the possibly confusing notation! Although they may look related, there is *no* connection between the symbols $\boldsymbol{K}$ and K. Sometimes there are not enough symbols to go around.) Evaluate (2.338) for the case $\boldsymbol{n} = \boldsymbol{e}_1$ and $\boldsymbol{a} = \boldsymbol{e}_2$ so that it becomes

$$\exp[(-i/2)\theta\sigma^1](\sigma^2)\exp[(i/2)\theta\sigma^1] = [R(\theta,\boldsymbol{e}_1)\boldsymbol{e}_2]\cdot\boldsymbol{\sigma}. \tag{8.2.592}$$

Verify that

$$[R(\pi/2,\boldsymbol{e}_1)\boldsymbol{e}_2] = \boldsymbol{e}_3. \tag{8.2.593}$$

Put another way, rotating $\boldsymbol{e}_2$ by $\theta = \pi/2$ about the $\boldsymbol{e}_1$ axis yields $\boldsymbol{e}_3$. See (3.7.205). Consequently, for $\theta = \pi/2$, (2.339) becomes

$$\exp[-i(\pi/4)\sigma^1](\sigma^2)\exp[i(\pi/4)\sigma^1] = \sigma^3. \tag{8.2.594}$$

As a sanity check, verify directly that (2.341) holds by evaluating the indicated exponential functions and carrying out the indicated multiplications. In particular, you should find for the exponential functions the results

$$\exp[\pm i(\pi/4)\sigma^1] = \cos(\pi/4)\sigma^0 \pm i\sin(\pi/4)\sigma^1 = (1/\sqrt{2})\sigma^0 \pm i(1/\sqrt{2})\sigma^1. \tag{8.2.595}$$

See (3.7.192).

Now verify that applying (7.3.378) to both sides of (2.341) yields the result

$$K\{\exp[-i(\pi/4)\sigma^1]\}K(\sigma^2)K\{\exp[i(\pi/4)\sigma^1]\} = K(\sigma^3). \tag{8.2.596}$$

Make the assignment

$$O = K\{\exp[i(\pi/4)\sigma^1]\} \tag{8.2.597}$$

and show that

$$O = (1/\sqrt{2})\begin{pmatrix} \sigma^0 & -\sigma^1 \\ \sigma^1 & \sigma^0 \end{pmatrix}. \tag{8.2.598}$$

Note that, as expected, O is real. Also, using (7.3.381), show that

$$O^{-1} = K\{\exp[-i(\pi/4)\sigma^1]\} \tag{8.2.599}$$

from which it follows that

$$O^{-1} = (1/\sqrt{2})\begin{pmatrix} \sigma^0 & \sigma^1 \\ -\sigma^1 & \sigma^0 \end{pmatrix}. \tag{8.2.600}$$

Verify, by comparing (2.345) and (2.347), that

$$O^{-1} = O^T. \tag{8.2.601}$$

Verify it follows from (2.343) that (2.335) has been achieved with

$$D = K(\sigma^3) = \begin{pmatrix} \sigma_3 & \mathbf{0} \\ \mathbf{0} & \sigma_3 \end{pmatrix} = \text{diag}(1,-1,1,-1). \tag{8.2.602}$$

As a final sanity check on our/your work, verify directly that (2.335) has been achieved using (2.327) for S, (2.345) for O, (2.347) for $O^{-1} = O^T$, and (2.349) for D.

Define matrices $M(k)$ by the rule

$$M(k) = O^T K(k) O = O^{-1} K(k) O. \tag{8.2.603}$$

Verify that, for $k \in SL(2, \mathbb{C}) = Sp(2, \mathbb{C})$, they satisfy the relation chain

$$\begin{aligned} &M^T(k) D M(k) = [O^T K(k) O]^T D [O^T K(k) O] = \\ &[O^T K^T(k)][O D O^T][K(k) O] = O^T [K^T(k) S K(k)] O = \\ &O^T S O = D. \end{aligned} \tag{8.2.604}$$

That is, upon comparing the beginning and end of (2.351), we see that there is the relation

$$M^T(k) D M(k) = D. \tag{8.2.605}$$

Verify also, using (7.3.380), that

$$M(I^{[2]}) = I^{[4]}. \tag{8.2.606}$$

Because D has two positive and two negative diagonal entries we conclude that, for $k \in SL(2, \mathbb{C}) = Sp(2, \mathbb{C})$, the matrices $M(k)$ provide a representation of $SO(2, 2, \mathbb{R})$. See Exercises 3.7.38 and 3.7.40.

Let P be the permutation operator that interchanges the two and three axes and leaves the other axes in peace. That is, P is a linear operator with the actions

$$P\boldsymbol{e}_1 = \boldsymbol{e}_1,\ P\boldsymbol{e}_2 = \boldsymbol{e}_3,\ P\boldsymbol{e}_3 = \boldsymbol{e}_2,\ P\boldsymbol{e}_4 = \boldsymbol{e}_4. \tag{8.2.607}$$

Correspondingly, P has the matrix representation

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{8.2.608}$$

Verify that P is symmetric, $P^T = P$; is an involution, $P^2 = I^{[4]}$; and is therefore orthogonal, $P^T = P^{-1}$. Define a linear operator $\hat{D}$ by the rule

$$\hat{D} = P^T D P. \tag{8.2.609}$$

Verify that $\hat{D}$ has the matrix representation

$$\hat{D} = \begin{pmatrix} I^{[2]} & \mathbf{0} \\ \mathbf{0} & -I^{[2]} \end{pmatrix} = \mathrm{diag}(1, 1, -1, -1). \tag{8.2.610}$$

Define matrices $\hat{M}(k)$ by the rule

$$\hat{M}(k) = P^T M(k) P. \tag{8.2.611}$$

Verify that, for $k \in SL(2,\mathbb{C}) = Sp(2,\mathbb{C})$, they satisfy the relation chain

$$\begin{aligned}&M^T(k)DM(k) = [O^TK(k)O]^TD[O^TK(k)O] =\\ &[O^TK^T(k)][ODO^T][K(k)O] = O^T[K^T(k)SK(k)]O =\\ &O^TSO = D.\end{aligned} \tag{8.2.612}$$

That is, upon comparing the beginning and end of (2.351), we see that there is the relation

$$M^T(k)DM(k) = D. \tag{8.2.613}$$

Verify also, using (7.3.380), that

$$M(I^{[2]}) = I^{[4]}. \tag{8.2.614}$$

Because D has two positive and two negative diagonal entries we conclude that, for $k \in SL(2,\mathbb{C}) = Sp(2,\mathbb{C})$, the matrices $M(k)$ provide a representation of $SO(2,2,\mathbb{R})$. See Exercises 3.7.38 and 3.7.40.

8.2.24. (Under Construction) Review Exercise 3.7.37 that relates the Lorentz group Lie algebra $so(3,1,\mathbb{R})$ to $so(4,\mathbb{R})$ when working over the complex field, and hence also to $su(2) \oplus su(2)$. See also Exercise 6.2.6.

In the case of the Lorentz group there are the 4×4 *Dirac gamma* matrices γ^μ, with $\mu = 1 \cdot\cdot 4$, that satisfy the anti-commutation rules

$$\{\gamma^\mu, \gamma^\nu\}_+ = \gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu = 2g^{\mu\nu}, \tag{8.2.615}$$

and transform under the action of the Lorentz group according to the rules

$$* = *. \tag{8.2.616}$$

For the case of $SO(4,\mathbb{R})$ there are analogous 4×4 matrices; call them Γ^μ. Define them as follows:

$$\Gamma^1 =, \tag{8.2.617}$$

$$\Gamma^2 =, \tag{8.2.618}$$

$$\Gamma^3 =, \tag{8.2.619}$$

$$\Gamma^4 = . \tag{8.2.620}$$

Note they are all real and involve 2×2 blocks featuring the Pauli matrices *. Verify that they satisfy the anti-commutation rules

$$\{\Gamma^\mu, \Gamma^\nu\}_+ = \delta^{\mu\nu}, \tag{8.2.621}$$

and transform under the action of the the rotation group $SO(4,\mathbb{R})$ according to the rules

$$* = *. \tag{8.2.622}$$

Show that there are six linearly independent matrices of the form

$$\Sigma^{\mu\nu} = \{\Gamma^\mu, \Gamma^\nu\} = \Gamma^\mu\Gamma^\nu - \Gamma^\nu\Gamma^\mu. \tag{8.2.623}$$

Show that these matrices form a basis for $so(4,\mathbb{R})$.

Something about $SO(n,\mathbb{R})$ and Clifford algebras.

8.2.25. (Under Construction) Exercise on the Möbius and Lorentz groups.

8.3 Questions of Order and other Miscellaneous Mysteries

Lie operators and Lie transformations have remarkable properties, and in many ways seem to lead lives of their own. The purpose of this section is to discuss various questions of operator ordering that are often confusing to the uninitiated, and sometimes puzzling to even the enlightened. We will also extend some previous results, and resolve some mysteries of sign that arose in previous sections.

8.3.1 Questions of Order and Map Multiplication

Suppose $\mathcal{M}_f$ is a symplectic map that sends the general point z in phase space to the point $\bar{z}$, and suppose $\mathcal{M}_g$ is another symplectic map that sends $\bar{z}$ to the point $\bar{\bar{z}}$. The reason for our naming convention using the subscripts f and g will become apparent shortly. See Figure 3.1

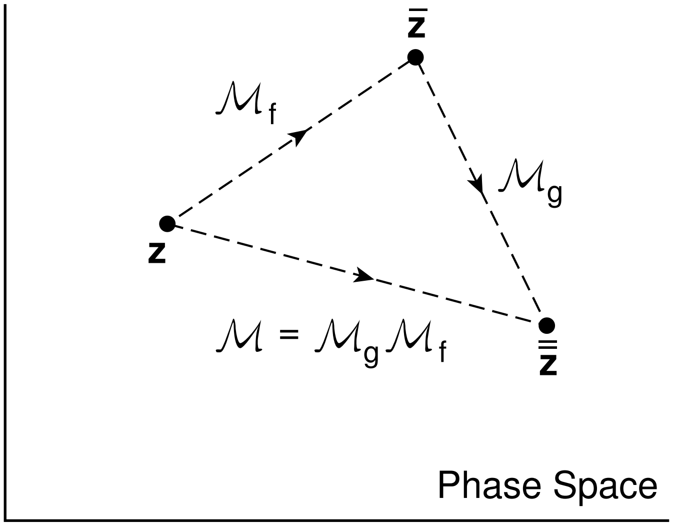


Figure 8.3.1: The composite action of two maps $\mathcal{M}_f$ and $\mathcal{M}_g$.

Equivalently, in the context of charged particle beam transport, we may think of a beam that first passes through beam line element f, the action of which is described by the map $\mathcal{M}_f$, and then through beam line element g whose action is described by the map $\mathcal{M}_g$. See Figure 3.2.

Now consider the composite mapping $\mathcal{M}$ which sends z to $\bar{\bar{z}}$ and which, following usual mathematical notation, would be written in the form

$$\mathcal{M} = \mathcal{M}_g\mathcal{M}_f, \tag{8.3.1}$$

$$\mathcal{M}_f : z \rightarrow \bar{z}, \tag{8.3.2}$$

$$\mathcal{M}_g : \bar{z} \rightarrow \bar{\bar{z}}, \tag{8.3.3}$$

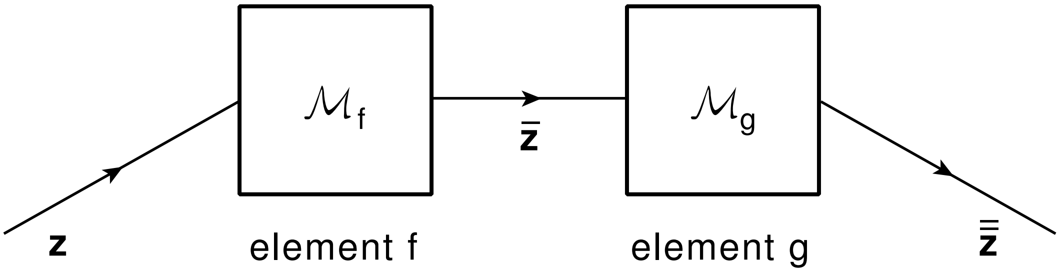


Figure 8.3.2: Successive passage of a trajectory with initial condition z through beam line elements f and g resulting in the intermediate condition $\bar{z}$ and final condition $\bar{\bar{z}}$.

$$\mathcal{M}_g \mathcal{M}_f : z \rightarrow \bar{\bar{z}} \; . \tag{8.3.4}$$

Note that, when reading (3.1) from left to right, the maps $\mathcal{M}_f$ and $\mathcal{M}_g$ occur in the *opposite* order from which they actually act. See Fig. 3.2. That is, $\mathcal{M}_f$ acts first, but appears last in (3.1); and $\mathcal{M}_g$ acts last, but appears first in (3.1). This order follows the standard mathematical convention for maps (including matrices as a special case), and is in accord with the ordering used earlier in Section 6.2 and Equations (6.4.5) and (6.4.7).

Suppose, for purposes of discussion, that both $\mathcal{M}_f$ and $\mathcal{M}_g$ can be written in exponential form using single exponents,

$$\mathcal{M}_f = \exp(: f :), \tag{8.3.5}$$

$$\mathcal{M}_g = \exp(: g :). \tag{8.3.6}$$

In this case (3.2) and (3.3) can be written in the more explicit form

$$\bar{z}(z) = \mathcal{M}_f z = \exp(: f(z) :)z, \tag{8.3.7}$$

$$\bar{\bar{z}}\,(\bar{z}) = \mathcal{M}_g \bar{z} = \exp(: g(\bar{z}) :)\bar{z}. \tag{8.3.8}$$

Also, if we regard $\bar{z}$ as a function of z, as done in (3.7), then (3.8) can also be written in the form

$$\bar{\bar{z}}\,(z) = \bar{\bar{z}}\,(\bar{z}(z)) = \exp(: g(\bar{z}(z)) :)\bar{z}(z). \tag{8.3.9}$$

[Note that the Poisson brackets implied in (3.9) can be evaluated using either the variables z or $\bar{z}$ with the same result. See (6.3.10), (6.3.11), and (6.3.20).] Finally, suppose we substitute (3.7) into (3.9). Doing so gives the result

$$\bar{\bar{z}}\,(z) = \exp(: g(\bar{z}(z)) :) \exp(: f(z) :)z. \tag{8.3.10}$$

This result is simply (3.4) written in explicit form.

Next suppose the identity operator $\mathcal{I}$, written in the form

$$\mathcal{I} = \exp(: f(z) :) \exp(- : f(z) :), \tag{8.3.11}$$

is inserted right after the equal sign in (3.10). This insertion brings (3.10) to the form

$$\bar{\bar{z}}\,(z) = \exp(: f(z) :) \exp(- : f(z) :) \exp(: g(\bar{z}(z)) :) \exp(: f(z) :)z. \tag{8.3.12}$$

Consider the quantity $\exp(-:f(z):)\exp(:g(\overline{z}(z)):)\exp(:f(z):)$ that appears in (3.12). According to (2.20) we have the result

$$\exp(-:f(z):)\exp(:g(\overline{z}(z)):)\exp(:f(z):) = \exp(:\exp(-:f(z):)g(\overline{z}(z)):). \tag{8.3.13}$$

According to (3.7) and (5.4.11) we have the result

$$g(\overline{z}(z)) = g(\exp(:f(z):)z) = \exp(:f(z):)g(z). \tag{8.3.14}$$

With this information in hand, we may rewrite (3.13) in the form

$$\begin{aligned}\exp(-:f(z):)&\exp(:g(\overline{z}(z)):)\exp(:f(z):)\\ &= \exp(:\exp(-:f(z):)g(\overline{z}(z)):)\\ &= \exp(:\exp(-:f(z):)\exp(:f(z):)g(z):)\\ &= \exp(:g(z):).\end{aligned} \tag{8.3.15}$$

Finally, use of (3.15) in (3.12) gives the remarkable result

$$\overline{\overline{z}}\,(z) = \exp(:f(z):)\exp(:g(z):)z. \tag{8.3.16}$$

Observe that (3.4) can be written in the form

$$\overline{\overline{z}}\,(z) = \mathcal{M}z = \mathcal{M}_g\mathcal{M}_f z, \tag{8.3.17}$$

while (3.16) can be written in the form

$$\overline{\overline{z}}\,(z) = \mathcal{M}z = \mathcal{M}_f\mathcal{M}_g z. \tag{8.3.18}$$

What is going on here to produce two seemingly contradictory results? The difference between (3.17) and (3.18) is as follows: In (3.17), as examination of (3.7), (3.8), and (3.10) shows, f is a function of the *initial* variable z while g is a function of the *intermediate* variable $\overline{z}$. By contrast in (3.18), as (3.16) shows, *both* f and g are functions of the *initial* variable z. What we have learned is that if a beam passes successively through beam line elements f and g, and in that order, then the map for the composite system is $\mathcal{M}_f\mathcal{M}_g$ where both f and g are taken to be functions of the initial variable z. We see that when the factors in a map (which is expressed as a product of factors all of the initial variable z) are read from left to right, they are encountered in the *same* order as they are encountered by the beam.

Strictly speaking, the last two sentences in the previous paragraph have been shown to be true for two maps with both maps assumed to be expressible in exponential form using single exponents as in (3.5) and (3.6). However, by similar arguments, analogous results can be shown to hold in general. For example, suppose $\mathcal{M}_f$, $\mathcal{M}_g$, and $\mathcal{M}_h$ are any three (analytic) maps. Then, from (7.8.1), $\mathcal{M}_f$ has a factorization of the form

$$\mathcal{M}_f = \exp(:f_1:)\exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_3:)\exp(:f_4:)\cdots; \tag{8.3.19}$$

and $\mathcal{M}_g$ and $\mathcal{M}_h$ have similar factorizations. Suppose a trajectory with *initial* condition z^i passes successively through the beam line elements f, g, and h described by the maps $\mathcal{M}_f$,

$\mathcal{M}_g$, and $\mathcal{M}_h$, respectively. Then the *final* condition z^f as a result of this passage is given by the relation

$$\begin{aligned} z^f(z^i) &= \mathcal{M}_f\mathcal{M}_g\mathcal{M}_h z^i \\ &= \exp(:f_1:)\exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_3:)\exp(:f_4:)\cdots\times \\ &\quad \exp(:g_1:)\exp(:g_2^c:)\exp(:g_2^a:)\exp(:g_3:)\exp(:g_4:)\cdots\times \\ &\quad \exp(:h_1:)\exp(:h_2^c:)\exp(:h_2^a:)\exp(:h_3:)\exp(:h_4:)\cdots z^i \end{aligned} \tag{8.3.20}$$

when all the homogeneous polynomials f_j, g_j, and h_j are taken to be functions of the initial variable z^i.

8.3.2 Questions of Order in the Linear Case

We are now prepared to revisit some questions of order that were passed over in earlier sections. In (7.2.3) and (7.2.8) we constructed polynomials f_2^a and f_2^c in such a way that

$$\exp(:f_2^a:)z_b = \sum_d P_{bd}z_d, \tag{8.3.21}$$

$$\exp(:f_2^c:)z_d = \sum_e O_{de}z_e. \tag{8.3.22}$$

In so doing, we made the z's transform as the components of a vector under the actions of the matrices P and O. Then we found the result

$$\exp(:f_2^c:)\exp(:f_2^a:)z_b = \sum_e (PO)_{be}z_e. \tag{8.3.23}$$

See (7.2.10) and (7.2.11). Here both f_2^c and f_2^a were quadratic polynomials in the variable z. Since P and O were defined in such a way that the z's transformed under their actions as the components of a vector, we expect to have the matrix product PO as the result of O acting first and then followed by P. On the other hand, since both $\exp(:f_2^c:)$ and $\exp(:f_2^a:)$ are functions of the initial variable z, we expect to have the operator product $\exp(:f_2^c:)$ $\exp(:f_2^a:)$ when $\exp(:f_2^c:)$ acts first and is then followed by the action of $\exp(:f_2^a:)$. But, as we see from (3.21) and (3.22), $\exp(:f_2^a:)$ corresponds to P and $\exp(:f_2^c:)$ corresponds to O. Thus, the orders on both sides of (3.23) are just as they should be.

Next consider the operator and matrix orders in (7.7.34). Here the operator and matrix orders are the *same* rather than reversed! But look at (7.7.26). We see that in this case the z's transformed under the action of the *transposed* matrix M^T. Thus the definition of M as given in (7.7.26) is different from the definitions of P and O as given by (3.21) and (3.22). One definition involved matrix transposition and the other did not. What (7.7.26) and (7.7.34) teach us is that the inclusion of a transpose in the definition of M makes it possible to have both operator and matrix orders the same.

To gain further insight into what is going on, it is useful to consider the general subject of linear operators and matrices. We begin our discussion by examining simple transformation properties of vectors under the action of linear operators. Consider a vector space V spanned by basis vectors $\boldsymbol{e}_\alpha$ that are orthonormal under some scalar product $(,)$. Let $\boldsymbol{\mathcal{L}}$ be a *linear* operator that sends V into itself. Suppose $\boldsymbol{\mathcal{L}}$ sends $\boldsymbol{e}_\alpha$ into $\boldsymbol{f}_\alpha$. Then we have a relation of the form

$$\boldsymbol{f}_\alpha = \boldsymbol{\mathcal{L}}\boldsymbol{e}_\alpha = \sum_\beta L_{\beta\alpha}\boldsymbol{e}_\beta. \tag{8.3.24}$$

The coefficients $L_{\beta\alpha}$ are given in terms of the scalar product by the matrix elements

$$L_{\beta\alpha} = (\boldsymbol{e}_\beta, \boldsymbol{f}_\alpha) = (\boldsymbol{e}_\beta, \boldsymbol{\mathcal{L}}\boldsymbol{e}_\alpha). \tag{8.3.25}$$

Let $\boldsymbol{A}$ be some vector in V. Since the $\boldsymbol{e}_\alpha$ form a basis, $\boldsymbol{A}$ must have an expansion of the form

$$\boldsymbol{A} = \sum_\alpha a_\alpha \boldsymbol{e}_\alpha. \tag{8.3.26}$$

Consider a vector $\boldsymbol{B}$ defined by

$$\boldsymbol{B} = \boldsymbol{\mathcal{L}}\boldsymbol{A}. \tag{8.3.27}$$

It must have an expansion of the form

$$\boldsymbol{B} = \sum_\beta b_\beta \boldsymbol{e}_\beta. \tag{8.3.28}$$

From (3.25) through (3.28) and the orthonormality condition we find that the components b_β are given by the relation

$$\begin{aligned} b_\beta &= (\boldsymbol{e}_\beta, \boldsymbol{B}) = (\boldsymbol{e}_\beta, \boldsymbol{\mathcal{L}}\boldsymbol{A}) = \sum_\alpha a_\alpha (\boldsymbol{e}_\beta, \boldsymbol{\mathcal{L}}\boldsymbol{e}_\alpha) \\ &= \sum_\alpha L_{\beta\alpha} a_\alpha. \end{aligned} \tag{8.3.29}$$

We note that the summation in (3.24) is over the first index in L, and that in (10.6) is over the second. Let us rewrite (3.24) in the form

$$\boldsymbol{f}_\beta = \boldsymbol{\mathcal{L}}\boldsymbol{e}_\beta = \sum_\alpha L_{\alpha\beta}\boldsymbol{e}_\alpha = \sum_\alpha (L^T)_{\beta\alpha}\boldsymbol{e}_\alpha. \tag{8.3.30}$$

(Note that the indices α and β are dummy indices, and can be changed at will.) Upon comparing (3.29) and (3.30), we see that if *components* are transformed by the matrix L, then *basis vectors* are transformed by the transpose matrix L^T, and vice versa.

8.3.3 Application to General Operators and General Monomials to Construct Matrix Representations

Let us apply what we have just learned to operators $\mathcal{M}$ of the general form

$$\mathcal{M} = \exp(:f_1:)\exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_3:)\exp(:f_4:)\cdots, \tag{8.3.31}$$

and monomials $G(\mu;\nu)$ of the form (7.3.1). [Here the f_j are homogeneous polynomials, and are not to be confused with the $\boldsymbol{f}_\alpha$ of (3.24), which are abstract vectors.] We will view these monomials as basis vectors for the vector space of all polynomials. For this purpose, it is convenient to subsume the indices μ, ν into a single index, which we will again call r. (For a concrete way of making a one-to-one correspondence between the indices μ, ν and the integers r, see Section 27.2.) We will thus work with a set of basis monomials G_r, and in accord with (7.3.8) we will assign them a scalar product $\langle , \rangle$ by the rule

$$\langle G_{r'}, G_r \rangle = \delta_{r'r}. \tag{8.3.32}$$

Now let $\mathcal{M}$ act on a general basis vector G_r. Then we find a result of the form

$$\mathcal{M} G_r = \sum_s M_{sr} G_s \tag{8.3.33}$$

where the coefficients M_{sr} are given in terms of the scalar product by the matrix elements

$$M_{sr} = \langle G_s, \mathcal{M} G_r \rangle. \tag{8.3.34}$$

Suppose $\mathcal{N}$ is another operator of the form (3.31). Then we also have the relation

$$\mathcal{N} G_r = \sum_s N_{sr} G_s \tag{8.3.35}$$

with the matrix N_{sr} given by

$$N_{sr} = \langle G_s, \mathcal{N} G_r \rangle. \tag{8.3.36}$$

Now consider the effect of the operator product $\mathcal{M}\mathcal{N}$ on G_r. Here all the Lie generators f_j appearing in $\mathcal{M}$ and $\mathcal{N}$, as well as all the functions G_r, are assumed to depend on the *same* set of variables z. Then, we find the result

$$\begin{aligned} \mathcal{M}\mathcal{N} G_r &= \mathcal{M} \sum_s N_{sr} G_s = \sum_s N_{sr} \mathcal{M} G_s \\ &= \sum_s N_{sr} \sum_t M_{ts} G_t = \sum_{st} M_{ts} N_{sr} G_t \\ &= \sum_t (MN)_{tr} G_t. \end{aligned} \tag{8.3.37}$$

We note that the ordering of the subscripts in (3.33) through (3.37) is analogous to that used in (7.3.37) through (7.3.40). Indeed, let $\mathcal{A}$ and $\mathcal{B}$ be any two *linear operators.* They could, for example, be Lie operators, or products of Lie operators, or sums of products of Lie operators, or infinite sums of products, etc., including Lie transformations and their products such as occur in (3.31). Define associated matrices $O(\mathcal{A})$ and $O(\mathcal{B})$ by rules of the form

$$O_{sr}(\mathcal{B}) = \langle G_s, \mathcal{B} G_r \rangle. \tag{8.3.38}$$

Then, since the G's form a basis (are a complete set), we have the results

$$\mathcal{B} G_r = \sum_s O_{sr}(\mathcal{B}) G_s, \tag{8.3.39}$$

$$\begin{aligned}\mathcal{A}\mathcal{B}G_r &= \mathcal{A}\sum_s O_{sr}(\mathcal{B})G_s = \sum_s O_{sr}(\mathcal{B})\mathcal{A}G_s \\ &= \sum_s O_{sr}(\mathcal{B})\sum_t O_{ts}(\mathcal{A})G_t = \sum_{st} O_{ts}(\mathcal{A})O_{sr}(\mathcal{B})G_t \\ &= \sum_t (O(\mathcal{A})O(\mathcal{B}))_{tr}G_t. \end{aligned} \tag{8.3.40}$$

From (3.38) and (3.40) we obtain the general relation

$$O(\mathcal{A}\mathcal{B}) = O(\mathcal{A})O(\mathcal{B}). \tag{8.3.41}$$

We have found a matrix representation of the algebra of linear operators acting on function space. It can be shown, in the case that these operators are Lie algebra or Lie group elements, that these matrices are related to the adjoint representation. See Section 8.9.[17]

Note that, for a 6-dimensional phase space and for polynomials of degree 0 through m, the matrices O are $[S(m,6)+1] \times [S(m,6)+1]$. See Section 7.10. For example, in the case $m=4$, 210×210 matrices are required. And, in the case $m=8$, 3003×3003 matrices are required.

8.3.4 Application to Linear Transformations of Phase Space

Let us also apply what we have learned to the subject of linear transformations of *phase space* into itself. Set up a Euclidean coordinate system in phase space with unit vectors $\boldsymbol{e}_a$ along the coordinate and momentum axes. Then the general point in phase space may be identified with the vector $\boldsymbol{z}$ from the origin to that point, and $\boldsymbol{z}$ may be written in the form

$$\boldsymbol{z} = \sum_a z_a \boldsymbol{e}_a \tag{8.3.42}$$

where the z_a are the usual coordinate variables. Suppose $\mathcal{L}$ is a linear transformation of phase space into itself, and suppose $\mathcal{L}$ sends the vector $\boldsymbol{z}$ to the vector $\overline{\boldsymbol{z}}$,

$$\overline{\boldsymbol{z}} = \mathcal{L}\boldsymbol{z}. \tag{8.3.43}$$

The vector $\overline{\boldsymbol{z}}$ must have an expansion of the form

$$\overline{\boldsymbol{z}} = \sum_b \overline{z}_b \boldsymbol{e}_b. \tag{8.3.44}$$

Correspondingly, in analogy to (3.25) through (3.29), the quantities $\overline{z}_b$ and z_a are related by the equation

$$\overline{z}_b = \sum_a L_{ba} z_a. \tag{8.3.45}$$

That is, the z's transform as the *components* of a vector, as is consistent with relations of the form (7.1.1) through (7.1.3), (7.6.1), (3.21), and (3.22).

[17] The fact that linear operators acting on function space can be represented by matrices is familiar to any student of Quantum Mechanics. In the context of differential equations and maps, the matrix representation can be realized by *Carleman linearization*, a construction suggested by Poincaré. See the references at the end of this chapter.

8.3.5 Dual role of the Phase-Space Coordinates z_a

So far we have regarded the z_a as the components of a vector as in (3.42). However, the z_a are also *functions* on phase space. Indeed the z_a are special cases of the functions G_r, and, according to (7.3.1) and (7.3.8), satisfy the orthonormality conditions

$$\langle z_a, z_b \rangle = \delta_{ab}. \tag{8.3.46}$$

Suppose $\mathcal{M}$ is of the form (7.7.28). Then, following (3.39), we find the result

$$\bar{z}_b(z) = \mathcal{M} z_b = \sum_a M^1_{ab} z_a \tag{8.3.47}$$

where the matrix M^1 is given by the relation

$$M^1_{ab} = \langle z_a, \mathcal{M} z_b \rangle. \tag{8.3.48}$$

In addition, according to (7.3.41) through (7.3.45) or relations of the form (3.41), we have the result

$$M^1 = \exp(F^1), \tag{8.3.49}$$

where

$$F^1_{ab} = \langle z_a, : f_2 : z_b \rangle. \tag{8.3.50}$$

Now observe that (3.47) can also be written in the form

$$\bar{z}_b(z) = \sum_a [(M^1)^T]_{ba} z_a. \tag{8.3.51}$$

Thus in view of our earlier discussion, see (3.30) for example, and noting that a and b are dummy indices, we conclude the convention used in (7.7.26) is equivalent to viewing the z_a as *basis vectors* (which is consistent with treating the G_r as basis vectors) rather than as components of a vector.

We have learned that the z_a play a dual role. If they are viewed as the components of a displacement vector as in (3.42), then it is appropriate to write their transformation law in the form (3.45) or (7.6.1). If they are viewed as functions, and therefore as special cases of the basis vectors (functions) G_r, then it may be more convenient to write their transformation law in the form (3.47) or, more generally, (3.33).

8.3.6 Extensions

We now turn to extensions of two results found previously. In our initial discussion of symplectic maps in Section 6.1, a symplectic map was defined as a mapping of phase space into itself that obeyed certain equivalent conditions such as (6.1.3) or (6.1.6) or (6.1.10). In Chapter 7 we learned that symplectic maps can be written in terms of Lie transformations, and obtained the factorizations (7.7.23) and (7.8.1). We also know from (5.4.13) and its generalizations that Lie transformations act on functions, and that the action of a Lie transformation on a function is determined once its action is known on phase space. Indeed, if a and b are any functions, then from (7.7.23), or (7.8.1), and (5.4.10) we have the results

$$\mathcal{M}a(z) = a(\mathcal{M}z), \;\; \mathcal{M}a(z)b(z) = \mathcal{M}a(z)\mathcal{M}b(z) = a(\mathcal{M}z)b(\mathcal{M}z). \tag{8.3.52}$$

Thus, we may also view symplectic maps as entities that act on functions. [Note that we have already encountered this idea in (6.3.6) and (7.1.11) when use is made of (7.7.11)]. Conversely, if we view symplectic maps as entities that act on functions, then, since the z_a are functions, we get the action of symplectic maps on phase space from (7.7.11).

The property (3.52) is a consequence of the fact that Lie transformations are isomorphisms with respect to (ordinary) multiplication. See Section 5.4. We also know from Section 5.4 that Lie transformations are isomorphisms with respect to Poisson bracket multiplication. See (5.4.14). It follows from (5.4.15) and its generalizations, and from (7.7.23) or (7.8.1), that we also have the result

$$\mathcal{M}[a, b] = [\mathcal{M}a, \mathcal{M}b]. \tag{8.3.53}$$

Again, this result should already be familiar. When combined with (3.48), it yields the result (6.3.20).

8.3.7 Sign Differences

The last question to be discussed in this section is the difference in sign between relations such as (5.5.1) and (7.2.3). The simple answer is that the sign in (5.5.1) was selected to achieve the correspondence (5.5.13), and that in (7.2.3) was selected to make (7.2.7) hold. But why should the signs turn out to be different? Our discussion of this topic may seem somewhat discursive. However, we shall learn some interesting concepts and facts along the way.

Exercise 3.7.33 studied how, given some matrix representation of a Lie algebra, one might find other similar or possibly different representations. Now let us carry out the analogous discussion for the corresponding Lie group. Suppose some set of matrices gives a respresentation of some group. To every representation matrix M we associate another matrix M' by the rule

$$M' = \bar{M} \tag{8.3.54}$$

where a bar denotes complex conjugation. Then these matrices satisfy the relation

$$M_1' M_2' = (M_1 M_2)', \tag{8.3.55}$$

and therefore also provide a representation of some group. If the matrices M are real, then nothing new has been found. However, if the matrices M are complex and the structure constants of the underlying Lie algebra cannot be made real by some appropriate basis choice, then one must determine whether the resulting group is the same as the original group. If the structure constants are real, then the group will be the same and the representation given by the matrices M' may be different than that given by the matrices M. For example, in the case of the group $SU(3)$, if the matrices M provide the representation $\Gamma(m, n)$, then the matrices M' provide the representation $\Gamma(n, m)$. See Section 5.8.

Suppose, instead of using (3.54) to define M', we use the rule

$$M' = (M^T)^{-1} = (M^{-1})^T. \tag{8.3.56}$$

Then the M' matrices defined in this way also satisfy (3.55), and also provide a representation of the group in question. As an example, consider the case where the matrices M are symplectic. The symplectic condition (3.1.2) can be written in the form

$$M' = (M^T)^{-1} = JMJ^{-1}. \tag{8.3.57}$$

From (3.57) we see that in this case the matrices M' and M are *similar*, and hence the representations of the group in question carried by M' and M are the *same*. As a second example, suppose the matrices M are unitary,

$$M^\dagger = (\bar{M})^T = M^{-1}. \tag{8.3.58}$$

Then we find the result

$$M' = (M^{-1})^T = \bar{M}. \tag{8.3.59}$$

In this case the definitions (3.54) and (3.56) coincide. What happens if the matrices M belong to $SU(2)$? Since these matrices are complex, we might hope that the "priming" operation (3.59) would give something new. However, since matrices in $SU(2)$ are 2×2 and have determinant $+1$, they must also be symplectic. See the comment after Exercise 3.1.3. Consequently, (3.57) must also hold, and we in fact find that both M' and M carry the *same* representation.

Let M^f be a symplectic matrix. Associate with M^f a symplectic map $\mathcal{M}(M^f)$ by the rule

$$\mathcal{M}(M^f)z_a = \sum_b [(M^f)^{-1}]_{ab} z_b. \tag{8.3.60}$$

Note that (3.60) is analogous to (7.7.26) except that M^T has been replaced by M^{-1}. Let M^g be another symplectic matrix, and make the definitions

$$\mathcal{M}_f = \mathcal{M}(M^f), \tag{8.3.61}$$

$$\mathcal{M}_g = \mathcal{M}(M^g). \tag{8.3.62}$$

Then if we regard $\mathcal{M}_f$ and $\mathcal{M}_g$ as composed of Lie transformations all involving the same variable z we find, in analogy to (7.7.33), the result

$$\begin{aligned}
\mathcal{M}_f \mathcal{M}_g z_a &= \mathcal{M}(M^f)\mathcal{M}(M^g) z_a = \mathcal{M}(M^f) \sum_b [(M^g)^{-1}]_{ab} z_b \\
&= \sum_{b,c} [(M^g)^{-1}]_{ab} [(M^f)^{-1}]_{bc} z_c \\
&= \sum_c [(M^g)^{-1}(M^f)^{-1}]_{ac} z_c = \sum_c [(M^f M^g)^{-1}]_{ac} z_c \\
&= \mathcal{M}(M^f M^g) z_a.
\end{aligned} \tag{8.3.63}$$

We see that the definition (3.60) makes it possible to have both operator and matrix orders the same just as the definition (7.7.26) did. This result is not surprising in view of (3.56) and (3.55).

From a group theory perspective, the advantage of (3.60) compared to (7.7.26) is that the computation of M^{-1} is a *group* operation whereas the computation of M^T is not.

Now that we have a relation that has both operator and matrix orders the same, we can compare their respective Lie algebras. Let S^f and S^g be two symmetric matrices and use them to define functions f_2 and g_2 as in (5.5.1) and (5.5.2),

$$f_2 = (1/2)\sum_{a,b} S^f_{ab} z_a z_b, \tag{8.3.64}$$

$$g_2 = (1/2)\sum_{a,b} S^g_{ab} z_a z_b. \tag{8.3.65}$$

Then we have results of the form

$$: f_2 : z = (-JS^f)z, \tag{8.3.66}$$

and hence

$$\exp(: f_2 :)z = \exp(-JS^f)z = [\exp(JS^f)]^{-1}z. \tag{8.3.67}$$

Upon comparing (3.60) and (3.67), we find the results

$$\mathcal{M}_f = \exp(: f_2 :) = \mathcal{M}[\exp(JS^f)], \tag{8.3.68}$$

$$\mathcal{M}_g = \exp(: g_2 :) = \mathcal{M}[\exp(JS^g)]. \tag{8.3.69}$$

Now consider the product $\exp(\epsilon : f_2 :)\exp(\epsilon : g_2 :)\exp(-\epsilon : f_2 :)\exp(-\epsilon : g_2 :)$ where ϵ is a small quantity. Then, as a consequence of (3.63), we find the relation

$$\begin{aligned} &\exp(\epsilon : f_2 :)\exp(\epsilon : g_2 :)\exp(-\epsilon : f_2 :)\exp(-\epsilon : g_2 :) \\ = \;&\mathcal{M}[\exp(\epsilon JS^f)\exp(\epsilon JS^g)\exp(-\epsilon JS^f)\exp(-\epsilon JS^g)]. \end{aligned} \tag{8.3.70}$$

The products occurring in (3.70) may be viewed as the group analog of what would be a commutator at the Lie algebraic level. Indeed, from (2.27) and (2.28) we find through terms of order ϵ^2 the result

$$\exp(\epsilon : f_2 :)\exp(\epsilon : g_2 :)\exp(-\epsilon : f_2 :)\exp(-\epsilon : g_2 :) = \exp(\epsilon^2 : [f_2, g_2] :). \tag{8.3.71}$$

Similarly, from (3.7.34) we find through terms of order ϵ^2 the result

$$\exp(\epsilon JS^f)\exp(\epsilon JS^g)\exp(-\epsilon JS^f)\exp(-\epsilon JS^g) = \exp(\epsilon^2\{JS^f, JS^g\}). \tag{8.3.72}$$

Now compare (3.70) through (3.72). Doing so gives through terms of order ϵ^2 the result

$$\exp(\epsilon^2 : [f_2, g_2] :) = \mathcal{M}[\exp(\epsilon^2\{JS^f, JS^g\})]. \tag{8.3.73}$$

We are ready for the final step. Let h_2 be the second-degree polynomial defined by the relation

$$h_2 = [f_2, g_2]. \tag{8.3.74}$$

Then, with S^h defined by

$$h_2 = (1/2)\sum_{a,b} S^h_{ab} z_a z_b, \tag{8.3.75}$$

we have the result

$$\mathcal{M}_h = \exp(: h_2 :) = \mathcal{M}[\exp(JS^h)]. \tag{8.3.76}$$

Now compare (3.73) and (3.76) to get the result

$$\mathcal{M}[\exp(\epsilon^2 JS^h)] = \mathcal{M}[\exp(\epsilon^2\{JS^f, JS^g\})]. \tag{8.3.77}$$

We see that for consistency we must have the relation

$$JS^h = \{JS^f, JS^g\}. \tag{8.3.78}$$

This relation is identical to that in (5.5.13). The moral of this rather long tale is that the sign in relations such as (5.5.1) was chosen so that, when (3.60) is used, it is possible to have relations of the form (3.68) and to have both operator and matrix orders the same in (3.63); and when the orders are the same it is easy to compare Lie algebras (exponents) as was done in (3.68) through (3.73). On the other hand, the sign in relations such as (7.2.3) was chosen to achieve relations such as (7.2.7) and (7.2.9), which are to be compared to (3.67).

Exercises

8.3.1. Verify (3.20).

8.3.2. From (3.38) show that $O(\mathcal{A}) + O(\mathcal{B}) = O(\mathcal{A} + \mathcal{B})$.

8.3.3. Verify (3.52) using (5.4.13) and (7.7.23) or (7.8.1).

8.3.4. Verify (3.53) using (5.4.15) and (7.7.23) or (7.8.1).

8.3.5. Given (3.56), verify (3.55).

8.3.6. Show that the map (3.60) is indeed symplectic.

8.3.7. Verify (3.66) and (3.67).

8.3.8. Verify (3.70).

8.3.9. Verify (3.71) and (3.72).

8.4 Lie Concatenation Formulas

As an application of the formulas and ideas developed so far, consider the problem of computing the product of two symplectic maps when each is expressed in factored product form. This problem arises in accelerator physics, for example, in the case that one knows the effect of each of two beam elements separately, and one wants to know the net effect when one beam element is followed by another. For simplicity, in this section we will consider maps that send the origin into itself, i.e. maps of the form (7.6.3). The most general case of maps that include leading and trailing translations, i.e. maps of the form (7.7.23), will be treated in the next chapter.

Let $\mathcal{M}_f$ and $\mathcal{M}_g$ denote the symplectic maps given by the expressions

$$\mathcal{M}_f = \exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_3:)\exp(:f_4:)\cdots, \tag{8.4.1}$$

$$\mathcal{M}_g = \exp(:g_2^c:)\exp(:g_2^a:)\exp(:g_3:)\exp(:g_4:)\cdots. \tag{8.4.2}$$

Also, let $\mathcal{M}_h$ be the product of $\mathcal{M}_f$ and $\mathcal{M}_g$,

$$\mathcal{M}_h = \mathcal{M}_f\mathcal{M}_g. \tag{8.4.3}$$

The problem is to find polynomials h_2^c, h_2^a, h_3, etc. such that

$$\mathcal{M}_h = \exp(:h_2^c:)\exp(:h_2^a:)\exp(:h_3:)\exp(:h_4:)\cdots. \tag{8.4.4}$$

That is, the problem is to express $\mathcal{M}_h$ as given by (4.3) in the factored product form (4.4). For simplicity, only expressions for $h_2^c, h_2^a, h_3, \cdots h_8$ will be found explicitly. Here, as described in Section 8.3, all polynomials f_j, g_j, and h_j are taken to be functions of the same variable z.

Before proceeding further, it is necessary to establish a few simple facts. Suppose g_2 is a quadratic polynomial written in the form

$$g_2 = -(1/2)\sum_{de} S_{de}z_d z_e = -(1/2)(z, Sz), \tag{8.4.5}$$

where S is some symmetric matrix. Suppose further that f_m is some homogeneous polynomial of degree m. Then $\exp(:g_2:)f_m$ is also a homogeneous polynomial of degree m. Indeed, we have the result

$$\exp(:g_2:)f_m(z) = f_m[\exp(:g_2:)z] = f_m(M^g z), \tag{8.4.6}$$

where M^g is the linear transformation defined by the equation

$$M^g = \exp(JS). \tag{8.4.7}$$

See (5.4.11) and Section 7.2. Suppose further that g_2^c and g_2^a are quadratic polynomials written in the forms

$$g_2^a = -(1/2)(z, S^a z), \tag{8.4.8}$$

$$g_2^c = -(1/2)(z, S^c z). \tag{8.4.9}$$

Then we have the result

$$\exp(:g_2^c:)\exp(:g_2^a:)f_m(z) = f_m[\exp(:g_2^c:)\exp(:g_2^a:)z] = f_m(R^g z), \tag{8.4.10}$$

where R^g is the linear transformation defined by the equations

$$R^g = P^g O^g, \tag{8.4.11}$$

$$P^g = \exp(JS^a), \tag{8.4.12}$$

$$O^g = \exp(JS^c). \tag{8.4.13}$$

See Section 7.2. Let us introduce the symplectic map $\mathcal{R}_g$ defined by the equation

$$\mathcal{R}_g = \exp(:g_2^c:)\exp(:g_2^a:). \tag{8.4.14}$$

What we have established is the relation

$$\mathcal{R}_g f_m(z) = f_m(R^g z). \tag{8.4.15}$$

Actually, (4.15) is not quite what will be needed. What we will need is the relation

$$(\mathcal{R}_g)^{-1} f_m(z) = f_m[(R^g)^{-1} z]. \tag{8.4.16}$$

This relation can be established in a similar fashion. Note that, thanks to the symplectic condition, the matrix $(R^g)^{-1}$ required in (4.16) is easily calculated using (3.1.9).

We are now ready to continue. Simply from its definition (4.3), $\mathcal{M}_h$ can be written in the form

$$\mathcal{M}_h = \mathcal{R}_f \exp(:f_3:)\exp(:f_4:)\cdots\mathcal{R}_g\exp(:g_3:)\exp(:g_4:)\cdots . \tag{8.4.17}$$

Here we have used (4.14) and an analogous definition for $\mathcal{R}_f$. Next, by insertion of a factor of $\mathcal{R}_g(\mathcal{R}_g)^{-1}$, (4.17) can be rewritten in the form

$$\mathcal{M}_h = \mathcal{R}_f\mathcal{R}_g(\mathcal{R}_g)^{-1}\exp(:f_3:)\exp(:f_4:)\cdots\mathcal{R}_g\exp(:g_3:)\exp(:g_4:)\cdots . \tag{8.4.18}$$

Evidently, comparison of (4.4) and (4.18) shows that h_2^c and h_2^a are determined by the equation

$$\mathcal{R}_h = \mathcal{R}_f\mathcal{R}_g. \tag{8.4.19}$$

Indeed, we have the result

$$R^h = R^g R^f, \tag{8.4.20}$$

where R^g is defined by (4.11) through (4.13), and R^f and R^h are defined by analogous relations. See Section 8.3.

Next define a function F of the Lie operators $:f_3:, :f_4:, \cdots$ by the relation

$$F(:f_3:, :f_4:, \cdots) = \exp(:f_3:)\exp(:f_4:)\cdots . \tag{8.4.21}$$

Evidently (4.18) contains the factor $(\mathcal{R}_g)^{-1}F\mathcal{R}_g$. As a consequence of (2.25) we have the result

$$(\mathcal{R}_g)^{-1}F(:f_3:, :f_4:, \cdots)\mathcal{R}_g = F(:f_3[(R^g)^{-1}z]:, :f_4[(R^g)^{-1}z]:, \cdots). \tag{8.4.22}$$

In order to simplify further expressions, introduce the notation

$$f_m^{tr}(z) = f_m[(R^g)^{-1}z], \tag{8.4.23}$$

which indicates that the homogeneous polynomial $f_m(z)$ of degree m has been *transformed* to the new homogeneous polynomial $f_m[(R^g)^{-1}z]$. With this notation, (4.22) can be written in the more compact form

$$(\mathcal{R}_g)^{-1}F(:f_3:,:f_4:,\cdots)\mathcal{R}_g = F(:f_3^{tr}:,:f_4^{tr}:,\cdots) = \exp(:f_3^{tr}:)\exp(:f_4^{tr}:)\cdots. \tag{8.4.24}$$

Putting together the work done so far, one finds that (4.18) can also be written in the form

$$\mathcal{M}_h = \mathcal{R}_h \exp(:f_3^{tr}:)\exp(:f_4^{tr}:)\cdots\exp(:g_3:)\exp(:g_4:)\cdots. \tag{8.4.25}$$

Upon comparing (4.4) and (4.25), we find the result

$$\exp(:h_3:)\exp(:h_4:)\cdots = \exp(:f_3^{tr}:)\exp(:f_4^{tr}:)\cdots\exp(:g_3:)\exp(:g_4:)\cdots. \tag{8.4.26}$$

We are now prepared to compute $h_3, h_4, \cdots$ in terms of $f_3^{tr}, f_4^{tr}, \cdots$ and $g_3, g_4, \cdots$. The tool for doing so will be the BCH formula as given by (2.27) and (2.28). We will also use the degree function defined in (7.6.13) and the relation (7.6.14). They are reproduced below for ready reference,

$$\deg(f_m) = m, \tag{8.4.27}$$

$$\deg([f_m, f_n]) = m + n - 2. \tag{8.4.28}$$

Consider the result of combining all exponents on the left side of (4.26) into one grand exponent h by repeated use of the BCH formula. Then, thanks to (4.28), it is relatively easy to pick out and collect various terms according to their degree. One finds the result

$$h = h_3 + h_4 + \{(1/2)[h_3, h_4] + h_5\} + \cdots. \tag{8.4.29}$$

Next, consider the result of combining all exponents on the right side of (4.26) into one grand exponent e. One finds the result

$$e = \{f_3^{tr} + g_3\} + \{(1/2)[f_3^{tr}, g_3] + f_4^{tr} + g_4\} + \cdots. \tag{8.4.30}$$

Now compare (4.29) and (4.30). By equating terms of like degree, we immediately obtain the results

$$h_3 = f_3^{tr} + g_3, \tag{8.4.31}$$

$$h_4 = f_4^{tr} + g_4 + [f_3^{tr}, g_3]/2. \tag{8.4.32}$$

With further work, it is possible to find the polynomials h_5, h_6, etc. Doing so one finds, for example, the results

$$h_5 = f_5^{tr} + g_5 - [g_3, f_4^{tr}] + \frac{1}{3} :g_3:^2 f_3^{tr} - \frac{1}{6} :f_3^{tr}:^2 g_3, \tag{8.4.33}$$

$$\begin{aligned} h_6 \;=\;& f_6^{tr} + g_6 - [g_3, f_5^{tr}] + \frac{1}{2} :g_3:^2 f_4^{tr} + \frac{1}{2}[f_4^{tr}, g_4] - \frac{1}{4}[f_4^{tr}, [f_3^{tr}, g_3]] \\ -\;& \frac{1}{4}[g_4, [f_3^{tr}, g_3]] + \frac{1}{24} :f_3^{tr}:^3 g_3 - \frac{1}{8} :g_3:^3 f_3^{tr} \\ +\;& \frac{1}{8}[f_3^{tr}, [g_3, [f_3^{tr}, g_3]]], \end{aligned} \tag{8.4.34}$$

$$
\begin{aligned}
h_7 \;=\;& f_7^{tr} + g_7 - :g_3: f_6^{tr} - :g_4: f_5^{tr} + \frac{1}{2} :g_3:^2 f_5^{tr} \\
+\;& \frac{1}{2} :f_4^{tr}::g_3: f_4^{tr} + :g_4::g_3: f_4^{tr} + \frac{1}{3} :g_3::f_4^{tr}::f_3^{tr}: g_3 \\
-\;& \frac{1}{6} :g_3:^3 f_4^{tr} - \frac{1}{6} :g_4::f_3^{tr}::g_3: f_3^{tr} - \frac{1}{6} :f_4^{tr}::f_3^{tr}::g_3: f_3^{tr} \\
-\;& \frac{1}{3} :g_4::g_3:^2 f_3^{tr} - \frac{1}{3}[:g_3: f_3^{tr}, :g_3: f_4^{tr}] \\
-\;& \frac{1}{120} :f_3^{tr}:^4 g_3 - \frac{1}{30} :g_3::f_3^{tr}:^3 g_3 - \frac{1}{20} :g_3:^2:f_3^{tr}:^2 g_3 \\
+\;& \frac{1}{30} :g_3:^4 f_3^{tr} + \frac{1}{30}[:f_3^{tr}: g_3, :f_3^{tr}:^2 g_3] + \frac{1}{15}[:g_3: f_3^{tr}, :g_3:^2 f_3^{tr}], \qquad (8.4.35)
\end{aligned}
$$

$$
\begin{aligned}
h_8 \;=\;& f_8^{tr} + g_8 - :g_3: f_7^{tr} - :g_4: f_6^{tr} - \frac{1}{2} :g_5: f_5^{tr} \\
+\;& \frac{1}{2} :g_3:^2 f_6^{tr} + \frac{1}{2} :f_5^{tr}::g_3: f_4^{tr} + :g_4::g_3: f_5^{tr} + \frac{1}{2} :g_5::g_3: f_4^{tr} + \frac{1}{3} :g_4:^2: f_4^{tr} \\
+\;& \frac{1}{6} :f_4^{tr}::g_4: f_4^{tr} - \frac{1}{6} :g_3:^3: f_5^{tr} - \frac{1}{12} :f_5^{tr}::f_3^{tr}::g_3: f_3^{tr} - \frac{1}{6} :g_3::f_5^{tr}::g_3: f_3^{tr} \\
-\;& \frac{1}{12} :g_5::f_3^{tr}::g_3: f_3^{tr} - \frac{1}{6} :g_5::g_3:^2 f_3^{tr} - \frac{1}{6}[:g_3: f_3^{tr}, :g_3: f_5^{tr}] \\
-\;& \frac{1}{6} :f_4^{tr}:^2: g_3: f_3^{tr} - \frac{1}{4} :g_3::f_4^{tr}::g_3: f_4^{tr} - \frac{1}{3} :g_4::f_4^{tr}::g_3: f_3^{tr} \\
-\;& \frac{1}{2} :g_4::g_3:^2 f_4^{tr} - \frac{1}{6}[:g_3: f_3^{tr}, :g_4: f_4^{tr}] - \frac{1}{6} :g_4:^2: g_3: f_3^{tr} \\
+\;& \frac{1}{24} :f_4^{tr}::f_3^{tr}:^2: g_3: f_3^{tr} + \frac{1}{8} :g_3::f_4^{tr}::f_3^{tr}::g_3: f_3^{tr} + \frac{1}{8} :g_3:^2: f_4^{tr}::g_3: f_3^{tr} \\
+\;& \frac{1}{24} :g_3:^4 f_4^{tr} - \frac{1}{24}[:g_3: f_4^{tr}, :f_3^{tr}::g_3: f_3^{tr}] - \frac{1}{12}[:g_3: f_4^{tr}, :g_3:^2 f_3^{tr}] \\
+\;& \frac{1}{24}[:g_3: f_3^{tr}, :f_4^{tr}::g_3: f_3^{tr}] + \frac{1}{8}[:g_3::f_3^{tr}:, :g_3:^2 f_4^{tr}] + \frac{1}{24} :g_4::f_3^{tr}:^2: g_3: f_3^{tr} \\
+\;& \frac{1}{8} :g_4::g_3::f_3^{tr}::g_3: f_3^{tr} + \ \frac{1}{8} :g_4::g_3:^3 f_3^{tr} : + \frac{1}{24}[:g_3: f_3^{tr}, :g_4::g_3: f_3^{tr}] \\
-\;& \frac{1}{720} :f_3^{tr}:^4: g_3: f_3^{tr} - \frac{1}{144} :g_3::f_3^{tr}:^3: g_3: f_3^{tr} - \frac{1}{144}[:g_3: f_3^{tr}, :f_3^{tr}:^2: g_3: f_3^{tr}] \\
-\;& \frac{1}{72} :g_3:^2: f_3^{tr}:^2: g_3: f_3^{tr} - \frac{1}{48}[:g_3: f_3^{tr}, :g_3::f_3^{tr}::g_3: f_3^{tr}] - \frac{1}{72} :g_3:^3: f_3^{tr}::g_3: f_3^{tr} \\
-\;& \frac{1}{48}[:g_3: f_3^{tr}, :g_3:^3 f_3^{tr}] - \frac{1}{144} :g_3:^5 f_3^{tr}. \qquad (8.4.36)
\end{aligned}
$$

Upon examining the expressions for h_4, h_5, h_6, etc. we see that they contain both what we will call *direct* terms and what we will call *feed-up* terms. For example, consider h_4 as given by (4.32). It contains the direct terms f_4^{tr} and g_4 which come from like terms in $\mathcal{M}_f$ and $\mathcal{M}_g$. It also contains the feed-up term $[f_3^{tr}, g_3]$ which comes from lower-order terms in $\mathcal{M}_f$ and $\mathcal{M}_g$. We see that low-order nonlinearities, when combined, can lead to higher-order nonlinearities.

There is also a way of getting relations such as (4.31) through (4.36) directly without use of the BCH formula. Suppose we expand all the exponentials appearing in (4.26). Doing so gives a relation of the form

$$\begin{aligned}&(1+:h_3:+:h_3:^2/2!+\cdots)(1+:h_4:+\cdots)\cdots\\&=(1+:f_3^{tr}:+:f_3^{tr}:^2/2!+\cdots)(1+:f_4^{tr}:+\cdots)\cdots\times\\&(1+:g_3:+:g_3:^2/2!+\cdots)(1+:g_4:+\cdots)\cdots\,.\end{aligned}\tag{8.4.37}$$

Next carry out the indicated multiplications and group terms according to the degree of the polynomial that would be produced if these terms were to act on z. We find the result

$$\begin{aligned}&1+:h_3:+(:h_3:^2/2!+:h_4:)+\cdots=1+(:f_3^{tr}:+:g_3:)\\&+(:f_3^{tr}::g_3:+:f_3^{tr}:^2/2!+:g_3:^2/2!+:f_4^{tr}:+:g_4:)+\cdots\,.\end{aligned}\tag{8.4.38}$$

Now equate terms of like degree to find results of the form

$$:h_3:=:f_3^{tr}:+:g_3:,\tag{8.4.39}$$

$$:h_3:^2/2!+:h_4:=:f_3^{tr}::g_3:+:f_3^{tr}:^2/2!+:g_3:^2/2!+:f_4^{tr}:+:g_4:\,.\tag{8.4.40}$$

Evidently (4.39) is equivalent to (4.31). Also, if we substitute (4.39) into (4.40) and rearrange terms, we find the result

$$\begin{aligned}:h_4:&=&&:f_4^{tr}:+:g_4:+:f_3^{tr}::g_3:-(1/2)(:f_3^{tr}::g_3:+:g_3::f_3^{tr}:)\\&=&&:f_4^{tr}:+:g_4:+(1/2)\{:f_3^{tr}:,:g_3:\}.\end{aligned}\tag{8.4.41}$$

We see, with the aid of (4.3.14), that (4.41) corresponds to (4.32). Similarly, if we retain more terms in (4.38), we can derive the relations (4.33) through (4.36), etc.

There is an equivalent but somewhat more elegant way of carrying out the same calculation. As before we expand all the exponentials appearing on the right side of (4.26); however, we retain the left side in factored product form. As a result we find the relation

$$\begin{aligned}&\exp(:h_3:)\exp(:h_4:)\cdots=(1+:f_3^{\rm tr}:+:f_3^{\rm tr}:^2/2!+\cdots)\times\\&(1+:f_4^{\rm tr}:+\cdots)\cdots\times(1+:g_3:+:g_3^2/2!+\cdots)(1+:g_4:+\cdots)\cdots\\&=1+(:f_3^{\rm tr}:+:g_3:)+(:f_3^{\rm tr}::g_3:+:f_3^{\rm tr}:^2/2!+:g_3:^2/2!+:f_4^{\rm tr}:+:g_4:)+\cdots\,.\end{aligned}\tag{8.4.42}$$

From (4.42) we infer, as before, the relation (4.39). But now we multiply both sides of (4.42) on the left by $\exp(-:h_3:)$ and make use (4.39) to find the result

$$\begin{aligned}&\exp(:h_4:)\cdots=\exp[-(:f_3^{\rm tr}:+:g_3:)]\times[1+(:f_3^{\rm tr}:+:g_3:)\\&+(:f_3^{\rm tr}::g_3:+:f_3^{\rm tr}:^2/2!+:g_3:^2/2!+:f_4^{\rm tr}:+:g_4:)+\cdots].\end{aligned}\tag{8.4.43}$$

Next we expand the exponential on the right side of (4.43) and carry out the indicated multiplications to get the relation

$$\exp(:h_4:)\cdots=1+[(:f_3^{\rm tr}::g_3:-:g_3::f_3^{\rm tr}:)/2+:f_4^{\rm tr}:+:g_4:]+\cdots\,.\tag{8.4.44}$$

At this point we are ready to repeat the process: From (4.44) we infer, again as before, the relation (4.41). Next we multiply both sides of (4.44) on the left by $\exp(-:h_4:)$, make use of (4.41), etc. This process is reminiscent of the factorization algorithm employed in Section 7.6. It is clear that it can be repeated indefinitely to find expressions for the $:h_m:$ for ever larger values of m. The only major problem (which also occurred before) is to write these expressions in commutator form so that the outer colons can be removed [using (5.3.14)] to obtain final results in the form (4.31) through (4.36), etc.

The problem of writing expressions in commutator form is solved by a result of *Dynkin*. Let $x_1, x_2, \cdots x_n$ be a collection of noncommuting variables. Suppose P is a polynomial in these variables of the form

$$P = \sum a_{i_1 i_2 \cdots i_k} x_{i_1} x_{i_2} \cdots x_{i_k}. \tag{8.4.45}$$

Note that each term contains k factors, not necessarily distinct, and therefore P is homogeneous of degree k. For each monomial $x_{i_1} x_{i_2} \cdots x_{i_k}$ form a related multiple commutator $(x_{i_1} x_{i_2} \cdots x_{i_k})^0$ by the rule

$$(x_{i_1} x_{i_2} \cdots x_{i_k})^0 = (1/k)\{\cdots\{x_{i_1}, x_{i_2}\}, x_{i_3}\}, \cdots x_{i_k}\}. \tag{8.4.46}$$

Also suppose it is known in principle that P can be written in terms of commutators (as is our situation thanks to the BCH formula). Then, Dynkin proved, *one* such commutator form for P is

$$P = \sum a_{i_1 i_2 \cdots i_k} (x_{i_1} x_{i_2} \cdots x_{i_k})^0. \tag{8.4.47}$$

Here it is helpful to work out an example. In the process of calculating h_5 based on (4.37) one finds the intermediate result

$$:h_5: = :f_5^{tr}: + :g_5: + :f_4^{tr}::g_3: - :g_3::f_4^{tr}: + P \tag{8.4.48}$$

with P given by the relation

$$\begin{aligned} P(:f_3^{tr}:, :g_3:) = & - :f_3^{tr}:^2:g_3:/6 + :f_3^{tr}::g_3::f_3^{tr}:/3 \\ & + :f_3^{tr}::g_3:^2/3 - :g_3::f_3^{tr}:^2/6 - (2/3):g_3::f_3^{tr}::g_3: \\ & + :g_3:^2:f_3^{tr}:/3. \end{aligned} \tag{8.4.49}$$

The problem is to put P in commutator form. Following (4.46) gives the results

$$\begin{gathered} (:f_3^{tr}:^2:g_3:)^0 = (1/3)\{\{:f_3^{tr}:, :f_3^{tr}:\}, :g_3:\} = 0, \\ (:f_3^{tr}::g_3::f_3^{tr}:)^0 = (1/3)\{\{:f_3^{tr}:, :g_3:\}, :f_3^{tr}:\}, \\ (:f_3^{tr}::g_3:^2)^0 = (1/3)\{\{:f_3^{tr}:, :g_3:\}, :g_3:\}, \\ (:g_3::f_3^{tr}:^2)^0 = (1/3)\{\{:g_3:, :f_3^{tr}:\}, :f_3^{tr}:\}, \\ (:g_3::f_3^{tr}::g_3:)^0 = (1/3)\{\{:g_3:, :f_3^{tr}:\}, :g_3:\}, \\ (:g_3:^2:f_3^{tr}:)^0 = (1/3)\{\{:g_3:, :g_3:\}, :f_3^{tr}:\} = 0. \end{gathered} \tag{8.4.50}$$

Consequently, according to Dynkin, P can be written in the commutator form

$$\begin{aligned} P = & (1/9)\{\{:f_3^{tr}:, :g_3:\}, :f_3^{tr}:\} + (1/9)\{\{:f_3^{tr}:, :g_3:\}, :g_3:\} \\ - & (1/18)\{\{:g_3:, :f_3^{tr}:\}, :f_3^{tr}:\} - (2/9)\{\{:g_3:, :f_3^{tr}:\}, :g_3:\}. \end{aligned} \tag{8.4.51}$$

Note that h_5 as given by (4.48) with P given by (4.51) bears only some resemblence to the h_5 in (4.33). However, repeated use of the antisymmetry condition (3.7.41) gives the results

$$\begin{gathered}\{\{:f_3^{tr}:,:g_3:\},:f_3^{tr}:\}=-\{:f_3^{tr}:,\{:f_3^{tr}:,:g_3:\}\},\\ \{\{:f_3^{tr}:,:g_3:\},:g_3:\}=\{:g_3:,\{:g_3:,:f_3^{tr}:\}\},\\ \{\{:g_3:,:f_3^{tr}:\},:f_3^{tr}:\}=\{:f_3^{tr}:,\{:f_3^{tr}:,:g_3:\}\},\\ \{\{:g_3:,:f_3^{tr}:\},:g_3:\}=-\{:g_3:,\{:g_3:,:f_3^{tr}:\}\}. \qquad (8.4.52)\end{gathered}$$

Inserting these results into P as given by (4.51) brings it to the form

$$P=(1/3)\{:g_3:,\{:g_3:,:f_3^{tr}:\}\}-(1/6)\{:f_3^{tr}:,\{:f_3^{tr}:,:g_3:\}\}, \qquad (8.4.53)$$

and h_5 as given by (4.48) with this P is the "colonized" version of (4.33).

This example illustrates both the use of Dynkin's theorem and a further complication. The complication is to determine when two expressions involving multiple Lie products are in fact equivalent when due account is taken of the antisymmetry condition (3.7.41) and/or the Jacobi condition (3.7.42). One method to compare expressions is to realize the Lie products in terms of commutators and then expand out all commutators to obtain a sum of monomials. If the two monomial sums agree term by term, then the two multiple Lie product expressions are equivalent. For example, expanding out P as given by (4.51) or (4.53) both produce (4.49). However, this expansion method is awkward since the expanded version may contain a very large number of terms. Another method is to employ some *basis* in standard form in which only certain multiple Lie product terms occur (with all other possible terms being brought to standard form by use of the antisymmetry and Jacobi conditions). Two multiple Lie product expressions are then equivalent if they agree term by term when re-expressed in some standard form basis. (Obviously their expanded commutator realizations will then also agree.) Two possible such bases are the *Hall* and *Chen-Fox-Lyndon-Shirshov* bases. See Appendix C.

There are yet another concerns when one is interested in numerical implementations. Because Lie multiplications (Poisson bracketing) are time consuming, it is desirable to minimize their number. For example, if only h_3 through h_5 were needed, the relations (4.32) and (4.33) could be rewritten and utilized in the form

$$h_4=f_4^{tr}+g_4-(1/2)[g_3,f_3^{tr}], \qquad (8.4.54)$$

$$h_5=f_5^{tr}+g_5+[g_3,-f_4^{tr}+(1/3)[g_3,f_3^{tr}]]+(1/6)[f_3^{tr},[g_3,f_3^{tr}]]. \qquad (8.4.55)$$

In this form a total of only three Poisson brackets is required. Also there is the sometimes conflicting desire to rearrange terms so that quantities already calculated can be reused to maximum benefit. Thus, the strategy might change if one wished to compute $h_6, h_7, \cdots$ as well. Finally, in actual practice, the quantities $g_3, g_4, \cdots$ may be *sparse* (most possible monomials in them having vanishing coefficients). In this case it is desirable to arrange the Poisson bracket terms in such a way that Poisson bracket routines designed to exploit sparseness can be employed. [The expansions (4.54) and (4.55) above have been arranged to exploit possible sparseness in g_3.] See Section 27.8.

We close this section by noting that there is yet another way of finding Lie concatenation formulas without needing the BCH coefficients, and it has the added advantage of immediately yielding results in Lie form. At present we do not have at our disposal all the tools required for its presentation. They will be developed in Chapter 10. See Section 10.5 where the subject of Lie concatenation is again discussed.

Exercises

8.4.1. Verify (4.16).

8.4.2. Verify (4.29) and (4.30).

8.4.3. Verify (4.31) through (4.33).

8.4.4. Starting with (4.37), verify (4.38) through (4.41).

8.4.5. Derive (4.48) and (4.49) from (4.37).

8.4.6. Expand out (4.51) and (4.53), and verify that both expansions produce (4.49).

8.4.7. Let us refer to a multiple commutator of the kind that appears in (4.46) as a *left nest*. Show that every left nest can be re-expressed as a *right nest*. That is show, by repeated use of the antisymmetry condition, that there is the relation

$$\begin{aligned}\{\cdots\{x_{i_1},x_{i_2}\},x_{i_3}\},\cdots x_{i_k}\} &= (-1)^{k-1}\{x_{i_k},\{x_{i_{k-1}},\{x_{i_{k-2}},\cdots\{x_{i_2},x_{i_1}\}\cdots\} \\ &= (-1)^{k-1}\#x_{i_k}\#\#x_{i_{k-1}}\#\ \#x_{i_{k-2}}\#\cdots\#x_{i_2}\#x_{i_1}. \qquad (8.4.56)\end{aligned}$$

8.5 Map Inversion and Reverse Factorization

Suppose the map $\mathcal{M}_f$ is written in the factored product form

$$\mathcal{M}_f = \mathcal{R}_f \exp(:f_3:)\exp(:f_4:)\cdots . \qquad (8.5.1)$$

Here, as in the previous section, $\mathcal{R}_f$ denotes the map

$$\mathcal{R}_f = \exp(:f_2^c:)\exp(:f_2^a:) \qquad (8.5.2)$$

that is associated with the linear transformation R^f given by the matrix relation

$$R^f = \exp(JS^a)\exp(JS^c). \qquad (8.5.3)$$

It follows immediately from (5.1) that the *inverse* of $\mathcal{M}_f$ has the representation

$$(\mathcal{M}_f)^{-1} = \cdots\exp(-:f_4:)\exp(-:f_3:)(\mathcal{R}_f)^{-1}. \qquad (8.5.4)$$

Although (5.4) gives a possible representation for the inverse of $\mathcal{M}_f$, it is in the form of a *reverse* factorization. We would also like to have a representation in the standard *forward* factorization. That is, we wish also to have a representation for the inverse of $\mathcal{M}_f$ in the form

$$(\mathcal{M}_f)^{-1} = \mathcal{R}_h \exp(: h_3 :) \exp(: h_4 :) \cdots . \tag{8.5.5}$$

See Section 7.8. This is easily accomplished with the aid of the concatenation formulas of the previous section. We simply write (5.4) and (5.5) in the form

$$\cdots [\exp(- : f_4 :)][\exp(- : f_3 :)][(\mathcal{R}_f)^{-1}] = \mathcal{R}_h \exp(: h_3 :) \exp(: h_4 :) \cdots \tag{8.5.6}$$

where we have used square brackets to indicate that the various maps are to be concatenated together. See Exercise 5.1. In particular, as needed in the next paragraph, we have the results

$$\mathcal{R}_h = (\mathcal{R}_f)^{-1}, \tag{8.5.7}$$

$$R^h = (R^f)^{-1}. \tag{8.5.8}$$

Note again, as a result of the symplectic condition, that the matrix $(R^f)^{-1}$ is easily calculated using (3.1.9).

The relation (5.5) also provides a procedure for reverse factorizing a map. Suppose we wish to represent $\mathcal{M}_f$ in reverse factorized form. That is, we wish to find generators g_m such that

$$\mathcal{M}_f = \mathcal{R}_f \exp(: f_3 :) \exp(: f_4 :) \cdots = \cdots \exp(: g_4 :) \exp(: g_3 :) \mathcal{R}_g. \tag{8.5.9}$$

Simply take the inverse of both sides of (5.9) and use (5.5) to get the relation

$$(\mathcal{R}_g)^{-1} \exp(- : g_3 :) \exp(- : g_4 :) \cdots = \mathcal{R}_h \exp(: h_3 :) \exp(: h_4 :) \cdots . \tag{8.5.10}$$

From (5.10) and (5.7) we find the desired results

$$\mathcal{R}_g = (\mathcal{R}_h)^{-1} = \mathcal{R}_f, \tag{8.5.11}$$

$$g_m = -h_m. \tag{8.5.12}$$

Exercises

8.5.1. Verify (5.7) and (5.8). Show that h_3, h_4, and h_5 are given by the formulas

$$h_3 =, \tag{8.5.13}$$

$$h_4 =, \tag{8.5.14}$$

$$h_5 = . \tag{8.5.15}$$

8.5.2. Verify (5.11) and (5.12).

8.6 Taylor and Hybrid Taylor-Lie Concatenation and Inversion

Section 8.4 treated the problem of concatenating two maps, both of which were in factored-product Lie form, to obtain their product, again in factored-product Lie form. We also know that maps can be written in Taylor form. See Section 7.5. For some applications it is useful to have concatenation procedures for which one or more of the maps is in Taylor form. Several possibilities arise, as illustrated in Figure 6.1. Of course, we can always pass back and forth between the Taylor and factored-product Lie forms (see Section 7.6 and Exercise 7.6.12) so that in principle we already have all needed results. However, it is also desirable to have procedures that work directly with Taylor maps. Four cases of particular interest are discussed below.

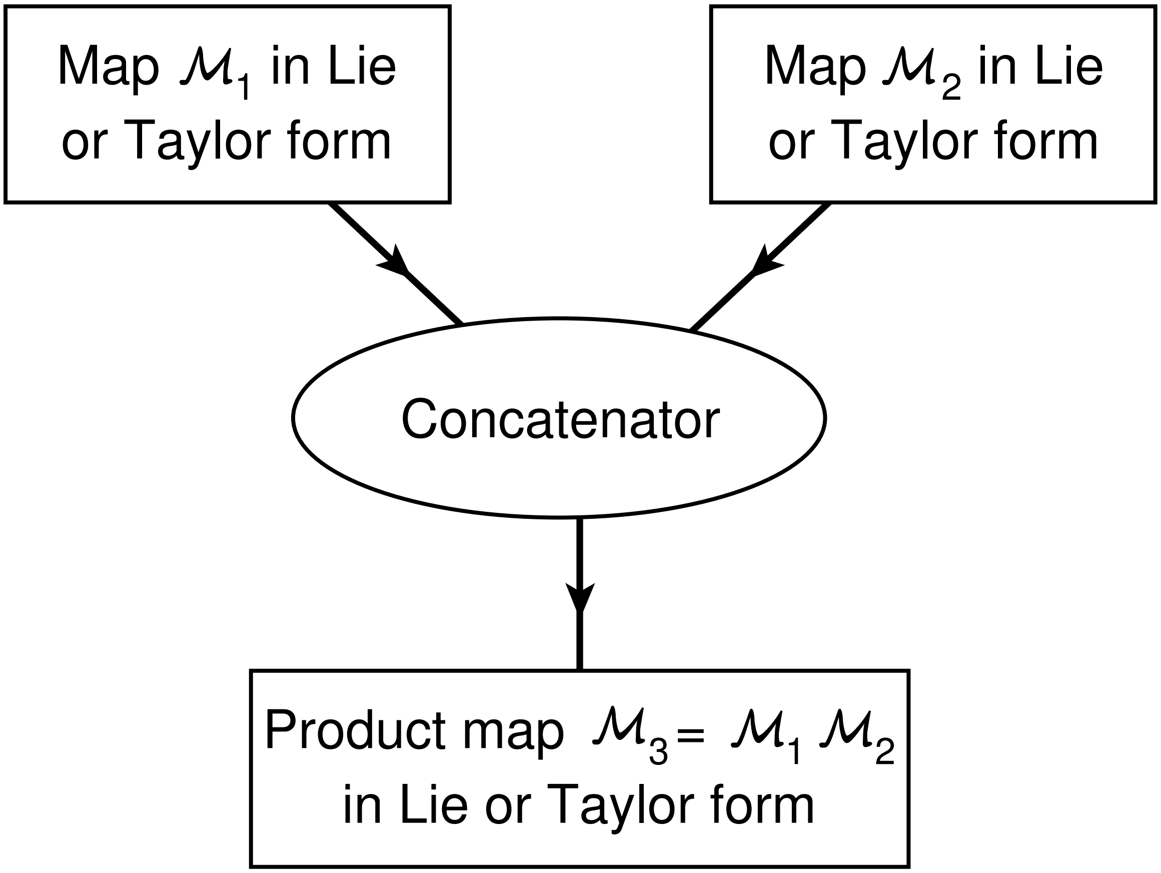

Map $\mathcal{M}_1$ in Lie or Taylor form
Map $\mathcal{M}_2$ in Lie or Taylor form
Concatenator
Product map $\mathcal{M}_3 = \mathcal{M}_1 \mathcal{M}_2$ in Lie or Taylor form

Figure 8.6.1: Various possibilities for the representation of maps in the operation of concatenation.

Let us begin with the case where both $\mathcal{M}_1$ and $\mathcal{M}_2$ are in Taylor form, and we desire as well to represent the product $\mathcal{M}_3 = \mathcal{M}_1\mathcal{M}_2$ in Taylor form. Suppose that $\mathcal{M}_1$ sends z to $\overline{z}$, and we express this fact in the form of a Taylor series that is truncated beyond terms of degree D,

$$\mathcal{M}_1 : z \to \overline{z} \tag{8.6.1}$$

with

$$\bar{z}_a = \bar{z}_a(z) = \sum_{m=1}^{D} g_a^1(m; z). \tag{8.6.2}$$

Here the $g_a^1(m; z)$ denote homogeneous polynomials of degree m in the variables z. Similarly, $\mathcal{M}_2$ sends $\bar{z}$ to $\bar{\bar{z}}$,

$$\mathcal{M}_2 : \bar{z} \rightarrow \bar{\bar{z}} \tag{8.6.3}$$

with

$$\bar{\bar{z}}_a = \bar{\bar{z}}_a(\bar{z}) = \sum_{m'=1}^{D} g_a^2(m'; \bar{z}). \tag{8.6.4}$$

What we desire is a representation for $\mathcal{M}_3$ of the form

$$\mathcal{M}_3 : z \rightarrow \bar{\bar{z}} \tag{8.6.5}$$

with

$$\bar{\bar{z}}_a = \bar{\bar{z}}_a(z) = \sum_{m''=1}^{D} g_a^3(m''; z). \tag{8.6.6}$$

Upon comparing (6.4) and (6.6) we see that the polynomials g_a^3 are given by the relations

$$g_a^3(m''; z) = P_{m''} \sum_{m'=1}^{D} g_a^2(m'; \bar{z}(z)). \tag{8.6.7}$$

Here $P_{m''}$ denotes a *projection* operator that retains only terms of degree m'' in the variables z.

To verify the truth of (6.7), we observe that the quantities $g_a^2(m', \bar{z})$ in (6.4) are linear combinations of monomials in the $\bar{z}$'s of degree m'. When these monomials are computed using (6.2), the results are linear combinations of monomials in the z's of degree as high as $m'D$. For example, second-order monomials in the $\bar{z}$'s are given by the relation

$$\bar{z}_c \bar{z}_d = \sum_{m'=1}^{D} g_c^1(m'; z) \sum_{m''=1}^{D} g_d^1(m''; z). \tag{8.6.8}$$

From these monomials we need to extract the terms of degree m in the z's in order to find their contribution to the $g_a^3(m; z)$,

$$P_m(\bar{z}_c \bar{z}_d) = \sum_{m'+m''=m} g_c^1(m'; z) g_d^1(m''; z). \tag{8.6.9}$$

We conclude that the operation of concatenating maps in Taylor form involves the multiplication of truncated Taylor series, the extraction of terms of various degrees from the resulting products, and the assembly of linear combinations of these terms to form the truncated Taylor expansion (6.6) for the resulting map $\mathcal{M}_3$. All these operations can in principle be carried out in a straight-forward manner to arbitrary order on a computer using various algorithms for *Truncated Power Series Algebra* (TPSA).

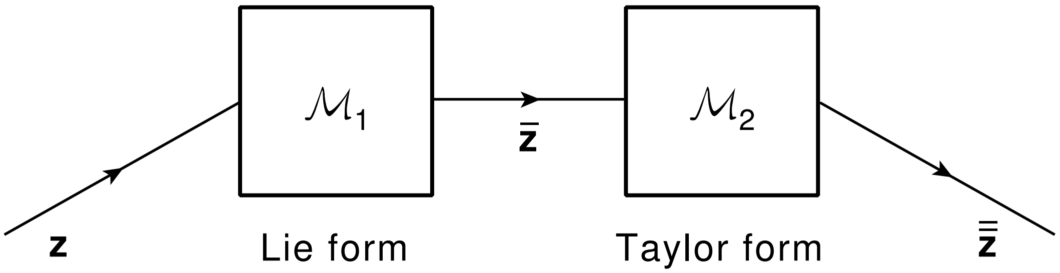


Figure 8.6.2: Product of a map in Lie form with a map in Taylor form.

In a second important case $\mathcal{M}_1$ is in Lie form, $\mathcal{M}_2$ is in Taylor form, and we desire or are content to know their product in Taylor form. See Figure 6.2. According to (6.4) we can express the action of $\mathcal{M}_2$ in Taylor form by writing the relations

$$\bar{\bar{z}}_a\,(\bar{z}) = T_a^D(\bar{z}) \tag{8.6.10}$$

where

$$T_a^D(\bar{z}) = \sum_{m'=1}^{D} g_a^2(m';\bar{z}). \tag{8.6.11}$$

We also have the relations

$$\bar{z}_a(z) = \mathcal{M}_1 z, \tag{8.6.12}$$

$$g_a^2(m';\bar{z}(z)) = g_a^2(m';\mathcal{M}_1 z) = \mathcal{M}_1 g_a^2(m';z). \tag{8.6.13}$$

Here we have used (5.4.13). Consequently, we have the result

$$\bar{\bar{z}}_a\,(z) = \mathcal{M}_1 T_a^D(z). \tag{8.6.14}$$

At this point we recognize that there are three common ways that $\mathcal{M}_1$ may be specified in Lie form. First, suppose that $\mathcal{M}_1$ is given in terms of a single exponent,

$$\mathcal{M}_1 = \exp(:h:), \tag{8.6.15}$$

where h has a homogeneous polynomial expansion of the form

$$h = h_2 + h_3 + \cdots + h_{D+1}. \tag{8.6.16}$$

[Note that, consistent with truncating maps beyond terms of degree D, we have truncated h beyond terms of degree $(D+1)$.] Maps of this kind arise from autonomous systems. See Sections 7.4 and 10.5. In this case we may expand exp$(:h:)$ to get the result

$$\bar{\bar{z}}_a\,(z) = \sum_{\ell=0}^{\infty}(1/\ell!) :h:^{\ell}\, T_a^D(z). \tag{8.6.17}$$

Correspondingly, we have from (6.6) the result

$$g_a^3(m,z) = P_m \sum_{m'=1}^{D}\sum_{\ell=0}^{\infty}(1/\ell!) :h:^{\ell}\, g_a^2(m';z). \tag{8.6.18}$$

In the circumstance that $h_2 = 0$, each sum over ℓ (for a given m and m') reduces to a finite sum because of (7.6.16). In the case that h_2 does not vanish, an infinite sum (for each value of m') is generally required. It can be shown, by an argument similar to that given in Section 10.5, that these sums always converge thanks to the $(1/\ell!)$ factor. However, all the caveats described in Section 4.1 concerning the use of Taylor series to evaluate the exponential function also apply here.

Suppose, as a second possibility, that $\mathcal{M}_1$ is given in the factored product form

$$\mathcal{M}_1 = \mathcal{R}_f \exp(:f_3:) \exp(:f_4:) \cdots \exp(:f_{D+1}:). \tag{8.6.19}$$

Let $\mathcal{N}_f$ be the nonlinear part of $\mathcal{M}_1$,

$$\mathcal{N}_f = \exp(:f_3:) \exp(:f_4:) \cdots \exp(:f_{D+1}:). \tag{8.6.20}$$

According to (6.14) we need to find the quantities

$$\mathcal{M}_1 T_a^D(z) = \mathcal{R}_f \mathcal{N}_f T_a^D(z). \tag{8.6.21}$$

Introduce the intermediate results $\tilde{T}_a^D$ and $\tilde{g}_a^3(m; z)$ defined by the equations

$$\tilde{T}_a^D(z) = \mathcal{N}_f T_a^D(z), \tag{8.6.22}$$

$$\tilde{g}_a^3(m; z) = P_m \sum_{m'=1}^{D} \mathcal{N}_f g_a^2(m'; z). \tag{8.6.23}$$

Then, by construction, we have the relation

$$\tilde{T}_a^D(z) = \sum_{m=1}^{D} \tilde{g}_a^3(m; z). \tag{8.6.24}$$

Next expand the exponentials appearing in (6.20). Doing so brings (6.23) to the form

$$\tilde{g}_a^3(m; z) = P_m \sum_{m'=1}^{D} \sum_{\ell_3=0}^{\infty} (1/\ell_3!) :f_3:^{\ell_3} \cdots \sum_{\ell_{D+1}=0}^{\infty} (1/\ell_{D+1}!) :f_{D+1}:^{\ell_{D+1}} g_a^2(m'; z). \tag{8.6.25}$$

Again because of (7.6.16), each of the sums over $\ell_3 \cdots \ell_{D+1}$ (for a given m and m') reduces to a finite sum. The remaining task is to take $\mathcal{R}_f$ into account. This is easily done. Again by construction we have the relation

$$g_a^3(m; z) = \mathcal{R}_f \tilde{g}_a^3(m; z). \tag{8.6.26}$$

Now use the analog of (8.4.15) to get the final result

$$g_a^3(m; z) = \tilde{g}_a^3(m; R^f z). \tag{8.6.27}$$

A third common possibility is that $\mathcal{M}_1$ arises in Lie form as a result of the use of some kind of Zassenhauss (symplectic integration) approximation. In this case $\mathcal{M}_1$ is typically a product of Lie transformations of the form

$$\mathcal{M}_1 = \exp(w_1 h :A:) \exp(w_2 h :B:) \cdots \exp(w_m h :A:) \tag{8.6.28}$$

where the w_j are various weights and h is the integration step size. See Section 10.8. Here the function A is typically a second-degree polynomial, and the function B has a homogeneous polynomial expansion consisting of terms of degree three and higher. If A is a second-degree polynomial, then $\exp(w_j h : A :)$ is a linear transformation that can be represented by some matrix R, and we can use methods analogous to (6.26) and (6.27). If B consists only of terms of degree three and higher, then $\exp(w_j h : B :)$ can be expanded in a Taylor series to give results analogous to (6.18) and for which only a finite number of ℓ values contribute.

We have seen how to find, in Taylor form, the product of a map in Lie form with a second map in Taylor form. Consider, as case three, the situation in which the two maps to be multiplied are both in Lie form, and we want their product in Taylor form. An obvious approach is to convert the second map from Lie to Taylor form, and then proceed as just described. This conversion is easily carried out as in Exercise 7.6.12. Equivalently, we may use the machinery just developed. The map $\mathcal{M}_2$ can always be written as

$$\mathcal{M}_2 = \mathcal{M}_2 \mathcal{I} \tag{8.6.29}$$

where $\mathcal{I}$ is the identity map. But the identity map has the immediate Taylor expansion

$$\mathcal{I} z_a = z_a. \tag{8.6.30}$$

Therefore, to find $\mathcal{M}_2$ in Taylor form, we simply concatenate $\mathcal{M}_2$ in Lie form with the identity map $\mathcal{I}$ in Taylor form.

As a generalization of this approach, suppose we wish to concatenate m maps in Lie form and obtain the net result in Taylor form,

$$\mathcal{M}_{\rm net} = \mathcal{M}_1 \mathcal{M}_2 \mathcal{M}_3 \cdots \mathcal{M}_m. \tag{8.6.31}$$

Rewrite the desired result in the form

$$\mathcal{M}_{\rm net} = \mathcal{M}_1(\mathcal{M}_2(\mathcal{M}_3 \cdots (\mathcal{M}_m \mathcal{I}) \cdots)), \tag{8.6.32}$$

and observe that each map $\mathcal{M}_j$ in Lie form is now to be concatenated with a map in Taylor form to produce a map again in Taylor form. Thus, after m concatenations [namely, $\mathcal{M}_m \mathcal{I}$, $\mathcal{M}_{m-1}(\mathcal{M}_m \mathcal{I})$, $\mathcal{M}_{m-2}(\mathcal{M}_{m-1}(\mathcal{M}_m \mathcal{I}))$, etc.] we obtain $\mathcal{M}_{\rm net}$ in Taylor form. At this stage we may, if desired, obtain $\mathcal{M}_{\rm net}$ in Lie form from $\mathcal{M}_{\rm net}$ in Taylor form by carrying out the steps of the Factorization Theorem of Section 7.6.

A fourth case of interest is that in which the two maps to be multiplied are both in Lie form, and we also want their product in Lie form. This case has already been discussed in Section 8.4 where the BCH formula was used to find the quantities $h_3, h_4, \cdots$ in the relation (4.26). The result was explicit formulas of the form (4.31) through (4.36). These formulas become ever more complicated as the order is increased.

If one is content with numerical results, which is often the case, then the h_m can be computed algorithmically for any order m, without recourse to the BCH series, by Taylor methods as described above. To be explicit, and with reference to (4.26), define variables $\overline{z}_a(z)$ by the relations

$$\overline{z}_a = \exp(: f_3^{tr} :) \exp(: f_4^{tr} :) \cdots \exp(: g_3 :) \exp(: g_4 :) \cdots z_a. \tag{8.6.33}$$

Then we have the result

$$\exp(:h_3:)\exp(:h_4:)\cdots z_a = \overline{z}_a(z). \tag{8.6.34}$$

Next let $\mathcal{T}^D$ be a *truncation* operator that acts on Taylor series. It is defined to be a linear operator that retains all terms in a Taylor series of degree less than or equal to D, and discards all terms of degree greater than D. With the aid of this operator we may define truncated Taylor series $T_a^D(z)$ by the relation

$$T_a^D(z) = \mathcal{T}^D\overline{z}_a = \mathcal{T}^D\exp(:f_3^{tr}:)\exp(:f_4^{tr}:)\cdots\exp(:g_3:)\exp(:g_4:)\cdots z_a. \tag{8.6.35}$$

Evidently, in view of (7.6.14), to compute the T_a^D it is only necessary to retain the factors containing $f_3^{tr}\cdots f_{D+1}^{tr}$ and $g_3\cdots g_{D+1}$ in (6.35). Moreover, only a finite number of terms need be retained in each exponential series. Therefore, for a fixed D, only a finite number of operations are required to evaluate (6.35) to find the T_a^D. Finally, examination of the proof of the Factorization Theorem of Section 7.6 shows that the desired quantities $h_3\cdots h_{D+1}$ can be found from the T_a^D by a finite number of operations. Moreover, unlike the case of the BCH series whose coefficients are very complicated, the only coefficients required are simply the factorials in the exponential series and those that arise in the use of (7.6.24).

In summary the virtue of the Taylor method just described is that, even when the maps to be concatenated are both in Lie factored product form and the desired product is also required in this form, results can be obtained to any desired degree $(D+1)$ in the Lie generators by means of a relatively simple algorithm. The price to be paid for this simplicity is increased computation [compared to that required for the direct formulas of the form (4.31) through (4.36)] and the increased (but temporary) storage associated with the intermediate truncated Taylor series T_a^D (see Section 7.9).

The last topic to be discussed in this section is the inversion of maps in Taylor form. By way of introduction, consider the simple quadratic equation

$$\overline{x} = \alpha x + \beta x^2. \tag{8.6.36}$$

This equation can immediately be solved to find x in terms of $\overline{x}$,

$$x = \{-\alpha \pm [\alpha^2 + 4\beta\overline{x}]^{1/2}\}/(2\beta). \tag{8.6.37}$$

The solution that vanishes when $\overline{x}=0$ has the expansion

$$\begin{aligned} x &= \{-\alpha + [\alpha^2+4\beta\overline{x}]^{1/2}\}/(2\beta) \\ &= [\alpha/(2\beta)]\{-1+[1+4\beta\overline{x}/\alpha^2]^{1/2}\} \\ &= [\alpha/(2\beta)]\{(1/2)(4\beta\overline{x}/\alpha^2) - (1/8)(4\beta\overline{x}/\alpha^2)^2 + (1/16)(4\beta\overline{x}/\alpha^2)^3 + \cdots\} \\ &= (1/\alpha)\overline{x} - (\beta/\alpha^3)\overline{x}^2 + (2\beta^2/\alpha^5)\overline{x}^3 + \cdots\ . \end{aligned} \tag{8.6.38}$$

Equation (6.36) may be viewed as a one-dimensional Taylor map that sends x to $\overline{x}$, and (6.38) is the Taylor expansion of its inverse.

Suppose we had not been able to solve (6.36) explicitly for x as in (6.37). Is there any other way to obtain the inverse series (6.38)? The answer is yes. The inverse series can also be found by a process of *recursion* or *iteration*: First rewrite (6.36) in the form

$$x = (1/\alpha)\overline{x} - (\beta/\alpha)x^2 = (1/\alpha)\overline{x} + n(x) \tag{8.6.39}$$

where $n(x)$ is the *nonlinear* term

$$n(x) = -(\beta/\alpha)x^2. \tag{8.6.40}$$

Now consider the recursion relation for functions $x^{(m)}(\overline{x})$ specified by the rule

$$x^{(m+1)}(\overline{x}) = (1/\alpha)\overline{x} + n[x^{(m)}(\overline{x})], \tag{8.6.41}$$

with the starting relation

$$x^{(1)}(\overline{x}) = (1/\alpha)\overline{x}. \tag{8.6.42}$$

Upon carrying out the indicated operations, we find the results

$$\begin{aligned} m=1: \quad x^{(1)} &= (1/\alpha)\overline{x}, \\ m=2: \quad x^{(2)} &= (1/\alpha)\overline{x} - (\beta/\alpha)[(1/\alpha)\overline{x}]^2 = (1/\alpha)\overline{x} - (\beta/\alpha^3)\overline{x}^2, \\ m=3: \quad x^{(3)} &= (1/\alpha)\overline{x} - (\beta/\alpha)[(1/\alpha)\overline{x} - (\beta/\alpha^3)\overline{x}^2]^2 \\ &= (1/\alpha)\overline{x} - (\beta/\alpha^3)\overline{x}^2 + (2\beta^2/\alpha^2)\overline{x}^3 + O(\overline{x}^4), \text{ etc.} \end{aligned} \tag{8.6.43}$$

Evidently m applications of the rule (6.41) reproduces the series (6.38) through terms of degree m. We also note the possible appearance, at each stage, of still higher degree terms that may not yet be correct. We may remind ourselves not to bother computing these terms by using the truncation operator $\mathcal{T}^m$. With the aid of this operator, the recursion relation (6.41) can be modified to take the more convenient form

$$x^{(m+1)}(\overline{x}) = (1/\alpha)\overline{x} + \mathcal{T}^{m+1}n[x^{(m)}(\overline{x})]. \tag{8.6.44}$$

The iteration method we have just used to invert the simple quadratic equation (6.34) can also be used to invert general Taylor maps. Let us rewrite the Taylor representation (6.2) for the map $\mathcal{M}_1$ in the form

$$\overline{z}_{b'}(z) = \sum_b R_{b'b}z_b + N_{b'}(z) \tag{8.6.45}$$

where the quantities $N_{b'}(z)$ are nonlinear terms of degree 2 and higher. Equation (6.43) can be partially solved to give the result

$$z_a = (R^{-1}\overline{z})_a + \tilde{N}_a(z)$$

where $\tilde{N}_a$ also contains terms only of degree 2 and higher, and is given by the relation

$$\tilde{N}_a = -\sum_{b'}(R^{-1})_{ab'}N_{b'}. \tag{8.6.46}$$

Note that we have assumed R^{-1} exists, as is required by the inverse function theorem for a map to have an inverse, and as will be the case for symplectic matrices. Now form the recursion relation

$$z_a^{(m+1)}(\overline{z}) = (R^{-1}\overline{z})_a + \mathcal{T}^{m+1}\tilde{N}_a[z^{(m)}(\overline{z})] \tag{8.6.47}$$

with the starting relation

$$z_a^{(1)}(\overline{z}) = (R^{-1}\overline{z})_a. \tag{8.6.48}$$

Application of this recursion relation D times produces the Taylor representation, through terms of degree D, for the map $\mathcal{M}_1^{-1}$.

Finally, we remark that the operations needed to carry out the recursion relation (6.47), as well as the Poisson brackets needed in procedures such as (6.18) and (6.35), can all be performed to arbitrary order on a computer programmed to handle TPSA.

Exercises

8.6.1. According to (4.31) and (4.32) the *direct* determination of h_3 and h_4 requires the computation of *one* Poisson bracket. How many Poisson brackets must be computed to find h_3 and h_4 by Taylor methods? Compare the amounts of work required for the direct and Taylor methods.

8.6.2. Prove that use of the recursion relation (6.47) does indeed produce the Taylor representation of the inverse of (6.45).

8.7 Working with Exponents

8.7.1 Formulas for Combining Exponents

The General Case

Sometimes, as will be shown later, it is useful to be able to write the product of two Lie transformations as a *single* Lie transformation. This is what the BCH formula (2.27) attempts to do. In general, there are no known convenient expressions for all the terms on the right side of (2.29). However, it is possible to sum the series completely with respect to s and the first few powers in t. One such result can be written in the form

$$h = sf + s : f : [1 - \exp(-s : f :)]^{-1}(tg) + O(t^2). \tag{8.7.1}$$

Here the operator expression involving $: f :$ is to be interpreted as the infinite series

$$s : f : [1 - \exp(-s : f :)]^{-1} = s : f : [1 - \sum_{m=0}^{\infty}(-s : f :)^m/m!]^{-1}$$

$$= s : f : [-\sum_{m=1}^{\infty}(-s : f :)^m/m!]^{-1} = 1 + (s/2) : f : +(s^2/12) : f :^2 + \cdots . \tag{8.7.2}$$

Equations (2.27) and (7.1) may be combined to give the result

$$\exp(s : f :)\exp(t : g :) = \exp[s : f : + : \{s : f : [1 - \exp(-s : f :)]^{-1}(tg)\} : +O(t^2)]. \tag{8.7.3}$$

See Appendix C where the $O(t^2)$ term is also worked out.

Suppose we succeed in writing a product of two or more Lie transformations as a single Lie transformation. Then, as shown in Section 7.1, the map corresponding to the product of Lie transformations has an invariant function. See (7.1.12) through (7.1.14). We will learn later that generically symplectic maps do *not* have invariant functions. Correspondingly, the series (7.1) is generally divergent. We recall from Section 7.7 that the Lie algebra $spm(2n, \mathbb{R})$ is *infinite* dimensional. Typically what happens in the infinite dimensional case is that inverses of the form $[1 - \exp(-s : f :)]^{-1}$ may fail to exist. Other difficulties can also arise. Put another way, the BCH series (3.7.34) may have no domain of convergence in the case of an infinite dimensional Lie algebra. See Section 38.7.

The Case of $Sp(2)$

There is one instructive case for which the sum of the BCH series is known exactly. That is the case of $Sp(2)$ or, more generally, $Sp(2, \mathbb{C})$= $SL(2, \mathbb{C})$. We will see that the result is quite complicated. Presumably yet more complicated formulas exist in the Platonic realm for the still more interesting cases of $Sp(4)$, $Sp(6)$, etc. But, to the author's knowledge, these formulas have not yet been brought down to Earth.

Given f_2 and g_2, there are the associated maps

$$\mathcal{M}_f = \exp(: f_2 :), \tag{8.7.4}$$

$$\mathcal{M}_g = \exp(:g_2:). \tag{8.7.5}$$

Our task is to find h_2 such that

$$\mathcal{M}_h = \exp(:h_2:) = \mathcal{M}_f \mathcal{M}_g. \tag{8.7.6}$$

According to (3.64), (3.65), and (3.76) there are symmetric matrices S^f, S^g, and S^h associated with f_2, g_2, and h_2 respectively. Also, according to Section 8.3, our task is equivalent to that of finding the matrix S^h such that

$$\exp(JS^h) = \exp(JS^f)\exp(JS^g). \tag{8.7.7}$$

In the case of $sp(2)$, we know that the vector space of matrices of the form JS is spanned by the matrices B^0, F, and G given by (5.6.7), (5.6.13), and (5.6.14). Upon comparison with the Pauli matrices (5.7.3), we find the results

$$B^0 = i\sigma^2, \tag{8.7.8}$$

$$F = -\sigma^1, \tag{8.7.9}$$

$$G = \sigma^3. \tag{8.7.10}$$

Let us define 3-vectors (with possibly complex components) $\boldsymbol{v}^f$, $\boldsymbol{v}^g$, and $\boldsymbol{v}^h$ by the rules

$$\boldsymbol{v}^f \cdot \boldsymbol{\sigma} = v_1^f\sigma^1 + v_2^f\sigma^2 + v_3^f\sigma^3 = JS^f, \text{ etc.} \tag{8.7.11}$$

With these results and definitions, the condition (7.7) in the $Sp(2)$ case is equivalent to the requirement

$$\exp(\boldsymbol{v}^h \cdot \boldsymbol{\sigma}) = \exp(\boldsymbol{v}^f \cdot \boldsymbol{\sigma})\exp(\boldsymbol{v}^g \cdot \boldsymbol{\sigma}). \tag{8.7.12}$$

The matrix $\exp(\boldsymbol{v}^h \cdot \boldsymbol{\sigma})$ can be found analytically using (5.7.40). We begin by noting that (5.7.40) can be written in the form

$$(\boldsymbol{u} \cdot \boldsymbol{\sigma})(\boldsymbol{v} \cdot \boldsymbol{\sigma}) = \sigma^0 \boldsymbol{u} \cdot \boldsymbol{v} + i(\boldsymbol{u} \times \boldsymbol{v}) \cdot \boldsymbol{\sigma}, \tag{8.7.13}$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ are any 3-vectors. Next, define the length of $\boldsymbol{v}$, denoted by v, by the rule

$$v = (\boldsymbol{v} \cdot \boldsymbol{v})^{1/2}. \tag{8.7.14}$$

Note that v may possibly be complex, and is specified only up to a sign. Using (7.13) and (7.14), we find the result

$$\begin{aligned} \exp(\boldsymbol{v} \cdot \boldsymbol{\sigma}) &= \cosh(\boldsymbol{v} \cdot \boldsymbol{\sigma}) + \sinh(\boldsymbol{v} \cdot \boldsymbol{\sigma}) \\ &= \sigma^0 \cosh v + \boldsymbol{v} \cdot \boldsymbol{\sigma}(\sinh v)/v. \end{aligned} \tag{8.7.15}$$

We observe that both the functions $\cosh v$ and $(\sinh v)/v$ are even in v, and therefore unaffected by the sign ambiguity in (7.14). At this point it is convenient to introduce the 3-vector $\boldsymbol{\tau}(\boldsymbol{v})$ defined by the equation

$$\boldsymbol{\tau}(\boldsymbol{v}) = \boldsymbol{v}(\tanh v)/v. \tag{8.7.16}$$

Equation (7.16) has as its inverse the relation

$$\boldsymbol{v}(\boldsymbol{\tau}) = \boldsymbol{\tau}(\tanh^{-1}\tau)/\tau \tag{8.7.17}$$

where

$$\tau = (\boldsymbol{\tau}\cdot\boldsymbol{\tau})^{1/2}. \tag{8.7.18}$$

Again observe that $(\tanh v)/v$ and $(\tanh^{-1}\tau)/\tau$ are even functions. With this definition, (7.15) can be written in the equivalent form

$$\exp(\boldsymbol{v}\cdot\boldsymbol{\sigma}) = [\cosh v][\sigma^0 + \boldsymbol{\tau}(\boldsymbol{v})\cdot\boldsymbol{\sigma}]. \tag{8.7.19}$$

Now use (7.19) and (7.13) in (7.12). Doing so gives the result

$$\begin{aligned} & \exp(\boldsymbol{v}^h\cdot\boldsymbol{\sigma}) = [\cosh v^h][\sigma^0 + \boldsymbol{\tau}(\boldsymbol{v}^h)\cdot\boldsymbol{\sigma}] \\ = \ & [\cosh v^f][\sigma^0 + \boldsymbol{\tau}(\boldsymbol{v}^f)\cdot\boldsymbol{\sigma}][\cosh v^g][\sigma^0 + \boldsymbol{\tau}(\boldsymbol{v}^g)\cdot\boldsymbol{\sigma}] \\ = \ & [\cosh v^f][\cosh v^g]\{\sigma^0[1 + \boldsymbol{\tau}(\boldsymbol{v}^f)\cdot\boldsymbol{\tau}(\boldsymbol{v}^g)] \\ + \ & [\boldsymbol{\tau}(\boldsymbol{v}^f) + \boldsymbol{\tau}(\boldsymbol{v}^g) + i\boldsymbol{\tau}(\boldsymbol{v}^f)\times\boldsymbol{\tau}(\boldsymbol{v}^g)]\cdot\boldsymbol{\sigma}\}. \end{aligned} \tag{8.7.20}$$

Use (5.7.41) to equate like terms on both sides of (7.20), and thereby find the relations

$$\cosh v^h = [\cosh v^f][\cosh v^g][1 + \boldsymbol{\tau}(\boldsymbol{v}^f)\cdot\boldsymbol{\tau}(\boldsymbol{v}^g)], \tag{8.7.21}$$

$$(\cosh v^h)\boldsymbol{\tau}(\boldsymbol{v}^h) = [\cosh v^f][\cosh v^g][\boldsymbol{\tau}(\boldsymbol{v}^f) + \boldsymbol{\tau}(\boldsymbol{v}^g) + i\boldsymbol{\tau}(\boldsymbol{v}^f)\times\boldsymbol{\tau}(\boldsymbol{v}^g)]. \tag{8.7.22}$$

Upon dividing (7.22) by (7.21) we obtain the final and remarkable result

$$\boldsymbol{\tau}(\boldsymbol{v}^h) = [\boldsymbol{\tau}(\boldsymbol{v}^f) + \boldsymbol{\tau}(\boldsymbol{v}^g) + i\boldsymbol{\tau}(\boldsymbol{v}^f)\times\boldsymbol{\tau}(\boldsymbol{v}^g)][1 + \boldsymbol{\tau}(\boldsymbol{v}^f)\cdot\boldsymbol{\tau}(\boldsymbol{v}^g)]^{-1}. \tag{8.7.23}$$

Given $\boldsymbol{v}^f$ and $\boldsymbol{v}^g$, (7.23) specifies $\boldsymbol{\tau}(\boldsymbol{v}^h)$ which, in turn by using (7.17), gives $\boldsymbol{v}^h$. Taken together, we will call (7.16), (7.17), and (7.23) the $Sp(2,\mathbb{C})$ BCH *function*.

Let us examine the singularity structure of the $Sp(2,\mathbb{C})$ BCH *function*, namely the relationship between $\boldsymbol{v}^f$, $\boldsymbol{v}^g$, and $\boldsymbol{v}^h$ implied by (7.16), (7.17), and (7.23). We see from (7.16) that $\boldsymbol{\tau}(\boldsymbol{v})$ is analytic in $\boldsymbol{v}$ for small $\boldsymbol{v}$, has poles when $v = i(\pi/2 + n\pi)$, and has an essential singularity at $v = \infty$. We also note that since $\boldsymbol{v}$ is possibly complex, $\boldsymbol{v}$ can tend toward infinity in various directions while v remains bounded. Thus the singularity structure at infinity is quite complicated. Near the origin $\boldsymbol{\tau}(\boldsymbol{v})$ has the convergent expansion

$$\boldsymbol{\tau}(\boldsymbol{v}) = \boldsymbol{v}(1 - v^2/3 + 2v^4/15 - 17v^6/315 + \cdots). \tag{8.7.24}$$

We see from (7.17) that $\boldsymbol{v}(\boldsymbol{\tau})$ is analytic in $\boldsymbol{\tau}$ for small $\boldsymbol{\tau}$ and has branch points at $\tau = \pm 1$. It also has a pole in τ at $\tau = 0$ on the Riemann sheets reached by circling these branch points. Moreover, there is a complicated singularity at infinity. Near the origin on the *principal* sheet $\boldsymbol{v}(\boldsymbol{\tau})$ is analytic and has the convergent expansion

$$\boldsymbol{v}(\boldsymbol{\tau}) = \boldsymbol{\tau}(1 + \tau^2/3 + \tau^4/5 + \tau^6/7 + \cdots). \tag{8.7.25}$$

Finally, we see that (7.23) has the denominator $[1 + \boldsymbol{\tau}(\boldsymbol{v}^f)\cdot\boldsymbol{\tau}(\boldsymbol{v}^g)]$ which can possibly vanish, but cannot vanish for small $\boldsymbol{v}^f$ and $\boldsymbol{v}^g$ because of (7.24). We conclude that the BCH series for

$Sp(2)$ converges for small $\boldsymbol{v}^f$ and $\boldsymbol{v}^g$, but presumably cannot converge everywhere because of the singularities just described. This result is consistent with our expectations. We know that any symplectic matrix can be written in the form (3.8.24). If it were always possible to combine the two exponents in (3.8.24) into one grand exponent using the BCH series, then (3.7.97) could be written in the form (3.7.36), which we know is false. Indeed, the reader will have the pleasure of showing in Exercise 7.9 that the offending singularity is the pole at $\tau = 0$ on a nonprincipal sheet of $\boldsymbol{v}(\boldsymbol{\tau})$.

What is the source of all these singularities? The fault does not lie with the group $Sp(2,\mathbb{C})$ itself. Indeed, if elements of $Sp(2,\mathbb{C})$ are parameterized by 2×2 possibly complex matrices, the operation of group element multiplication is simply matrix multiplication, and entries in the product of two matrices are *entire* functions of the entries in the matrices being multiplied.[18] Rather the fault lies in the use of canonical coordinates of the first kind, which is what the Ansatz (7.12) essentially does. See Section 7.9. And the use of canonical coordinates of the first kind depends on the properties of the exponential map. See Section 3.8. Thus, the source of singularities in this case can be traced back to the (not globally possible/successful) use of the exponential map for $Sp(2,\mathbb{C})$. Seeking the impossible results in singularities.

8.7.2 Nature of Single Exponential Form

Let us explore further what elements of $Sp(2,\mathbb{R})$ can be written in single exponential form. In the case of two-dimensional phase space, the most general (real) f_2 can be written in the form

$$f_2 = -(bp^2 + 2aqp + cq^2)/2, \tag{8.7.26}$$

where a, b, c are (real) parameters. We define an associated symplectic matrix $R(a,b,c)$ by the rule

$$\exp(:f_2:)z_d = \sum_e R_{de} z_e. \tag{8.7.27}$$

Our goal is to find an explicit expression for $R(a,b,c)$ in terms of the quantities a, b, c. So doing amounts to finding the exponential map from $sp(2,\mathbb{R})$ to $Sp(2,\mathbb{R})$.[19]

Direct calculation gives the result

$$:f_2: z_d = -(1/2) : (bp^2 + 2aqp + cq^2) : z_d = \sum_e F_{de} z_e \tag{8.7.28}$$

where F is the Hamiltonian matrix

$$F = \begin{pmatrix} a & b \\ -c & -a \end{pmatrix}. \tag{8.7.29}$$

We readily verify that F has the property

$$F^2 = \Delta I. \tag{8.7.30}$$

[18] An entire function is a function that is analytic everywhere except at infinity.

[19] As announced, we will seek results for the case of $Sp(2,\mathbb{R})$. Partial results are also known for the more complicated case of $Sp(4,\mathbb{R})$. In particular, there is an explicit formula for matrices of the form $\exp(JS^a)$. See the references at the end of this chapter.

Here Δ is the *discriminant* of the quadratic form (7.26),

$$\Delta = a^2 - bc. \tag{8.7.31}$$

We know from Section 7.2 or (7.3.41) that R is given by the relation

$$R = \exp(F) = \cosh(F) + \sinh(F). \tag{8.7.32}$$

The term $\cosh(F)$ has the expansion

$$\begin{aligned} \cosh(F) &= F^0 + F^2/2! + F^4/4! + \cdots \\ &= I(1 + \Delta/2! + \Delta^2/4! + \cdots) \\ &= I\cosh(\Delta^{1/2}). \end{aligned} \tag{8.7.33}$$

Here use has been made of (7.30). For $\sinh(F)$ we find the result

$$\begin{aligned} \sinh(F) &= F + F^3/3! + F^5/5! + \cdots \\ &= F(I + F^2/3! + F^4/5! + \cdots) \\ &= F(1 + \Delta/3! + \Delta^2/5! + \cdots) \\ &= (F/\Delta^{1/2})(\Delta^{1/2} + \Delta^{3/2}/3! + \Delta^{5/2}/5! + \cdots) \\ &= F[\sinh(\Delta^{1/2})]/\Delta^{1/2}. \end{aligned} \tag{8.7.34}$$

Note that both $\cosh(\Delta^{1/2})$ and $\{[\sinh(\Delta^{1/2})]/\Delta^{1/2}\}$ are even functions of $\Delta^{1/2}$, and thus do not depend on which root we take in computing $\Delta^{1/2}$. In fact, they are *analytic* functions of Δ and hence of a, b, c. Putting everything in (7.32) together gives for R the result

$$R = \begin{pmatrix} \cosh(\Delta^{1/2}) + a[\sinh(\Delta^{1/2})]/\Delta^{1/2} & b[\sinh(\Delta^{1/2})]/\Delta^{1/2} \\ -c[\sinh(\Delta^{1/2})]/\Delta^{1/2} & \cosh(\Delta^{1/2}) - a[\sinh(\Delta^{1/2})]/\Delta^{1/2} \end{pmatrix}. \tag{8.7.35}$$

Let us compute the eigenvalues of R. It has the characteristic polynomial

$$P(\lambda) = \det(R - \lambda I) = \lambda^2 - 2\lambda\cosh(\Delta^{1/2}) + 1. \tag{8.7.36}$$

This polynomial has the roots

$$\lambda = \exp(\Delta^{1/2})\ ,\ \exp(-\Delta^{1/2}). \tag{8.7.37}$$

Note that if a, b, c are real, then so is Δ. It follows that $\Delta^{1/2}$ is real if $\Delta \geq 0$, and pure imaginary if $\Delta < 0$. Correspondingly, the eigenvalues of R are real if $\Delta > 0$, and have the *hyperbolic* configuration shown in Case 1 of Figure 3.4.1. If $\Delta < 0$, then the eigenvalues of R are on the unit circle corresponding to the *elliptic* configuration shown in Case 3 of Figure 3.4.1. The possibilities $\Delta = 0$ and $\Delta = -\pi^2$ are discussed further in Exercise 7.11, and correspond to the *parabolic* and *inversion parabolic* configurations shown in Cases 4 and 5 of Figure 3.4.1. We note that the *inversion hyperbolic* configuration shown in Case 2 does *not* occur. It follows that such symplectic matrices cannot be written in single exponential form with real exponents. [The use of complex exponents may be possible in some cases. See Exercises 2.16 and 7.12. But we know that even this expedient fails for the matrices M and N given by (3.7.134) and (3.7.135).] We have proved earlier that all (real) symplectic matrices, including the inversion hyperbolic case, can be written in the product form (3.8.26) with real exponents. Also, all the coefficients in the BCH series are real. It follows that the BCH series must *diverge* if we try to combine the exponents in (3.8.26) for the inversion hyperbolic case: if the series converged, the resulting single exponent would be real, and we have seen that a single real exponent never gives an inversion hyperbolic symplectic matrix.

Exercises

8.7.1. Verify the expansion (7.2).

8.7.2. Verify (7.8) through (7.10).

8.7.3. Verify (7.13).

8.7.4. Verify (7.15).

8.7.5. Given (7.16), verify (7.17).

8.7.6. Verify (7.19) and (7.20).

8.7.7. Verify (7.21) through (7.23).

8.7.8. Verify the singularity statements made about $\boldsymbol{\tau}(\boldsymbol{v})$ and $\boldsymbol{v}(\boldsymbol{\tau})$, and verify (7.24) and (7.25).

8.7.9. Review Exercises 3.7.11 and 5.9.3. Consider the matrix $N = -M$. [Note that this N is not that given by (3.7.105).] Show that N can be written in the form (3.7.36) with S given by the equation

$$S = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}. \tag{8.7.38}$$

Consider the polar decompositions of M and N given by (3.8.1). Show that both M and N have the same P given by (5.9.16). Find the O matrices for M and N. Show that they have the form

$$O = \exp(i\theta\sigma^2), \tag{8.7.39}$$

and find θ in each case. Following (3.8.15) and (3.8.23) find the matrix S^a associated with P and the matrices S^c associated with the matrices O. Compute the corresponding vectors $\boldsymbol{v}^a$, $\boldsymbol{v}^c$, $\boldsymbol{\tau}^a = \boldsymbol{\tau}(\boldsymbol{v}^a)$, and $\boldsymbol{\tau}^c = \boldsymbol{\tau}(\boldsymbol{v}^c)$. Make the identifications $\boldsymbol{v}^a = \boldsymbol{v}^f$, $\boldsymbol{\tau}^a = \boldsymbol{\tau}^f$, $\boldsymbol{v}^c = \boldsymbol{v}^g$, and $\boldsymbol{\tau}^c = \boldsymbol{\tau}^g$. See (3.8.24) and (7.12). Consider the vector $\boldsymbol{v}^h(\theta)$ defined by the relation

$$\exp[\boldsymbol{v}^h(\theta) \cdot \boldsymbol{\sigma}] = \exp(\boldsymbol{v}^f \cdot \boldsymbol{\sigma}) \exp(i\theta\sigma^2). \tag{8.7.40}$$

Use (7.23) to find $\boldsymbol{\tau}^h(\theta) = \boldsymbol{\tau}^h(\boldsymbol{v}^h(\theta))$. Starting from $\theta = 0$, follow the quantities θ, $\boldsymbol{\tau}^h(\theta)$, and $\boldsymbol{v}^h(\theta)$ to the θ value for N. Repeat this same process, again starting at $\theta = 0$, and continuing to the θ value for M. Show that $\boldsymbol{v}^h(\theta)$ is well defined for the θ value corresponding to N and produces (7.38), and that $\boldsymbol{v}^h(\theta)$ is singular at the θ value corresponding to M.

8.7.10. Verify (7.28) through (7.37).

8.7.11. Show that $R = -I$ when $\Delta = -\pi^2$. What happens when $\Delta = -4\pi^2$? Show that R takes the form

$$R = \begin{pmatrix} 1+a & b \\ -c & 1-a \end{pmatrix} \tag{8.7.41}$$

in the case $\Delta = 0$. Show that this R can be diagonalized only if $b = c = 0$. Hint: When $\Delta = 0$, $R = I + F$ and, according to (7.30), $F^2 = 0$. Show that such an F can be diagonalized only if $F = 0$.

8.7.12. Consider the inversion hyperbolic symplectic matrix M given by

$$M = \begin{pmatrix} -\lambda & 0 \\ 0 & -1/\lambda \end{pmatrix} \tag{8.7.42}$$

where λ is real and positive. We know that M cannot be written in single exponent form with a real exponent. But can it be written in single exponent form with a complex exponent? Let S be the symmetric matrix given by

$$S = \begin{pmatrix} 0 & a \\ a & 0 \end{pmatrix}. \tag{8.7.43}$$

Verify the results

$$JS = \begin{pmatrix} a & 0 \\ 0 & -a \end{pmatrix} \tag{8.7.44}$$

and

$$N = \exp(JS) = \begin{pmatrix} \exp(a) & 0 \\ 0 & \exp(-a) \end{pmatrix}. \tag{8.7.45}$$

Evidently M can be written in single exponent form if one can satisfy the relation

$$\exp(a) = -\lambda. \tag{8.7.46}$$

Define a quantity α by the rule

$$\alpha = \log(\lambda). \tag{8.7.47}$$

Show that a solution to (7.46) is

$$a = \alpha + \pi i. \tag{8.7.48}$$

You have shown that, although the inversion hyperbolic symplectic matrix M cannot be written in single exponent form with a real exponent, it can be written in single exponent form with a complex exponent.

8.7.13. For some purposes it is useful to have an $SU(2)$ version of the BCH formula (7.23). Recall the 2×2 matrices K^j defined in Exercise 3.7.30 and manufactured from the Pauli matrices by the rules

$$K^j = (-i/2)\sigma^j. \tag{8.7.49}$$

Verify that they satisfy the multiplication rules

$$K^j K^k = (-1/4)\delta_{jk} I + (1/2)\sum_{\ell} \epsilon_{jk\ell} K^{\ell}, \tag{8.7.50}$$

and recall that these rules can be summarized in the "vector" form

$$(\boldsymbol{a} \cdot \boldsymbol{K})(\boldsymbol{b} \cdot \boldsymbol{K}) = -(1/4)(\boldsymbol{a} \cdot \boldsymbol{b})I + (1/2)(\boldsymbol{a} \times \boldsymbol{b}) \cdot \boldsymbol{K}. \tag{8.7.51}$$

See (3.7.176).

Verify that any matrix $u \in SU(2)$ can be written in the form

$$u(\boldsymbol{v}) = \exp(\boldsymbol{v} \cdot \boldsymbol{K}). \tag{8.7.52}$$

Show that the infinite series implied by (7.52) can be summed to give the explicit result

$$\exp(\boldsymbol{v}\cdot\boldsymbol{K}) = I\cos(v/2) + (\boldsymbol{v}\cdot\boldsymbol{K})(2/v)\sin(v/2) \tag{8.7.53}$$

where

$$v = (\boldsymbol{v}\cdot\boldsymbol{v})^{1/2}. \tag{8.7.54}$$

Define a vector $\boldsymbol{\tau}(\boldsymbol{v})$ by the rule

$$\boldsymbol{\tau}(\boldsymbol{v}) = \boldsymbol{v}(2/v)\tan(v/2). \tag{8.7.55}$$

Show that (7.55) has the inverse

$$\boldsymbol{v}(\boldsymbol{\tau}) = \boldsymbol{\tau}(2/\tau)\tan^{-1}(\tau/2) \tag{8.7.56}$$

where

$$\tau = (\boldsymbol{\tau}\cdot\boldsymbol{\tau})^{1/2}. \tag{8.7.57}$$

Show that (7.53) can also be written in the form

$$\exp(\boldsymbol{v}\cdot\boldsymbol{K}) = \cos(v/2)[I + (\boldsymbol{\tau}\cdot\boldsymbol{K})]. \tag{8.7.58}$$

Given two vectors $\boldsymbol{v}^a$ and $\boldsymbol{v}^b$, your task is to find a third vector $\boldsymbol{v}^c$ such that

$$\exp(\boldsymbol{v}^a\cdot\boldsymbol{K})\exp(\boldsymbol{v}^b\cdot\boldsymbol{K}) = \exp(\boldsymbol{v}^c\cdot\boldsymbol{K}). \tag{8.7.59}$$

Show that there is the formula

$$\boldsymbol{\tau}(\boldsymbol{v}^c) = [\boldsymbol{\tau}(\boldsymbol{v}^a) + \boldsymbol{\tau}(\boldsymbol{v}^b) + (1/2)\boldsymbol{\tau}(\boldsymbol{v}^a)\times\boldsymbol{\tau}(\boldsymbol{v}^b)][1 - (1/4)\boldsymbol{\tau}(\boldsymbol{v}^a)\cdot\boldsymbol{\tau}(\boldsymbol{v}^b)]^{-1}. \tag{8.7.60}$$

8.7.14. Review Exercise 8.7.13. Determine the analytic behavior of $\boldsymbol{v}^c$ as a function of $\boldsymbol{v}^a$ and $\boldsymbol{v}^b$.

8.8 Zassenhaus or Factorization Formulas

The BCH formula (3.7.33) and (3.7.34) attempts to combine two exponents into one. There are related formulas, called *Zassenhaus* formulas, that attempt the reverse: They try to write a single exponent term as a product of several such terms. One simple such formula is the relation

$$\exp(sA + tB) = \exp(sA)\exp(tB)\exp(Z), \tag{8.8.1}$$

where Z has the expansion

$$Z = -(st/2)[A,B] + (s^2t/6)[A,[A,B]] - (st^2/3)[B,[B,A]] + O(s^3t, s^2t^2, st^3). \tag{8.8.2}$$

It will be seen in Sections 10.8 through 10.10 that Zassenhaus formulas are useful in constructing symplectic integrators and computing maps.

Equation (8.1) writes a single exponent term as a product of *three* such terms. It may also be desirable to write a single exponent term as a product of *two* such terms, and to attempt to sum some of the infinite series that occur. Consider (7.3), which gives a formula for combining two exponentials into one grand exponential. Sometimes, as in for example the construction of a factored product decomposition, it is useful to be able to turn the process around. Define a quantity h by writing

$$s : f : [1 - \exp(-s : f :)]^{-1} g = h. \tag{8.8.3}$$

Observe that (8.3) may be solved for the quantity g to give the relation

$$g = \{[1 - \exp(-s : f :)]/[s : f :]\}h. \tag{8.8.4}$$

Here the operator expression appearing on the right of (8.4) is interpreted to be the series

$$\begin{aligned} [1 - \exp(-s : f :)]/[s : f :] &= -\sum_{m=1}^{\infty}(-s)^m : f :^m /[m! s : f :] \\ &= \sum_{m=1}^{\infty}(-s)^{m-1} : f :^{m-1} /m!. \end{aligned} \tag{8.8.5}$$

Now insert (8.3) into (7.3). One finds, upon reading right to left, the result

$$\exp[s : f : +t : h : +O(t^2)] = \exp(s : f :) \exp(t : g :). \tag{8.8.6}$$

Finally, the term of $O(t^2)$ can be taken from the left to the right side of (7.6) to produce the relation

$$\exp[s : f : +t : h :] = \exp(s : f :) \exp(t : g :) \exp[: O(t^2) :]. \tag{8.8.7}$$

Equation (8.7) gives a formula for writing the exponential of the sum of two exponents as a product of two exponentials.

It is worth remarking that the operation described by (8.4), which is required for evaluating (8.7), can be written in a more compact form. First, observe the formal integral identity

$$[1 - \exp(-s : f :)]/[s : f :] = \int_0^1 d\tau \exp(-\tau s : f :). \tag{8.8.8}$$

Let us define the function iex(w), called the *integrated* exponential function, by the rule

$$\text{iex}(w) = \int_0^1 d\tau \exp(\tau w) = (e^w - 1)/w = \sum_{m=0}^{\infty} w^m/(m+1)!. \tag{8.8.9}$$

Evidently iex is an *entire* analytic function. (Like the exponential function, it has no singularities in the complex plane except at infinity, and its Taylor series has an infinite radius of convergence.) By using the identity (8.8) and the definition (8.9), the relation (8.4) can be written in the forms

$$g = \int_0^1 d\tau \exp(-\tau s : f :)h = \text{iex}(-s : f :)h. \tag{8.8.10}$$

But now (5.4.11) can be employed to give the final integral formula

$$g(z) = \int_0^1 d\tau h[\exp(-\tau s : f :)z]. \tag{8.8.11}$$

In summary, we have the operator identity

$$\begin{aligned} &\exp(s : f : + t : h :) = \exp(s : f :) \exp[\mathrm{iex}(-s\#f\#)(t : h :)] \exp[: O(t^2) :] \\ &= \exp(s : f :) \exp[: \mathrm{iex}(-s : f :)(th) :] \exp[: O(t^2) :]. \end{aligned} \tag{8.8.12}$$

We also note that in (8.4), unlike in (7.3), no inverses of the form $[1 - \exp(-s : f :)]^{-1}$ are involved. Instead, we have benign relations like (8.11). Correspondingly, because they presuppose the existence of a single exponent form, Zassenhaus formulas can have better convergence properties than the BCH formula. Finally, we remark that the next few terms in (8.12), those terms proportional to powers of t^2, t^3, $\cdots$, can also be found explicitly. See Exercises 10.3.* and 10.4.*.

Exercises

8.8.1. Derive (8.1) and (8.2) from (3.7.33) and (3.7.34).

8.8.2. Verify the expansion (8.5); derive (8.6) from (7.3) and (8.3).

8.8.3. Verify the integral identity (8.8), and the integral identity and expansion (8.9). Show that the Taylor series for iex(w) has an infinite radius of convergence.

8.8.4. Derive the formula

$$\exp(s : f : + t : h :) = \exp[: O(t^2) :] \exp[\mathrm{iex}(s\#f\#)(t : h :)] \exp(s : f :). \tag{8.8.13}$$

8.9 Ideals, Quotients, and Gradings

We know that the Lie algebra of all Lie operators, which we have called $ispm(2n, \mathbb{R})$, is infinite dimensional. Correspondingly $ISpM(2n, \mathbb{R})$, the group of symplectic maps, is infinite dimensional. Indeed, the factorization (7.7.23) gives a representation of the general analytic symplectic map. We see that the specification of a symplectic map generally requires an infinite number of parameters. This fact produces an awkward situation for human beings and computers, which can only work with a finite number of quantities (and often only with finite precision).

An optimistic perspective on the experimental and theoretical situation, for example in the field of accelerator physics, might be stated as follows: We know that a beam transport system, accelerator, storage ring, or any portion thereof may be described by a symplectic transfer map. However, because we cannot measure or control electromagnetic fields exactly, we are unsure of and unable to control exactly what this map is. Also, since it is impossible to perform computations with an infinite number of variables and to infinite precision, it is necessary to develop various approximation schemes. Thus, we are able to study computationally (and probably theoretically) the detailed properties of only a subset of all symplectic maps. The hope is that if two symplectic maps are in some sense nearly the same, then their behavior [including, in some cases, long-term (repeated iteration) behavior] will be nearly the same. If this were not true from an experimental standpoint, then it would be impossible to build satisfactory storage rings, etc. If this were not true from a theoretical standpoint, then it would be impossible to design storage rings, etc., with any assurance that their actual performance would be satisfactory.

As just described, it is necessary to develop some sort of approximation scheme to treat symplectic maps in a practical way. In this section we will describe truncation schemes that maintain a Lie algebraic structure. We already know from Section 8.4 that the rules for multiplying symplectic maps can be expressed entirely in Lie algebraic terms. Thus, if the truncation scheme maintains a Lie algebraic structure, it follows that maps may either be truncated and then multiplied, or multiplied and then truncated. The results from both procedures are guaranteed to be the same.

For example, consider the Lie algebra spanned by the homogeneous polynomials f_2, f_3, f_4, $\cdots$. Evidently, this Lie algebra is infinite dimensional. Let D be some integer. Suppose we decide to retain only the polynomials f_2, f_3, f_4, $\cdots$ f_{D-1}, and discard all polynomials f_m with $m \geq D$. Correspondingly, in the map $\mathcal{M}$ given by (7.6.3) we drop from the product all f_m with $m \geq D$. Is this a consistent procedure? The answer is yes. As we will see, the discarding of all f_m with $m \geq D$ amounts mathematically to working with a *quotient* Lie algebra and its corresponding quotient group. We note that since $: f_m : z_b$ consists of terms of degree $(m-1)$, the decision to drop the f_m with $m \geq D$ amounts physically to neglecting all aberrations of degree $(D-1)$ and higher.

As a second example, suppose ϵ is a (presumed small) parameter, and consider the quantities $f^{(0)}, \epsilon f^{(1)}, \epsilon^2 f^{(2)}, \epsilon^3 f^{(3)}, \cdots$, where $f^{(0)}, f^{(1)}, f^{(2)}, f^{(3)} \cdots$ are *arbitrary* functions. The quantities $f^{(0)}, \epsilon f^{(1)}, \epsilon^2 f^{(2)}, \epsilon^3 f^{(3)}, \cdots$ also form (with the Poisson bracket as a Lie product) an infinite dimensional Lie algebra. Suppose, as a kind of perturbation theory, we decide to discard all $\epsilon^m f^{(m)}$ with $m > D$ where again D is some integer. Is this a consistent procedure? The answer again is yes. We will see that expanding in powers of ϵ is equivalent to introducing a *grading* into the Lie algebra, and that truncating the expansion is equivalent to using the grading to produce a quotient structure.

With this motivation as background, we are ready to develop some mathematical tools. The first concept we will need is that of an *ideal.* Let L be a Lie algebra, and let L' be a subalgebra of L. For L' to be a subalgebra means that the elements of L' must be in L, and must form a Lie algebra in their own right. That is, by themselves they must satisfy the properties 1 through 5 (as given in Section 3.7) required of a Lie algebra. Let x be any element in L and let x' be any element in L'. Suppose the elements of L' have the property

$$[x, x'] \in L' \text{ for all } x \in L,\ x' \in L'. \tag{8.9.1}$$

That is, no element of L' can be sent beyond L' by taking Lie products with arbitrary elements in L. In this case L' is said to be an *invariant* subalgebra. And if L' is a genuine invariant *sub*algebra, i.e. neither zero nor the full Lie algebra L, it is called an *ideal.*[20]

[20] Here is an opportunity for three more definitions: Recall that a Lie algebra is called *simple* if it has no ideals. Recall also that a Lie algebra or subalgebra is called *Abelian* if the Lie product of any two elements

Suppose a Lie algebra L has a subalgebra L'. Then L' can be used to set up an equivalence relation among the elements of L. Let x_1 and x_2 be any two elements in L. We say that x_2 is *equivalent* to x_1 (and write $x_2 \sim x_1$) if their difference $(x_2 - x_1)$ is in L',

$$x_2 \sim x_1 \Leftrightarrow (x_2 - x_1) \in L'. \tag{8.9.2}$$

(Here the symbol $\Leftrightarrow$ is used to indicate logical implication in both directions.) This equivalence relation can be used to partition the elements of L into *disjoint* equivalence classes. Let the symbols $\{x\}$ denote all the elements of L that are equivalent to some element x. In a Lie algebraic (actually, vector space) context, the collection of these equivalence classes is called a quotient space, and is customarily denoted by the symbols L/L'. See Exercise 9.1.

To get a feeling for this construction, let 0 be the *zero* element in L and consider the set of elements $\{0\}$, the set of elements in L that are equivalent to 0. We see from (9.2) that the set $\{0\}$ is identical to the set of elements L'. Consequently, in the quotient space construction, all elements in L' are identified with (are equivalent to) the zero element in L. That is, we have the logical relation

$$x' \in L' \Leftrightarrow \{x'\} = \{0\}. \tag{8.9.3}$$

Moreover, suppose $x_2 \sim x_1$. Then by (9.2) we have a relation of the form

$$x_2 = x_1 + x' \text{ with } x' \in L'. \tag{8.9.4}$$

Thus, if x_1 and x_2 are equivalent, they differ only by an element that has been identified with zero.

As defined so far, the quotient space L/L' is simply a collection of equivalence classes. We now give it a vector space structure by a simple but ingenious (and, at first sight, confusing) construction. We begin by noting the logical implication

$$x_2 \sim x_1 \Rightarrow ax_2 \sim ax_1, \tag{8.9.5}$$

where a is any scalar. This result follows from (9.4) by noting that ax' belongs to L' if x' belongs to L'. (Remember that L' is an algebra.) Next, suppose that the elements x_1, x_2 and y_1, y_2 satisfy the equivalence relations

$$x_2 \sim x_1 \ , \ y_2 \sim y_1. \tag{8.9.6}$$

Then it follows from (9.4) and its y analog that we have the relation

$$x_2 + y_2 = x_1 + y_1 + x' + y'. \tag{8.9.7}$$

in it vanishes. Colloquially, we say that all elements in an Abelian Lie algebra or subalgebra *commute*. A Lie algebra is called *semisimple* if it has no Abelian ideals. By this definition, a simple Lie algebra is also semisimple. That is, simple Lie algebras form a subset of the set of semisimple Lie algebras. Suppose a Lie algebra L is semisimple but not simple. Then it can be shown that L is the *direct sum* of two or more simple Lie algebras. By direct sum it is meant the all the elements of any simple Lie subalgebra in the sum commute with all the elements of any other simple Lie subalgebra in the sum.

But if x' and y' belong to L', then so must the sum $(x'+y')$. (Again, remember that L' is an algebra.) Thus, we also have the logical implication

$$x_2 \sim x_1 \text{ and } y_2 \sim y_1 \Rightarrow (x_2+y_2) \sim (x_1+y_1). \tag{8.9.8}$$

Now we are ready to give L/L' a *vector space* structure. First, we have to define scalar multiplication. Consider some equivalence class. Then, since equivalence classes are disjoint, each equivalence class may be labelled by any one of its members. Select some member of the equivalence class under consideration, and call it x_1. Then the equivalence class may be given the label $\{x_1\}$. Now let a be any scalar. We define scalar multiplication acting on the element $\{x_1\}$ of L/L' by the rule

$$a\{x_1\} = \{ax_1\}. \tag{8.9.9}$$

Thus, by this definition, scalar multiplication sends equivalence classes into each other. But suppose x_2 also belongs to $\{x_1\}$, and that we had used x_2 to label $\{x_1\}$ instead of x_2. Would this different choice affect the definition (9.9)? It would not. With the choice of x_2 as a label we would have the definition

$$a\{x_2\} = \{ax_2\}. \tag{8.9.10}$$

But, by (9.5), we have the relation

$$\{ax_2\} = \{ax_1\} \tag{8.9.11}$$

because an equivalence class is uniquely defined by any of its members. Thus, scalar multiplication is uniquely defined by the rule (9.9).

Next we define vector addition. Let $\{x_1\}$ and $\{y_1\}$ be two equivalence classes labelled by two members x_1 and y_1. We define addition by the rule

$$\{x_1\} + \{y_1\} = \{(x_1+y_1)\}. \tag{8.9.12}$$

By this definition, addition sends a pair of equivalence classes into some third (not necessarily different) equivalence class. Again, there is the question of uniqueness under the choice of labelling. However, thanks to (9.8), the definition (9.12) is in fact independent of labelling. Note that as a special case of (9.12) we have the relation

$$\{x\} + \{0\} = \{x\}. \tag{8.9.13}$$

That is, the equivalence class $\{0\}$ plays the role of the zero vector in L/L'.

We have given L/L' a vector space structure. What is the dimension of L/L'? Suppose that L' has a basis $b_1, b_2, \cdots$, and that a basis for L is constructed by taking the vectors $b_1, b_2, \cdots$ supplemented by the additional linearly independent vectors $v_1, v_2, \cdots v_n$. Then, we have the dimensional relations

$$\dim L = \dim L' + n \text{ , or } n = \dim L - \dim L'. \tag{8.9.14}$$

Suppose x is any vector in L. Then x has the unique decomposition

$$x = x' + \sum_{i=1}^{n} \xi_i v_i, \tag{8.9.15}$$

where x' is the portion of x spanned by the basis vectors $b_1, b_2, \cdots$. Now form equivalence classes of both sides of (9.15). From the definition (9.12) and (9.3), (9.9), and (9.13) we find the result

$$\{x\} = \{x'\} + \left\{ \sum_{i=1}^{n} \xi_i v_i \right\} = \sum_{i=1}^{n} \xi_i \{v_i\}. \tag{8.9.16}$$

It is easily verified that the vectors $\{v_i\}$ are linearly independent. See Exercise 7.2. Consequently, the quotient space L/L' has dimension n,

$$\dim(L/L') = n = \dim L - \dim L'. \tag{8.9.17}$$

So far we have assumed that L' is a subalgebra. Now make the further supposition that L' is an ideal. In this case we can give the quotient space a *Lie algebraic* structure. We have already seen that the quotient space can be given a vector space structure. What remains is to define a Lie product. Let $\{x_1\}$ and $\{y_1\}$ be two equivalence classes labelled by two members x_1 and y_1. We define a *quotient space* Lie product, denoted by the symbols $[\ ,\ ]_{qs}$, by the rule

$$[\{x_1\}, \{y_1\}]_{qs} = \{[x_1, y_1]\}. \tag{8.9.18}$$

By this definition the quotient space Lie product sends a pair of equivalence classes into some third (not necessarily different) equivalence class. As before there is the question of uniqueness under the choice of labelling. Suppose we use instead labels x_2 and y_2 that satisfy (9.6). Then from (9.4) and its y counterpart we find the result

$$\begin{aligned} [\{x_2\}, \{y_2\}]_{qs} &= \{[x_1 + x', y_1 + y']\} \\ &= \{[x_1, y_1] + [x', y_1] + [x_1, y'] + [x', y']\} \\ &= \{[x_1, y_1]\} + \{[x', y_1]\} + \{[x_1, y']\} + \{[x', y']\} \\ &= [\{x_1\}, \{y_1\}]_{qs} + \{[x', y_1]\} + \{[x_1, y']\} + \{[x', y']\}. \end{aligned} \tag{8.9.19}$$

But, since L' is assumed to be an ideal, all the quantities $[x', y_1]$, $[x_1, y']$, and $[x', y']$ must be in L'. See (9.1). It follows from (9.3) and (9.13) that we have the relation

$$\{[x', y_1]\} + \{[x_1, y']\} + \{[x', y']\} = \{0\} + \{0\} + \{0\} = \{0\}. \tag{8.9.20}$$

Consequently, upon combining (9.19) with (9.20) and again using (9.3), we find the result

$$[\{x_2\}, \{y_2\}]_{qs} = [\{x_1\}, \{y_1\}]_{qs}. \tag{8.9.21}$$

Thus, the quotient space Lie product is uniquely defined by (9.18).

We claim that the addition rule (9.12) and Lie product rule (9.18) together satisfy requirements 1 through 5 for a Lie algebra as given in Section 3.7. For example, if $\{x\}, \{y\}$, and $\{z\}$ are any three equivalence classes, we have from (9.18) the relation

$$[\{x\}, [\{y\}, \{z\}]_{qs}]_{qs} = [\{x\}, \{[y, z]\}]_{qs} = \{[x, [y, z]]\}. \tag{8.9.22}$$

The Jacobi condition requirement follows immediately from (9.22). Verification of the remaining requirements is left as an exercise for the reader. We conclude that if L' is an ideal, the elements of the quotient space L/L' can be viewed as elements of an n dimensional Lie algebra. This Lie algebra is called the *quotient* Lie algebra.

The discussion of the quotient Lie algebra L/L' that we have just worked through may seem overly abstract. It can be made more concrete by using structure constants. See (3.7.40). We know from our previous discussion that L is spanned by the basis vectors $b_1, b_2, b_3 \cdots$ and $v_1, v_2, \cdots v_n$. There are therefore three kinds of Lie products: $[b_i, b_j]$, $[v_i, b_j]$, and $[v_i, v_j]$. Correspondingly, the structure constants are of six kinds: ${}^1c, {}^2c, \cdots {}^6c$. Consider first the Lie products $[b_i, b_j]$. Their results can be written in the form

$$[b_i, b_j] = \sum_k {}^1c_{ij}^k b_k + \sum_k {}^2c_{ij}^k v_k. \tag{8.9.23}$$

If L' is a Lie subalgebra spanned by the b_i, as we have assumed, then the structure constants ${}^2c$ must vanish,

$${}^2c_{ij}^k = 0. \tag{8.9.24}$$

Consider next the Lie products $[v_i, b_j]$. Their results can be written in the form

$$[v_i, b_j] = \sum_k {}^3c_{ij}^k b_k + \sum_k {}^4c_{ij}^k v_k. \tag{8.9.25}$$

If L' is an ideal, as we have also assumed, then the structure constants ${}^4c$ must also vanish,

$${}^4c_{ij}^k = 0. \tag{8.9.26}$$

See (9.1). Finally, the Lie products $[v_i, v_j]$ can be written in the form

$$[v_i, v_j] = \sum_k {}^5c_{ij}^k b_k + \sum_k {}^6c_{ij}^k v_k. \tag{8.9.27}$$

Now form equivalence classes of both sides of (9.27). Then, using (9.3), (9.13), and (9.18), we find the result

$$[\{v_i\}, \{v_j\}]_{qs} = \sum_k {}^6c_{ij}^k \{v_k\}. \tag{8.9.28}$$

We also know from (9.16) that the $\{v_i\}$ span L/L'. From (9.28) we conclude that the ${}^6c_{ij}^k$ are the structure constants of L/L'.

The next topic we need to discuss is that of *quotient groups.* We will see that for every quotient Lie algebra there is a corresponding quotient Lie group. To understand this connection we begin by describing the concept of a quotient group. Suppose G is a group, and suppose G' is a subgroup of G. We use the subgroup G' to set up an equivalence relation in G. Let g_1 and g_2 be any two elements in G. We say that g_2 is equivalent to g_1 (and again use the notation $g_2 \sim g_1$) if there exists a g' in G' such that $g_1^{-1}g_2 = g'$ or, put another way, $g_2 = g_1 g'$:

$$g_2 \sim g_1 \Leftrightarrow g_1^{-1}g_2 = g' \in G' \Leftrightarrow g_2 = g_1 g'. \tag{8.9.29}$$

This equivalence relation can be used to partition the elements of G into disjoint equivalence classes. These equivalence classes are called the (left) *cosets* of G with respect to G'. The collection of all of these cosets is called the *coset space*, and is customarily denoted by the symbols G/G'. See Exercise 5.12.7. If g is an element in G, we use the notation $\{g\}$ to denote all the elements in G that are equivalent to g. Suppose e is the identity element in G. Then it is easily checked that

$$\{g'\} = \{e\}, \tag{8.9.30}$$

where g' is any element in G'.

We next assume that G' is a *normal* or *invariant* subgroup of G. Suppose g is any element of G, and g' is any element of G'. The subgroup G' is called invariant or normal if there is the relation

$$g^{-1}g'g \in G' \text{ for all } g \in G, g' \in G'. \tag{8.9.31}$$

If G' is normal, the collection of equivalence classes (coset space) G/G' can be made into a *group*. This group is called the *quotient* group.

To show that G/G' can be given a group structure, we must set up a rule for multiplying equivalence classes (cosets) in such a way that rules analogous to those given for matrices in Section 3.6 are satisfied. Suppose $\{g_1\}$ and $\{h_1\}$ are two equivalence classes labelled by representative elements g_1 and h_1 in G. We define their product, denoted by the symbols $\{g_1\}\{h_1\}$, to be the equivalence class given by the rule

$$\{g_1\}\{h_1\} = \{g_1h_1\}. \tag{8.9.32}$$

As a special case of (9.32) we find the results

$$\{g_1\}\{e\} = \{g_1e\} = \{g_1\}, \tag{8.9.33}$$

$$\{e\}\{h_1\} = \{eh_1\} = \{h_1\}. \tag{8.9.34}$$

Also, we define $\{g_1\}^{-1}$ by the rule

$$\{g_1\}^{-1} = \{g_1^{-1}\}. \tag{8.9.35}$$

Then, upon combining (9.32) and (9.35), we find the results

$$\{g_1\}\{g_1\}^{-1} = \{g_1g_1^{-1}\} = \{e\}, \tag{8.9.36}$$

$$\{g_1\}^{-1}\{g_1\} = \{g_1^{-1}g_1\} = \{e\}. \tag{8.9.37}$$

Both (9.32) and (9.35) are rules that send equivalence classes to equivalence classes. Of course, as usual, we must verify that the definitions (9.32) and (9.35) are in fact independent of the choice of representative elements selected to label the equivalence classes $\{g_1\}$ and $\{h_1\}$. For example, suppose we decide to designate the equivalence classes $\{g_1\}$ and $\{h_1\}$ by the representatives g_2 and h_2 so that we have the alternate labels $\{g_2\}$ and $\{h_2\}$. Of course g_2 and g_1 are related by (9.29), and h_2 and h_1 are related by an analogous equation. Then we find from (9.32) the result

$$\{g_2\}\{h_2\} = \{g_2h_2\} = \{g_1g'h_1h'\} = \{g_1h_1h_1^{-1}g'h_1h'\} = \{g_1h_1\} = \{g_1\}\{h_1\}. \tag{8.9.38}$$

Here we have used (9.31) and the fact that G' is a group to deduce that $h_1^{-1}g'h_1h'$ is in G'. We conclude that the equivalence class product is uniquely defined by (9.32). Similarly, it can be shown that the equivalence class inverse is uniquely defined by (9.35). See Exercise 9.4. Thus, the quotient group is uniquely defined.

We are now ready to see the connection between quotient Lie algebras and quotient Lie groups. Suppose, for simplicity, that the Lie algebra L is realized as a set of linear operators, with the Lie product being a commutator. Then, as described in Section 3.7, there is (at least locally near the identity) an associated Lie group G obtained by exponentiating L. Also, suppose L has a subalgebra L'. Exponentiating L' gives a Lie subgroup G'. Suppose that L' is an ideal. Then we will discover that G' is normal. Also, since L' is an ideal, we can form the quotient Lie algebra L/L'. Correspondingly, since G' is normal, we can form the quotient group G/G'. We will discover that L/L' is the Lie algebra of G/G'.

To see how this comes about, suppose ℓ is an element of L, and ℓ' is an element of L'. Upon exponentiation we get elements g and g' of G and G', respectively,

$$g = \exp(\ell) \quad , \quad g' = \exp(\ell'). \tag{8.9.39}$$

Now form the combination $g^{-1}g'g$. We find from (9.39) the result

$$g^{-1}g'g = \exp(-\ell)\exp(\ell')\exp(\ell). \tag{8.9.40}$$

Next use the adjoint operator $\#\ell\#$ and a relation of the form (2.16) to rewrite (9.40) in the form

$$g^{-1}g'g = \exp[\exp(-\#\ell\#)\ell']. \tag{8.9.41}$$

Since L' is assumed to be an ideal, we have from (8.1) the result

$$\#\ell\#\ell' = [\ell, \ell'] \in L', \tag{8.9.42}$$

from which it follows that $\exp(-\#\ell\#)\ell'$ is also in L',

$$\exp(-\#\ell\#)\ell' \in L'. \tag{8.9.43}$$

But, from (9.43) it follows that

$$g^{-1}g'g = \exp[\exp(-\#\ell\#)\ell'] \in G'. \tag{8.9.44}$$

Consequently G' is normal, as advertised. The converse can also be proved: If G' with Lie algebra L' is a normal subgroup of a Lie group G with Lie algebra L, then L' is an ideal in L.

To complete our demonstration, we must show that L/L' is the Lie algebra of G/G'. Suppose that x_1 is some element of L, and that it is used to label the equivalence class $\{x_1\}$, which is an element of L/L'. We define $\exp\{x_1\}$, which is supposed to be an element of G/G', by the rule

$$\exp(\{x_1\}) = \{\exp(x_1)\}. \tag{8.9.45}$$

[Note that the { } on the left side of (9.45) refers to the Lie algebraic equivalence class, and that on the right side refers to the group equivalence class.] As a special case of (9.45) we have the relation

$$\exp(\{0\}) = \{\exp(0)\} = \{e\}. \tag{8.9.46}$$

Of course, as usual, we must check that our definition does not depend on the choice of equivalence class labels. Suppose we label $\{x_1\}$ by x_2 where $x_2 \sim x_1$. Then, we find the result

$$\begin{aligned} \exp(\{x_2\}) &= \{\exp(x_2)\} = \{\exp(x_1 + x')\} \\ &= \{\exp(x_1)\exp(-x_1)\exp(x_1 + x')\}. \end{aligned} \tag{8.9.47}$$

Here we have used (9.4). Now use the BCH series (3.7.33) and (3.7.34) to combine the exponents $(-x_1)$ and $(x_1 + x')$. According to (3.7.34), the first thing we must do is add them. We find the result

$$(-x_1) + (x_1 + x') = x' \in L'. \tag{8.9.48}$$

Next, according to (3.7.34), we must find their commutator. Doing so gives the result

$$[(-x_1), (x_1 + x')] = [(-x_1), x'] \in L'. \tag{8.9.49}$$

Here, because L' is an ideal, we have been able to use (9.1). Finally, we must compute an infinite number of higher-order commutators. See (3.7.34) and Appendix C. Examination of the contents of these commutators shows that each of them has a term of the form (9.49) buried inside, and we know this term is in L'. But since L' is an ideal, (9.1) shows that all further commutators will also be in L'. We have learned that all terms that arise when we combine the exponents in $\exp(-x_1)\exp(x_1 + x')$ are in L'. Consequently, the product $\exp(-x_1)\exp(x_1 + x')$ must be in G',

$$\exp(-x_1)\exp(x_1 + x') = g' \in G'. \tag{8.9.50}$$

It follows from (9.47), (9.50), (9.29), and (9.45) that we have the result

$$\begin{aligned} \exp(\{x_2\}) &= \{\exp(x_1)\exp(-x_1)\exp(x_1 + x'\} \\ &= \{\exp(x_1)g'\} = \{\exp(x_1)\} = \exp(\{x_1\}). \end{aligned} \tag{8.9.51}$$

Thus, the definition (9.45) is indeed independent of the choice of equivalence class labels.

The last thing we must show is that products of the form $\exp(\{x_1\})\exp(\{y_1\})$ can be computed from a knowledge only of the quotient Lie algebra L/L'. Suppose $\{x_1\}$ and $\{y_1\}$ are two elements of L/L'. We then find from (9.45) and (9.32) the result

$$\begin{aligned} \exp(\{x_1\})\exp(\{y_1\}) &= \{\exp(x_1)\}\{\exp(y_1)\} \\ &= \{\exp(x_1)\exp(y_1)\}. \end{aligned} \tag{8.9.52}$$

Let us use the BCH series to combine the exponents x_1 and y_1 into one grand exponent z_1. Then we have the relation

$$\exp(x_1)\exp(y_1) = \exp(z_1). \tag{8.9.53}$$

Consequently, we find from (9.52) and (9.53) the result

$$\exp(\{x_1\})\exp(\{y_1\}) = \{\exp(z_1)\} = \exp(\{z_1\}). \tag{8.9.54}$$

From the BCH formula (3.7.34) we know that z_1 is given by the series

$$\begin{aligned} z_1 = x_1 + y_1 &+ (1/2)[x_1, y_1] + (1/12)[x_1, [x_1, y_1]] \\ &+ (1/12)[y_1, [y_1, x_1]] + \cdots . \end{aligned} \tag{8.9.55}$$

Now, form equivalence classes of both sides of (9.55). By making repeated use of (9.12) and (9.18) we find from (9.55) the result

$$\begin{aligned} \{z_1\} &= \{x_1\} + \{y_1\} + (1/2)[\{x_1\}, \{y_1\}]_{qs} + (1/12)[\{x_1\}, [\{x_1\}, \{y_1\}]_{qs}]_{qs} \\ &+ (1/12)[\{y_1\}, [\{y_1\}, \{x_1\}]_{qs}]_{qs} + \cdots . \end{aligned} \tag{8.9.56}$$

We see from (9.54) and (9.56) that the group multiplication rules for the quotient group G/G' are indeed determined by the quotient Lie algebra L/L'.

So far in this section our discussion has been devoted to the general concepts of quotient Lie algebras and their associated quotient Lie groups. We now turn to applying these concepts to $ispm(2n, \mathbb{R})$, the Lie algebra of the group of all symplectic maps acting on a $2n$ dimensional phase space. As mentioned at the beginning of this chapter, we will first restrict our attention to those symplectic maps that send the origin into itself. The general case will be treated at the end of this section. From Section 7.6 we know that the Lie algebra of maps that send the origin into itself is spanned by the Lie operators $: f_2 :, : f_3 :, \cdots$. Let us call this Lie algebra L_2.

Let D be some integer satisfying $D > 2$, and let L_D be the set of Lie operators spanned by all $: f_m :$ with $m \geq D$. From (5.3.14) and (7.6.14) we find the result

$$\{: f_m :, : f_n :\} =: [f_m, f_n] :=: \mathcal{P}_{m+n-2} :, \tag{8.9.57}$$

where we have used the notation $\mathcal{P}_\ell$ to denote the space spanned by all f_ℓ. Observe that if $m \geq D$ and $n \geq D$ (with $D > 2$), then $(m + n - 2) \geq D$. Thus, if $: f_m :$ and $: f_n :$ are in L_D, then so is their Lie product (9.57). It follows that L_D is a subalgebra of L_2.

As a special case of (9.57) we have the result

$$\{: f_2 :, : f_n :\} =: [f_2, f_n] :=: \mathcal{P}_n : . \tag{8.9.58}$$

We see from (9.57) and (9.58) that if $: f_n :$ is in L_D, then so is $\{: f_m :, : f_n :\}$ for all $: f_m :$ in L_2. It follows that L_D is an ideal in L_2.

Suppose we form the quotient algebra L_2/L_D. From our discussion of quotient algebras, we know this construction is equivalent to discarding all $: f_m :$ with $m \geq D$, and retaining only the $: f_\ell :$ with $\ell = 2, 3, \cdots (D - 1)$. We also discard all Lie products $\{: f_m :, : f_n :\}$ when $(m + n - 2) \geq D$. We have seen that dropping all $: f_m :$ with $m \geq D$ is equivalent to ignoring all aberrations of degree $(D - 1)$ and higher. The result of this construction, the quotient algebra L_2/L_D, is a *finite*-dimensional Lie algebra whose dimension equals the number of monomials in the phase-space variables z of degrees $2, 3, \cdots (D-1)$. The number of monomials of degrees $1, 2, 3 \cdots (D - 1)$ is given by $S(D - 1, d)$. (Note that D as defined in this section differs by 2 from that defined in Section 7.9.) Also, we know that the number of monomials of degree 1 (in a d-dimensional phase space) is d. Thus, we conclude that the dimension of L_2/L_D is given by the relation

$$\dim(L_2/L_D) = S(D - 1, d) - d. \tag{8.9.59}$$

This dimension is tabulated in Table 9.1 below for the cases of $d = 4$, $d = 6$, and $d = 8$ and various values of D. Finally, there is a finite-dimensional quotient group corresponding to L_2/L_D. The elements of this group are all symplectic maps of the form

$$\begin{aligned}\mathcal{M} = &\exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_3:)\exp(:f_4:)\cdots\exp(:f_{D-1}:)\times\\ &\exp(:h_D:)\exp(:h_{D+1}:)\cdots,\end{aligned} \tag{8.9.60}$$

Table 8.9.1: Values of $\dim(L_2/L_D)$.

D	dim for $d = 4$	dim for $d = 6$	dim for $d = 8$
4	30	77	156
5	65	203	486
6	121	455	1278
7	205	917	2994
8	325	1709	6426
9	490	2996	12,861
10	710	4998	24,301
11	996	8001	43,749
12	1360	12,369	75,573
13	1815	18,557	125,961

where the f's are specified and the homogeneous polynomial Lie operators $:h_D:$, $:h_{D+1}:$, $\cdots$ can be anything since all such elements are in L'_D and their exponentials are in the normal subgroup G'. We note that (9.60) is a relation of the form $g_2 = g_1 g'$. See (9.29). Consequently, we may view all members of the quotient group as belonging to the equivalence classes $\{\mathcal{M}_f\}$ where

$$\mathcal{M}_f = \exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_3:)\exp(:f_4:)\cdots\exp(:f_{D-1}:). \tag{8.9.61}$$

Since the Lie algebra L_2/L_D is finite dimensional, we might hope to be able to represent it by finite dimensional matrices. This is indeed possible. We will consider first the whole Lie algebra L_1 and show that it has a representation by infinite dimensional matrices.

We can take as a basis for L_1 the Lie operators $: G_t :$. Then, in accord with (3.38), these Lie operators have the associated matrices

$$O_{sr}(: G_t :) = \langle G_s, : G_t : G_r \rangle. \tag{8.9.62}$$

But we also have the result

$$: G_t : G_r = [G_t, G_r] = \sum_{s'} c_{tr}^{s'} G_{s'}, \tag{8.9.63}$$

where the coefficients $c_{tr}^{s'}$ are the structure constants of the underlying Poisson bracket Lie algebra. See (3.7.43). It follows from (9.62) and (9.63) that we have the relation

$$O_{sr}(: G_t :) = c_{tr}^{s}. \tag{8.9.64}$$

We see that the matrix representation of L_1 is determined by the structure constants. Evidently these matrices are infinite dimensional.

Let us pursue the relation (9.64) a bit further. From (5.3.14), (3.41), and (9.63) we have the result

$$\begin{aligned} \{O(: G_t :), O(: G_{t'} :)\} &= O(\{: G_t :, : G_{t'} :\}) = O(: [G_t, G_{t'}] :) \\ &= O(: \sum_{t''} c_{tt'}^{t''} G_{t''} :) = \sum_{t''} c_{tt''}^{t''} O(: G_{t''} :). \end{aligned} \tag{8.9.65}$$

Next take r, s matrix elements of both sides of (9.65). Doing so and expanding the commutator on the left side of (9.65) gives the result

$$\sum_{t''} O_{rt''}(: G_t :) O_{t''s}(: G_{t'} :) - O_{rt''}(: G_{t'} :) O_{t''s}(: G_t :) = \sum_{t''} c_{tt'}^{t''} O_{rs}(: G_{t''} :). \tag{8.9.66}$$

Finally, use (9.64) in (9.66) to find the relation

$$\sum_{t''} c_{tt''}^{r} c_{t's}^{t''} - c_{t't''}^{r} c_{ts}^{t''} - c_{tt'}^{t''} c_{t''s}^{r} = 0. \tag{8.9.67}$$

It is easily verified with the aid of (3.7.44) that (9.67) is equivalent to (3.7.45), which is in turn a consequence of the Jacobi identity. [That the Jacobi identity should be involved should come as no surprise since (5.3.14), which was used in (9.65), is a consequence of the Jacobi identity.] Notice that the steps we have been following are quite general and hold, in fact, for any Lie algebra. We have learned from (9.65) that matrices defined in terms of the structure constants by the relation (9.64) provide a representation of the underlying Lie algebra. Indeed, a moment's reflection reveals that our present discussion is a recapitulation of that given at the end of Section 3.7. That is, we have found the *adjoint* representation of L_1. Finally, we observe that if we exponentiate the matrices (9.64), we get a representation, again called the adjoint representation, of the group. It follows, in the case of the group of all symplectic maps, that the matrix representation given by (3.34) is just the adjoint representation. The matrices in this representation are also infinite dimensional.

We now turn to the case of the Lie algebra L_2/L_D. As in Section 8.6, define a *truncation* (by degree) operator $\mathcal{T}(>m)$ on the basis functions G_r by the rules

$$\deg(G_r) \leq m \Rightarrow \mathcal{T}(>m)G_r = G_r, \tag{8.9.68}$$

$$\deg(G_r) > m \Rightarrow \mathcal{T}(>m)G_r = 0, \tag{8.9.69}$$

and extend $\mathcal{T}(>m)$ to all functions by requiring that it be a *linear* operator. Given any $D > 2$ and any Lie operator $:f:$ in L_2, we define an associated linear operator $\mathcal{L}_D(:f:)$ by the rule

$$\mathcal{L}_D(:f:) = \mathcal{T}(>D-2):f:\mathcal{T}(>D-2). \tag{8.9.70}$$

We note that since (by hypothesis) $:f:$ is in L_2, it can only preserve or raise the degree of any monomial on which it acts. Therefore the operator $\mathcal{T}(>D-2)$ on the far right of (9.70) is actually redundant. We also define $\mathcal{L}_D$ for other linear operators, for example products of Lie operators in L_2, by rules of the form

$$\mathcal{L}_D(:f::g:) = \mathcal{T}(>D-2):f::g:\mathcal{T}(>D-2). \tag{8.9.71}$$

Again, strictly speaking, the $\mathcal{T}(>D-2)$ on the far right of (9.71) is redundant since $:f:$ and $:g:$ are assumed to be in L_2.

This definition has several important properties: Suppose $:f:$ is also in L_D. Then, from (9.57) and (9.68) through (9.70), we find the result

$$:f:\in L_D \Rightarrow \mathcal{L}_D(:f:) = 0. \tag{8.9.72}$$

Next suppose $:f:$ and $:g:$ are in L_2. Then, from (9.57) and (9.68) through (9.71), we have the product rule

$$\begin{aligned} :f:,:g:\in L_2 \;&\Rightarrow\; \mathcal{L}_D(:f:)\mathcal{L}_D(:g:) \\ &=\; \mathcal{T}(>D-2):f:\mathcal{T}(>D-2)\mathcal{T}(>D-2):g:\mathcal{T}(>D-2) \\ &=\; \mathcal{T}(>D-2):f::g:\mathcal{T}(>D-2) = \mathcal{L}_D(:f::g:). \end{aligned} \tag{8.9.73}$$

Here we have used the fact that, since $:f:$ and $:g:$ are in L_2, the truncation operator product $\mathcal{T}(>D-2)\mathcal{T}(>D-2)$ that occurs in the intermediate expression in (9.73) is redundant. From (9.73) it follows that the operators $\mathcal{L}_D(:f:)$ for $:f:$ in L_2 form a Lie algebra. Indeed, we have the result

$$\begin{aligned} :f:,:g:\in L_2 \Rightarrow \{\mathcal{L}_D(:f:), \mathcal{L}_D(:g:)\} &= \mathcal{L}_D(\{:f:,:g:\}) \\ &= \mathcal{L}_D(:[f,g]:). \end{aligned} \tag{8.9.74}$$

Finally we find that all the $\mathcal{L}_D(:f:)$, for fixed D and $:f:$ in L_2, have only a finite number of nonzero matrix elements. Suppose either G_r or G_s have degree greater than $(D-2)$. Then from (9.57) and (9.68) through (9.70) we find the result

$$\begin{aligned} O_{sr}(\mathcal{L}_D(:f:)) &= \langle G_s, \mathcal{T}(>D-2):f:\mathcal{T}(>D-2)G_r\rangle \\ &= 0 \text{ when } \deg(G_r) > D-2 \text{ or } \deg(G_s) > D-2. \end{aligned} \tag{8.9.75}$$

Recall (3.32) and the fact that each monomial has a definite degree.

We now claim that the matrices $O(\mathcal{L}_D(: f :))$ provide a faithful representation of the quotient algebra L_2/L_D. By construction we have the relation

$$O(\mathcal{L}_D(: f : + : g :)) = O(\mathcal{L}_D(: f :)) + O(\mathcal{L}_D(: g :)), \tag{8.9.76}$$

and from (3.41), (9.73), (9.74) we find the relations

$$O(\mathcal{L}_D(: f :))O(\mathcal{L}_D(: g :)) = O(\mathcal{L}_D(: f :: g :)), \tag{8.9.77}$$

$$\{O(\mathcal{L}_D(: f :)), O(\mathcal{L}_D(: g :)\} = O(\mathcal{L}_D(: [f, g] :)). \tag{8.9.78}$$

Consequently, the matrices $O(\mathcal{L}_D(: f :))$ provide a representation of L_2. Also, according to (9.72), all elements in the ideal L_D are mapped to the zero matrix. Therefore, in view of (9.76), the matrix $O(\mathcal{L}_D(: f :))$ depends only on the equivalence class to which $: f :$ belongs. Thus, the matrices $O(\mathcal{L}_D(: f :))$ provide a representation of the quotient algebra L_2/L_D. Finally, it remains to be shown that this representation is *faithful*. That is, given a matrix $O(\mathcal{L}_D(: f :))$, we need to be able to determine the equivalence class to which $: f :$ belongs. Suppose that f is written in the form

$$f = f_2 + f_3 + \cdots f_{D-1}. \tag{8.9.79}$$

Compute the action of $\mathcal{L}_D(: f :)$ on z_a. We find the result

$$\begin{aligned} \mathcal{L}_D(: f :)z_a &= \mathcal{L}_D(: f_2 :)z_a + \mathcal{L}_D(: f_3 :)z_a + \cdots + \mathcal{L}_D(: f_{D-1} :)z_a \\ &= \mathcal{T}(> D-2)(: f_2 : z_a + : f_3 : z_a + \cdots + : f_{D-1} : z_a) \\ &= \mathcal{T}(> D-2)[g_a(1, z) + g_a(2, z) + \cdots + g_a(D-2, z)] \\ &= g_a(1, z) + g_a(2, z) + \cdots g_a(D-2, z). \end{aligned} \tag{8.9.80}$$

Here we have used the notation

$$g_a(m, z) = : f_{m+1} : z_a = [f_{m+1}, z_a], \tag{8.9.81}$$

and observe that the $g_a(m, z)$ are homogeneous polynomials of degree m. As is evident from (9.80), the polynomials $g_a(m, z)$ can be determined from a knowledge of the matrix elements

$$\begin{aligned} O_{ra} &= \langle G_r, \mathcal{L}_D(: f :)z_a\rangle \\ &= \langle G_r, g_a(1, z)\rangle + \langle G_r, g_a(2, z)\rangle + \cdots + \langle G_r, g_a(D-2, z)\rangle \end{aligned} \tag{8.9.82}$$

with the G_r having degree

$$\deg(G_r) \le D - 2. \tag{8.9.83}$$

Since the $g_a(m, z)$ are now known, we can determine the f_m from (9.81). See (7.6.24).

Is the matrix representation of L_2/L_D just described the adjoint representation? It is not. If it were, the matrix elements of $O(\mathcal{L}_D(: f :))$ would be related to the structure constants of the quotient algebra, which are the ${}^6c^s_{tr}$ where both the subscripts s and r refer to basis elements for the quotient algebra. See (9.28) and (9.64). But examination of (9.82) shows that the functions z_a were used to compute matrix elements, and the Lie operators $:z_a:$ are not in L_2, and therefore not even candidates for the quotient algebra L_2/L_D.

Consider the "truncated" analog of (9.61) written in the form

$$\begin{aligned}\mathcal{M}_f^{\mathcal{T}} &= \exp(\mathcal{L}_D(:f_2^c:))\exp(\mathcal{L}_D(:f_2^a:))\exp(\mathcal{L}_D(:f_3:))\times \\ &\quad \exp(\mathcal{L}_D(:f_4:))\cdots\exp(\mathcal{L}_D(:f_{D-1}:)).\end{aligned} \tag{8.9.84}$$

Also, arrange the labelling of the basis functions G_r so that G_0 corresponds to the constant function 1 (the monomial of degree zero), the G_a (with $a = 1\cdots d$) correspond to the linear monomials z_a, and the subsequent monomials G_r [with $r = d+1\cdots S(D-2,d)$] correspond to the monomials of degrees $2, 3, \cdots (D-2)$. Take matrix elements of both sides of (9.84) using the basis G_r with $r = 1, 2\cdots S(D-2,d)$. Doing so gives the result

$$\begin{aligned}M_f^{\mathcal{T}} &= \exp(O(\mathcal{L}_D(:f_2^c:)))\exp(O(\mathcal{L}_D(:f_2^a:)))\exp(O(\mathcal{L}_D(:f_3:)))\times \\ &\quad \exp(O(\mathcal{L}_D(:f_4:)))\cdots\exp(O(\mathcal{L}_D(:f_{D-1}:))).\end{aligned} \tag{8.9.85}$$

Here use has been made of relations of the form (9.76) and (9.77). We see that (9.85) provides a $S(D-2,d)\times S(D-2,d)$ matrix representation of the quotient Lie *group* associated with the quotient Lie algebra L_2/L_D. Moreover, this representation is faithful. To see the truth of this assertion, consider the matrix elements

$$(M_f^{\mathcal{T}})_{ra} = \langle G_r, \mathcal{M}_f^{\mathcal{T}} G_a\rangle = \langle G_r, \mathcal{L}_D(\mathcal{M}_f)G_a\rangle, \tag{8.9.86}$$

with

$$r \in [1, S(D-2,d)] \text{ and } a \in [1,d]. \tag{8.9.87}$$

From these matrix elements we can determine the coefficients in the Taylor series (7.6.1) through terms of degree $(D-2)$, and from these coefficients we can determine the polynomials $f_2^c, f_2^a, f_3, \cdots f_{D-1}$.

The last concept to be discussed in this section is gradings. For our purposes, a *grading* of a vector space V is a decomposition of V into a direct sum of subspaces along with a function *gr* (called the *grading* function) that assigns an integer (called the *grade*) to all the elements of any subspace. For example, we may take as our vector space V the set of all analytic functions $f(z)$ on phase space. Any such function can be decomposed into homogeneous polynomials by writing

$$f = f_0 + f_1 + f_2 + \cdots, \tag{8.9.88}$$

and these polynomials are in the subspaces we have called $\mathcal{P}_m$. In this case we may define the function *gr* by the rule

$$\mathrm{gr}(f_m) = \deg(f_m) = m. \tag{8.9.89}$$

Elements of V that satisfy (9.89) are called *homogeneous*. Suppose there is some multiplication rule, $\circ$, that makes V into an algebra. See Section 3.7. Suppose also that this multiplication rule, and the direct sum decomposition, are such that the product of any two homogeneous elements is also homogeneous. The multiplication rule and grading function are said to be *compatible* if, for all homogeneous elements, we have the relation

$$\mathrm{gr}(f_m \circ g_n) = \mathrm{gr}(f_m) + \mathrm{gr}(g_n). \tag{8.9.90}$$

For example, in the case of functions, we may take for $\circ$ the operation of ordinary function multiplication. Then, if we use the definition (9.89), we find the result

$$\mathrm{gr}(f_m \circ g_n) = \deg(f_m g_n) = m + n = \mathrm{gr}(f_m) + \mathrm{gr}(g_n), \tag{8.9.91}$$

which shows that ordinary function multiplication and the grading function (9.89) are compatible.

Suppose we use Poisson bracket multiplication for $\circ$ instead of ordinary multiplication. Then (4.28) shows that (9.89) is not compatible with Poisson bracket multiplication. However, if we define *gr* by the rule

$$\mathrm{gr}(f_m) = m - 2, \tag{8.9.92}$$

we find the result

$$\begin{aligned}\mathrm{gr}([f_m, g_n]) &= \mathrm{gr}(\mathcal{P}_{m+n-2}) = m + n - 2 - 2 \\ &= m - 2 + n - 2 = \mathrm{gr}(f_m) + \mathrm{gr}(g_n).\end{aligned} \tag{8.9.93}$$

Thus, the grading function (9.92) is compatible with Poisson bracket multiplication. A Lie algebra equipped with a grading (compatible with the Lie product) is called a *graded* Lie algebra.

Given a graded Lie algebra, it is easy to find subalgebras and ideals. Consider the case of analytic functions defined on phase space. Introduce the notation

$$f^{n-2}(z) = f_n(z) \tag{8.9.94}$$

to indicate, in accord with (9.92), that homogeneous polynomials of degree n have grade $(n-2)$. Equivalently, we have the relation

$$\mathrm{gr}(f^m) = m. \tag{8.9.95}$$

We also introduce the notation

$$\mathcal{P}^{n-2} = \mathcal{P}_n \tag{8.9.96}$$

to indicate the subspace of polynomials of degree n and grade $(n-2)$. Finally, we extend the concept of grade to Lie operators by the rule

$$\mathrm{gr}(: f^m :) = \mathrm{gr}(f^m) = m. \tag{8.9.97}$$

Now consider the space of all Lie operators spanned by basis elements of the form $: f^0 :$, $: f^1 :, : f^2 :, \cdots$. Then, because of the relation

$$\begin{aligned}\mathrm{gr}(\{: f^m :, : f^n\}) &= \mathrm{gr}(: [f^m, f^n] :) = \mathrm{gr}([f^m, f^n]) \\ &= \mathrm{gr}(f^m) + \mathrm{gr}(f^n) = m + n,\end{aligned} \tag{8.9.98}$$

we see that

$$\{: f^m :, : f^n :\} \in : \mathcal{P}^{m+n} :, \tag{8.9.99}$$

and hence this space is a Lie algebra. This is just the Lie algebra that we found and called L_2 earlier, and that we now will also call L^0. Similarly, we can use arguments based on grading to show that $L^m = L_{m+2}$ (with $m > 0$) is a subalgebra of $L^0 = L_2$, and also an ideal in L^0. (Indeed, the arguments we used earlier for this purpose were actually grading arguments without being identified as such.)

Note that the Lie algebra $L^{-1} = L_1$ also has L^m as a subalgebra. However L^m is not an ideal in L^{-1} since L^{-1} contains $: f^{-1} :=: f_1 :$ which, according to (9.99), can lower the grade of elements in L^m until they are no longer in L^m. We have seen that L^0 can be approximated by using the finite dimensional quotient algebras $L^0/L^m = L_2/L_{m+2}$. Is there some way that we can approximate L^{-1} in a consistent Lie algebraic way even though L^m is not an ideal in L^{-1}? We would certainly like to do so since maps of the form (7.8.1) are of interest when the origin is not mapped into itself. The answer to the question just posed is yes provided we are willing to treat the f_1 as being in some sense *small.* Fortunately this circumstance is the one usually encountered since, as we will see in Chapter 14, f_1 terms are usually associated with misalignment, misplacement, and mispowering errors, and these errors are generally small.

Before considering the inclusion of small f_1 terms, let us return to the case described at the beginning of this section. Let ϵ be a (presumably small) parameter, and define V to be the vector space spanned by all phase-space functions of the form $f^{(0)}$, $\epsilon f^{(1)}$, $\epsilon^2 f^{(2)} \cdots$. Here the $f^{(n)}$ are *arbitrary* analytic functions not to be confused with the f^n defined earlier. Assign a grade to the subspaces $\epsilon^n f^{(n)}$ and the associated Lie operators $: \epsilon^n f^{(n)} :$ by the rules

$$\mathrm{gr}(\epsilon^n f^{(n)}) = n, \tag{8.9.100}$$

$$\mathrm{gr}(: \epsilon^n f^{(n)} :) = n. \tag{8.9.101}$$

This grading function is evidently compatible with Lie multiplication,

$$\begin{aligned}\mathrm{gr}([\epsilon^m f^{(m)}, \epsilon^n f^{(n)}]) &= \mathrm{gr}(\epsilon^{m+n}[f^{(m)}, f^{(n)}]) \\ &= m + n = \mathrm{gr}(\epsilon^m f^{(m)}) + \mathrm{gr}(\epsilon^n f^{(n)}),\end{aligned} \tag{8.9.102}$$

$$\begin{aligned}\mathrm{gr}(\{: \epsilon^m f^{(m)} : \ , : \epsilon^n f^{(n)} :\}) = \mathrm{gr}(: [\epsilon^m f^{(m)}, \epsilon^n f^{(n)}] :) \\ = \ m + n = \mathrm{gr}(: \epsilon^m f^{(m)} :) + \mathrm{gr}(: \epsilon^n f^{(n)} :).\end{aligned} \tag{8.9.103}$$

Here we have used the fact that $[f^{(m)}, f^{(n)}]$ is again an arbitrary analytic phase-space function. Thus we may use this grading to construct subalgebras, ideals, and quotient algebras. Suppose we define L^n to be the vector space spanned by $\epsilon^n f^{(n)}$, $\epsilon^{n+1} f^{(n+1)}$, $\epsilon^{n+2} f^{(n+2)} \cdots$. Then it is easily verified that the L^n are Lie algebras. Moreover, any L^n for $n > 0$ is an ideal in L^0, and each L^0/L^n is a quotient algebra. Finally, it is evident that working with the quotient algebra L^0/L^n is equivalent to doing perturbation theory in ϵ and retaining only those terms that carry powers of ϵ of the form ϵ^0, $\epsilon^1, \cdots \epsilon^{n-1}$. What we have learned is that *finite* order perturbation theory is a consistent Lie algebraic procedure.

We now turn to the inclusion of small f_1 terms. Again let ϵ be a parameter. Consider now the vector space V spanned by all functions of the form $\epsilon^m f_n$. Assign a grade to these subspaces and their associated Lie operators $: \epsilon^m f_n :$ by the rules

$$\mathrm{gr}(\epsilon^m f_n) = m + n - 2, \tag{8.9.104}$$

$$\mathrm{gr}(:\epsilon^m f_n:) = m + n - 2. \tag{8.9.105}$$

With this definition we find that $\epsilon^2 f_0$, ϵf_1, and f_2 have grade 0; $\epsilon^3 f_0$, $\epsilon^2 f_1$, ϵf_2, and f_3 have grade 1; $\epsilon^4 f_0$, $\epsilon^3 f_1$, $\epsilon^2 f_2$, ϵf_3, and f_4 have grade 2; etc. From the previous discussion it is easy to see that the grading functions (9.104) and (9.105) are compatible with Lie multiplication. Consequently, we may use it as before to construct subalgebras, ideals, and quotient algebras. Let ${}^{\epsilon}L^{\ell}$ denote the vector space of functions spanned by elements of the form $\epsilon^m f_n$ with $(m + n - 2) \geq \ell$ [that is, $\mathrm{gr}(\epsilon^m f_n) \geq \ell$]. Then, following arguments given previously, it is easy to check that the ${}^{\epsilon}L^{\ell}$ (with $\ell \geq 0$) are Lie algebras, ${}^{\epsilon}L^{\ell}$ with $\ell > 0$ is an ideal in ${}^{\epsilon}L^0$, and each ${}^{\epsilon}L^0/{}^{\epsilon}L^{\ell}$ is a quotient algebra. We also see that ${}^{\epsilon}L^0$ contains the element ϵf_1, which can be interpreted as being a small f_1. Finally, we observe that the quotient algebra ${}^{\epsilon}L^0/{}^{\epsilon}L^{\ell}$, for fixed ℓ, is finite dimensional. (These dimensions are listed below in Table 9.2 for the cases of four and six-dimensional phase spaces. In computing these dimensions we set $\epsilon = 1$ and ignored all terms of the form $\epsilon^m f_0$.) Consequently, we have discovered a systematic and Lie algebraically consistent approximation scheme that includes small f_1 terms. This approximation scheme will be studied extensively in the next chapter.

Table 8.9.2: Values of $\dim({}^{\epsilon}L^0/{}^{\epsilon}L^{\ell})$.

ℓ	dim for $d = 4$	dim for $d = 6$
1	14	27
2	34	83
3	69	209
4	125	461
5	209	923
6	329	1715
7	494	3002
8	714	5004
9	1000	8007
10	1364	12,375
11	1819	18,563

Exercises

8.9.1. Review Exercise 5.12.7. Let L be a vector space having a vector subspace L'. Show that (9.2) defines (satisfies the properties of) an equivalence relation among the elements of L.

8.9.2. Verify that the vectors $\{v_i\}$ used in (9.16) are linearly independent. Hint: Assume there exist scalars α_i such that

$$\sum_{i=1}^{n} \alpha_i \{v_i\} = \{0\}.$$

Show that there exists a vector $x' \epsilon L'$ such that

$$\sum_{i=1}^{n} \alpha_i v_i = x'.$$

Show that since $x' \epsilon L'$, it must have an expansion of the form

$$x' = \sum_j \beta_j b_j.$$

Now show that all the coefficients α_i and β_j must vanish.

8.9.3. Verify the relation

$$[\{0\}, \{x\}]_{qs} = \{0\}.$$

Show that the addition rule (9.12) and Lie product rule (9.18) together satisfy requirements 1 through 5 for a Lie algebra as given in Section 3.7.

8.9.4. Verify (9.30). Verify (9.38) in detail. Show that the definition (9.35) is independent of equivalence class labeling if G' is normal. Verify that the definition (9.32) satisfies the associative property

$$\{f_1\}(\{g_1\}\{h_1\}) = (\{f_1\}\{g_1\})\{h_1\}.$$

8.9.5. If G is a group with a subgroup G', the subgroup G' can be used to produce equivalence classes in G in two possibly different ways. First, there is the equivalence relation $\sim$ defined by

$$g_2 \sim g_1 \Leftrightarrow g_1^{-1} g_2 \in G'.$$

Second, there is another equivalence relation, let us denote it by the symbol $\leftrightarrow$, defined by

$$g_2 \leftrightarrow g_1 \Leftrightarrow g_2 g_1^{-1} \in G'.$$

We know that $\sim$ decomposes G into *left* coset equivalence classes. Show that $\leftrightarrow$ is indeed an equivalence relation, and that it decomposes G into *right* coset equivalence classes. Suppose g is any element of G. Let $\{g\}_\ell$ denote the set of all elements in G that are equivalent to g using the relation $\sim$, and let $\{g\}_r$ denote the set of all elements in G that are equivalent to g using the relation $\leftrightarrow$. Show that

$$\{e\}_\ell = \{e\}_r.$$

Next assume that G' is normal. In this case show that

$$\{g\}_\ell = \{g\}_r$$

for any g in G. Thus, left and right cosets are the same if G' is normal.

8.9.6. The *center* of a group G consists of those elements of G that commute with all elements of G. See Exercise 7.2.13. Show that the center of a group is a special case of an invariant (normal) subgroup. Review Exercise 5.11.1.

8.9.7. Verify (9.43).

8.9.8. Show that if G' with Lie algebra L' is a normal subgroup of a Lie group G with Lie algebra L, then L' is an ideal in L.

8.9.9. Starting with (9.64), verify (9.67) and show that it is equivalent to (3.7.42).

8.9.10. Find and describe the adjoint representation of L_2. What is its dimension? Find and describe the adjoint reprsentation of L_2/L_D. What is its dimension?

8.9.11. Verify (9.73) and (9.74).

8.9.12. Verify (9.77) and (9.78).

8.9.13. Verify (9.85).

8.9.14. Consider a 2-dimensional phase space with canonical variables q, p. Referring to (9.70), find the matrix elements of $\mathcal{L}_4(: q^3 :)$ and $\exp(\mathcal{L}_4(: q^3 :))$. See (9.75), (9.84), and (9.85).

8.9.15. Read Exercise 9.14. Find the matrix representations for the Lie algebras L_2/L_3 and L_2/L_4 in the case of a 2-dimensional phase space.

8.9.16. Verify (9.93).

8.9.17. Use the grading (9.104) and (9.105). Show that it is compatible with Lie multiplication.

8.9.18. Let ${}^{\epsilon}L^{\ell}$ denote the vector space of functions spanned by elements of the form $\epsilon^m f_n$ with $(m + n - 2) \geq \ell$. Show that the ${}^{\epsilon}L^{\ell}$ with $\ell \geq 0$ are Lie algebras, ${}^{\epsilon}L^{\ell}$ with $\ell > 0$ is an ideal in ${}^{\epsilon}L^0$, and each ${}^{\epsilon}L^0/{}^{\epsilon}L^{\ell}$ is a quotient Lie algebra. For a given ℓ and assuming a d-dimensional phase space, show that the dimension of ${}^{\epsilon}L^0/{}^{\epsilon}L^{\ell}$ is given in terms of (7.10.4) by the relation

$$\dim({}^{\epsilon}L^0/{}^{\epsilon}L^{\ell}) = S(\ell + 1, d). \tag{8.9.106}$$

In computing these dimensions, set $\epsilon = 1$ and ignore all terms of the form $\epsilon^m f_0$. Verify Table 9.2.

8.9.19. Review Exercise 8.2.12. This exercise is a continuation of that exercise. Our task is to show that the $L^{\alpha} = L(K^{\alpha})$ form a basis for $so(6, \mathbb{R})$. This task could be carried out by computing all the L^{α} and verifying that they are indeed linearly independent. Instead, we will use an approach that is less tedious but also more abstract.

Suppose, to the contrary, that the $L^{\alpha} = L(K^{\alpha})$ do not form a basis, and are therefore not linearly independent. Then there are constants λ_{α}, not all zero, such that

$$\sum_{\alpha} \lambda_{\alpha} L^{\alpha} = 0. \tag{8.9.107}$$

Use the constants λ_{α} to form an $su(4)$ element, call it $K(\lambda)$, by the rule

$$K(\lambda) = \sum_{\alpha} \lambda_{\alpha} K^{\alpha}. \tag{8.9.108}$$

We know that $K(\lambda) \neq 0$ because the K^{α} are linearly independent and not all the λ_{α} are zero. Show that

$$L[K(\lambda)] = \sum_{\alpha} \lambda_{\alpha} L(K^{\alpha}) = \sum_{\alpha} \lambda_{\alpha} L^{\alpha} = 0. \tag{8.9.109}$$

Next, let $\mathcal{I}$ be the set of all elements in $su(4)$ of the form

$$K(\sigma) = \sum_{\alpha} \sigma_\alpha K^\alpha \tag{8.9.110}$$

such that

$$L[K(\sigma)] = 0. \tag{8.9.111}$$

Show that $\mathcal{I}$ is a Lie subalgebra of $su(4)$. That is, show that $\mathcal{I}$ is a linear vector space and verify that

$$L[\{K(\sigma), K(\sigma')\}] = \{L[K(\sigma)], L[K(\sigma')]\} = 0 \tag{8.9.112}$$

so that $\{K(\sigma), K(\sigma')\} \in \mathcal{I}$ if $K(\sigma) \in \mathcal{I}$ and $K(\sigma') \in \mathcal{I}$. Finally, show that

$$L[\{K^\beta, K(\sigma)\}] = \{L(K^\beta), L[K(\sigma)]\} = 0 \tag{8.9.113}$$

for any β and any $K(\sigma) \in \mathcal{I}$. It follows that $\mathcal{I}$ is an *invariant* subalgebra of $su(4)$. Also, we know that $\mathcal{I}$ is not empty because, by hypothesis, it contains $K(\lambda)$. Nor is it all of $su(4)$ because, for example, inspection of (2.164) shows that $L(K^1) \neq 0$. Therefore $\mathcal{I}$ is an *ideal*. But, this is a contradiction because $su(4)$ is supposed to be *simple*, i.e. have no ideals. See Section 3.7.6.

Bibliography

Relations between Orthogonal and Unitary Groups

[1] J. D. Louck and H. W. Galbraith, “Application of Orthogonal and Unitary Group Methods to the N-Body Problem”, *Reviews of Modern Physics* **44**, 540-601 (1972).

The Lorentz Group

[2] R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics, and All That*, W. A. Benjamin (1964).

Clifford Algebras, Spinors, and Lie Theory

[3] P. Lounesto, *Clifford algebras and spinors* (2nd ed.), Cambridge University Press (2001).

[4] R. Delanghe, F. Sommen, and V. Souček, *Clifford Algebra and Spinor-Valued Functions: A Function Theory For The Dirac Operator*, Springer Science (1992).

[5] E. Meinrenken, *Clifford Algebras and Lie Theory*, Springer Verlag (2013).

Connection between Linear Operators and Matrices

[6] J.M. Finn, “Integrals of Canonical transformations and Normal Forms for Mirror Machine Hamiltonians”, University of Maryland College Park Physics Department Ph.D. thesis (1974).

[7] K. Kowalski and W-H. Steeb, *Nonlinear Dynamical Systems and Carleman Linearization*, World Scientific (1991).

[8] K. Kowalski, *Methods of Hilbert Spaces in the Theory of Nonlinear Dynamical Systems*, World Scientific (1994).

[9] P. Gralewicz and K. Kowalski, “Continuous time evolution from iterated maps and Carleman linearization”, *Chaos, Solitons, and Fractals* **14**, 563-572 (2002).

Transforming Expressions to Commutator Form

[10] E.B. Dynkin, *Selected Papers of E.B. Dynkin with Commentary*, American Mathematical Society (2000).

[11] A. Dragt and E. Forest, "Computation of nonlinear behavior of Hamiltonian systems using Lie algebraic methods", J. Math. Phys. **24**, p. 2734 (1983). See Section 11 and Appendix.

Single Exponent Form

[12] S. Habib and R. Ryne, "Symplectic Calculation of Lyapunov Exponents", arXiv:acc-phys/9411001v1, (1994).

[13] I. Gjaja, "Closed-Form Expressions for the Noncompact Part of $Sp(2n)$", arXiv:chao-dyn/9602013v1, (1996).

Zassenhaus Formulas

[14] R.M. Wilcox, "Exponential Operators and Parameter Differentiation in Quantum Physics", J. Math. Phys. **8**, p. 962 (1967).

[15] F. Casas, A. Murua, and M. Nadinic, "Efficient computation of the Zassenhaus formula", available on arXiv (2012) at http://arxiv.org/abs/1204.0389.

[16] F. Bayen, "On the convergence of the Zassenhaus formula", *Lett. Math. Phys.* **3**, 161-167 (1979).

Operational Calculus for Noncommuting Operators

[17] G. Johnson and M. Lapidus *The Feynman Integral and Feynman's Operational Calculus*, Oxford University Press (2003).

[18] G. Johnson, M. Lapidus, and L. Nielsen *Feynman's Operational Calculus and Beyond: Noncommutativity and Time-Ordering*, Oxford University Press (2015).

Ideals and Quotient Groups

[19] N. Jacobson, *Lectures in Abstract Algebra, Vol. I - Basic Concepts*, D. Van Nostrand (Princeton, 1951).

Gradings

[20] I. Dorfman, *Dirac Structure and Integrability of Nonlinear Evolution Equations*, John Wiley and Sons (Chichester, 1993).

Chapter 9

Inclusion of Translations in the Calculus

9.1 Introduction

In Chapter 8 we dealt with, among other things, the restricted problem of concatenating maps, all of which had the property of sending the origin into itself. In this chapter we consider the general case. Let $\mathcal{M}_f$ and $\mathcal{M}_g$ denote the general symplectic maps given by the expressions

$$\mathcal{M}_f = \exp(:f_1:)\exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_3:)\exp(:f_4:)\cdots, \tag{9.1.1}$$

$$\mathcal{M}_g = \exp(:g_1:)\exp(:g_2^c:)\exp(:g_2^a:)\exp(:g_3:)\exp(:g_4:)\cdots. \tag{9.1.2}$$

Also, let $\mathcal{M}_h$ be the product of $\mathcal{M}_f$ and $\mathcal{M}_g$,

$$\mathcal{M}_h = \mathcal{M}_f\mathcal{M}_g. \tag{9.1.3}$$

Given $\mathcal{M}_f$ and $\mathcal{M}_g$, our problem will be to find polynomials h_1, h_2^c, h_2^a, h_3, h_4, etc. such that

$$\mathcal{M}_h = \exp(:h_1:)\exp(:h_2^c:)\exp(:h_2^a:)\exp(:h_3:)\exp(:h_4:)\cdots. \tag{9.1.4}$$

For future use it is convenient to use the notation

$$\mathcal{R}_f = \exp(:f_2^c:)\exp(:f_2^a:),\ \text{etc.} \tag{9.1.5}$$

In this notation, our goal is to write $\mathcal{M}_h$ in the form

$$\mathcal{M}_h = \exp(:h_1:)\mathcal{R}_h\exp(:h_3:)\exp(:h_4:)\cdots. \tag{9.1.6}$$

Section 9.2 treats the special case where both the maps $\mathcal{M}_f$ and $\mathcal{M}_g$ produce only constant and linear terms when acting on z_a. This is the case of $ISp(2n,\mathbb{R})$ where all the nonlinear generators $f_3, f_4, \cdots$ and $g_3, g_4, \cdots$ are assumed to be zero. See Section 6.2 and Exercise 7.7.2. In this case we will be able to solve all possible problems to our hearts' content.

Subsequent sections will treat the general case where the nonlinear generators are also present. In this case, following the discussion in Section 8.9, we will find it necessary to introduce a grading in which f_1 and g_1 are treated as being small.

9.2 The Inhomogeneous Symplectic Group $ISp(2n,\mathbb{R})$

The *inhomogeneous* symplectic group $ISp(2n,\mathbb{R})$ consists of all transformations of phase space of the form

$$\bar{z}_a = \delta_a + \sum_b R_{ab} z_b \tag{9.2.1}$$

where the δ_a are arbitrary constants and R is a symplectic matrix. It is closely related to the *Jacobi* group. See Exercises 6.2.2, 7.7.2, and 7.7.3. We already know from Section 7.7 that there is a map $\mathcal{M}$ such that

$$\bar{z} = \mathcal{M} z, \tag{9.2.2}$$

and $\mathcal{M}$ has the factorization

$$\mathcal{M} = \exp(:f_2^c:)\exp(:f_2^a:)\exp(:g_1:). \tag{9.2.3}$$

Indeed, f_2^a and f_2^c are determined by R using (7.6.17), (7.2.2), (7.2.3), and (7.2.8); and g_1 is given in terms of δ by (7.7.3). Lie operators of the form $:f_1$ and $:f_2:$ provide a basis for $isp(2n,\mathbb{R})$, the Lie algebra of $ISp(2n,\mathbb{R})$. The purpose of this section is to state and prove various rearrangement, factorization, and concatenation formulas for the inhomogeneous symplectic group.

9.2.1 Rearrangement Formula

We begin with a *rearrangement* formula. Let us rewrite (2.3) in the form

$$\mathcal{M}_f = \mathcal{R}_f \exp(:g_1:) \tag{9.2.4}$$

with $\mathcal{R}_f$ being the linear map defined by (1.5). In terms of this notation we have the results

$$\mathcal{R}_f z = R_f z, \tag{9.2.5}$$

$$\exp(:g_1:)z = z + \delta, \tag{9.2.6}$$

$$\bar{z} = \mathcal{M}_f z = \mathcal{R}_f \exp(:g_1:)z = \mathcal{R}_f(z+\delta) = \delta + R_f z. \tag{9.2.7}$$

Next let δ' be another set of constants. Using (7.7.3), define a first-degree polynomial f_1 such that

$$\exp(:f_1:)z = z + \delta'. \tag{9.2.8}$$

Consider the map $\exp(:f_1:)\mathcal{R}_f$. It has the action

$$\exp(:f_1:)\mathcal{R}_f z = \exp(:f_1:)R_f z = R_f(z+\delta') = R_f\delta' + R_f z. \tag{9.2.9}$$

Upon comparing (2.7) and (2.9) we see that there will be the equality

$$\mathcal{R}_f \exp(:g_1:) = \exp(:f_1:)\mathcal{R}_f \tag{9.2.10}$$

provided there is the relation

$$R_f\delta' = \delta. \tag{9.2.11}$$

We will call the relation specified by (2.10) and (2.11) a rearrangement formula.

There is another way to obtain the rearrangement formula. Insert the identity factor $\mathcal{R}_f^{-1}\mathcal{R}_f$ on the right side of (2.4) to get the result

$$\mathcal{M}_f = \mathcal{R}_f \exp(: g_1 :)\mathcal{R}_f^{-1}\mathcal{R}_f. \tag{9.2.12}$$

Next make use of (8.2.25) to write

$$\mathcal{R}_f \exp(: g_1 :)\mathcal{R}_f^{-1} = \exp(: \mathcal{R}_f g_1 :). \tag{9.2.13}$$

Now define a first-degree polynomial f_1 by the rule

$$f_1 = \mathcal{R}_f g_1. \tag{9.2.14}$$

We have found the result

$$\mathcal{M}_f = \mathcal{R}_f \exp(: g_1 :) = \exp(: f_1 :)\mathcal{R}_f \tag{9.2.15}$$

with f_1 and g_1 related by (2.14).

What remains is to make (2.15) more explicit. According to the work of Section 7.7 there are the relations

$$f_1(z) = (J\delta', z), \tag{9.2.16}$$

$$g_1(z) = (J\delta, z). \tag{9.2.17}$$

Consequently, (2.15) can be rewritten in the form

$$(J\delta', z) = f_1(z) = \mathcal{R}_f g_1 = .\mathcal{R}_f(J\delta, z) = (J\delta, R_f z) = (R_f^T J\delta, z). \tag{9.2.18}$$

Upon comparing the left and right sides of (2.18) we conclude that

$$J\delta' = R_f^T J\delta. \tag{9.2.19}$$

Now multiply both sides of (2.19) by $-R_f J$. Doing so to the left side of (2.19) yields the result

$$-R_f JJ\delta' = R_f\delta'. \tag{9.2.20}$$

And doing so to the right side of (2.19) yields the result

$$-R_f J R_f^T J\delta = -JJ\delta = \delta. \tag{9.2.21}$$

Upon comparing the right sides of (2.20) and (2.21) we see that (2.11) has been recovered.

9.2.2 Factorization Formula

The next result to obtain is a *factorization* formula. Given any two homogeneous polynomials h_1 and h_2 (of degrees 1 and 2, respectively), there exist related polynomials f_1, f_2 that satisfy the *factorization formula*

$$\exp(: h_1 + h_2 :) = \exp(: f_2 :)\exp(: f_1 :). \tag{9.2.22}$$

There are at least two ways to prove (2.22). The first amounts to combining the two exponents on the right side of (2.22), and then matching terms with the left side. Consider all Lie products made from f_1 and f_2. We observe that all Lie products containing two or more f_1 factors, for example $[f_1,[f_1,f_2]]$, must give only constant terms, and hence their associated Lie operators vanish. Let s and t be small parameters. According to formula (8.7.3), we have the result

$$\begin{aligned}&\exp(s:f_2:)\exp(t:f_1:)=\\&\exp[s:f_2:+:\{s:f_2:[1-\exp(-s:f_2:)]^{-1}(tf_1)\}:+:O(t^2):],\end{aligned}\tag{9.2.23}$$

where the notation $:O(t^2):$ indicates Lie operators that contain at least *two* factors of t. But the presence of two factors of t requires the presence of two factors of f_1 and hence, by the previous observation these Lie operators must all vanish. It follows that (2.23) is *exact* for f_1 and f_2. Now set $s=t=1$ in (2.23) and combine the result so obtained with (2.22) to get the relation

$$\exp(:h_1+h_2:)=\exp[:f_2:+:\{:f_2:[1-\exp(-:f_2:)]^{-1}f_1\}:].\tag{9.2.24}$$

Upon comparing like terms in (2.6), we find the relations

$$f_2=h_2,\tag{9.2.25}$$

$$\begin{aligned}h_1&=:f_2:[1-\exp(-:f_2:)]^{-1}f_1\\&=:h_2:[1-\exp(-:h_2:)]^{-1}f_1.\end{aligned}\tag{9.2.26}$$

Finally, (2.26) may be solved for f_1 to give the relation

$$\begin{aligned}f_1&=\{[1-\exp(-:h_2:)]/[:h_2:]\}h_1\\&=\text{iex}(-:h_2:)h_1.\end{aligned}\tag{9.2.27}$$

See (8.8.9). We note that f_1 and f_2 are well defined by (2.27) and (2.25) for all h_1 and h_2.

The converse question is more difficult: given f_1 and f_2, does (2.26) always define an h_1? Or equivalently, in view of (2.27), does $[\text{iex}(-:f_2:)]^{-1}f_1$ always exist? If not, then it is not possible to write every inhomogeneous symplectic group element in terms of a single exponential. See Exercise 2.3 for a discussion of this question.

There is a second derivation of (2.25) and (2.27) that is worth knowing. Consider the functions $\overline{z}_a$ defined by the relation

$$\overline{z}_a=\exp(:sh_2+th_1:)z_a.\tag{9.2.28}$$

As before, s and t are parameters. Expanding (2.28) in a power series gives the result

$$\overline{z}_a=\sum_{n=0}^{\infty}[(:sh_2+th_1:)^n/n!]z_a.\tag{9.2.29}$$

Next expand the powers in (2.29). For $n \geq 1$ we have the result

$$\begin{aligned}(: sh_2 + th_1 :)^n &= (s : h_2 : + t : h_1 :)^n \\ &= s^n : h_2 :^n + ts^{n-1} \sum_{m=0}^{n-1} : h_2 :^m : h_1 :: h_2 :^{n-m-1} + O(t^2). \end{aligned} \quad (9.2.30)$$

Here we have kept track of the fact that $: h_1 :$ and $: h_2 :$ may not commute. Observe that the terms in (2.30) proportional to t^2 must have two factors of $: h_1 :$ and are therefore of the form

$$t^2 \text{ terms} \sim (: h_2 :)^\alpha : h_1 : (: h_2 :)^\beta : h_1 : (: h_2 :)^\gamma \quad (9.2.31)$$

where α, β, γ satisfy the relations

$$\begin{aligned}&\alpha + \beta + \gamma = n - 2, \\ &\alpha \geq 0 \ , \ \beta \geq 0 \ , \ \gamma \geq 0. \end{aligned} \quad (9.2.32)$$

From (7.6.16) it is easy to verify the relation

$$(: h_2 :)^\alpha : h_1 : (: h_2 :)^\beta : h_1 : (: h_2 :)^\gamma z_a = 0. \quad (9.2.33)$$

Similarly, analogous results hold for terms having three or more factors of $: h_1 :$. It follows that all the $O(t^2)$ terms in (2.30 annihilate the z_a. Thus, we find the exact result

$$\begin{aligned}&\exp(: sh_2 + th_1 :)z_a = \\ &\left\{ \sum_{n=0}^{\infty} s^n : h_2 :^n /n! + t \sum_{n=1}^{\infty} (s^{n-1}/n!) \sum_{m=0}^{n-1} : h_2 :^m : h_1 :: h_2 :^{n-m-1} \right\} z_a. \end{aligned} \quad (9.2.34)$$

The first term on the right side of (2.34) sums to the exponential function,

$$\sum_{n=0}^{\infty} s^n : h_2 :^n /n! = \exp(s : h_2 :). \quad (9.2.35)$$

The second term sums to a relation involving the exponential and integrated exponential functions,

$$\sum_{n=1}^{\infty} \sum_{m=0}^{n-1} (1/n!) : sh_2 :^m : th_1 :: sh_2 :^{n-m-1} = \exp(s : h_2 :)\text{iex}(-\#sh_2\#) : th_1 : . \quad (9.2.36)$$

See Appendix C. As a consequence of relations of the form (8.2.22), we have the result

$$\text{iex}(-\#sh_2\#) : th_1 := : \text{iex}(- : sh_2 :)th_1 : . \quad (9.2.37)$$

Putting (2.34) through (2.37) together gives the relation

$$\exp(: sh_2 + th_1 :)z_a = \exp(s : h_2 :)[1 + : \text{iex}(- : sh_2 :)th_1 :]z_a. \quad (9.2.38)$$

Suppose we define f_1 by writing

$$f_1 = \mathrm{iex}(- : sh_2 :)h_1. \tag{9.2.39}$$

Again from (7.6.16) we have the relation

$$[1 + t : f_1 :]z_a = \exp(t : f_1 :)z_a. \tag{9.2.40}$$

We conclude that (2.38) can be rewritten in the form

$$\exp(: sh_2 + th_1 :)z_a = \exp(s : h_2 :)\exp(t : f_1 :)z_a. \tag{9.2.41}$$

Since the operator factors on both sides of (2.23) are manifestly group elements (symplectic maps), we get the group (map) factorization relation

$$\exp(: sh_2 + th_1 :) = \exp(s : h_2 :)\exp(t : f_1 :) \tag{9.2.42}$$

with f_1 defined by (2.39). Now put $s = t = 1$ in (2.39) and (2.42) and make the definition (2.25) to recover (2.22), (2.25), and (2.27).

9.2.3 Concatenation Formulas

We have found the rearrangement formula given by (2.10) and (2.11) and the factorization formula given by (2.22), (2.25, and (2.27). We now turn to the easier subjects of concatenation formulas. Let $\mathcal{M}_f$ and $\mathcal{M}_g$ be the inhomogeneous symplectic group maps

$$\mathcal{M}_f = \exp(: f_1 :)\exp(: f_2^c :)\exp(: f_2^a :) = \exp(: f_1 :)\mathcal{R}_f, \tag{9.2.43}$$

$$\mathcal{M}_g = \exp(: g_1 :)\exp(: g_2^c :)\exp(: g_2^a :) = \exp(: g_1 :)\mathcal{R}_g. \tag{9.2.44}$$

Also, let $\mathcal{M}_h$ be the product of $\mathcal{M}_f$ and $\mathcal{M}_g$ as in (1.3). We wish to find polynomials h_1, h_2^c, and h_2^a such that

$$\mathcal{M}_h = \mathcal{M}_f\mathcal{M}_g = \exp(: h_1 :)\exp(: h_2^c :)\exp(: h_2^a :) = \exp(: h_1 :)\mathcal{R}_h. \tag{9.2.45}$$

From (2.43) and (2.44) we have the result

$$\mathcal{M}_f\mathcal{M}_g = \exp(: f_1 :)\mathcal{R}_f\exp(: g_1 :)\mathcal{R}_g. \tag{9.2.46}$$

Insert an identity factor of the form $\mathcal{R}_f^{-1}\mathcal{R}_f$ into (2.46) to find the relation

$$\mathcal{M}_f\mathcal{M}_g = \exp(: f_1 :)\mathcal{R}_f\exp(: g_1 :)\mathcal{R}_f^{-1}\mathcal{R}_f\mathcal{R}_g. \tag{9.2.47}$$

From (2.13) we conclude that (2.47) can be rewritten in the form

$$\begin{aligned}\mathcal{M}_f\mathcal{M}_g &= \exp(: f_1 :)\exp(: \mathcal{R}_f g_1 :)\mathcal{R}_f\mathcal{R}_g \\ &= \exp(: f_1 + \mathcal{R}_f g_1 :)\mathcal{R}_f\mathcal{R}_g.\end{aligned} \tag{9.2.48}$$

Here we have used the fact that the Lie operators associated with first-degree polynomials commute. (See Exercise 2.4.) Now compare (2.45) and (2.48) to get the concatenation formulas

$$h_1 = f_1 + \mathcal{R}_f g_1, \tag{9.2.49}$$

$$\mathcal{R}_h = \mathcal{R}_f\mathcal{R}_g. \tag{9.2.50}$$

We note that (2.50) is identical to (8.4.19), as expected, and consequently we also have the corresponding matrix relation (8.4.20) as before.

Exercises

9.2.1. Read (2.22) from right to left. That is, assume that f_1 and f_2 are two given homogeneous polynomials (of degrees 1 and 2, respectively), and we wish to find h_1 and h_2. From Exercise 7.7.2 we know that $: f_1 :$ and $: f_2 :$ generate a Lie algebra. Also, we know from the BCH formula (3.7.33) and (3.7.34) that all terms that occur when we combine the exponents on the right side of (2.22) must be in this Lie algebra. Show that the most general such term is of the form $: h_1 + h_2 :$ where h_1 and h_2 have degrees 1 and 2, respectively. Show from (8.7.3) that h_1 as defined by the first line in (2.26) does indeed have degree 1, so that terms of like degree have indeed been equated.

9.2.2. The purpose of this exercise is to convert (2.27) into an explicit matrix equation. Since h_1 is a given degree 1 polynomial, it can be written in the form

$$h_1 = \sum_a h_1^a z_a \tag{9.2.51}$$

where the h_1^a are known coefficients. From (7.6.16) the action of $: h_2 :$ on the z_a is a relation of the form

$$: h_2 : z_a = \sum_{a'} H_{a'a} z_{a'}. \tag{9.2.52}$$

The matrix H is given in terms of the scalar product (7.3.8) by the relation

$$H_{a'a} = (z_{a'}, : h_2 : z_a). \tag{9.2.53}$$

Define a matrix O in terms of H by the rule

$$O = \text{iex}(-H) = \sum_{m=0}^{\infty} (-H)^m/(m+1)!. \tag{9.2.54}$$

Show that the series (2.54) converges for any matrix H, and therefore O is well defined. Next verify the relation

$$\text{iex}(- : h_2 :) z_a = \sum_{a'} O_{a'a} z_{a'}. \tag{9.2.55}$$

Suppose we write f_1 in the form

$$f_1 = \sum_{a'} f_1^{a'} z_{a'}. \tag{9.2.56}$$

Show that (2.27) implies the relation

$$f_1^{a'} = \sum_a O_{a'a} h_1^a. \tag{9.2.57}$$

What remains is to determine more explicitly the matrix H. Following (7.2.3), let us write h_2 in the form

$$h_2 = -(1/2) \sum_{de} S_{de} z_d z_e \tag{9.2.58}$$

where S is a symmetric matrix. Following (7.2.4), show that

$$H = (JS)^T = -SJ = J(JSJ) = JS' \tag{9.2.59}$$

where

$$S' = JSJ. \tag{9.2.60}$$

Show that S' is symmetric.

9.2.3. This exercise examines the following question: given f_2^c, f_2^a, and f_1, when do there exist h_1 and h_2 such that there is the relation

$$\exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_1:) = \exp(:h_1 + h_2:)? \tag{9.2.61}$$

Thanks to Section 8.7 we know that such a relation is not always possible even when f_1 is absent (and its presence never helps). So we should phrase the question more narrowly: given f_2 and f_1, when do there exist an h_1 and h_2 such that (2.22) holds? Even this question is too broad. Given an $\mathcal{R}$ such as in (1.5), we are in effect given a symplectic matrix R and the requirement

$$\mathcal{R}z_a = \sum_b R_{ab} z_b. \tag{9.2.62}$$

There may be *many* f_2 such that

$$\mathcal{R} = \exp(:f_2:) \tag{9.2.63}$$

and (2.62) holds. Give an explicit example of this fact. [Hint: look at (7.2.23).] Thus, the question should be made still narrower: Given f_1 and a symplectic matrix R with the property that there exists some f_2 such that (2.62) and (2.63) hold, when do there exist h_1 and h_2 such that

$$\mathcal{R}\exp(:f_1:) = \exp(:h_1 + h_2:)? \tag{9.2.64}$$

We will begin with the case of 2-dimensional phase space. Following (8.7.25), let us write h_2 in the form

$$h_2 = -(bp^2 + 2aqp + cq^2)/2. \tag{9.2.65}$$

With this definition, show from (8.7.27) and (8.7.28) that the H matrix of (2.53) is given by the relation

$$H = F^T = \begin{pmatrix} a & -c \\ b & -a \end{pmatrix}. \tag{9.2.66}$$

Like F, the matrix H has the property

$$H^2 = \Delta I \tag{9.2.67}$$

where Δ is the discriminant (8.7.30). We are now prepared to compute O as given by (2.54). From (2.54) and (8.7.9) derive the formal result

$$\begin{aligned} O &= \text{iex}(-H) = \sum_{m=0}^{\infty}(-H)^m/(m+1)! = -[\exp(-H) - I]/H \\ &= -[\cosh(H) - \sinh(H) - I]/H = [\sinh(H)]/H - [\cosh(H) - I]/H. \end{aligned} \tag{9.2.68}$$

Show using (2.67) that the series in (2.68) have the sums

$$[\sinh(H)]/H = I[\sinh(\Delta^{1/2})]/\Delta^{1/2}, \tag{9.2.69}$$

$$-[\cosh(H) - I]/H = -H[\cosh(\Delta^{1/2}) - 1]/\Delta. \tag{9.2.70}$$

Thus, show that O has the explicit matrix form

$$O = \tag{9.2.71}$$
$$\begin{pmatrix} [\sinh(\Delta^{1/2})]/\Delta^{1/2} - a[\cosh(\Delta^{1/2}) - 1]/\Delta & c[\cosh(\Delta^{1/2}) - 1]/\Delta \\ -b[\cosh(\Delta^{1/2}) - 1]/\Delta & [\sinh(\Delta^{1/2})]/\Delta^{1/2} + a[\cosh(\Delta^{1/2}) - 1]/\Delta \end{pmatrix}.$$

Note that O is well defined for all values of a, b, c as expected. Verify that O has the determinant

$$\det(O) = 2[\cosh(\Delta^{1/2}) - 1]/\Delta. \tag{9.2.72}$$

It follows that O is not invertible when $\Delta = -4\pi^2, -16\pi^2, \cdots$. Indeed, inspection of (2.71) shows that O vanishes identically for these Δ values! Give a geometrical argument showing that this must be the case. From looking at (2.57) we conclude that, given f_1, h_1 cannot be determined when O is not invertible. Therefore, we are tempted to conclude that our single-exponent goal (2.64) cannot be accomplished in this case. However, we see from (8.7.34) that R is the *identity* matrix when $\Delta = -4\pi^2, -16\pi^2, \cdots$. In this case we may take $f_2 = 0$ and $h_2 = f_2 = 0$. Show that O is then the identity matrix, and the single-exponent goal is trivially achieved. We conclude, despite our fears to the contrary, that the single-exponent goal (2.64) can always be achieved in the 2-dimensional phase-space case. What about the cases of higher dimensional phase spaces, say 4 and 6-dimensional phase spaces? These cases appear to be more difficult. They will be treated after we have developed a theory of normal forms for quadratic polynomials.

9.2.4. Verify that Lie operators associated with first-degree polynomials commute. That is, let f_1 and g_1 be any two first-degree polynomials. Show that

$$\{: f_1 :, : g_1 :\} =: [f_1, g_1] := 0 \tag{9.2.73}$$

using (5.3.14), (5.3.21), and (7.6.14). From Exercise 7.7.2 we know that Lie operators of the form $: f_1 :$ and $: f_2 :$ comprise the Lie algebra $isp(2n,\mathbb{R})$. Show that the subalgebra composed of Lie operators of the form $: f_1 :$ comprises a subalgebra that is an ideal in $isp(2n,\mathbb{R})$. Show that $isp(2n,\mathbb{R})$ is not semisimple. See Section 8.9. It is in fact the *semi-direct sum* of the Lie subalgebra generated by Lie operators of the form $: f_2 :$, namely the subalgebra $sp(2n,\mathbb{R})$, and the Lie subalgebra generated by Lie operators of the form $: f_1 :$, namely the Lie subalgebra of the translation group. The sum is called *semi-direct* because Lie operators of the form $: f_2 :$ and do not commute with Lie operators of the form $: f_1 :$. Instead, Lie operators of the form $: f_2 :$ have a nontrivial action on first-order polynomials and, correspondingly, on Lie operators of the form $: f_1 :$. See Section 25.2.1.

9.2.5. Equation (2.22) gives a reverse factorization. Consider the problem of making the forward factorization

$$\exp(: h_1 + h_2 :) = \exp(: g_1 :) \exp(: g_2 :). \tag{9.2.74}$$

Show that in this case one has the results

$$g_2 = h_2, \tag{9.2.75}$$

$$g_1 = \text{iex}(:h_2:)h_1. \tag{9.2.76}$$

Hint: Either invert both sides of (2.22), or use (2.22), (2.10), and (2.11).

9.2.6. Review (2.1) through (2.3) and (7.7.3). Consider the general $ISp(2n,\mathbb{R})$ element $\mathcal{M}$ given by (2.3). When $g_1 = 0$ we know that $\mathcal{M}$ has the origin as a fixed point. The purpose of this exercise is to show that when $g_1 \neq 0$ the map $\mathcal{M}$ generally still has a fixed point. Suppose that $\mathcal{M}$ has a fixed point z^f. Show from (2.1) that there must then be the relation

$$z^f = \delta + Rz^f, \tag{9.2.77}$$

from which it follows that z^f is uniquely defined by the relation

$$z^f = -(R-I)^{-1}\delta \tag{9.2.78}$$

provided the matrix $(R-I)$ has an inverse. For $(R-I)$ to have an inverse it must be the case that

$$\det(R-I) \neq 0. \tag{9.2.79}$$

Consequently, provided R does not have $+1$ as an eigenvalue, $\mathcal{M}$ has a fixed point z^f given by (2.78), and this fixed point is unique. In particular, if the eigenvalues of R lie on the unit circle so that tunes are defined, no tune may be integer.

Provide an example of a symplectic matrix R, with some eigenvalue equal $+1$, for which $\mathcal{M}$ does not have a fixed point for some $g_1 \neq 0$. For example, study in detail the 2×2 case for which

$$R = \begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix} \tag{9.2.80}$$

where λ is an arbitrary parameter. Show that, when $\delta_2 \neq 0$, $\mathcal{M}$ has no fixed point. Show that, when $\delta_2 = 0$ and $\lambda \neq 0$, $\mathcal{M}$ has a whole line of fixed points and therefore does not have a unique fixed point. What happens when $\lambda = 0$?

Given $\mathcal{M}$ and any degree one polynomial h_1, define a map $\mathcal{N}$ by the relation

$$\mathcal{N} = \exp(:h_1:)\mathcal{M}\exp(-:h_1:). \tag{9.2.81}$$

Show that, when

$$h_1 = (z, Jz^f), \tag{9.2.82}$$

there is the relation

$$\mathcal{N} = \mathcal{R}. \tag{9.2.83}$$

Thus, show that generally $\mathcal{M}$ has the factorization

$$\mathcal{M} = \exp(-:h_1:)\mathcal{R}\exp(:h_1:) \tag{9.2.84}$$

with h_1 given by (2.82), z^f being the fixed point of $\mathcal{M}$, and $\mathcal{R}$ being the linear part of $\mathcal{M}$. However, not all $\mathcal{M}$ can be written in the form (2.84) because we know from the work of the previous paragraph that there are some $\mathcal{M}$ that do not have a fixed point.

9.2.7. Consider a generating function of the form

$$g(u) = (k, u) + (1/2)(u, Wu) \tag{9.2.85}$$

where k is a vector with $2n$ entries and W is a general $2n \times 2n$ symmetric matrix. That is, g consists of arbitrary linear and quadratic parts. Using (6.7.21), find the associated symplectic map $\mathcal{M}$ for any Darboux matrix α, and show that $\mathcal{M}$ is an element of $ISp(2n, \mathbb{R})$.

9.3 Lie Concatenation in the General Nonlinear Case

We now turn to the general case $ISpM(2n, \mathbb{R})$ where we have to take into account the presence of the nonlinear generators $f_3, f_4 \cdots$ and $g_3, g_4 \cdots$. In this case, the Lie algebras are infinite dimensional and, as described in Section 8.9, we will introduce a quotient-space structure in order to produce an approximation scheme. This quotient-space structure will be based on the grading given by (8.9.78) and (8.9.79). As we will see, it amounts to treating the quantities f_1 and g_1 in (1.1) and (1.2) as being small.

For purposes of illustration, we will begin our discussion with the case in which we retain only f_3 and f_4 in (1.1) and g_3 and g_4 in (1.2). Correspondingly, we will only retain h_3 and h_4 in (1.3). In addition, we will let ϵ be a parameter which we will initially regard as small, but will eventually set equal to one. Now consider terms of the form $\epsilon^m f_n$, and corresponding terms for the g's and h's. We assign these terms a grade following (8.9.78) and (8.9.79),

$$\text{grade } 0: \quad \epsilon^2 f_0, \epsilon f_1, f_2; \tag{9.3.1}$$

$$\text{grade } 1: \quad \epsilon^3 f_0, \epsilon^2 f_1, \epsilon f_2, f_3; \tag{9.3.2}$$

$$\text{grade } 2: \quad \epsilon^4 f_0, \epsilon^3 f_1, \epsilon^2 f_2, \epsilon f_3, f_4. \tag{9.3.3}$$

We have already seen that these elements span the quotient Lie algebra ${}^{\epsilon}L^0/{}^{\epsilon}L^3$, and we will work within this quotient Lie algebra and its corresponding quotient group. Note that since we will be working with Lie operators, and the Lie operator for a constant function vanishes, terms of the form $\epsilon^m f_0$ actually play no role.

Following the notation (2.26) and (8.4.14), let us rewrite (1.1), (1.2), and (1.4) in the form

$$\mathcal{M}_f = \exp(:\epsilon f_1:)\mathcal{R}_f \exp(:f_3:)\exp(:f_4:), \tag{9.3.4}$$

$$\mathcal{M}_g = \exp(:\epsilon g_1:)\mathcal{R}_g \exp(:g_3:)\exp(:g_4:), \tag{9.3.5}$$

$$\mathcal{M}_h = \exp(:\epsilon h_1:)\mathcal{R}_h \exp(:h_3:)\exp(:h_4:). \tag{9.3.6}$$

Here we have replaced f_1 by ϵf_1, etc. Upon comparing (3.4) through (3.6), we see that performing the multiplication (1.3) requires that all first-order exponents be moved to the extreme left. We will do this in steps. Let us write the product (1.3) in the form

$$\mathcal{M}_h = \exp(:\epsilon f_1:)\mathcal{R}_f \exp(:f_3:)\exp(:f_4:)\exp(:\epsilon g_1:)\mathcal{R}_g \exp(:g_3:)\exp(:g_4:). \tag{9.3.7}$$

The first step will be to move the first-order exponent, g_1 in this case, to the left of f_4. We begin with the relation

$$\begin{aligned} \exp(:f_4:)\exp(:\epsilon g_1:) &= \exp(:\epsilon g_1:)\exp(-:\epsilon g_1:)\exp(:f_4:)\exp(:\epsilon g_1:) \\ &= \exp(:\epsilon g_1:)\exp(:\exp(-:\epsilon g_1:)f_4:). \end{aligned} \tag{9.3.8}$$

Here use has been made of (8.2.20). Now we use the relation

$$\begin{aligned} : \exp(- : \epsilon g_1 :) f_4 :=: & \left[\sum_{m=0}^{\infty} (: -\epsilon g_1 :^m /m!) f_4 \right] : \\ = & \ : [f_4 - \epsilon : g_1 : f_4 + (\epsilon^2/2!) : g_1 :^2 f_4 - (\epsilon^3/3!) : g_1 :^3 f_4] : . \end{aligned} \tag{9.3.9}$$

Note that the apparently infinite series in (3.9) actually terminates because of (7.6.16). Also, use has been made of (5.3.21).

Let us rewrite (3.8) and (3.9) in the form

$$\exp(: f_4 :) \exp(: \epsilon g_1 :) = \exp(: \epsilon g_1 :) \exp(: j_1^{(2)} + j_2^{(2)} + j_3^{(2)} + j_4^{(2)} :) \tag{9.3.10}$$

where the $j_i^{(2)}$ are the homogeneous polynomials

$$j_1^{(2)} = -(\epsilon^3/3!) : g_1 :^3 f_4, \tag{9.3.11}$$

$$j_2^{(2)} = (\epsilon^2/2!) : g_1 :^2 f_4, \tag{9.3.12}$$

$$j_3^{(2)} = -\epsilon : g_1 : f_4, \tag{9.3.13}$$

$$j_4^{(2)} = f_4. \tag{9.3.14}$$

Here a subscript on a j indicates the *degree* of the polynomial. Observe from (3.3) that all the j polynomials have *grade* two. Hence they also all carry a superscript 2 in parentheses to indicate this fact. Next we use the factorization theorem to write the product representation

$$\exp(: j_1^{(2)} + j_2^{(2)} + j_3^{(2)} + j_4^{(2)} :) = \exp(: k_1 :) \exp(: k_2 :) \exp(: k_3 :) \exp(: k_4 :). \tag{9.3.15}$$

Without loss of generality we may require that the k_i are in the Lie algebra generated by the $j_i^{(2)}$ and are also in ${}^{\epsilon}L^0/{}^{\epsilon}L^3$. It follows that we have the relations

$$k_i = j_i^{(2)}. \tag{9.3.16}$$

Now combine (3.10) and (3.15) to obtain the result

$$\begin{aligned} \exp(: f_4 :) \ & \exp(: \epsilon g_1 :) = \exp(: \epsilon g_1 :) \exp(: k_1 :) \exp(: k_2 :) \exp(: k_3 :) \exp(: k_4 :) \\ & = \exp(: \epsilon g_1 + k_1 :) \exp(: k_2 :) \exp(: k_3 :) \exp(: k_4 :). \end{aligned} \tag{9.3.17}$$

Observe that in relations such as (3.11) through (3.14), powers of ϵ are correlated with powers of $: g_1 :$. Thus, we may simply view the introduction of ϵ as a way of counting powers of g_1. Correspondingly, after obtaining final results, we may set $\epsilon = 1$ to obtain, under the assumption that g_1 itself is small, a set of formulas that make systematic expansions in the size of g_1. The final result of this process for the work done thus far is a formula that can be written as

$$\exp(: f_4 :) \exp(: g_1 :) = \exp(: h_1^4 :) \exp(: h_2^4 :) \exp(: h_3^4 :) \exp(: h_4^4 :) \tag{9.3.18}$$

where the h_i^4 are given by the relations

$$h_1^4 = g_1 - (1/3!) : g_1 :^3 f_4, \tag{9.3.19}$$

$$h_2^4 = (1/2!) : g_1 :^2 f_4, \tag{9.3.20}$$

$$h_3^4 = - : g_1 : f_4, \tag{9.3.21}$$

$$h_4^4 = f_4. \tag{9.3.22}$$

Here the subscript on h denotes its degree, and the superscript 4 indicates that it is associated with f_4. (Unlike the notation used earlier, the superscript on h is *not* a grade.) Note that in moving g_1 to the left of f_4 we have *generated* the third and second-degree polynomials h_3^4 and h_2^4 as well as the additional first-degree term in h_1^4. This generation of lower-order terms is a nonlinear *feed-down* effect. It shows, for example, that a misplaced octupole can produce sextupole, quadrupole, and steering-like effects.

We have seen how to move a first-order exponent to the left of f_4. The second step is to move such an exponent to the left of f_3. Let us now call the first-order exponent $\tilde{g}_1$. Then, in analogy to (3.8) and (3.9) find the relations

$$\exp(: f_3 :)\exp(: \epsilon\tilde{g}_1 :) = \exp(: \epsilon\tilde{g}_1 :)\exp(: \exp(- : \epsilon\tilde{g}_1 :)f_3 :), \tag{9.3.23}$$

$$: \exp(- : \epsilon\tilde{g}_1 :)f_3 :=: [f_3 - \epsilon : \tilde{g}_1 : f_3 + (\epsilon^2/2!) : \tilde{g}_1 :^2 f_3] : . \tag{9.3.24}$$

As before, we rewrite these relations in the form

$$\exp(: f_3 :)\exp(: \epsilon\tilde{g}_1 :) = \exp(: \epsilon\tilde{g}_1 :)\exp(: \tilde{j}_1^{(1)} + \tilde{j}_2^{(1)} + \tilde{j}_3^{(1)} :), \tag{9.3.25}$$

where the $\tilde{j}_i^{(1)}$ are now the homogeneous polynomials

$$\tilde{j}_1^{(1)} = (\epsilon^2/2!) : \tilde{g}_1 :^2 f_3, \tag{9.3.26}$$

$$\tilde{j}_2^{(1)} = -\epsilon : \tilde{g}_1 : f_3, \tag{9.3.27}$$

$$\tilde{j}_3^{(1)} = f_3. \tag{9.3.28}$$

We note that all the terms on the left sides of the relations (3.26) through (3.28) are of grade one. Hence all the $\tilde{j}$'s carry, within parentheses, a superscript of 1. Again, as before, we use the factorization theorem to write the representation

$$\exp(: \tilde{j}_1^{(1)} + \tilde{j}_2^{(1)} + \tilde{j}_3^{(1)} :) = \exp(: \tilde{k}_1 :)\exp(: \tilde{k}_2 :)\exp(: \tilde{k}_3 :)\exp(: \tilde{k}_4 :). \tag{9.3.29}$$

At this point there are two new features to the calculation: First, we have included a fourth-degree polynomial $\tilde{k}_4$ on the right side of (3.29) even though the highest degree polynomial on the left of (3.29), namely $\tilde{j}_3^{(1)}$, is of degree three. This is done for the sake of consistency since our calculation is being carried out in the quotient algebra ${}^{\epsilon}L^0/{}^{\epsilon}L^3$, which contains fourth-degree generators. Second, since the $\tilde{k}_i$ are in the Lie algebra generated by the $\tilde{j}_i^{(1)}$ and also in ${}^{\epsilon}L^0/{}^{\epsilon}L^3$, they may contain terms of grade 1 and grade 2. Therefore we make the decompositions

$$\tilde{k}_1 = \tilde{k}_1^{(1)} + \tilde{k}_1^{(2)}, \tag{9.3.30}$$

$$\tilde{k}_2 = \tilde{k}_2^{(1)} + \tilde{k}_2^{(2)}, \tag{9.3.31}$$

$$\tilde{k}_3 = \tilde{k}_3^{(1)} + \tilde{k}_3^{(2)}, \tag{9.3.32}$$

$$\tilde{k}_4 = \tilde{k}_4^{(2)}. \tag{9.3.33}$$

[Note that according to (3.1) through (3.3) there is no fourth-degree polynomial of grade 1.] Now use the BCH formula and the decompositions (3.30) through (3.33) to combine all exponents on the right side of (3.29) into one grand exponent. We find, through terms of grade 2, the result

$$\exp(:\tilde{k}_1:)\exp(:\tilde{k}_2:)\exp(:\tilde{k}_3:)\exp(:\tilde{k}_4:) = \exp(:\ell_1 + \ell_2 + \ell_3 + \ell_4:) \tag{9.3.34}$$

where the ℓ_i are given by the relations

$$\ell_1 = \tilde{k}_1^{(1)} + \tilde{k}_1^{(2)} + [\tilde{k}_1^{(1)}, \tilde{k}_2^{(1)}]/2, \tag{9.3.35}$$

$$\ell_2 = \tilde{k}_2^{(1)} + \tilde{k}_2^{(2)} + [\tilde{k}_1^{(1)}, \tilde{k}_3^{(1)}]/2, \tag{9.3.36}$$

$$\ell_3 = \tilde{k}_3^{(1)} + \tilde{k}_3^{(2)} + [\tilde{k}_2^{(1)}, \tilde{k}_3^{(1)}]/2, \tag{9.3.37}$$

$$\ell_4 = \tilde{k}_4^{(2)}. \tag{9.3.38}$$

Upon comparing (3.29) and (3.34) we find the results

$$\ell_i = \tilde{j}_i \text{ for } i = 1 \text{ to } 3, \tag{9.3.39}$$

$$\ell_4 = 0. \tag{9.3.40}$$

We thus have the relations

$$\tilde{j}_1^{(1)} = \tilde{k}_1^{(1)} + \tilde{k}_1^{(2)} + [\tilde{k}_1^{(1)}, \tilde{k}_2^{(1)}]/2, \tag{9.3.41}$$

$$\tilde{j}_2^{(1)} = \tilde{k}_2^{(1)} + \tilde{k}_2^{(2)} + [\tilde{k}_1^{(1)}, \tilde{k}_3^{(1)}]/2, \tag{9.3.42}$$

$$\tilde{j}_3^{(1)} = \tilde{k}_3^{(1)} + \tilde{k}_3^{(2)} + [\tilde{k}_2^{(1)}, \tilde{k}_3^{(1)}]/2, \tag{9.3.43}$$

$$0 = \tilde{k}_4^{(2)}, \tag{9.3.44}$$

and these relations must be solved for the $\tilde{k}_i^{(1)}$ and $\tilde{k}_i^{(2)}$. To carry out this task, we equate terms of like grade in (3.41) through (3.44) to find the results

$$\tilde{k}_1^{(1)} = \tilde{j}_1^{(1)}, \tag{9.3.45}$$

$$\tilde{k}_2^{(1)} = \tilde{j}_2^{(1)}, \tag{9.3.46}$$

$$\tilde{k}_3^{(1)} = \tilde{j}_3^{(1)}, \tag{9.3.47}$$

$$\tilde{k}_1^{(2)} = -[\tilde{k}_1^{(1)}, \tilde{k}_2^{(1)}]/2 = -[\tilde{j}_1^{(1)}, \tilde{j}_2^{(1)}]/2, \tag{9.3.48}$$

$$\tilde{k}_2^{(2)} = -[\tilde{k}_1^{(1)}, \tilde{k}_3^{(1)}]/2 = -[\tilde{j}_1^{(1)}, \tilde{j}_3^{(1)}]/2, \tag{9.3.49}$$

$$\tilde{k}_3^{(2)} = -[\tilde{k}_2^{(1)}, \tilde{k}_3^{(1)}]/2 = -[\tilde{j}_2^{(1)}, \tilde{j}_3^{(1)}]/2, \tag{9.3.50}$$

$$\tilde{k}_4^{(2)} = 0. \tag{9.3.51}$$

Now put the results (3.26) through (3.28), (3.30) through (3.33), and (3.45) through (3.51) together to get the relations

$$\tilde{k}_1 = (\epsilon^2/2) : \tilde{g}_1 :^2 f_3 - (\epsilon^3/4)[: \tilde{g}_1 :^2 f_3, : \tilde{g}_1 : f_3], \tag{9.3.52}$$

$$\tilde{k}_2 = -\epsilon : \tilde{g}_1 : f_3 - (\epsilon^2/4)[: \tilde{g}_1 :^2 f_3, f_3], \tag{9.3.53}$$

$$\tilde{k}_3 = f_3 + (\epsilon/2)[: \tilde{g}_1 : f_3, f_3], \tag{9.3.54}$$

$$\tilde{k}_4 = 0. \tag{9.3.55}$$

Finally, as before, these relations should be evaluated with $\epsilon = 1$. The net result is the formula

$$\exp(: f_3 :) \exp(: \tilde{g}_1 :) = \exp(: h_1^3 :) \exp(: h_2^3 :) \exp(: h_3^3 :) \exp(: h_4^3 :) \tag{9.3.56}$$

where the h_i^3 are given by the relations

$$h_1^3 = \tilde{g}_1 + (1/2) : \tilde{g}_1 :^2 f_3 - (1/4)[: \tilde{g}_1 :^2 f_3, : \tilde{g}_1 : f_3], \tag{9.3.57}$$

$$h_2^3 = - : \tilde{g}_1 : f_3 - (1/4)[: \tilde{g}_1 :^2 f_3, f_3], \tag{9.3.58}$$

$$h_3^3 = f_3 + (1/2)[: \tilde{g}_1 : f_3, f_3], \tag{9.3.59}$$

$$h_4^3 = 0. \tag{9.3.60}$$

We note that moving $\tilde{g}_1$ past f_3 is more difficult than moving g_1 past f_4! This greater difficulty occurs because, as is evident from (3.57) through (3.60), the feed-down terms are more complicated.

We have seen how to move a first-degree exponent past f_4 and f_3, and how to calculate (within the quotient algebra ${}^{\epsilon}L^0/{}^{\epsilon}L^3$) the feed-down terms left in its wake. The penultimate step is to move the first-degree exponent past $\mathcal{R}_f$ and then combine it with f_1. But this we know how to do exactly using the concatenation formulas (2.34) through (2.38) for the inhomogeneous symplectic group. Finally, we have to combine the results obtained so far with the remaining factors $\mathcal{R}_g \exp(: g_3 :) \exp(: g_4 :)$ in (3.7). This we also know how to do exactly using the concatenation formulas of Section 8.4. Thus, we have explored in some detail how to perform concatenation within the quotient group generated by the Lie algebra ${}^{\epsilon}L^0/{}^{\epsilon}L^3$.

Two tasks remain. The first is to find a convenient way of evaluating the symplectic matrices associated with the feed-down linear transformations of the form $\exp(: h_2^n :)$. This subject has already been treated in Chapter 4.

The second task is to find results for the larger quotient algebras ${}^{\epsilon}L^0/{}^{\epsilon}L^{\ell}$ with $\ell > 3$. A suitable Mathematica program for this purpose is presented in Appendix E. Essentially two problems must be solved to carry out the second task. First, we need formulas of the kind (3.15) and (3.29). Given a set of graded polynomials $j_i^{(n)}$, we need formulas for the k_m that appear in the product representation

$$\exp(: j_1^{(n)} + j_2^{(n)} + j_3^{(n)} + \cdots :) = \exp(: k_1 :) \exp(: k_2 :) \exp(: k_3 :) \cdots . \tag{9.3.61}$$

Second, let us write a relation of the form

$$\exp(:f_n:)\exp(:g_1:)=\exp(:h_1^n:)\exp(:h_2^n:)\exp(:h_3^n:)\cdots, \tag{9.3.62}$$

where, as before, we have used the notation h_i^n to denote the polynomial of degree i that results from moving $:g_1:$ past $:f_n:$. For this relation we need formulas for the h_i^n in terms of g_1 and f_n. We summarize below the results we have already found for the quotient algebra ${}^{\epsilon}L^0/{}^{\epsilon}L^3$ and, as a more complicated example found using the program in Appendix E, the results for the quotient algebra ${}^{\epsilon}L^0/{}^{\epsilon}L^5$.

Formulas for the k_i in (3.61) in the quotient algebra ${}^{\epsilon}L^0/{}^{\epsilon}L^3$.

n=2

$$k_i=j_1^{(2)},\ i=1\text{ to }4. \tag{9.3.63}$$

n=1

$$k_1=j_1^{(1)}-[j_1^{(1)},j_2^{(1)}]/2, \tag{9.3.64}$$

$$k_2=j_2^{(1)}-[j_1^{(1)},j_3^{(1)}]/2, \tag{9.3.65}$$

$$k_3=j_3^{(1)}-[j_2^{(1)},j_3^{(1)}]/2, \tag{9.3.66}$$

$$k_4=0. \tag{9.3.67}$$

Formulas for the h_i^n in (3.62) in the quotient algebra ${}^{\epsilon}L^0/{}^{\epsilon}L^3$.

$$h_1^4=g_1-(1/3!):g_1:^3 f_4, \tag{9.3.68}$$

$$h_2^4=(1/2!):g_1:^2 f_4, \tag{9.3.69}$$

$$h_3^4=-:g_1:f_4, \tag{9.3.70}$$

$$h_4^4=f_4. \tag{9.3.71}$$

$$h_1^3=g_1+(1/2!):g_1:^2 f_3-(1/4)[:g_1:^2 f_3,:g_1:f_3], \tag{9.3.72}$$

$$h_2^3=-:g_1:f_3-(1/4)[:g_1:^2 f_3,f_3], \tag{9.3.73}$$

$$h_3^3=f_3+(1/2)[:g_1:f_3,f_3], \tag{9.3.74}$$

$$h_4^3=0. \tag{9.3.75}$$

Formulas for the k_i in (3.61) in the quotient algebra ${}^{\epsilon}L^0/{}^{\epsilon}L^5$.

n=4

$$k_i=j_i^{(4)},i=1,6. \tag{9.3.76}$$

n=3

$$k_1=j_1^{(3)}, \tag{9.3.77}$$

$$k_2=j_2^{(3)}, \tag{9.3.78}$$

$$k_3=j_3^{(3)}, \tag{9.3.79}$$

$$k_4=j_4^{(3)}, \tag{9.3.80}$$

$$k_5=j_5^{(3)}, \tag{9.3.81}$$

$$k_6 = 0. \tag{9.3.82}$$

n=2

$$k_1 = j_1^{(2)} + [j_2^{(2)}, j_1^{(2)}]/2, \tag{9.3.83}$$

$$k_2 = j_2^{(2)} + [j_3^{(2)}, j_1^{(2)}], \tag{9.3.84}$$

$$k_3 = j_3^{(2)} + [j_3^{(2)}, j_2^{(2)}]/2 + [j_4^{(2)}, j_1^{(2)}], \tag{9.3.85}$$

$$k_4 = j_4^{(2)} + [j_4^{(2)}, j_2^{(2)}], \tag{9.3.86}$$

$$k_5 = [j_4^{(2)}, j_3^{(2)}], \tag{9.3.87}$$

$$k_6 = 0. \tag{9.3.88}$$

n=1

$$\begin{aligned} k_1 &= j_1^{(1)} + [j_2^{(1)}, j_1^{(1)}]/2 - [j_1^{(1)}, j_3^{(1)}, j_1^{(1)}]/6 + [j_2^{(1)}, j_2^{(1)}, j_1^{(1)}]/6 \\ &- [j_1^{(1)}, j_3^{(1)}, j_2^{(1)}, j_1^{(1)}]/8 - [j_2^{(1)}, j_1^{(1)}, j_3^{(1)}, j_1^{(1)}]/24 \\ &+ [j_2^{(1)}, j_2^{(1)}, j_2^{(1)}, j_1^{(1)}]/24, \end{aligned} \tag{9.3.89}$$

$$\begin{aligned} k_2 &= j_2^{(1)} + [j_3^{(1)}, j_1^{(1)}]/2 - [j_2^{(1)}, j_3^{(1)}, j_1^{(1)}]/12 + [j_3^{(1)}, j_2^{(1)}, j_1^{(1)}]/6 \\ &- [j_2^{(1)}, j_3^{(1)}, j_2^{(1)}, j_1^{(1)}]/24 - [j_3^{(1)}, j_1^{(1)}, j_3^{(1)}, j_1^{(1)}]/24 \\ &+ [j_3^{(1)}, j_2^{(1)}, j_2^{(1)}, j_1^{(1)}]/24, \end{aligned} \tag{9.3.90}$$

$$\begin{aligned} k_3 &= j_3^{(1)} + [j_3^{(1)}, j_2^{(1)}]/2 - [j_2^{(1)}, j_3^{(1)}, j_2^{(1)}]/6 + [j_3^{(1)}, j_3^{(1)}, j_1^{(1)}]/6 \\ &+ [j_2^{(1)}, j_2^{(1)}, j_3^{(1)}, j_2^{(1)}]/24 - [j_2^{(1)}, j_3^{(1)}, j_3^{(1)}, j_1^{(1)}]/8 + [j_3^{(1)}, j_2^{(1)}, j_3^{(1)}, j_1^{(1)}]/24 \\ &+ [j_3^{(1)}, j_3^{(1)}, j_2^{(1)}, j_1^{(1)}]/24, \end{aligned} \tag{9.3.91}$$

$$k_4 = -[j_3^{(1)}, j_3^{(1)}, j_2^{(1)}]/12 + [j_3^{(1)}, j_2^{(1)}, j_3^{(1)}, j_2^{(1)}]/24 - [j_3^{(1)}, j_3^{(1)}, j_3^{(1)}, j_1^{(1)}]/24, \tag{9.3.92}$$

$$k_5 = [j_3^{(1)}, j_3^{(1)}, j_3^{(1)}, j_2^{(1)}]/24, \tag{9.3.93}$$

$$k_6 = 0. \tag{9.3.94}$$

Formulas for the h_i^n in (3.62) in the quotient algebra ${}^{\epsilon}L^0/{}^{\epsilon}L^5$.

$$h_1^6 = g_1 - (1/120) : g_1 :^5 f_6, \tag{9.3.95}$$

$$h_2^6 = (1/24) : g_1 :^4 f_6, \tag{9.3.96}$$

$$h_3^6 = -(1/6) : g_1 :^3 f_6, \tag{9.3.97}$$

$$h_4^6 = (1/2) : g_1 :^2 f_6, \tag{9.3.98}$$

$$h_5^6 = - : g_1 : f_6, \tag{9.3.99}$$

$$h_6^6 = f_6; \tag{9.3.100}$$

$$h_1^5 = g_1 + (1/24) : g_1 :^4 f_5, \tag{9.3.101}$$

$$h_2^5 = -(1/6) : g_1 :^3 f_5, \tag{9.3.102}$$

$$h_3^5 = (1/2) : g_1 :^2 f_5, \tag{9.3.103}$$

$$h_4^5 = - : g_1 : f_5, \tag{9.3.104}$$

$$h_5^5 = f_5, \tag{9.3.105}$$

$$h_6^5 = 0; \tag{9.3.106}$$

$$h_1^4 = g_1 + j_1^4 + [j_2^4, j_1^4]/2, \tag{9.3.107}$$

$$h_2^4 = j_2^4 + [j_3^4, j_1^4]/2, \tag{9.3.108}$$

$$h_3^4 = j_3^4 + [j_3^4, j_2^4]/2 + [j_4^4, j_1^4]/2, \tag{9.3.109}$$

$$h_4^4 = j_4^4 + [j_4^4, j_2^4]/2, \tag{9.3.110}$$

$$h_5^4 = [j_4^4, j_3^4]/2, \tag{9.3.111}$$

$$h_6^4 = 0, \tag{9.3.112}$$

where

$$j_1^4 = - : g_1 :^3 f_4/6, \tag{9.3.113}$$

$$j_2^4 =: g_1 :^2 f_4/2, \tag{9.3.114}$$

$$j_3^4 = - : g_1 : f_4, \tag{9.3.115}$$

$$j_4^4 = f_4; \tag{9.3.116}$$

$$\begin{aligned} h_1^3 \;=\;& g_1 + j_1^3 + [j_2^3, j_1^3]/2 - [j_1^3, j_3^3, j_1^3]/6 + [j_2^3, j_2^3, j_1^3]/6 - [j_1^3, j_3^3, j_2^3, j_1^3]/8 \\ -\;& [j_2^3, j_1^3, j_3^3, j_1^3]/24 + [j_2^3, j_2^3, j_2^3, j_1^3]/24, \end{aligned} \tag{9.3.117}$$

$$\begin{aligned} h_2^3 \;=\;& j_2^3 + [j_3^3, j_1^3]/2 - [j_2^3, j_3^3, j_1^3]/12 + [j_3^3, j_2^3, j_1^3]/6 - [j_2^3, j_3^3, j_2^3, j_1^3]/24 \\ -\;& [j_3^3, j_1^3, j_3^3, j_1^3]/24 + [j_3^3, j_2^3, j_2^3, j_1^3]/24, \end{aligned} \tag{9.3.118}$$

$$\begin{aligned} h_3^3 \;=\;& j_3^3 + [j_3^3, j_2^3]/2 - [j_2^3, j_3^3, j_2^3]/6 + [j_3^3, j_3^3, j_1^3]/6 + [j_2^3, j_2^3, j_3^3, j_2^3]/24 \\ -\;& [j_2^3, j_3^3, j_3^3, j_1^3]/8 + [j_3^3, j_2^3, j_3^3, j_1^3]/24 + [j_3^3, j_3^3, j_2^3, j_1^3]/24, \end{aligned} \tag{9.3.119}$$

$$h_4^3 = -[j_3^3, j_3^3, j_2^3]/12 + [j_3^3, j_2^3, j_3^3, j_2^3]/24 - [j_3^3, j_3^3, j_3^3, j_1^3]/24, \tag{9.3.120}$$

$$h_5^3 = [j_3^3, j_3^3, j_3^3, j_2^3]/24, \tag{9.3.121}$$

$$h_6^3 = 0, \tag{9.3.122}$$

where

$$j_1^3 =: g_1 :^2 f_3/2, \tag{9.3.123}$$

$$j_2^3 = - : g_1 : f_3, \tag{9.3.124}$$

$$j_3^3 = f_3. \tag{9.3.125}$$

Here for multiple Poisson brackets we have used the notation

$$[a_1, a_2, a_3] = [a_1, [a_2, a_3]], \tag{9.3.126}$$

$$[a_1, a_2, a_3, a_4] = [a_1, [a_2, [a_3, a_4]]]. \tag{9.3.127}$$

Note that for ${}^\epsilon L^0/{}^\epsilon L^5$ the feed-down terms are quite complicated. Also, note that the terms h_5^4, h_4^3, and h_5^3 are nonzero. Consequently there can also be, in effect, nonlinear feed-up terms due to translations in phase space. Finally, we see that the ${}^\epsilon L^0/{}^\epsilon L^3$ formulas (3.68) through (3.75) are special cases of the ${}^\epsilon L^0/{}^\epsilon L^5$ formulas for h_i^4 and h_i^3 in which higher-power terms in g_1 are neglected.

We close this section with an observation that will be of relevance for the work of the next section. Observe that the relations (3.25) and (3.29) can be written in the form

$$\exp(:f_3:)\exp(:\epsilon g_1:) = \exp(:k_1^{(0)}:)\exp(:k_1^{(1)} + k_1^{(2)}:)\exp(:k_2^{(1)} + k_2^{(2)}:)\exp(:k_3^{(1)} + k_3^{(2)}:). \tag{9.3.128}$$

Here, for convenience, we have dropped the tildes and we have defined the $k_i^{(n)}$ by the relations

$$k_1^{(0)} = \epsilon g_1, \tag{9.3.129}$$

$$k_1^{(1)} = (\epsilon^2/2) : g_1 :^2 f_3, \tag{9.3.130}$$

$$k_1^{(2)} = -(\epsilon^3/4)[: g_1 :^2 f_3, : g_1 : f_3], \tag{9.3.131}$$

$$k_2^{(1)} = -\epsilon : g_1 : f_3, \tag{9.3.132}$$

$$k_2^{(2)} = -(\epsilon^2/4)[: g_1 :^2 f_3, f_3], \tag{9.3.133}$$

$$k_3^{(1)} = f_3, \tag{9.3.134}$$

$$k_3^{(2)} = (\epsilon/2)[: g_1 : f_3, f_3]. \tag{9.3.135}$$

See (3.1) through (3.3) and (3.52) through (3.55). We see that several of the exponents contain terms of different grades. We will therefore refer to expressions of the form (3.128) as *mixed* grade factorizations.

From the Lie algebraic perspective of working within ${}^\epsilon L^0/{}^\epsilon L^3$ we could as well sought relations of the form

$$\begin{aligned}\exp(:f_3:)\exp(:\epsilon g_1:) = {}& \exp(:\hat{k}_1^{(0)}:)\exp(:\hat{k}_1^{(1)}:)\exp(:\hat{k}_1^{(2)}:) \times \\ & \exp(:\hat{k}_2^{(0)}:)\exp(:\hat{k}_2^{(1)}:)\exp(:\hat{k}_2^{(2)}:) \times \\ & \exp(:\hat{k}_3^{(1)}:)\exp(:\hat{k}_3^{(2)}:)\exp(:\hat{k}_4^{(2)}:).\end{aligned} \tag{9.3.136}$$

Here we have used hats to indicate that the $\hat{k}_i^{(n)}$ may differ from the $k_i^{(n)}$. (They are actually the same for the algebra ${}^\epsilon L^0/{}^\epsilon L^3$, but they may differ for relations analagous to (3.136) in the case of algebras having a larger maximum grade.) Note that in the factorization (3.136) each exponent has a single grade. [And, according to (3.1) through (3.3), each exponent carries a single power of ϵ.] We will call factorizations of this kind *single* grade factorizations.

Once again ϵ serves as a counting parameter and, having obtained a single grade factorization of the form (3.136), we may set $\epsilon = 1$. Evidently, since we are working within a quotient algebra based on grade, setting $\epsilon = 1$ in either a mixed grade or a single grade factorization gives equivalent results.

Exercises

9.3.1. Verify the relations (3.8) through (3.22).

9.3.2. Verify the relations (3.30) through (3.38). Note that Poisson brackets between grade one and grade two polynomials, and Poisson brackets between two grade two polynomials, vanish in the quotient algebra ${}^{\epsilon}L^0/{}^{\epsilon}L^3$.

9.3.3. Verify the relations (3.39) through (3.55).

9.3.4.

9.3.5.

9.4 Canonical Treatment of Translations

The Lie concatenation formulas for maps that send the origin into itself, see Sections 8.4 and 10.12, were relatively easy to derive. By contrast, the concatenation formulas just derived in the previous section, where translations were included, seem much more complicated. In this section we will show how the translation case can be handled using a concatenator for the simpler origin-preserving case. This will be done by *enlarging* the $2n$-dimensional phase space to include the extra variables q_{n+1} and p_{n+1}. For this reason, the method will be referred to as a *canonical* treatment of translations.

9.4.1 Preliminaries

Since the origin-preserving concatenation formulas do not depend on the phase-space dimension, the only costs associated with increased phase-space dimension are those of increased storage and slower execution when a concatenator is realized as computer code. The advantages of this approach are simplicity and reliability. If one has a reliable origin-preserving concatenator, one can construct from it a self-checking concatenator for general maps.

Those who have been meticulous to do the Exercises in this book will recognize that Exercise 7.7.2 showed that the Jacobi Lie algebra $j(2n,\mathbb{R})$ is homomorphic to the inhomogeneous symplectic Lie algebra $isp(2n,\mathbb{R})$ of Section 9.2, and Exercise 7.7.3 treated, among other things, the relation between $j(2n,\mathbb{R})$ and the symplectic group Lie algebra $sp[2(n+1),\mathbb{R}]$ for a phase space having two additional dimensions. We begin our discussion here by further elaborating on this theme. Subsequently we will treat $ISpM(2n,\mathbb{R})$ and $ispm(2n,\mathbb{R})$, the full nonlinear case of all symplectic maps.

Let us use the symbol $\hat{z}$ to denote the coordinates in the $(2n+2)$-dimensional enlarged phase space,

$$\hat{z} = (q_1 \cdots q_n, q_{n+1}, p_1 \cdots p_n, p_{n+1}). \tag{9.4.1}$$

We will also use the notation $\hat{\mathcal{M}}_{\hat{f}}$ to denote a symplectic (and origin-preserving) map acting on the enlarged phase space. For what follows we will want to consider special maps $\hat{\mathcal{M}}_{\hat{h}}$ on the enlarged phase space that have the property

$$\hat{\mathcal{M}}_{\hat{h}} q_{n+1} = q_{n+1}. \tag{9.4.2}$$

Such maps obviously form a group. Moreover, they have a factorization of the form

$$\hat{\mathcal{M}}_{\hat{h}} = \hat{\mathcal{R}}_{\hat{h}} \exp(:\hat{h}_3:)\exp(:\hat{h}_4:)\cdots \tag{9.4.3}$$

where the linear part $\hat{\mathcal{R}}_{\hat{h}}$ has the property

$$\hat{\mathcal{R}}_{\hat{h}} q_{n+1} = q_{n+1}, \tag{9.4.4}$$

and the $\hat{h}_m(\hat{z})$ that describe the nonlinear part are independent of p_{n+1},

$$\partial \hat{h}_m(\hat{z})/\partial p_{n+1} = 0. \tag{9.4.5}$$

To see that this is so, equate terms of like degree on both sides of (4.2). From (7.6.14) there is the result

$$[\exp(:\hat{h}_3:)\exp(:\hat{h}_4:)\cdots]q_{n+1} = q_{n+1} + O(\hat{z}^2). \tag{9.4.6}$$

Therefore (4.2) requires the relation

$$q_{n+1} = \hat{\mathcal{M}}_{\hat{h}} q_{n+1} = \hat{\mathcal{R}}_{\hat{h}}[q_{n+1} + O(\hat{z}^2)] = \hat{\mathcal{R}}_{\hat{h}} q_{n+1} + O(\hat{z}^2), \tag{9.4.7}$$

and equating terms of first degree yields (4.4).

Now that (4.4) is established, multiply both sides of (4.2) by $\hat{\mathcal{R}}_{\hat{h}}^{-1}$ to find the result

$$\hat{\mathcal{R}}_{\hat{h}}^{-1}\hat{\mathcal{M}}_{\hat{h}} q_{n+1} = \hat{\mathcal{R}}_{\hat{h}}^{-1} q_{n+1} = q_{n+1} \tag{9.4.8}$$

from which it follows that

$$[\exp(:\hat{h}_3:)\exp(:\hat{h}_4:)\cdots]q_{n+1} = q_{n+1}. \tag{9.4.9}$$

Evaluate both sides of (4.9) through terms of degree 2. Doing so gives the relation

$$q_{n+1} + :\hat{h}_3: q_{n+1} + O(\hat{z}^3) = q_{n+1}, \tag{9.4.10}$$

and equating terms of like degree gives the relation

$$0 =: \hat{h}_3 : q_{n+1} = [\hat{h}_3, q_{n+1}] = -\partial \hat{h}_3/\partial p_{n+1}, \tag{9.4.11}$$

which establishes (4.5) for the case $m = 3$. From (4.11) we also conclude that

$$\exp(:\hat{h}_3:)q_{n+1} = q_{n+1}. \tag{9.4.12}$$

Finally, multiply both sides of (4.9) by $\exp(-:\hat{h}_3:)$ to find the result

$$[\exp(:\hat{h}_4:)\exp(:\hat{h}_5:)\cdots]q_{n+1} = \exp(-:\hat{h}_3:)q_{n+1} = q_{n+1}. \tag{9.4.13}$$

Expanding both sides of (4.13) and equating terms of like degree shows that (4.5) also holds for the case $m = 4$, etc.

Let us explore the consequences of (4.4) in detail. For simplicity, consider the case of a phase space that is initially two dimensional ($n = 1$) and is enlarged to become four dimensional. The results for general n can easily be inferred from what we will find for the

$n = 1$ case. Suppose $\hat{R}^{\hat{h}}$ is the matrix associated with $\hat{\mathcal{R}}_{\hat{h}}$. In the 4×4 case we have decided to consider, and in view of (4.4) with $n = 1$, it has the general form

$$\hat{R}^{\hat{h}} = \left(\begin{array}{cccc} \hat{R}^{\hat{h}}_{11} & \hat{R}^{\hat{h}}_{12} & \hat{R}^{\hat{h}}_{13} & \hat{R}^{\hat{h}}_{14} \\ \hat{R}^{\hat{h}}_{21} & \hat{R}^{\hat{h}}_{22} & \hat{R}^{\hat{h}}_{23} & \hat{R}^{\hat{h}}_{24} \\ 0 & 0 & 1 & 0 \\ \hat{R}^{\hat{h}}_{41} & \hat{R}^{\hat{h}}_{42} & \hat{R}^{\hat{h}}_{43} & \hat{R}^{\hat{h}}_{44} \end{array} \right). \tag{9.4.14}$$

Here we have used the ordering

$$\hat{z} = (q_1, p_1; q_2, p_2). \tag{9.4.15}$$

However, since $\hat{R}_{\hat{h}}$ is a symplectic map, $\hat{R}^{\hat{h}}$ must be a symplectic matrix. Enforcing the symplectic condition (3.1.2) or (3.1.10) gives, among others, the relations

$$\hat{R}^{\hat{h}}_{a4} = \delta_{a4}. \tag{9.4.16}$$

It follows that $\hat{R}^{\hat{h}}$ has the more specific form

$$\hat{R}^{\hat{h}} = \left(\begin{array}{cccc} \hat{R}^{\hat{h}}_{11} & \hat{R}^{\hat{h}}_{12} & \hat{R}^{\hat{h}}_{13} & 0 \\ \hat{R}^{\hat{h}}_{21} & \hat{R}^{\hat{h}}_{22} & \hat{R}^{\hat{h}}_{23} & 0 \\ 0 & 0 & 1 & 0 \\ \hat{R}^{\hat{h}}_{41} & \hat{R}^{\hat{h}}_{42} & \hat{R}^{\hat{h}}_{43} & 1 \end{array} \right). \tag{9.4.17}$$

Introduce the 2×2 matrix $\bar{R}^{\hat{h}}$ defined by the upper-left 2×2 block in $\hat{R}^{\hat{h}}$,

$$\bar{R}^{\hat{h}} = \left(\begin{array}{cc} \hat{R}^{\hat{h}}_{11} & \hat{R}^{\hat{h}}_{12} \\ \hat{R}^{\hat{h}}_{21} & \hat{R}^{\hat{h}}_{22} \end{array} \right), \tag{9.4.18}$$

and let $\breve{R}^{\hat{h}}$ be the 4×4 matrix

$$\breve{R}^{\hat{h}} = \left(\begin{array}{cc} \bar{R}^{\hat{h}} & 0 \\ 0 & I \end{array} \right). \tag{9.4.19}$$

Also, define quantities $\alpha^{\hat{h}}_2$ and $\delta^{\hat{h}}_a$ for $a = 1$ to 2 by the rules

$$\alpha^{\hat{h}}_2 = \hat{R}^{\hat{h}}_{43}/2, \tag{9.4.20}$$

$$\delta^{\hat{h}}_1 = -\hat{R}^{\hat{h}}_{42}, \tag{9.4.21}$$

$$\delta^{\hat{h}}_2 = \hat{R}^{\hat{h}}_{41}, \tag{9.4.22}$$

and define associated polynomials h_1 and $\hat{h}^2_1$ by the rules

$$h_1(z) = \delta^{\hat{h}}_2 q_1 - \delta^{\hat{h}}_1 p_1 = \delta^{\hat{h}}_2 z_1 - \delta^{\hat{h}}_1 z_2 = (z, (\delta^{\hat{h}})^*), \tag{9.4.23}$$

$$\hat{h}^2_1(\hat{z}) = q_2 h_1(z). \tag{9.4.24}$$

Here we have used the notation of Section 7.7, and z denotes the original phase-space variables,

$$z = (q_1, p_1).$$

The superscript indicates that $\hat{h}_1^2$ is homogeneous of degree *two* in the variables $\hat{z}$, and the subscript indicates that it is homogeneous of degree *one* with respect to the variables z.

Note that h_1 has the property

$$: h_1 : z_a = \delta_a^{\hat{h}}. \tag{9.4.25}$$

Correspondingly, $\hat{h}_1^2$ has the properties

$$: \hat{h}_1^2 : z_a =: q_2 h_1(z) : z_a = [q_2 h_1(z), z_a] = q_2[h_1(z), z_a] = q_2 : h_1 : z_a = q_2 \delta_a^{\hat{h}}, \tag{9.4.26}$$

$$: \hat{h}_1^2 :^m z_a = 0 \text{ for } m > 1, \tag{9.4.27}$$

$$: \hat{h}_1^2 : q_2 = [q_2 h_1(z), q_2] = 0, \tag{9.4.28}$$

$$: \hat{h}_1^2 : p_2 = [q_2 h_1(z), p_2] = h_1(z), \tag{9.4.29}$$

$$: \hat{h}_1^2 :^m p_2 = 0 \text{ for } m > 1. \tag{9.4.30}$$

Then, with these definitions, we assert that $\hat{\mathcal{R}}_{\hat{h}}$ has the unique factorization

$$\hat{\mathcal{R}}_{\hat{h}} = \hat{\mathcal{F}}_{\hat{h}} \hat{\mathcal{R}}_{\hat{h}_1^2} \check{\mathcal{R}}_{\hat{h}} \tag{9.4.31}$$

where

$$\hat{\mathcal{F}}_{\hat{h}} = \exp(: \alpha_2^{\hat{h}} q_2^2 :), \tag{9.4.32}$$

$$\hat{\mathcal{R}}_{\hat{h}_1^2} = \exp(: \hat{h}_1^2 :), \tag{9.4.33}$$

and $\check{\mathcal{R}}_{\hat{h}}$ is a linear symplectic map whose associated matrix $\check{R}^{\hat{h}}$ is given by (4.19).

If correct, the operator assertion (4.31) is equivalent to the matrix assertion

$$\hat{R}^{\hat{h}} = \check{R}^{\hat{h}} \hat{R}^{\hat{h}_1^2} \hat{F}^{\hat{h}} \tag{9.4.34}$$

where $\hat{R}^{\hat{h}_1^2}$ and $\hat{F}^{\hat{h}}$ are the matrices associated with $\hat{\mathcal{R}}_{\hat{h}_1^2}$ and $\hat{\mathcal{F}}_{\hat{h}}$. Let us find these matrices. From (4.26) through (4.30) we see that $\hat{\mathcal{R}}_{\hat{h}_1^2}$ has the property

$$\hat{\mathcal{R}}_{\hat{h}_1^2} z_a = z_a + q_2 \delta_a^{\hat{h}}, \tag{9.4.35}$$

$$\hat{\mathcal{R}}_{\hat{h}_1^2} q_2 = q_2, \tag{9.4.36}$$

$$\hat{\mathcal{R}}_{\hat{h}_1^2} p_2 = p_2 + h_1(z) = p_2 + (z, (\delta^{\hat{h}})^*). \tag{9.4.37}$$

It follows that the matrix $\hat{R}^{\hat{h}_1^2}$ is given by the relation

$$\hat{R}^{\hat{h}_1^2} = \begin{pmatrix} 1 & 0 & \delta_1^{\hat{h}} & 0 \\ 0 & 1 & \delta_2^{\hat{h}} & 0 \\ 0 & 0 & 1 & 0 \\ \delta_2^{\hat{h}} & -\delta_1^{\hat{h}} & 0 & 1 \end{pmatrix}. \tag{9.4.38}$$

Finding $\hat{F}^{\hat{h}}$ is easier. A simple calculation gives the result

$$\hat{F}^{\hat{h}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \hat{R}^{\hat{h}}_{43} & 1 \end{pmatrix}. \tag{9.4.39}$$

Let us solve (4.34) for $\check{R}^{\hat{h}}$ to find the relation

$$\check{R}^{\hat{h}} = \hat{R}^{\hat{h}}(\hat{F}^{\hat{h}})^{-1}(\hat{R}^{\hat{h}_1^2})^{-1}. \tag{9.4.40}$$

Carrying out the indicated multiplications gives the result

$$\check{R}^{\hat{h}} = \begin{pmatrix} \hat{R}^{\hat{h}}_{11} & \hat{R}^{\hat{h}}_{12} & \epsilon_1 & 0 \\ \hat{R}^{\hat{h}}_{21} & \hat{R}^{\hat{h}}_{22} & \epsilon_2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{9.4.41}$$

where

$$\epsilon_1 = \hat{R}^{\hat{h}}_{13} + \hat{R}^{\hat{h}}_{11}\hat{R}^{\hat{h}}_{42} - \hat{R}^{\hat{h}}_{12}\hat{R}^{\hat{h}}_{41} = \hat{R}^{\hat{h}}_{13} - \hat{R}^{\hat{h}}_{11}\delta^{\hat{h}}_1 - \hat{R}^{\hat{h}}_{12}\delta^{\hat{h}}_2, \tag{9.4.42}$$

$$\epsilon_2 = \hat{R}^{\hat{h}}_{23} + \hat{R}^{\hat{h}}_{21}\hat{R}^{\hat{h}}_{42} - \hat{R}^{\hat{h}}_{22}\hat{R}^{\hat{h}}_{41} = \hat{R}^{\hat{h}}_{23} - \hat{R}^{\hat{h}}_{21}\delta^{\hat{h}}_1 - \hat{R}^{\hat{h}}_{22}\delta^{\hat{h}}_2. \tag{9.4.43}$$

Next, because $\hat{R}^{\hat{h}}$, $(\hat{F}^{\hat{h}})^{-1}$, and $(\hat{R}^{\hat{h}_1^2})^{-1}$ are symplectic matrices ($\mathcal{R}_{\hat{h}}$, $\mathcal{F}_{\hat{h}}$, and $\mathcal{R}_{\hat{h}_1^2}$ are symplectic maps), $\check{R}^{\hat{h}}$ must be a symplectic matrix. The symplectic condition (3.1.10) yields for $\check{R}^{\hat{h}}$ as given by (4.41) the relations

$$\epsilon_a = 0 \text{ for } a = 1 \text{ to } 2, \tag{9.4.44}$$

$$\hat{R}^{\hat{h}}_{11}\hat{R}^{\hat{h}}_{22} - \hat{R}^{\hat{h}}_{12}\hat{R}^{\hat{h}}_{21} = 1. \tag{9.4.45}$$

From (4.44) we conclude that the ϵ_a entries in (4.41) must vanish, and therefore (4.34) is correct with $\check{R}^{\hat{h}}$ given by (4.19). The relation (4.45) is the condition that $\bar{R}^{\hat{h}}$ and hence $\check{R}^{\hat{h}}$ be symplectic matrices. Finally (4.42) and (4.43), when combined with (4.44), show that the matrix $\hat{R}^{\hat{h}}$ in (4.17) must have the form

$$\hat{R}^{\hat{h}} = \begin{pmatrix} \hat{R}^{\hat{h}}_{11} & \hat{R}^{\hat{h}}_{12} & (\bar{R}^{\hat{h}}\delta^{\hat{h}})_1 & 0 \\ \hat{R}^{\hat{h}}_{21} & \hat{R}^{\hat{h}}_{22} & (\bar{R}^{\hat{h}}\delta^{\hat{h}})_2 & 0 \\ 0 & 0 & 1 & 0 \\ \delta^{\hat{h}}_2 & -\delta^{\hat{h}}_1 & \hat{R}^{\hat{h}}_{43} & 1 \end{pmatrix}. \tag{9.4.46}$$

With what we have learned we are now prepared to show how general maps (including translations) for the case of $2n$-dimensional phase space can be imbedded in the set of $(2n+2)$-dimensional origin preserving maps. We will first consider maps with no nonlinear part, and then move on to the general case.

9.4.2 Case of Maps with No Nonlinear Part

Enlarging

Let $\mathcal{M}_f$ be an inhomogeneous symplectic group map acting on the original $2n$-dimensional phase space z. As in (2.31), it may be written in the form

$$\mathcal{M}_f = \exp(: f_1 :)\mathcal{R}_f. \tag{9.4.47}$$

Define a function $\hat{f}_1^2(\hat{z})$ by the rule

$$\hat{f}_1^2(\hat{z}) = (q_{n+1})f_1(z). \tag{9.4.48}$$

As before, the superscript indicates that $\hat{f}_1^2$ is homogeneous of degree *two* in the variables $\hat{z}$, and the subscript indicates that it is homogeneous of degree *one* with respect to the variables z. Now define a map $\hat{\mathcal{M}}_{\hat{f}}$ on the enlarged phase space by the rule

$$\hat{\mathcal{M}}_{\hat{f}} = \exp(: \hat{f}_1^2 :)\check{\mathcal{R}}_{\hat{f}}. \tag{9.4.49}$$

Here $\check{\mathcal{R}}_{\hat{f}}$ is a linear map with the associated matrix $\check{R}_{\hat{f}}$ given by

$$\check{R}^{\hat{f}} = \begin{pmatrix} R^f & 0 \\ 0 & I \end{pmatrix} \tag{9.4.50}$$

where R^f is the matrix associated with $\mathcal{R}_f$, I denotes the 2×2 identity matrix acting on the q_{n+1}, p_{n+1} space, and the 0's denote rectangular matrices of zeroes. Evidently, we have mapped an element of the inhomogeneous symplectic group in $2n$ dimensions into an element of the homogeneous (origin-preserving) symplectic group in $(2n+2)$ dimensions. This process will be called *enlarging*.

What is the effect of $\hat{\mathcal{M}}_{\hat{f}}$ on the enlarged phase space? Evidently we immediately have the relation

$$\hat{\mathcal{M}}_{\hat{f}} q_{n+1} = q_{n+1}. \tag{9.4.51}$$

To explore matters further suppose, in analogy with (4.25), that f_1 has the property

$$: f_1 : z_a = \delta_a^f. \tag{9.4.52}$$

See (7.7.1) through (7.7.6). Then $\hat{f}_1^2$ has the properties

$$\begin{aligned} : \hat{f}_1^2 : z_a &= \; : (q_{n+1})f_1(z) : z_a = [(q_{n+1})f_1(z), z_a] \\ &= (q_{n+1})[f_1(z), z_a] = (q_{n+1}) : f_1 : z_a = (q_{n+1})\delta_a^f, \end{aligned} \tag{9.4.53}$$

$$: \hat{f}_1^2 :^m z_a = 0 \text{ for } m > 1, \tag{9.4.54}$$

$$: \hat{f}_1^2 : (q_{n+1}) = [(q_{n+1})f_1(z), (q_{n+1})] = 0, \tag{9.4.55}$$

$$: \hat{f}_1^2 : (p_{n+1}) = [(q_{n+1})f_1(z), (p_{n+1})] = f_1(z), \tag{9.4.56}$$

$$: \hat{f}_1^2 :^m (p_{n+1}) = 0 \text{ for } m > 1. \tag{9.4.57}$$

Let $\hat{\mathcal{R}}_{\hat{f}_1^2}$ denote the map

$$\hat{\mathcal{R}}_{\hat{f}_1^2} = \exp(:\hat{f}_1^2:). \tag{9.4.58}$$

From (4.53) through (4.57) we see that $\hat{\mathcal{R}}_{\hat{f}_1^2}$ has the property

$$\hat{\mathcal{R}}_{\hat{f}_1^2} z_a = z_a + (q_{n+1})\delta_a^f, \tag{9.4.59}$$

$$\hat{\mathcal{R}}_{\hat{f}_1^2} q_{n+1} = q_{n+1}, \tag{9.4.60}$$

$$\hat{\mathcal{R}}_{\hat{f}_1^2} p_{n+1} = p_{n+1} + f_1(z) = p_{n+1} + (z, (\delta^f)^*). \tag{9.4.61}$$

See also (7.7.3).

It follows from (4.59) through (4.61) that the action of $\hat{\mathcal{R}}_{\hat{f}_1^2}$ can be represented by a matrix $\hat{R}^{\hat{f}_1^2}$. In the simplest case that the original phase space is two dimensional, this matrix is 4×4 and is given by the relation

$$\hat{R}^{\hat{f}_1^2} = \begin{pmatrix} 1 & 0 & \delta_1^f & 0 \\ 0 & 1 & \delta_2^f & 0 \\ 0 & 0 & 1 & 0 \\ \delta_2^f & -\delta_1^f & 0 & 1 \end{pmatrix}. \tag{9.4.62}$$

Here we have used the ordering (4.15) and the J of (3.2.11). Evidently (4.62) is analogous to (4.38).

Finally, let us find the effect of $\hat{\mathcal{M}}_{\hat{f}}$. According to (4.49) and (4.58), it can be written in the form

$$\hat{\mathcal{M}}_{\hat{f}} = \hat{\mathcal{R}}_{\hat{f}_1^2} \check{\mathcal{R}}_{\hat{f}}. \tag{9.4.63}$$

It follows that the action of $\hat{\mathcal{M}}_{\hat{f}}$ can be represented by the matrix $\hat{R}^{\hat{f}}$ given by the relation

$$\hat{R}^{\hat{f}} = \check{R}^{\hat{f}} \hat{R}^{\hat{f}_1^2}. \tag{9.4.64}$$

See (8.4.19) and (8.4.20). Carrying out the indicated multiplication gives (in the 4×4 case) the result

$$\hat{R}^{\hat{f}} = \begin{pmatrix} R_{11}^f & R_{12}^f & (R^f\delta^f)_1 & 0 \\ R_{21}^f & R_{22}^f & (R^f\delta^f)_2 & 0 \\ 0 & 0 & 1 & 0 \\ \delta_2^f & -\delta_1^f & 0 & 1 \end{pmatrix}. \tag{9.4.65}$$

which is analogous to (4.46).

Shrinking

We have defined enlarged maps and have studied their effect on the enlarged phase space. Let us now explore how they behave under multiplication. Suppose $\mathcal{M}_f$ and $\mathcal{M}_g$ are any two inhomogeneous symplectic group maps, and $\hat{\mathcal{M}}_{\hat{f}}$ and $\hat{\mathcal{M}}_{\hat{g}}$ are their enlargements. We can form corresponding maps $\mathcal{M}_h$ and $\hat{\mathcal{M}}_{\hat{h}}$ by the products

$$\mathcal{M}_h = \mathcal{M}_f \mathcal{M}_g, \tag{9.4.66}$$

$$\hat{\mathcal{M}}_{\hat{h}} = \hat{\mathcal{M}}_{\hat{f}}\hat{\mathcal{M}}_{\hat{g}}. \tag{9.4.67}$$

The product (4.66) involves the concatenation of maps that include translations, and its calculation entails the derivation and use of complicated (and only partially known) feed-down formulae as described in the previous section. By contrast, the product (4.67) is for origin-preserving maps in the enlarged 8-dimensional phase space. Its computation involves only the use of far simpler universal dimension-independent origin-preserving concatenation rules. What we wish to learn is whether $\mathcal{M}_h$ can be deduced from a knowledge of $\hat{\mathcal{M}}_{\hat{h}}$. The process of constructing $\mathcal{M}_h$ from $\hat{\mathcal{M}}_{\hat{h}}$ will be called *shrinking*. See Figure 9.4.1 for a pictorial presentation of this question.

To answer this question, let us compute $\hat{R}^{\hat{h}}$, the matrix corresponding to $\hat{\mathcal{M}}_{\hat{h}}$. It is given by the relation

$$\begin{aligned}
\hat{R}^{\hat{h}} = \hat{R}^{\hat{g}}\hat{R}^{\hat{f}} \quad &= \begin{pmatrix} R^g_{11} & R^g_{12} & (R^g\delta^g)_1 & 0 \\ R^g_{21} & R^g_{22} & (R^g\delta^g)_2 & 0 \\ 0 & 0 & 1 & 0 \\ \delta^g_2 & -\delta^g_1 & 0 & 1 \end{pmatrix}\begin{pmatrix} R^f_{11} & R^f_{12} & (R^f\delta^f)_1 & 0 \\ R^f_{21} & R^f_{22} & (R^f\delta^f)_2 & 0 \\ 0 & 0 & 1 & 0 \\ \delta^f_2 & -\delta^f_1 & 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} R^h_{11} & R^h_{12} & (R^h\delta^f + R^g\delta^g)_1 & 0 \\ R^h_{21} & R^h_{22} & (R^h\delta^f + R^g\delta^g)_2 & 0 \\ 0 & 0 & 1 & 0 \\ * & * & * & 1 \end{pmatrix}.
\end{aligned} \tag{9.4.68}$$

Here,

$$R^h = R^g R^f. \tag{9.4.69}$$

As before, for simplicity, we have treated the case where the original phase space is two dimensional, and the enlarged phase space is four dimensional. Again, the result in this case is sufficient to deduce the result in any dimension. Finally, we have not computed the starred entries in the bottom row of $\hat{R}^{\hat{h}}$. See Exercise 4.10.

We observe, as a consequence of (4.69), that the matrix R^h can be read off from the upper-left corner of $\hat{R}^{\hat{h}}$. Also, upon comparison of (4.68) with (4.46), we expect the upper two entries of the penultimate column of $\hat{R}^{\hat{h}}$ to be the entries of the vector $(R^h\delta^h)$. Therefore, from (4.46) and (4.68), we get the relation

$$R^h\delta^h = R^h\delta^f + R^g\delta^g. \tag{9.4.70}$$

In view of (4.69), the relation (4.70) can also be written in the form

$$\delta^h = \delta^f + (R^f)^{-1}\delta^g, \tag{9.4.71}$$

from which it follows that

$$(z, J\delta^h) = (z, J\delta^f) + (z, J(R^f)^{-1}\delta^g). \tag{9.4.72}$$

But from the symplectic condition (3.1.2) there is the relation

$$J(R^f)^{-1} = (R^f)^T J. \tag{9.4.73}$$

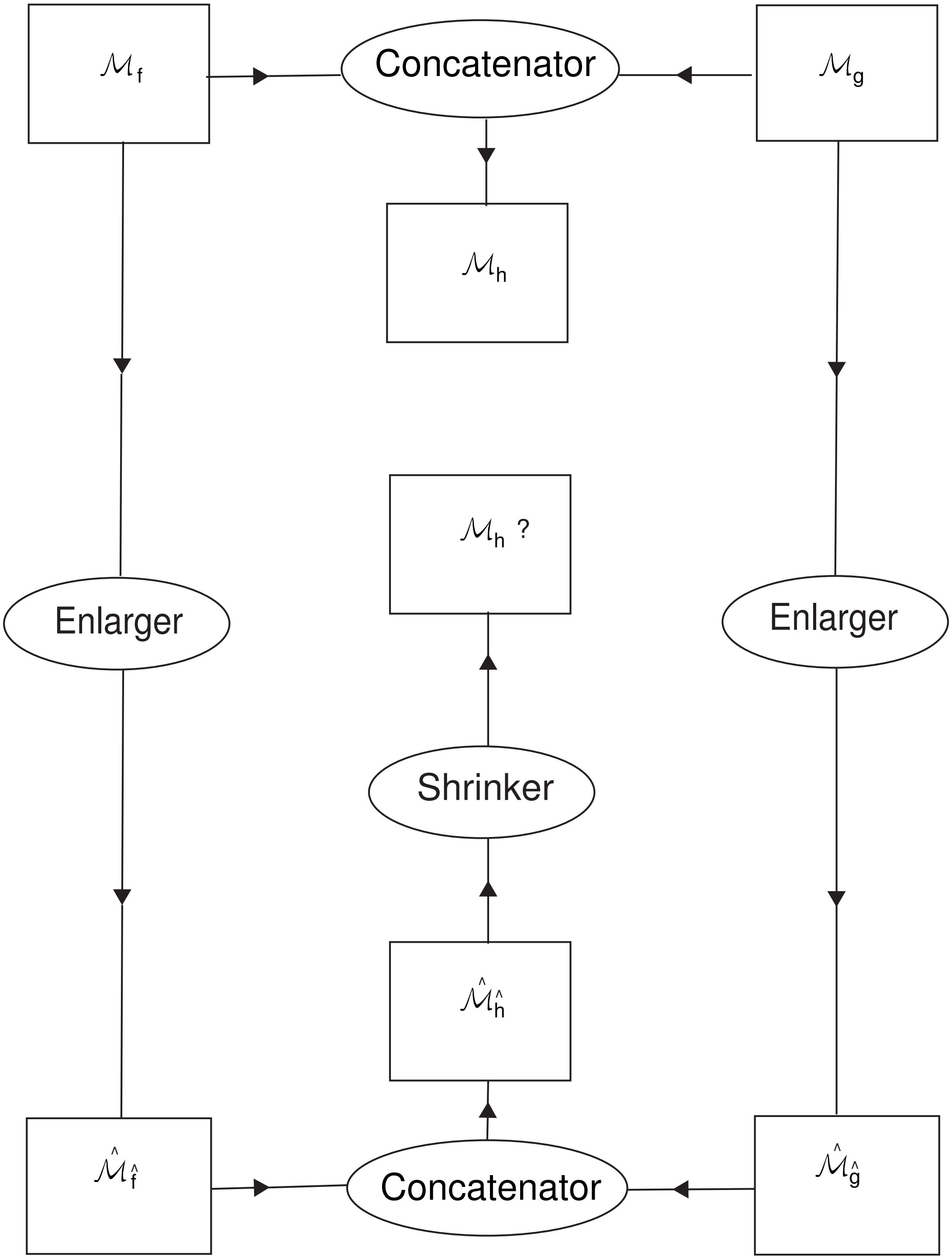

$\mathcal{M}_{\mathrm{f}}$
Concatenator
$\mathcal{M}_{\mathrm{g}}$
$\mathcal{M}_{\mathrm{h}}$
$\mathcal{M}_{\mathrm{h}}$?
Enlarger
Enlarger
Shrinker
$\hat{\mathcal{M}}_{\hat{\mathrm{h}}}$
$\hat{\mathcal{M}}_{\hat{\mathrm{f}}}$
Concatenator
$\hat{\mathcal{M}}_{\hat{\mathrm{g}}}$

Figure 9.4.1: Concatenation of origin-preserving maps in an enlarged phase space to find equivalent results for maps, including translations, in the original phase space. The concatenator depicted at the top of the figure works with the usual phase space. When translations are taken into account, it involves the use of complicated feed-down formulae as illustrated in Section 9.3. The concatenator at the bottom of the figure works in an enlarged phase space, and employs the far-simpler concatenation rules for origin preserving maps.

Consequently, (4.72) can also be written in the form

$$(z, J\delta^h) = (z, J\delta^f) + (R^f z, J\delta^g). \tag{9.4.74}$$

Finally, use of (7.7.3) gives the result

$$h_1(z) = f_1(z) + g_1(R^f z) \tag{9.4.75}$$

or, equivalently,

$$h_1(z) = f_1(z) + \mathcal{R}_f g_1(z). \tag{9.4.76}$$

But (4.75), along with (4.69), are the rules (2.37) and (2.38) for concatenating inhomogeneous symplectic group maps. We conclude that, in the case of inhomogeneous symplectic group maps, the map $\mathcal{M}_h$ can indeed be deduced from $\hat{\mathcal{M}}_{\hat{h}}$.

9.4.3 Case of General Maps

Enlarging

We now turn to the general case $ISpM(2n, \mathbb{R})$ for maps $\mathcal{M}_f$ and $\mathcal{M}_g$ of the form (1.1) and (1.2). The enlargement process will be carried out as before to yield the maps $\hat{\mathcal{M}}_{\hat{f}}$ and $\hat{\mathcal{M}}_{\hat{g}}$. For example, the map $\hat{\mathcal{M}}_{\hat{f}}$ is given by

$$\hat{\mathcal{M}}_{\hat{f}} = \exp(:\hat{f}_1^2:)\check{\mathcal{R}}_{\hat{f}}\exp(:\hat{f}_3^3:)\exp(:\hat{f}_4^4:)\cdots \tag{9.4.77}$$

where $\hat{f}_1^2$ and $\check{\mathcal{R}}_{\hat{f}}$ are given by (4.48) and (4.50) as before, and

$$\hat{f}_m^m(\hat{z}) = f_m(z),\ m = 3, 4, \cdots . \tag{9.4.78}$$

Next form the product map

$$\hat{\mathcal{M}}_{\hat{h}} = \hat{\mathcal{M}}_{\hat{f}}\hat{\mathcal{M}}_{\hat{g}}. \tag{9.4.79}$$

Since $\hat{\mathcal{M}}_{\hat{h}}$ sends the origin into itself, it has a factorization of the form

$$\hat{\mathcal{M}}_{\hat{h}} = \hat{\mathcal{R}}_{\hat{h}}\hat{\mathcal{N}}_{\hat{h}}. \tag{9.4.80}$$

The linear map $\hat{\mathcal{R}}_{\hat{h}}$ will be described by a matrix $\hat{R}^{\hat{h}}$, and from the relation

$$\hat{\mathcal{R}}_{\hat{h}} = \hat{\mathcal{R}}_{\hat{f}}\hat{\mathcal{R}}_{\hat{g}} \tag{9.4.81}$$

we have the rule

$$\hat{R}^{\hat{h}} = \hat{R}^{\hat{g}}\hat{R}^{\hat{f}}. \tag{9.4.82}$$

The nonlinear map $\hat{\mathcal{N}}_{\hat{h}}$ will have a representation of the form

$$\hat{\mathcal{N}}_{\hat{h}} = \exp(:\hat{h}_3:)\exp(:\hat{h}_4:)\cdots \tag{9.4.83}$$

with the $\hat{h}_m$ given by the relations of the form (8.4.31) through (8.4.36) already found in Section 8.4. Our task now is to extract $\mathcal{M}_h$ from $\hat{\mathcal{M}}_{\hat{h}}$.

Let us first examine $\hat{\mathcal{R}}_{\hat{h}}$ and its associated matrix $\hat{R}^{\hat{h}}$. We know that $\hat{\mathcal{M}}_{\hat{h}}$ has the property (4.2) since by construction the maps $\hat{\mathcal{M}}_{\hat{f}}$ and $\hat{\mathcal{M}}_{\hat{g}}$ have this property, and such maps form a group. Consequently $\hat{\mathcal{R}}_{\hat{h}}$ satisfies (4.4). It follows, in analogy with (4.46) when written out for the full 8×8 case, that the matrix $\hat{R}^{\hat{h}}$ has the form

$$\hat{R}^{\hat{h}} = \left(\begin{array}{cccccccc} R^h_{11} & R^h_{12} & R^h_{13} & R^h_{14} & R^h_{15} & R^h_{16} & (R^h\delta^h)_1 & 0 \\ R^h_{21} & R^h_{22} & R^h_{23} & R^h_{24} & R^h_{25} & R^h_{26} & (R^h\delta^h)_2 & 0 \\ R^h_{31} & R^h_{32} & R^h_{33} & R^h_{34} & R^h_{35} & R^h_{36} & (R^h\delta^h)_3 & 0 \\ R^h_{41} & R^h_{42} & R^h_{43} & R^h_{44} & R^h_{45} & R^h_{46} & (R^h\delta^h)_4 & 0 \\ R^h_{51} & R^h_{52} & R^h_{53} & R^h_{54} & R^h_{55} & R^h_{56} & (R^h\delta^h)_5 & 0 \\ R^h_{61} & R^h_{62} & R^h_{63} & R^h_{64} & R^h_{65} & R^h_{66} & (R^h\delta^h)_6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \delta^h_2 & -\delta^h_1 & \delta^h_4 & -\delta^h_3 & \delta^h_6 & -\delta^h_5 & \hat{R}^{\hat{h}}_{87} & 1 \end{array} \right). \tag{9.4.84}$$

We can read off the entries in δ^h from the bottom row of (4.84), and from these entries construct $h_1(z)$. Specifically, if we write

$$h_1(z) = (z, (\delta^h)^*), \tag{9.4.85}$$

then we have the relation

$$[(\delta^h)^*]_b = (\hat{R}^{\hat{h}})_{8b} \text{ for } b \in [1, 6]. \tag{9.4.86}$$

Again see Section 7.7.

Next let $\check{\mathcal{R}}^{\hat{h}}$ be the symplectic map described by the matrix $\check{R}^{\hat{h}}$ with

$$\check{R}^{\hat{h}} = \begin{pmatrix} R^h & 0 \\ 0 & I \end{pmatrix}. \tag{9.4.87}$$

Here R^h is the 6×6 matrix obtained earlier, and I denotes the 2×2 identity matrix. It follows that $\hat{\mathcal{R}}_{\hat{h}}$ can be rewritten in the form

$$\hat{\mathcal{R}}_{\hat{h}} = \exp(\alpha^h_2 : q^2_{n+1} :) \exp(: q_{n+1} h_1 :) \check{\mathcal{R}}_{\hat{h}} \tag{9.4.88}$$

with $n = 3$. Here α^h_2 is given by the relation

$$\alpha^h_2 = R^h_{87}/2. \tag{9.4.89}$$

The quantity α^h_2 is not presently of direct interest to us, but if desired it can be computed from the entries in $\hat{R}^{\hat{f}}$ and $\hat{R}^{\hat{g}}$. See Exercise 4.10.

We next turn to the nonlinear part $\hat{\mathcal{N}}_{\hat{h}}$. We know from (4.5) that the $\hat{h}_m$ are independent of p_{n+1}. Therefore the $\hat{h}_m$ must have expansions of the form

$$\hat{h}_m(\hat{z}) = h^m_m(z) + (q_{n+1}) h^m_{m-1}(z) + \cdots + (q_{n+1})^m h^m_0(z). \tag{9.4.90}$$

Here the superscript m on the quantity h^m_ℓ indicates that the quantity is associated with $\hat{h}_m$, and the subscript ℓ indicates that the quantity is homogeneous of degree ℓ in the variables z. Let us explore the consequences of this expansion. In Section 9.3 we employed an expansion in powers of ϵ where ϵ was a measure of the smallness of the first-degree generators. See, for example, relations of the kind (3.10) through (3.14). From this perspective, the expansion (4.90) is an expansion in powers of q_{n+1} with q_{n+1} playing the role of ϵ. Compare also (3.1) and (4.48). (See also Exercise 4.11.) Moreover, the use of the standard concatenator for origin preserving maps in the enlarged phase space produces power series expansions in the quantity q_{n+1} automatically!

Shrinking by Concatenation

Equally pleasant is the fact that this concatenator can be used to construct a *shrinker*. Since the quantities $[(q_{n+1}^{m-\ell})h_\ell^m]$ form a basis for the (p_{n+1} independent) polynomials of degree m in the enlarged phase space, $\hat{\mathcal{N}}_{\hat{h}}$ must have a factorization of the form

$$\begin{aligned}\hat{\mathcal{N}}_{\hat{h}} = \quad & \exp(:\alpha_3 q_{n+1}^3:)\exp(:\alpha_4 q_{n+1}^4:)\exp(:\alpha_5 q_{n+1}^5:)\cdots\times\\ & \exp(:q_{n+1}^2\tilde{h}_1^3:)\exp(:q_{n+1}^3\tilde{h}_1^4:)\exp(:q_{n+1}^4\tilde{h}_1^5:)\cdots\times\\ & \exp(:q_{n+1}\tilde{h}_2^3:)\exp(:q_{n+1}^2\tilde{h}_2^4:)\exp(:q_{n+1}^3\tilde{h}_2^5:)\cdots\times\\ & \exp(:\tilde{h}_3^3:)\exp(:q_{n+1}\tilde{h}_3^4:)\exp(:q_{n+1}^2\tilde{h}_3^5:)\cdots\times\\ & \exp(:\tilde{h}_4^4:)\exp(:q_{n+1}\tilde{h}_4^5:)\exp(:q_{n+1}^2\tilde{h}_4^6:)\cdots\times\\ & \exp(:\tilde{h}_5^5:)\exp(:q_{n+1}\tilde{h}_5^6:)\exp(:q_{n+1}^2\tilde{h}_5^7:)\cdots,\end{aligned} \tag{9.4.91}$$

where the quantities $\tilde{h}_\ell^m$ are yet to be determined. In a moment we will see that the $\tilde{h}_\ell^m$ can be computed using the concatenator. But first we observe that (4.91) is a *single grade* factorization with q_{n+1} playing the role of ϵ. See the end of Section 9.3. We may therefore set $q_{n+1} = 1$ in (4.91) and (4.88) to obtain $\mathcal{M}_h$ from $\hat{\mathcal{M}}_{\hat{h}}$. So doing gives the result

$$\begin{aligned}\mathcal{M}_h = \quad & \exp(:h_1^2:)\mathcal{R}_h\exp(:\tilde{h}_1^3+\tilde{h}_1^4+\tilde{h}_1^5+\cdots)\times\\ & \exp(:\tilde{h}_2^3:)\exp(:\tilde{h}_2^4:)\exp(:\tilde{h}_2^5:)\cdots\times\\ & \exp(:\tilde{h}_3^3:)\exp(:\tilde{h}_3^4:)\exp(:\tilde{h}_3^5:)\cdots\times\\ & \exp(:\tilde{h}_4^4:)\exp(:\tilde{h}_4^5:)\exp(:\tilde{h}_4^6:)\cdots\times\\ & \exp(:\tilde{h}_5^5:)\exp(:\tilde{h}_5^6:)\exp(:\tilde{h}_5^7:)\cdots.\end{aligned} \tag{9.4.92}$$

Here $\mathcal{R}_h$ is a linear map whose associated matrix is R^h, and

$$h_1^2(z) = h_1(z). \tag{9.4.93}$$

The second two factors in (4.92) can be rearranged using the results of Section 9.2,

$$\mathcal{R}_h\exp(:\tilde{h}_1^3+\tilde{h}_1^4+\tilde{h}_1^5+\cdots) = \exp(:\check{h}_1:)\mathcal{R}_h \tag{9.4.94}$$

where

$$\check{h}_1 = \mathcal{R}_h(\tilde{h}_1^3+\tilde{h}_1^4+\tilde{h}_1^5+\cdots). \tag{9.4.95}$$

Consequently, $\mathcal{M}_h$ can also be written in the form

$$\begin{aligned}\mathcal{M}_h = \quad & \exp(:h_1^2 + \check{h}_1:) \times \\ & \mathcal{R}_h \exp(:\tilde{h}_2^3:)\exp(:\tilde{h}_2^4:)\exp(:\tilde{h}_2^5:)\cdots \times \\ & \exp(:\tilde{h}_3^3:)\exp(:\tilde{h}_3^4:)\exp(:\tilde{h}_3^5:)\cdots \times \\ & \exp(:\tilde{h}_4^4:)\exp(:\tilde{h}_4^5:)\exp(:\tilde{h}_4^6:)\cdots \times \\ & \exp(:\tilde{h}_5^5:)\exp(:\tilde{h}_5^6:)\exp(:\tilde{h}_5^7:)\cdots . \end{aligned} \tag{9.4.96}$$

Finally, the factors appearing in each of the lines in (4.96) beyond the first line may be combined using the concatenator for origin preserving maps in the original $2n$-dimensional phase space to obtain $\mathcal{M}_h$ in the final form (1.6). We have constructed a shrinker based on the assumption that the terms $\tilde{h}_\ell^m$ appearing in (4.91) can be found.

Illustration for the quotient algebra L^0/L^3

What remains to be shown is how the $\tilde{h}_\ell^m$ can be computed from the $\hat{h}_m$ in (4.83) and (4.90) using the concatenator for origin preserving maps in the enlarged $(2n+2)$-dimensional phase space. Since the procedure requires several steps, it is best illustrated first for a relatively simple example. Suppose we are working in the quotient algebra L^0/L^3. Then $\hat{\mathcal{N}}_{\hat{h}}$ has the form

$$\begin{aligned}\hat{\mathcal{N}}_{\hat{h}} = \quad & \exp(:h_3^3 + q_{n+1}h_2^3 + q_{n+1}^2 h_1^3 + q_{n+1}^3 h_0^3:) \times \\ & \exp(:h_4^4 + q_{n+1}h_3^4 + q_{n+1}^2 h_2^4 + q_{n+1}^3 h_1^4 + q_{n+1}^4 h_0^4:). \end{aligned} \tag{9.4.97}$$

Since the generators have *no* p_{n+1} dependence, they are in involution with powers of q_{n+1}, and these powers may be removed to the far right so that we may also write

$$\begin{aligned}\hat{\mathcal{N}}_{\hat{h}} = \quad & \exp(:h_3^3 + q_{n+1}h_2^3 + q_{n+1}^2 h_1^3:) \times \\ & \exp(:h_4^4 + q_{n+1}h_3^4 + q_{n+1}^2 h_2^4 + q_{n+1}^3 h_1^4:) \times \\ & \exp(:q_{n+1}^3 h_0^3:)\exp(:q_{n+1}^4 h_0^4:). \end{aligned} \tag{9.4.98}$$

Now we are ready to begin.

Isolation of linear in z generators

The *linear* in z generator $q_{n+1}^2 h_1^3$, which produces a translation in the $2n$-dimensional phase space, may be isolated by writing the identity

$$\hat{\mathcal{N}}_{\hat{h}} = \exp(:q_{n+1}^2 h_1^3:)[\exp(-:q_{n+1}^2 h_1^3:)\hat{\mathcal{N}}_{\hat{h}}] \tag{9.4.99}$$

and making the definition

$$^1\hat{\mathcal{N}}_{\hat{h}} = [\exp(-:q_{n+1}^2 h_1^3:)\hat{\mathcal{N}}_{\hat{h}}]. \tag{9.4.100}$$

Upon manipulating exponents using the BCH theorem we find that $^1\hat{\mathcal{N}}_{\hat{h}}$ has the form

$$\begin{aligned}{}^1\hat{\mathcal{N}}_{\hat{h}} = \quad & \exp(:{}^1h_3^3 + q_{n+1}\,{}^1h_2^3:) \times \\ & \exp(:{}^1h_4^4 + q_{n+1}\,{}^1h_3^4 + q_{n+1}^2\,{}^1h_2^4 + q_{n+1}^3\,{}^1h_1^4:) \times \\ & \exp(:q_{n+1}^3\,{}^1h_0^3:)\exp(:q_{n+1}^4\,{}^1h_0^4:), \end{aligned} \tag{9.4.101}$$

and we conclude that

$$\tilde{h}_1^3 = h_1^3. \tag{9.4.102}$$

Here the superscript 1 in ${}^1\hat{\mathcal{N}_{\hat{h}}}$ indicates that one isolation step has been taken; and the superscript 1 in ${}^1h_\ell^m$ indicates that one isolation step has been taken and that the ${}^1h_\ell^m$ may differ from the previous h_ℓ^m.

Next the linear in z generator $q_{n+1}^3\ {}^1h_1^4$, which which also produces a translation in the $2n$-dimensional phase space, may be isolated by writing the identity

$${}^1\hat{\mathcal{N}_{\hat{h}}} = \exp(:q_{n+1}^3\ {}^1h_1^4:)[\exp(-:q_{n+1}^3\ {}^1h_1^4:)\ {}^1\hat{\mathcal{N}_{\hat{h}}}] \tag{9.4.103}$$

and making the definition

$${}^2\hat{\mathcal{N}_{\hat{h}}} = [\exp(-:q_{n+1}^3\ {}^1h_1^4:)\ {}^1\hat{\mathcal{N}_{\hat{h}}}]. \tag{9.4.104}$$

Again, upon manipulating exponents using the BCH theorem, we find that ${}^2\hat{\mathcal{N}_{\hat{h}}}$ has the form

$$\begin{aligned}{}^2\hat{\mathcal{N}_{\hat{h}}} = \quad & \exp(:{}^2h_3^3 + q_{n+1}\ {}^2h_2^3:) \times \\ & \exp(:{}^2h_4^4 + q_{n+1}\ {}^2h_3^4 + q_{n+1}^2\ {}^2h_2^4:) \times \\ & \exp(:q_{n+1}^3\ {}^2h_0^3:)\exp(:q_{n+1}^4\ {}^2h_0^4:),\end{aligned} \tag{9.4.105}$$

and we conclude that

$$\tilde{h}_1^4 = {}^1h_1^4. \tag{9.4.106}$$

Here the superscript 2 in ${}^2\hat{\mathcal{N}_{\hat{h}}}$ indicates that a second isolation step has been taken; and the superscript 2 in ${}^2h_\ell^m$ indicates that a second isolation step has been taken and that the ${}^2h_\ell^m$ may differ from the previous ${}^1h_\ell^m$.

Inspection of (4.105) indicates that all linear in z generators have now been isolated away. We are ready to begin isolating the *quadratic* in z generators.

Isolation of quadratic in z generators

The quadratic in z generator $q_{n+1}\ {}^2h_2^3$, which produces a linear transformation in the $2n$-dimensional phase space, may be isolated by writing the identity

$${}^2\hat{\mathcal{N}_{\hat{h}}} = \exp(:q_{n+1}\ {}^2h_2^3:)[\exp(-:q_{n+1}\ {}^2h_2^3:)\ {}^2\hat{\mathcal{N}_{\hat{h}}}] \tag{9.4.107}$$

and making the definition

$${}^3\hat{\mathcal{N}_{\hat{h}}} = [\exp(-:q_{n+1}\ {}^2h_2^3:)\ {}^2\hat{\mathcal{N}_{\hat{h}}}]. \tag{9.4.108}$$

Now use of the BCH theorem shows that ${}^3\hat{\mathcal{N}_{\hat{h}}}$ has the form

$$\begin{aligned}{}^3\hat{\mathcal{N}_{\hat{h}}} = \quad & \exp(:{}^3h_3^3:) \times \\ & \exp(:{}^3h_4^4 + q_{n+1}\ {}^3h_3^4 + q_{n+1}^2\ {}^3h_2^4:) \times \\ & \exp(:q_{n+1}^3\ {}^3h_0^3:)\exp(:q_{n+1}^4\ {}^3h_0^4:),\end{aligned} \tag{9.4.109}$$

and we conclude that

$$\tilde{h}_2^3 = {}^2h_2^3. \tag{9.4.110}$$

Next the quadratic in z generator $q_{n+1}^2\ {}^3h_2^4$, which also produces a linear transformation in the $2n$-dimensional phase space, may be isolated by writing the identity

$$ {}^3\hat{\mathcal{N}}_{\hat{h}} = \exp(:q_{n+1}^2\ {}^3h_2^4:)[\exp(-:q_{n+1}^2\ {}^3h_2^4:)\ {}^3\hat{\mathcal{N}}_{\hat{h}}] \tag{9.4.111}$$

and making the definition

$$ {}^4\hat{\mathcal{N}}_{\hat{h}} = [\exp(-:q_{n+1}^2\ {}^3h_2^4:)\ {}^3\hat{\mathcal{N}}_{\hat{h}}]. \tag{9.4.112}$$

Now use of the BCH theorem shows that ${}^4\hat{\mathcal{N}}_{\hat{h}}$ has the form

$$\begin{aligned} {}^4\hat{\mathcal{N}}_{\hat{h}} = \quad & \exp(:{}^4h_3^3:)\times \\ & \exp(:{}^4h_4^4 + q_{n+1}\ {}^4h_3^4:)\times \\ & \exp(:q_{n+1}^3\ {}^4h_0^3:)\exp(:q_{n+1}^4\ {}^4h_0^4:), \end{aligned} \tag{9.4.113}$$

and we conclude that

$$\tilde{h}_2^4 = {}^3h_2^4. \tag{9.4.114}$$

Inspection of (4.113) indicates that all quadratic in z generators have now been isolated away. We are ready to begin isolating the *cubic* in z generators.

Isolation of cubic in z generators

The cubic in z generator ${}^4h_3^3$, which produces a quadratic plus higher-order transformation in the $2n$-dimensional phase space, may be isolated by writing the identity

$$ {}^4\hat{\mathcal{N}}_{\hat{h}} = \exp(:{}^4h_3^3:)[\exp(-:{}^4h_3^3:)\ {}^4\hat{\mathcal{N}}_{\hat{h}}] \tag{9.4.115}$$

and making the definition

$$ {}^5\hat{\mathcal{N}}_{\hat{h}} = [\exp(-:{}^4h_3^3:)\ {}^4\hat{\mathcal{N}}_{\hat{h}}]. \tag{9.4.116}$$

Now use of the BCH theorem shows that ${}^5\hat{\mathcal{N}}_{\hat{h}}$ has the form

$$\begin{aligned} {}^5\hat{\mathcal{N}}_{\hat{h}} = \quad & \exp(:{}^5h_4^4 + q_{n+1}\ {}^5h_3^4:)\times \\ & \exp(:q_{n+1}^3\ {}^5h_0^3:)\exp(:q_{n+1}^4\ {}^5h_0^4:), \end{aligned} \tag{9.4.117}$$

and we conclude that

$$\tilde{h}_3^3 = {}^4h_3^3. \tag{9.4.118}$$

Next the cubic in z generator $q_{n+1}\ {}^5h_3^4$, which also produces a quadratic plus higher-order transformation in the $2n$-dimensional phase space, may be isolated by writing the identity

$$ {}^5\hat{\mathcal{N}}_{\hat{h}} = \exp(:q_{n+1}\ {}^5h_3^4:)[\exp(-:q_{n+1}\ {}^5h_3^4:)\ {}^5\hat{\mathcal{N}}_{\hat{h}}] \tag{9.4.119}$$

and making the definition

$$ {}^6\hat{\mathcal{N}}_{\hat{h}} = [\exp(-:q_{n+1}\ {}^5h_3^4:)\ {}^5\hat{\mathcal{N}}_{\hat{h}}]. \tag{9.4.120}$$

Now use of the BCH theorem shows that ${}^6\hat{\mathcal{N}}_{\hat{h}}$ has the form

$$\begin{aligned} {}^6\hat{\mathcal{N}}_{\hat{h}} = \quad & \exp(:{}^6h_4^4:)\times \\ & \exp(:q_{n+1}^3\,{}^6h_0^3:)\exp(:q_{n+1}^4\,{}^6h_0^4:), \end{aligned} \tag{9.4.121}$$

and we conclude that

$$\tilde{h}_3^4 = {}^5h_3^4. \tag{9.4.122}$$

Inspection of (4.121) indicates that all cubic in z generators have now been isolated away. We are ready to begin isolating the *quartic* in z generators.

Isolation of quartic in z generators

The remaining quartic in z generator ${}^6h_4^4$, which produces a cubic plus higher-order transformation in the $2n$-dimensional phase space, may be isolated by writing the identity

$${}^6\hat{\mathcal{N}}_{\hat{h}} = \exp(:{}^6h_4^4:)[\exp(-:{}^6h_4^4:)\,{}^6\hat{\mathcal{N}}_{\hat{h}}] \tag{9.4.123}$$

and making the definition

$${}^7\hat{\mathcal{N}}_{\hat{h}} = [\exp(-:{}^6h_4^4:)\,{}^6\hat{\mathcal{N}}_{\hat{h}}]. \tag{9.4.124}$$

Now use of the BCH theorem shows that ${}^7\hat{\mathcal{N}}_{\hat{h}}$ has the form

$${}^7\hat{\mathcal{N}}_{\hat{h}} = \exp(:q_{n+1}^3\,{}^7h_0^3:)\exp(:q_{n+1}^4\,{}^7h_0^4:), \tag{9.4.125}$$

and we conclude that

$$\tilde{h}_4^4 = {}^6h_4^4. \tag{9.4.126}$$

Overview

Inspection of (4.125) indicates that, in the case of L^0/L^3, we have achieved our goal. Namely, by repeated isolating and concatenating, we eventually achieved the factorization (4.91). All z-dependent generators have been isolated away, and all that remains are generators that depend solely on q_{n+1}. These remaining generators have no effect on the the $2n$-dimensional phase space variables z, and therefore ${}^7\hat{\mathcal{N}}_{\hat{h}}$ acts as the identity map $\mathcal{I}$ on the z phase space.

We note that while these steps are somewhat difficult to characterize analytically (which is to be expected because the results of Section 9.3 were complicated), they are easy to implement numerically.

Is there a pattern in what we have done? Review of the steps we have taken shows that we have extracted the $\tilde{h}_\ell^m$ in a particular order. Let r be a *running* index. Table 4.1 below shows the order in which we have extracted the $\tilde{h}_\ell^m$. Here we have defined the difference d by the relation

$$d = m - \ell \tag{9.4.127}$$

so that q_{n+1} and $\tilde{h}_\ell^m$ occur in the combination $q_{n+1}^d\,\tilde{h}_\ell^m$.

Let us make the definition

$${}^0\hat{\mathcal{N}}_{\hat{h}} = \hat{\mathcal{N}}_{\hat{h}}. \tag{9.4.128}$$

Table 9.4.1: Order in which the $\tilde{h}_\ell^m$ are to be extracted for the case L^0/L^3.

r	ℓ	m	d
1	1	3	2
2	1	4	3
3	2	3	1
4	2	4	2
5	3	3	0
6	3	4	1
7	4	4	0

Then we may view the whole process as being recursive. At any given stage a map ${}^{r-1}\hat{\mathcal{N}}_{\hat{h}}$ and a pair of indices $\ell(r)$ and $m(r)$ are provided as input, and a map ${}^{r}\hat{\mathcal{N}}_{\hat{h}}$ and polynomial $\tilde{h}_\ell^m$ are produced as output. See Figure 4.2. The polynomial $\tilde{h}_\ell^m$ is determined by examination of ${}^{r-1}\hat{\mathcal{N}}_{\hat{h}}$ and given by the rule

$$\tilde{h}_\ell^m = {}^{r-1}h_\ell^m. \tag{9.4.129}$$

The map ${}^{r}\hat{\mathcal{N}}_{\hat{h}}$ is given by carrying out the concatenation

$${}^{r}\hat{\mathcal{N}}_{\hat{h}} = \exp(- : q_{n+1}^d \tilde{h}_\ell^m :) \times {}^{r-1}\hat{\mathcal{N}}_{\hat{h}}. \tag{9.4.130}$$

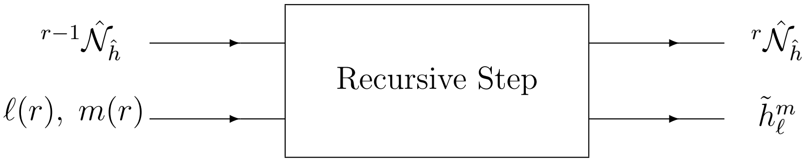

${}^{r-1}\hat{\mathcal{N}}_{\hat{h}}$
$\ell(r),\ m(r)$
Recursive Step
${}^{r}\hat{\mathcal{N}}_{\hat{h}}$
$\tilde{h}_\ell^m$

Figure 9.4.2: A recursive step that takes a map ${}^{r-1}\hat{\mathcal{N}}_{\hat{h}}$ and a pair of indices $\ell(r)$ and $m(r)$ as input, and produces a map ${}^{r}\hat{\mathcal{N}}_{\hat{h}}$ and polynomial $\tilde{h}_\ell^m$ as output.

Shrinking in the General Case

It is now a simple matter to generalize to higher-order cases. For example, Table 4.2 shows the extraction order to be used when working with maps through 7th order, the order that is available using the concatenation rules provided by (8.4.31) through (8.4.36).

Finally, we note that the procedure is self checking. For the final value of r the corresponding map ${}^{r}\hat{\mathcal{N}}_{\hat{h}}$ will have the property that all its generators have no z dependence; they can depend only on q_{n+1}. We have already seen this for the map ${}^{7}\hat{\mathcal{N}}_{\hat{h}}$ when working with third-order maps. See (4.125). The same will be true of the map ${}^{33}\hat{\mathcal{N}}_{\hat{h}}$ when working with maps through 7th order. Conversely, if all the factors in (4.91) are concatenated together, the result must be the original map $\hat{\mathcal{N}}_{\hat{h}}$.

Table 9.4.2: Order in which the $\tilde{h}_{\ell}^{m}$ are to be extracted for the case L^0/L^7.

r	ℓ	m	d
1	1	3	2
2	1	4	3
3	1	5	4
4	1	6	5
5	1	7	6
6	1	8	7
7	2	3	1
8	2	4	2
9	2	5	3
10	2	6	4
11	2	7	5
12	2	8	6
13	3	3	0
14	3	4	1
15	3	5	2
16	3	6	3
17	3	7	4
18	3	8	5
19	4	4	0
20	4	5	1
21	4	6	2
22	4	7	3
23	4	8	4
24	5	5	0
25	5	6	1
26	5	7	2
27	5	8	3
28	6	6	0
29	6	7	1
30	6	8	2
31	7	7	0
32	7	8	1
33	8	8	0

Exercises

9.4.1. Verify that maps $\hat{\mathcal{M}}$ satisfying (4.2) form a group.

9.4.2. Prove (4.5) in detail.

9.4.3. Verify that the symplectic condition requires the relations (4.16).

9.4.4. Verify (4.25) through (4.30).

9.4.5. Verify (4.38) and (4.39).

9.4.6. Verify (4.41) through (4.43).

9.4.7. Verify (4.44) through (4.46).

9.4.8. Verify (4.53) through (4.62).

9.4.9. Verify (4.65).

9.4.10. Consider two *linear* symplectic maps $\hat{\mathcal{M}}_{\hat{f}}$ and $\hat{\mathcal{M}}_{\hat{g}}$ of the factored form

$$\hat{\mathcal{M}}_{\hat{f}} = \exp(:\alpha_2^{\hat{f}} q_{n+1}^2:)\exp(:q_{n+1}f_1(z):)\check{\mathcal{R}}_{\hat{f}}, \tag{9.4.131}$$

$$\hat{\mathcal{M}}_{\hat{g}} = \exp(:\alpha_2^{\hat{g}} q_{n+1}^2:)\exp(:q_{n+1}g_1(z):)\check{\mathcal{R}}_{\hat{g}}, \tag{9.4.132}$$

where $\check{\mathcal{R}}_{\hat{f}}$ and $\check{\mathcal{R}}_{\hat{g}}$ leave the q_{n+1}, p_{n+1} subspace invariant,

$$\check{\mathcal{R}}_{\hat{f}} q_{n+1} = q_{n+1}, \text{ etc.}, \tag{9.4.133}$$

$$\check{\mathcal{R}}_{\hat{f}} p_{n+1} = p_{n+1}, \text{ etc.} \tag{9.4.134}$$

Show that (4.106) and (4.107) plus the symplectic condition require that the associated matrices $\check{R}^{\hat{f}}$ etc. be of the general form (4.50). Let $\hat{\mathcal{M}}_{\hat{h}}$ be the product of $\hat{\mathcal{M}}_{\hat{f}}$ and $\hat{\mathcal{M}}_{\hat{g}}$,

$$\begin{aligned}\hat{\mathcal{M}}_{\hat{h}} &= \hat{\mathcal{M}}_{\hat{f}}\hat{\mathcal{M}}_{\hat{g}} = \exp(:\alpha_2^{\hat{f}} q_{n+1}^2:)\exp(:q_{n+1}f_1(z):)\check{\mathcal{R}}_{\hat{f}} \times \\ &\quad \exp(:\alpha_2^{\hat{g}} q_{n+1}^2:)\exp(:q_{n+1}g_1(z):)\check{\mathcal{R}}_{\hat{g}}.\end{aligned} \tag{9.4.135}$$

Show, by manipulating the various factors involved, that $\hat{\mathcal{M}}_{\hat{h}}$ can be re-expressed in the factored form

$$\begin{aligned}\hat{\mathcal{M}}_{\hat{h}} &= \exp\{: q_{n+1}^2 : (\alpha_2^{\hat{f}} + \alpha_2^{\hat{g}} + [f_1(z), \check{\mathcal{R}}_{\hat{f}} g_1(z)]/2) :\} \times \\ &\quad \exp\{: q_{n+1}(f_1(z) + \check{\mathcal{R}}_{\hat{f}} g_1(z)) :\}\check{\mathcal{R}}_{\hat{f}}\check{\mathcal{R}}_{\hat{g}}.\end{aligned} \tag{9.4.136}$$

Thus, if we write $\hat{\mathcal{M}}_{\hat{h}}$ as

$$\hat{\mathcal{M}}_{\hat{h}} = \exp(:\alpha_2^{\hat{h}} q_{n+1}^2:)\exp(:q_{n+1}h_1(z):)\check{\mathcal{R}}_{\hat{h}}, \tag{9.4.137}$$

there are the relations

$$\alpha_2^{\hat{h}} = \alpha_2^{\hat{f}} + \alpha_2^{\hat{g}} + [f_1(z), \check{\mathcal{R}}_{\hat{f}} g_1(z)]/2, \tag{9.4.138}$$

$$h_1(z) = f_1(z) + \check{\mathcal{R}}_{\hat{f}} g_1(z), \tag{9.4.139}$$

$$\check{\mathcal{R}}_{\hat{h}} = \check{\mathcal{R}}_{\hat{f}}\check{\mathcal{R}}_{\hat{g}}. \tag{9.4.140}$$

Carry out the indicated multiplication in (4.68) and verify that your results are equivalent to those found above.

9.4.11. Problem about the relation between the inhomogeneous symplectic map Lie algebra ${}^{\epsilon}L^0/{}^{\epsilon}L^{\ell}$ in $2n$ dimensions and the subgroup of the homogeneous symplectic map Lie algebra L^0/L^{ℓ} in $2n+2$ dimensions produced by all p_{n+1} independent generators.

9.4.12. Verify that a factorization of the form (4.91) is possible.

9.4.13. Verify the relations (4.93) through (4.96).

9.4.14. Verify that $\hat{\mathcal{N}}_{\hat{h}}$ has the form (4.97) when one works in the quotient algebra L^0/L^3.

9.4.15. Verify (4.98) through (4.103).

9.4.16. Verify that extracting the $\tilde{h}_{\ell}^{m}$ in the order given by Table 4.1 or 4.2 never results, during the extraction process, in the "reappearance" in the subsequent maps ${}^{r}\hat{\mathcal{N}}_{\hat{h}}$ of any of the previously removed terms.

9.5 Map Inversion and Reverse and Mixed Factorizations

Much of the discussion of this section is analogous to that of Section 8.5. Suppose, as in (1.1), that the map $\mathcal{M}_f$ is written in the standard factored product form

$$\mathcal{M}_f = \exp(:f_1:)\mathcal{R}_f \exp(:f_3:)\exp(:f_4:)\cdots\,. \tag{9.5.1}$$

Here, as in Section 8.5, $\mathcal{R}_f$ denotes the map

$$\mathcal{R}_f = \exp(:f_2^c:)\exp(:f_2^a:) \tag{9.5.2}$$

with the associated matrix R^f given by the relation

$$R^f = \exp(JS^a)\exp(JS^c). \tag{9.5.3}$$

It follows immediately from (5.1) that the *inverse* of $\mathcal{M}_f$ has the representation

$$(\mathcal{M}_f)^{-1} = \cdots \exp(-:f_4:)\exp(-:f_3:)(\mathcal{R}_f)^{-1}\exp(-:f_1:). \tag{9.5.4}$$

As before we observe that (5.4) gives a representation for the inverse of $\mathcal{M}_f$ in the form of a reverse factorization, and that we would also like to have a representation in the standard forward factorization

$$(\mathcal{M}_f)^{-1} = \exp(:h_1:)\mathcal{R}_h \exp(:h_3:)\exp(:h_4:)\cdots\,. \tag{9.5.5}$$

See Section 7.8. This is easily accomplished with the aid of the concatenation formulas of the previous section. We simply write (5.4) and (5.5) in the form

$$\begin{aligned}&\cdots[\exp(-:f_4:)][\exp(-:f_3:)][(\mathcal{R}_f)^{-1}][\exp(-:f_1:)] =\\ &\exp(:h_1:)\mathcal{R}_h \exp(:h_3:)\exp(:h_4:)\cdots\end{aligned} \tag{9.5.6}$$

where we have used square brackets to indicate that the various maps are to be concatenated together. Note that in this case (8.5.7) no longer holds because of the feed-down terms produced by moving the f_1 factor in (5.6) to the left.

The relation (5.5) also provides a procedure for reverse factorizing a map. Suppose we wish to represent $\mathcal{M}_f$ in reverse factorized form. That is, we wish to find generators g_m such that

$$\begin{aligned}\mathcal{M}_f = \quad &\exp(:f_1:)\mathcal{R}_f \exp(:f_3:)\exp(:f_4:)\cdots =\\ \cdots \quad &\exp(:g_4:)\exp(:g_3:)\mathcal{R}_g \exp(:g_1:).\end{aligned} \tag{9.5.7}$$

Simply take the inverse of both sides of (5.7) and use (5.5) to get the relation

$$\begin{aligned}&\exp(-:g_1:)(\mathcal{R}_g)^{-1}\exp(-:g_3:)\exp(-:g_4:)\cdots =\\ &\exp(:h_1:)\mathcal{R}_h \exp(:h_3:)\exp(:h_4:)\cdots\,.\end{aligned} \tag{9.5.8}$$

From (5.8) we find the desired results

$$\mathcal{R}_g = (\mathcal{R}_h)^{-1}, \tag{9.5.9}$$

$$g_m = -h_m. \tag{9.5.10}$$

We close this section with a remark about mixed factorizations. See Section 7.8. Suppose, for example, we desire a mixed factorization for $\mathcal{M}_f$ of the form

$$\mathcal{M}_f = \mathcal{R}_{f'} \exp(: f_3' :) \exp(: f_4' :) \cdots \exp(: f_1'). \tag{9.5.11}$$

That is, we wish to move the f_1 term in (5.1) to the far right, but keep the remaining factors in ascending order. Comparison of (5.7) and (5.11) gives the result

$$f_1' = g_1, \tag{9.5.12}$$

and the relation

$$\mathcal{R}_{f'} \exp(: f_3' :) \exp(: f_4' :) \cdots = \cdots \exp(: g_4 :) \exp(: g_3 :)\mathcal{R}_g. \tag{9.5.13}$$

The remaining factors $\mathcal{R}_{f\prime}$, $: f_3' :$, $: f_4' : \cdots$ can be gotten by applying the concatenation formulas of Section 8.4 to the right side of (5.13). That is, we write (5.13) in the form

$$\mathcal{R}_{f'} \exp(: f_3' :) \exp(: f_4' :) = \cdots [\exp(: g_4 :)][\exp(: g_3 :)][\mathcal{R}_g]. \tag{9.5.14}$$

where the square brackets indicate that the various maps are to be concatenated together.

There is a relation between f_1 and f_1' that could have been examined in the previous section, or even in Section 7.7, but can just as conveniently be discussed here. See Exercises 5.1 and 5.2. Let us write (5.1) and (5.11) in the forms

$$\mathcal{M}_f = \exp(: f_1 :)\mathcal{S}_f, \tag{9.5.15}$$

$$\mathcal{M}_f = \mathcal{S}_{f'} \exp(: f_1' :), \tag{9.5.16}$$

with

$$\mathcal{S}_f = \mathcal{R}_f \exp(: f_3 :) \exp(: f_4 :) \cdots , \tag{9.5.17}$$

$$\mathcal{S}_{f'} = \mathcal{R}_{f'} \exp(: f_3' :) \exp(: f_4' :) \cdots . \tag{9.5.18}$$

Following Section 7.7, let us also write f_1 and f_1' in the forms

$$f_1(z) = -(\delta, Jz), \tag{9.5.19}$$

$$f_1'(z) = -(\delta', Jz). \tag{9.5.20}$$

Then we have the relations

$$\delta_a' = \mathcal{S}_f z_a\big|_{z=\delta}, \tag{9.5.21}$$

$$\delta_a = (\mathcal{S}_f)^{-1} z_a\big|_{z=\delta'}. \tag{9.5.22}$$

To see the truth of (5.21) and (5.22), apply $\mathcal{M}_f$ in both its representations (5.15) and (5.16) to the origin. Consider first the representation (5.16). We know that by construction, see (5.18), the map $\mathcal{S}_{f'}$ sends the origin into itself. Therefore, since maps act in the order in which they occur when read from left to right (see Section 8.3), the first factor in (5.16) acts on the origin and leaves it in peace. Also, from Section 7.7, we know that $\exp(:f_1':)$ sends the origin into δ'. Consequently, we find the result

$$\mathcal{M}_f z_a|_{z=0} = \delta'_a. \tag{9.5.23}$$

Consider next the representation (5.15). The first factor, $\exp(:f_1:)$, sends the origin into δ. Subsequently $\mathcal{S}_f$ acts on δ to give the net result

$$\mathcal{M}_f z_a|_{z=0} = \mathcal{S}_f z_a|_{z=\delta}\,. \tag{9.5.24}$$

Upon comparing (5.23) and (5.24) we see that (5.21) and (5.22) are indeed correct. For another set of similar relations, see Exercise 5.3.

Exercises

9.5.1. Derive (5.21) starting with (7.7.13) and (7.7.23).

9.5.2. Consider a 2-dimensional phase space and suppose that $\mathcal{S}_{f'}$ has the simple form

$$\mathcal{S}_{f'} = \exp(:q^4:). \tag{9.5.25}$$

Working within the quotient group generated by the Lie algebra L^0/L^3, use (3.18) and (3.22) to verify (5.21).

9.5.3. Verify the relations

$$\mathcal{M}_f z_a\,|_{z=-\delta} = 0\ , \tag{9.5.26}$$

$$(\mathcal{M}_f)^{-1} z_a\,|_{z=0} = -\delta_a\ . \tag{9.5.27}$$

Derive the relations

$$-\delta'_a = \mathcal{S}_{f'} z_a|_{z=-\delta}\,, \tag{9.5.28}$$

$$-\,\delta_a = (\mathcal{S}_{f'})^{-1}\, z_a\big|_{z=-\delta'}\,. \tag{9.5.29}$$

Hint: Apply $\mathcal{M}_f$ as given by (5.15) and (5.16) to the phase-space point $(-\delta)$.

9.6 Taylor and Hybrid Taylor-Lie Concatenation and Inversion

This section extends the results of Section 8.6 to the case where the map to be treated may have constant terms. Again all the possibilities illustrated in Figure 8.6.1 may arise, and we will discuss those of greatest interest.

Suppose, as described in the beginning of Section 8.6, that both the maps $\mathcal{M}_1$ and $\mathcal{M}_2$ are in Taylor form, and we also desire to represent their product in Taylor form. We consider first the best of all possible circumstances. In that circumstance $\mathcal{M}_1$ sends the phase-space point z^0 to the intermediate point $\bar{z}^0$,

$$\mathcal{M}_1 : z^0 \rightarrow \bar{z}^0, \tag{9.6.1}$$

and we assume that $\mathcal{M}_1$ has a known Taylor expansion about z^0. Also, $\mathcal{M}_2$ sends the intermediate point $\bar{z}^0$ to the final point $\bar{\bar{z}}^0$,

$$\mathcal{M}_2 : \bar{z}^0 \rightarrow \bar{\bar{z}}^0, \tag{9.6.2}$$

and we assume that $\mathcal{M}_2$ has a known Taylor expansion about $\bar{z}^0$. What we desire is a Taylor expansion about the point z^0 for the product map $\mathcal{M}_3$ that sends z^0 immediately to $\bar{\bar{z}}^0$,

$$\mathcal{M}_3 : z^0 \rightarrow \bar{\bar{z}}^0. \tag{9.6.3}$$

This desire is easily met. Introduce the deviation variables ζ_a by writing

$$z_a = z_a^0 + \zeta_a. \tag{9.6.4}$$

Then, by assumption, $\mathcal{M}_1$ has a known truncated Taylor expansion of the form

$$\bar{z}_a = \bar{z}_a(z) = \bar{z}_a^0 + \sum_{m=1}^{D} g_a^1(m;\zeta). \tag{9.6.5}$$

Introduce as well the deviation variables $\bar{\zeta}_a$ by writing

$$\bar{z}_a = \bar{z}_a^0 + \bar{\zeta}_a. \tag{9.6.6}$$

Then $\mathcal{M}_2$ is assumed to have the known truncated Taylor expansion

$$\bar{\bar{z}}_a = \bar{\bar{z}}_a\,(\bar{z}) = \bar{\bar{z}}_a^0 + \sum_{m'=1}^{D} g_a^2(m';\bar{\zeta}). \tag{9.6.7}$$

Now use (6.5) and (6.6) to write the relation

$$\bar{\zeta}_a = \bar{\zeta}_a(\zeta) = \sum_{m=1}^{D} g_a^1(m;\zeta). \tag{9.6.8}$$

Also introduce the deviation variables $\overline{\overline{\zeta}}_a$, defined by

$$\overline{\overline{z}}_a = \overline{\overline{z}}^0_a + \overline{\overline{\zeta}}_a, \tag{9.6.9}$$

to rewrite (6.7) in the form

$$\overline{\overline{\zeta}}_a = \overline{\overline{\zeta}}_a (\overline{\zeta}) = \sum_{m'=1}^{D} g^2_a(m';\overline{\zeta}). \tag{9.6.10}$$

Then $\mathcal{M}_3$ has the expansion

$$\overline{\overline{z}}_a = \overline{\overline{z}}_a (z) = \overline{\overline{z}}^0_a + \overline{\overline{\zeta}}_a (\zeta) = \overline{\overline{z}}^0_a + \sum_{m''=1}^{D} g^3_a(m'';\zeta) \tag{9.6.11}$$

where the polynomials g^3_a are given by the relations

$$g^3_a(m'';\zeta) = P_{m''} \sum_{m'=1}^{D} g^2_a(m';\overline{\zeta}(\zeta)). \tag{9.6.12}$$

As before, $P_{m''}$ denotes a projection operator that retains only terms of degree m'' in the variables ζ. Also, all the required operations can again be carried out using TPSA. We see that the relations (6.8), (6.10), and (6.12) are completely analogous to the relations (8.6.2), (8.6.4), and (8.6.7) for the case of no translations. Indeed, with the use of deviation variables, all the methods of Section 8.6 can be employed. For example, a deviation variable map of the form (6.8) can be inverted by the recursion method.

Often the optimal circumstance we have just treated does not hold. It may be that $\mathcal{M}_1$ sends z^0 to $\overline{z}^0$ and $\mathcal{M}_2$ sends $\overline{z}^0$ to $\overline{\overline{z}}^0$ as before and as described by (6.1) and (6.2). However, it may happen that $\mathcal{M}_2$ does not have a known Taylor expansion about $\overline{z}^0$. Instead, we assume that $\mathcal{M}_2$ has a known Taylor expansion about a point $\overline{z}'$ that is near $\overline{z}^0$. With the introduction of suitable deviation variables if necessary, and without loss of generality, we may consider truncated Taylor series of the form

$$\overline{z}_a = \overline{z}_a(z) = \sum_{m=0}^{D} g^1_a(m;z) \tag{9.6.13}$$

and

$$\overline{\overline{z}}_a = \overline{\overline{z}}_a (\overline{z}) = \sum_{m'=0}^{D} g^2_a(m';\overline{z}). \tag{9.6.14}$$

The relations (8.6.2) and (8.6.4) have simply been modified so that all summations over m and m' begin with 0 instead of 1, and we assume that the constant term $g^1_a(0;z)$ is small. See Exercise 6.*. Finally, we make the expansion

$$\overline{\overline{z}}_a = \overline{\overline{z}}_a (z) = \sum_{m''=0}^{D} g^3_a(m'';z) \tag{9.6.15}$$

and find that the polynomials g_a^3 are given by the relations

$$g_a^3(m'';z) = P_{m''} \sum_{m'=0}^{D} g_a^2(m';\bar{z}(z)). \tag{9.6.16}$$

Suppose the maps $\mathcal{M}_1$ and $\mathcal{M}_2$ are symplectic. Then, in the terminology of Section 7.5, the truncated Taylor series (6.13) and (6.14) are symplectic D-jets (about the origin). It is important to remark at at this juncture that the concatenation of two symplectic D-jets does not generally yield a *symplectic* D-jet if $\mathcal{M}_1$ has nonvanishing constant terms $g_a^1(0;z)$ so that $\mathcal{M}_1$ does not send the origin into itself. Correspondingly, the factorization theorems of Sections 7.6 and 7.7 generally do not apply to the D-jet (6.15) for $\mathcal{M}_3$. The problem is that truncation of the Taylor expansion of a symplectic map, in this case the map $\mathcal{M}_2$, generally violates the symplectic condition, and this violation can feed down to low orders in the presence of translations. See, for example, Exercise 6.*.

There is a second point that we should also recognize. Suppose that the Taylor expansion for $\mathcal{M}_2$ is not truncated. That is, consider letting $D \to \infty$ in (6.14) and (6.16). Then it may happen that the series (5.16) diverges. This will happen if the point $\bar{z}_a(0)$ lies outside the convergence domain of the homogeneous polynomial expansion for $\mathcal{M}_2$. See Exercise 1.4.4 and Chapter 26. We conclude that translations must be handled with care.

Let $\mathcal{J}_3$ denote the D-jet (6.15). We expect that if the translation part of $\mathcal{M}_1$ is small, then $\mathcal{J}_3$ will be nearly symplectic. It should therefore be possible to construct a D-jet that is symplectic and near $\mathcal{J}_3$ in the sense that the two jets differ only by appropriate powers of the small translation terms. Indeed, this is what the method of Section 9.3 accomplishes when maps are represented in Lie form. That is, suppose the two maps $\mathcal{M}_1$ and $\mathcal{M}_2$ are written in Lie form and are concatenated using the method of Section 9.3, and suppose that the resulting map is then expanded as a Taylor series about the origin and truncated beyond terms of degree D. This resulting D-jet will be symplectic, and will be near $\mathcal{J}_3$.

Given a D-jet that is nearly a symplectic jet, there are many procedures for constructing nearby jets that are symplectic. The method just described is only one such procedure. Another convenient procedure is to employ methods analogous to those used in Section 7.6 to prove the factorization theorem.

Let $\mathcal{J}_3'$ be the jet obtained from $\mathcal{J}_3$ by removing its translation part. That is, $\mathcal{J}_3'$ sends z to $\bar{\bar{z}}'$ according to the rule

$$\bar{\bar{z}}_a' = \sum_{m=1}^{D} g_a^3(m;z), \tag{9.6.17}$$

and thus sends the origin into itself. Examine the matrix (linear) part of $\mathcal{J}_3'$ described by the terms $g_a^3(1;z)$. These terms correspond to a matrix that is nearly symplectic. Replace this matrix by a matrix that is exactly symplectic using one of the matrix symplectification methods of Chapter 4. Call this matrix R, and let $\mathcal{R}$ be its corresponding linear symplectic map. Finally, let $\mathcal{J}_3''$ be the jet that results from replacing the $g_a^3(1;z)$ terms in (6.17) by $(Rz)_a$.

Next apply $\mathcal{R}^{-1}$ to $\mathcal{J}_3''$ to get a result of the form

$$(\mathcal{R}^{-1}\mathcal{J}_3''z)_a = z_a + r_a(>1), \tag{9.6.18}$$

which is analogous to (7.6.19). As before, the remainder term $r_a(>1)$ will have a quadratic piece and still higher degree terms,

$$r_a(>1) = \hat{g}_a(2;z) + r_a(>2). \tag{9.6.19}$$

Because $\mathcal{J}_3$ is not a symplectic jet, the quadratic piece will generally not satisfy the analog of (7.6.23). However, we may still define an f_3 by the rule

$$f_3 = -(1/3)\sum_{ab} \hat{g}_a(2;z) J_{ab} z_b, \tag{9.6.20}$$

which is analogous to (7.6.26). (Indeed, Section 17.11 shows that this prescription is unique.) From this f_3 we produce the quadratic polynomials $\tilde{g}_a(2;z)$ by the rules

$$\tilde{g}_a(2;z) =: f_3 : z_a. \tag{9.6.21}$$

Because $\mathcal{J}_3$ is nearly symplectic, the polynomials $\hat{g}_a(2;z)$ and $\tilde{g}_a(2;z)$ will be nearly the same. We therefore may replace $\hat{g}_a(2;z)$ by $\tilde{g}_a(2,z)$ and, in so doing, obtain a nearby map that is more nearly symplectic.

It should now be as clear to the reader as it is to the writer that the steps just described can be applied repeatedly to yield a sequence of homogeneous polynomials $f_3, f_4 \cdots f_{D+1}$. Also, by Section 7.7, there is an f_1 polynomial that will reproduce the translation part $g_a^3(0;z)$ in (6.15). Consequently, we have found the approximate result

$$\mathcal{M}_3 \simeq \mathcal{R} \exp(: f_3 :) \cdots \exp(: f_{D+1} :) \exp(: f_1 :). \tag{9.6.22}$$

Finally, the map on the right side of (6.22) may be expanded in a Taylor series and truncated beyond terms of degree D. Doing so yields a symplectic D-jet that is close to $\mathcal{J}_3$.

After this pleasant digression, let us return to the subject of map concatenation. In analogy to the discussion of Section 8.6, the next topic to be treated is the case where $\mathcal{M}_1$ is in Lie form and $\mathcal{M}_2$ is in Taylor form. See Figure 8.6.2. In this case the definition (8.6.11) must be extended to become

$$T_a^D(\bar{z}) = \sum_{m'=0}^{D} g_a^2(m';\bar{z}) \tag{9.6.23}$$

to include the possibility that $\mathcal{M}_2$ may have a translation part. The remaining relations (8.6.10) and (8.6.12) through (8.6.14) continue to hold. In particular, we still have the result

$$\bar{\bar{z}}_a(z) = \mathcal{M}_1 T_a^D(z). \tag{9.6.24}$$

Again, there are three common ways that $\mathcal{M}_1$ may be specified in Lie form. First suppose, as before, that $\mathcal{M}_1$ is given in terms of a single exponent,

$$\mathcal{M}_1 = \exp(: h :), \tag{9.6.25}$$

where h now has a homogeneous polynomial expansion of the form

$$h = h_1 + h_2 + \cdots h_{D+1}. \tag{9.6.26}$$

We still have the relation (8.6.17) and the result

$$g_a^3(m;z) = P_m \sum_{m'=0}^{D} \sum_{\ell=0}^{\infty} (1/\ell!) :h:^\ell g_a^2(m';z). \tag{9.6.27}$$

As before, there are caveats about the rate at which the sum over ℓ converges. Moreover, as illustrated for a special example in Section 10.5, the sum over ℓ may also fail to converge. As in Section 10.5, this possible divergence is not due to any defect in the method of direct Taylor summation, but rather indicates that $\mathcal{M}_1$ may fail to exist, and shows that Hamiltonians for which $h_1 \neq 0$ must be treated with care.

Next we suppose, as a second possibility, that $\mathcal{M}_1$ is given in the factored product form

$$\mathcal{M}_1 = \exp(:f_1:)\mathcal{R}_f \exp(:f_3:) \cdots \exp(:f_{D+1}:). \tag{9.6.28}$$

Handling this possibility is straight forward. Suppose that $\exp(:f_1:)$ has the effect

$$\exp(:f_1:)z_a = z_a + k_a. \tag{9.6.29}$$

Then the results (8.6.20) through (8.6.25) continue to hold except that the sums over m and m' begin at 0 instead of 1, and (8.6.27) is modified to become

$$g_a^3(m;z) = P_m \sum_{m'=0}^{D} \tilde{g}_a[m'; R^f(z+k)]. \tag{9.6.30}$$

The third possibility is that $\mathcal{M}_1$ arises as a result of some symplectic integration approximation and is therefore given as a product of Lie transformations of the form

$$\mathcal{M}_1 = \exp[(w_1 h:A:) \exp(w_2 h:B:) \cdots \exp(w_m h:A:). \tag{9.6.31}$$

As before, B typically has a homogeneous polynomial expansion consisting of terms of degree three and higher. However, if $\mathcal{M}_1$ has a translation part, A will contain terms of degree one as well as a second-degree terms. In this case we make use of (2.4) or (2.62) to factorize the terms of the form $\exp(w_j h:A:)$. With this accomplished, we may proceed as before using the tools already developed.

At this point we should remark that if $\mathcal{M}_1$ has a translation part, then the D-jet produced for $\mathcal{M}_3$ in each of the three possibilities just described will again not be a symplectic D-jet, and for the same reason as before. Also, nearby symplectic D-jets can again be constructed. For example, the procedure based on the methods of the factorization theorem will work as before.

The last topic to be discussed in this section is the inversion of maps in Taylor form including the possibility of translations. When translations are included, (8.6.40) takes the form

$$\overline{z}_{b'} = k_{b'} + \sum_b R_{b'b} z_b + N_{b'}(z). \tag{9.6.32}$$

This equation can be partially solved to give the result

$$z_a = [R^{-1}(\overline{z} - k)]_a + \tilde{N}_a(z) \tag{9.6.33}$$

where $\tilde{N}_a$ is again given by (8.6.41), and therefore contains terms only of degree 2 and higher. Now form the recursion relation

$$z_a^{(m+1)}(\overline{z}) = [R^{-1}(\overline{z} - k)]_a + \mathcal{T}^{m+1}\tilde{N}_a[z^{(m)}(\overline{z})] \tag{9.6.34}$$

with the starting relation

$$z_a^{(1)}(\overline{z}) = [R^{-1}(\overline{z} - k)]_a. \tag{9.6.35}$$

Here the translation quantities k_b are to be treated as small, and a monomial in *all* the variables z_a and k_b is regarded as having degree d if the sum of *all* the exponents in the monomial adds up to d. Correspondingly, the operator $\mathcal{T}^d$ in (6.34) is now defined in terms of this total degree. Application of the recursion relation (6.34) D times produces a D-jet representation for the map $\mathcal{M}_1^{-1}$.

As the reader should expect by now, if $\mathcal{M}_1$ has a translation part (as we have assumed), then the D-jet for $\mathcal{M}_1^{-1}$ obtained in this way will generally not be symplectic. But again, nearby symplectic D-jets can be constructed from this D-jet.

Finally, we remark that the concatenation and inversion methods of this section can, if desired, be employed in the formulas of Section 9.4 to compute reverse and mixed factorizations.

Exercises

9.6.1.

9.7 The Lie Algebra of the Group of all Symplectic Maps Is Simple

Section 8.9 described what it means for a Lie algebra to be simple. In this section we will show that $ispm(2n, \mathbb{R})$, the Lie algebra of the group of all symplectic maps, is simple.

Bibliography

[1] L.M. Healy and A.J. Dragt, Concatentation of Lie Algebraic Maps, in *Lie Methods in Optics II*, K.B. Wolf, Ed., Lecture Notes in Physics **352**, Springer-Verlag (1989).

Chapter 10

Computation of Transfer Maps

Much of the material in the previous chapters dealt with the general problem of representing and manipulating symplectic maps. This chapter, along with some that follow, deals with the *computation* of transfer maps. For the most part we will deal with the symplectic case, but there are ready extensions to the general case that can be found by replacing Hamiltonian vector fields by general vector fields.

10.1 Equation of Motion

10.1.1 Background and Derivation

Let $H(z,t)$ be a general, possibly time-dependent, Hamiltonian. We know from Theorem 6.4.1 that following the flow specified by H produces a symplectic transfer map $\mathcal{M}(t)$. Let z^i denote a general initial condition. Then we have the relations

$$z(t) = \mathcal{M}(t)z^i, \tag{10.1.1}$$

$$\mathcal{M}(t^i) = \mathcal{I}. \tag{10.1.2}$$

Our goal is to find an equation of motion for $\mathcal{M}$.

Suppose $g(z)$ is any function of the phase-space variables z (but not explicitly of the time t). By (8.3.52) we have the relation

$$g(z) = g(\mathcal{M}z^i) = \mathcal{M}g(z^i). \tag{10.1.3}$$

Now differentiate both sides of (1.3) along the flow specified by H. We find the result

$$\dot{g}(z) = \dot{\mathcal{M}}g(z^i). \tag{10.1.4}$$

But from (1.7.4) we also have the relation

$$\dot{g}(z) = [g(z), H(z,t)]. \tag{10.1.5}$$

With the aid of (1.1), (8.3.52), and (8.3.53) this relation can be rewritten in the form

$$\begin{aligned}
\dot{g}(z) &= [g(z), H(z,t)] = [g(\mathcal{M}z^i), H(\mathcal{M}z^i,t)] \\
&= [\mathcal{M}g(z^i), \mathcal{M}H(z^i,t)] = \mathcal{M}[g(z^i), H(z^i,t)] \\
&= \mathcal{M}[-H(z^i,t), g(z^i)] = \mathcal{M} : -H(z^i,t) : g(z^i).
\end{aligned} \tag{10.1.6}$$

Now compare (1.4) and (1.6). Doing so gives the result

$$\dot{\mathcal{M}} g(z^i) = \mathcal{M} : -H(z^i, t) : g(z^i). \tag{10.1.7}$$

However, g is an arbitrary function. We conclude that (1.7) is equivalent to the *operator* equation of motion

$$\dot{\mathcal{M}} = \mathcal{M} : -H : . \tag{10.1.8}$$

We note that this result agrees with the result (7.4.9) that was obtained earlier for the special case of autonomous Hamiltonians.

10.1.2 Perturbation/Splitting Theory and Reverse Factorization

In some cases the Hamiltonian can be split into the sum of two terms so that it can be written in the form

$$H(z, t) = H_0(z, t) + H_1(z, t). \tag{10.1.9}$$

Often the motion governed by H_0 can be determined and the effect of H_1 may be viewed as a perturbation. Let $\mathcal{M}_0$ be the map produced by H_0. That is, $\mathcal{M}_0$ satisfies the equation of motion

$$\dot{\mathcal{M}}_0 = \mathcal{M}_0 : -H_0 : \tag{10.1.10}$$

with the initial condition

$$\mathcal{M}_0(t^i) = \mathcal{I}. \tag{10.1.11}$$

For the $\mathcal{M}$ produced by H let us write the representation

$$\mathcal{M} = \mathcal{M}_1 \mathcal{M}_0 \tag{10.1.12}$$

where the map $\mathcal{M}_1$ remains to be determined. We will call the Ansatz (1.12) a *reverse* factorization.

What is the equation of motion for $\mathcal{M}_1$? From (1.8), (1.9), and (1.12) we find the result

$$\begin{aligned} \dot{\mathcal{M}} &= \dot{\mathcal{M}}_1 \mathcal{M}_0 + \mathcal{M}_1 \dot{\mathcal{M}}_0 = \mathcal{M} : -H := \mathcal{M}_1 \mathcal{M}_0 : -H : \\ &= \mathcal{M}_1 \mathcal{M}_0 : -H_0 : + \, \mathcal{M}_1 \mathcal{M}_0 : -H_1 : . \end{aligned} \tag{10.1.13}$$

Use of (1.10) gives the relation

$$\mathcal{M}_1 \dot{\mathcal{M}}_0 = \mathcal{M}_1 \mathcal{M}_0 : -H_0 :, \tag{10.1.14}$$

and consequently (1.13) can be reduced to the relation

$$\dot{\mathcal{M}}_1 \mathcal{M}_0 = \mathcal{M}_1 \mathcal{M}_0 : -H_1 : \tag{10.1.15}$$

from which it follows that

$$\dot{\mathcal{M}}_1 = \mathcal{M}_1 \mathcal{M}_0 : -H_1 : \mathcal{M}_0^{-1}. \tag{10.1.16}$$

Given H_1 and $\mathcal{M}_0$, let us define an *interaction* Hamiltonian H_1^{int} by the rule

$$H_1^{\text{int}}(z^i, t) = H_1(\mathcal{M}_0 z^i, t). \tag{10.1.17}$$

[We note that, because $\mathcal{M}_0(t)$ is time dependent, in general $H_1^{\rm int}$ will depend on time even if H_1 happens to be time independent.] Then, as a consequence of (8.2.25), we have the result

$$\mathcal{M}_0 : -H_1 : \mathcal{M}_0^{-1} =: -H_1^{\rm int} : . \tag{10.1.18}$$

Upon combining (1.16) and (1.18) we find the equation of motion

$$\dot{\mathcal{M}}_1 = \mathcal{M}_1 : -H_1^{\rm int} : . \tag{10.1.19}$$

Finally, we observe from (1.2) and (1.11) that $\mathcal{M}_1$ also has the initial condition

$$\mathcal{M}_1(t^i) = \mathcal{I}. \tag{10.1.20}$$

10.1.3 Perturbation/Splitting Theory and Forward Factorization

For the $\mathcal{M}$ produced by H let us write, instead of (1.12), the representation

$$\mathcal{M} = \mathcal{M}_0\mathcal{N}_1 \tag{10.1.21}$$

where the map $\mathcal{N}_1$ remains to be determined. We will call the Ansatz (1.21) a *forward* factorization.

Note that the $\mathcal{N}_1$ in (1.21) and the $\mathcal{M}_1$ in (1.12) are generally different. Indeed, we may rewrite (1.12) in the form

$$\mathcal{M} = \mathcal{M}_1\mathcal{M}_0 = \mathcal{M}_0\mathcal{M}_0^{-1}\mathcal{M}_1\mathcal{M}_0 \tag{10.1.22}$$

from which it follows that

$$\mathcal{N}_1 = \mathcal{M}_0^{-1}\mathcal{M}_1\mathcal{M}_0. \tag{10.1.23}$$

We see that if a reverse factorization has been found so that both $\mathcal{M}_0$ and $\mathcal{M}_1$ are known, then (1.21) and (1.23) provide the associated forward factorization.

10.2 Series (Dyson) Solution

Readers familiar with Quantum Mechanics will recognize a similarity between equations (1.8), (1.17), and (1.19) and analogous quantum mechanical results. This is to be expected because Quantum Mechanics and Classical Mechanics have closely related Lie algebraic structures. Because of this similiarity, many mathematical tools originally developed for Quantum Mechanics can also be applied in Classical Mechanics. Indeed, these tools could have (and, in retrospect, should have) been developed first in the context of Classical Mechanics.

One such tool is *Neumann* iteration. Suppose both sides of (1.8) are integrated with respect to time from the initial time t^i to the variable time t. So doing, and making use of (1.2), gives the integral equation

$$\mathcal{M}(t) = \mathcal{I} + \int_{t^i}^{t} dt'\mathcal{M}(t') : -H(t') : . \tag{10.2.1}$$

(Here, for notational simplicity, we have suppressed the fact that H also depends on z^i.) Now iterate (2.1) by substituting the right side back into the integral. If this is done once, we obtain the result

$$\mathcal{M}(t) = \mathcal{I} + \int_{t^i}^{t} dt' : -H(t') : + \int_{t^i}^{t} dt' \int_{t^i}^{t'} dt'' \mathcal{M}(t'') : -H(t'') :: -H(t') : . \tag{10.2.2}$$

Evidently, repeated iteration gives the result

$$\mathcal{M}(t) = \mathcal{I} + \int_{t^i}^{t} dt' : -H(t') : + \int_{t^i}^{t} dt' \int_{t^i}^{t'} dt'' : -H(t'') :: -H(t') : + \cdots . \tag{10.2.3}$$

Note that (2.3) is in effect an expansion in powers of H, and that the factors in the integrands occur in *chronological* order – with earlier times preceding later times. We conclude that $\mathcal{M}$ can be expressed as an infinite sum of multiple *time-ordered* integrals over the Lie operators $: -H(t) :$. In the context of Quantum Mechanics, the analog of the series (2.3) is the *Dyson* series. Also note that Neumann iteration is the map counterpart of Picard iteration. Compare (1.3.8) and (1.3.9) with (2.1) through (2.3).

The series (2.3) bears a certain resemblance to the exponential series. Consider m-dimensional Euclidean space. It is easily verified that the volume of the region $t^i \le t_m \le t_{m-1} \le \cdots \le t_2 \le t_1 \le t$ is related to the volume of the region $[t^i, t] \times [t^i, t] \times [t^i, t] \cdots$ (m factors) by the proportionality constant $(m!)$. Indeed, we have the relation

$$\begin{aligned} \int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2 \int_{t^i}^{t_2} dt_3 \cdots \int_{t^i}^{t_{m-1}} dt_m &= (1/m!) \int_{t^i}^{t} dt_1 \int_{t^i}^{t} dt_2 \cdots \int_{t^i}^{t} dt_m \\ &= (1/m!) \left[\int_{t^i}^{t} dt' \right]^m . \end{aligned} \tag{10.2.4}$$

Consequently, we may write the identity

$$\begin{aligned} &\int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2 \cdots \int_{t^i}^{t_{m-1}} dt_m : -H(t_m) :: -H(t_{m-1}) : \cdots : -H(t_1) : \\ &= (1/m!) \int_{t^i}^{t} dt_1 \int_{t^i}^{t} dt_2 \cdots \int_{t^i}^{t} dt_m T : -H(t_m) :: -H(t_{m-1}) : \cdots : -H(t_1) : \\ &= (1/m!) T \left[\int_{t^i}^{t} dt' : -H(t') : \right]^m . \end{aligned} \tag{10.2.5}$$

Here the *time-ordering* symbol T indicates that the factors in the operator product $: -H(t_m) : \cdots : -H(t_1) :$ are to be rearranged so that operators with earlier times precede those with later times. Finally, with the aid of (2.5), we may write the series (2.3) in the form

$$\mathcal{M}(t) = \mathcal{I} + T \sum_{m=1}^{\infty} (1/m!) \left[\int_{t^i}^{t} dt' : -H(t') : \right]^m = T \exp \left[\int_{t^i}^{t} dt' : -H(t') : \right] . \tag{10.2.6}$$

The right side of (2.6) is often called a *time-ordered exponential.* However, as neat as this expression may appear to be, in reality it is simply the series (2.3).

We close this section by noting that our discussion of the series solution to (1.8) applies equally well to (1.19). Consequently, $\mathcal{M}_1$ has the series solution

$$\mathcal{M}_1(t) = T\exp\left[\int_{t^i}^{t} dt' : -H_1^{\rm int}(t') :\right]. \tag{10.2.7}$$

This result for $\mathcal{M}_1$ may be viewed as an expansion in powers of $H_1^{\rm int}$.

Exercises

10.2.1. Verify (2.4) by performing the indicated integrations on each side.

10.2.2. Suppose $F(z,t)$ and $G(z,t)$ are two Hamiltonians that are in *involution.* That is, we assume

$$[F(z,t), G(z,t')] = 0 \text{ for all } t, t'. \tag{10.2.8}$$

Correspondingly, there will be the Lie operator commutation relation

$$\{: F(z,t) :, : G(z,t') :\} = 0 \text{ for all } t, t'. \tag{10.2.9}$$

Let $\mathcal{M}_F(t^{\rm in}, t^{\rm fin})$ and $\mathcal{M}_G(t^{\rm in}, t^{\rm fin})$ be the maps generated by F and G, respectively. Define a sum Hamiltonian $H(z,t)$ by writing

$$H(z,t) = F(z,t) + G(z,t), \tag{10.2.10}$$

and let $\mathcal{M}_H(t^{\rm in}, t^{\rm fin})$ be the map generated by H. Your task is to show that

$$\mathcal{M}_H(t^{\rm in}, t^{\rm fin}) = \mathcal{M}_F(t^{\rm in}, t^{\rm fin})\mathcal{M}_G(t^{\rm in}, t^{\rm fin}) = \mathcal{M}_G(t^{\rm in}, t^{\rm fin})\mathcal{M}_F(t^{\rm in}, t^{\rm fin}). \tag{10.2.11}$$

Begin by making the Ansatz

$$\mathcal{M}_H(t^{\rm in}, t) = \mathcal{M}_?(t^{\rm in}, t)\mathcal{M}_G(t^{\rm in}, t) \tag{10.2.12}$$

where the map $\mathcal{M}_?(t^{\rm in}, t)$ remains to be determined. Verify that taking the time derivative of both sides of (2.12) produces the result

$$\begin{aligned} \dot{\mathcal{M}}_H &= \dot{\mathcal{M}}_?\mathcal{M}_G + \mathcal{M}_?\dot{\mathcal{M}}_G = \mathcal{M}_H : -H := \mathcal{M}_?\mathcal{M}_G : -H : \\ &= \mathcal{M}_?\mathcal{M}_G : -F : + \mathcal{M}_?\mathcal{M}_G : -G : . \end{aligned} \tag{10.2.13}$$

Next find the equation of motion for $\mathcal{M}_?$. To do so, verify that

$$\mathcal{M}_?\dot{\mathcal{M}}_G = \mathcal{M}_?\mathcal{M}_G : -G :, \tag{10.2.14}$$

and consequently (2.13) can be reduced to the relation

$$\dot{\mathcal{M}}_?\mathcal{M}_G = \mathcal{M}_?\mathcal{M}_G : -F : . \tag{10.2.15}$$

At this point imagine that the series solution (2.3) is used to represent $\mathcal{M}_G$. Verify that doing so gives the result

$$\mathcal{M}_G(t^{\rm in},t) = \mathcal{I} + \int_{t^{\rm in}}^{t} dt' : -G(t') : + \int_{t^{\rm in}}^{t} dt' \int_{t^{\rm in}}^{t'} dt'' : -G(t'') :: -G(t') : + \cdots . \quad (10.2.16)$$

Employ the assumption (2.9) and the representation (2.16) to conclude that

$$\mathcal{M}_G : -F :=: -F : \mathcal{M}_G, \quad (10.2.17)$$

from which it follows that (2.15) can be rewritten in the form

$$\dot{\mathcal{M}}_? \mathcal{M}_G = \mathcal{M}_? : -F : \mathcal{M}_G, \quad (10.2.18)$$

and consequently

$$\dot{\mathcal{M}}_? = \mathcal{M}_? : -F : . \quad (10.2.19)$$

Finally, observe from (2.12) that $\mathcal{M}_?$ has the initial condition

$$\mathcal{M}_?(t^{\rm in}, t^{\rm in}) = \mathcal{I}. \quad (10.2.20)$$

It follows from (2.19) and (2.20) that

$$\mathcal{M}_?(t^{\rm in}, t^{\rm fin}) = \mathcal{M}_F(t^{\rm in}, t^{\rm fin}), \quad (10.2.21)$$

and insertion of (2.21) into the Ansatz (2.12) proves the first part of (2.11).

In an analogous way, prove the second part of (2.11) by making the Ansatz

$$\mathcal{M}_H(t^{\rm in}, t) = \mathcal{M}_?(t^{\rm in}, t)\mathcal{M}_F(t^{\rm in}, t). \quad (10.2.22)$$

10.3 Exponential (Magnus) Solution

The series solutions (2.6) and (2.7) for the transfer map, or equivalently (2.3), have the defect that they are not manifestly symplectic. Indeed, if the series are truncated, the resulting maps are generally not symplectic. Moreover, maps in series form are somewhat difficult to concatenate. For these reasons, it is also useful to have solutions in exponential form. Possibilities include the single exponential form and various factored product forms. Here we consider the single exponential form.

Let us seek a solution to the equation of motion (1.8), or (1.19), of the form

$$\mathcal{M}(t) = \exp(: F(z^i, t) :). \quad (10.3.1)$$

We know that in general there is no such solution. Indeed, even in the simplest linear case of $Sp(2)$, maps cannot generally be written in single exponent form. See Section 8.7. Nevertheless we will pursue the assumption (3.1) to see where it leads. We will find an expansion for F in terms of powers of H, and the convergence of this expansion will determine the validity of the assumption.

Let us differentiate both sides of (3.1). Doing so gives the relation

$$\dot{\mathcal{M}} = \exp(:F:)\mathrm{iex}(-\#F\#) : \dot{F} := \mathcal{M}\mathrm{iex}(-\#F\#) : \dot{F} : . \tag{10.3.2}$$

Here we have used results from Appendix C. Now insert the equation of motion (1.8) into (3.2) to get the relation

$$: -H := \mathrm{iex}(-\#F\#) : \dot{F} : . \tag{10.3.3}$$

This relation can be solved to produce an equation of motion for the Lie operator $:F:$,

$$: \dot{F} := [\mathrm{iex}(-\#F\#)]^{-1} : -H : . \tag{10.3.4}$$

According to Appendix C, the operator $[\mathrm{iex}(-\#F\#)]^{-1}$ has an expansion in powers of $\#F\#$ of the form

$$[\mathrm{iex}(-\#F\#)]^{-1} = \sum_{m=0}^{\infty} b_m(\#F\#)^m. \tag{10.3.5}$$

Consequently, (3.4) can be rewritten in the form

$$: \dot{F} := \sum_{m=0}^{\infty} b_m(\#F\#)^m : -H : = : [\sum_{m=0}^{\infty} b_m : F :^m (-H)] : . \tag{10.3.6}$$

Here we have also used (8.2.2). Now remove the outside colons from both sides of (3.6). So doing gives an equation of motion for F,

$$\dot{F} = \sum_{m=0}^{\infty} b_m : F :^m (-H) = [\mathrm{iex}(- : F :)]^{-1}(-H), \tag{10.3.7}$$

which could have been deduced directly by "decolonizing" (3.4). This equation is to be solved with the initial condition

$$F(t^i) = 0, \tag{10.3.8}$$

which follows from (1.2).

Let us try to solve (3.7) by perturbation theory. Replace H by ϵH, and assume F has an expansion of the form

$$F = \sum_{n=1}^{\infty} \epsilon^n F_n. \tag{10.3.9}$$

(This procedure is equivalent to the introduction of a grading, and the expansion obtained is often called the *Magnus* expansion. See Section 8.9.) Put the expansion (3.9) into (3.7) to get the result

$$\sum_{n=1}^{\infty} \epsilon^n \dot{F}_n = \sum_{m=0}^{\infty} b_m(: \sum_{n=1}^{\infty} \epsilon^n F_n :)^m(-\epsilon H). \tag{10.3.10}$$

Now equate powers of ϵ. So doing gives the results

$$\dot{F}_1 = -H, \tag{10.3.11}$$

$$\dot{F}_2 = (1/2) : F_1 : (-H), \tag{10.3.12}$$

$$\dot{F}_3 = (1/12) : F_1 :^2 (-H) + (1/2) : F_2 : (-H), \tag{10.3.13}$$

$$\dot{F}_4 = (1/2) : F_3 : (-H) + (1/12)(: F_1 :: F_2 : + : F_2 :: F_1 :)(-H), \tag{10.3.14}$$

$$\dot{F}_n = \text{ something involving } H \text{ and the } : F_m : \text{ with } m < n. \tag{10.3.15}$$

Here we have used the values for the coefficients b_m given in Appendix C. The equations (3.11) through (3.15) are to be solved with the initial conditions

$$F_n(t^i) = 0. \tag{10.3.16}$$

The equations for the $\dot{F}_n$ can be integrated numerically or solved by quadrature. Evidently (3.11) with the initial condition (3.16) has the solution

$$F_1(t) = \int_{t^i}^{t} dt_1[-H(t_1)]. \tag{10.3.17}$$

Now substitute (3.17) into (3.12) to get the result

$$\begin{aligned} \dot{F}_2(t_1) &= (1/2)\int_{t^i}^{t_1} dt_2 : -H(t_2) : [-H(t_1)] \\ &= (1/2)\int_{t^i}^{t_1} dt_2[-H(t_2), -H(t_1)]. \end{aligned} \tag{10.3.18}$$

This equation can be integrated, again with the initial condition (3.16), to give the result

$$F_2(t) = (1/2)\int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2[-H(t_2), -H(t_1)]. \tag{10.3.19}$$

Let us introduce the short-hand notation $-j$ for the quantity $-H(t_j)$. With this notation, (3.19) can be written in the more compact form

$$F_2(t) = (1/2)\int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2[-2, -1]. \tag{10.3.20}$$

Similarly, again using this notation, we find the result

$$\begin{aligned} &F_3(t) = (1/6)\int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2 \int_{t^i}^{t_2} dt_3 \times \\ &\{2[-3,[-2,-1]] - [-2,[-3,-1]]\} \\ &= (1/6)\int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2 \int_{t^i}^{t_2} dt_3 \times \\ &\{[-1,[-2,-3]] + [-3,[-2,-1]]\}. \end{aligned} \tag{10.3.21}$$

Here $2[-3,[-2,-1]]$ is short hand for $2[-H(t_3),[-H(t_2),-H(t_1)]]$, etc. And for F_4 we find the result

$$\begin{aligned} F_4(t) = \quad & (1/12)\int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2 \int_{t^i}^{t_2} dt_3 \int_{t^i}^{t_3} dt_4 \times \\ & \{+3[-4,[-3,[-2,-1]]] - [-4,[-2,[-3,-1]]] \\ & -[-3,[-4,[-2,-1]]] - [-3,[-2,[-4,-1]]] \\ & -[-2,[-4,[-3,-1]]] + [-2,[-3,[-4,-1]]]\}. \end{aligned} \tag{10.3.22}$$

See Exercise 3.3.

At this point at least three comments are in order. First, we recognize from the structure of the equations (3.11) through (3.15) that the quantity ϵ serves only as a "counting" parameter that counts powers of H, and therefore we may now set $\epsilon = 1$. The net result is an expansion of F in powers of H. Next we see from (3.15) that the F_n can be determined successively, and consequently a solution by numerical methods or by quadrature is always possible. Third, we see that all the quantities on the right side of the equations (3.11) through (3.15) lie in the Lie algebra generated by the $H(z^i, t)$ for different values of t. (That is, all the quantities consist of H and Poisson brackets involving only factors of H.) In particular, if all the $H(z^i, t)$ are in *involution*,

$$[H(z^i, t), H(z^i, t')] = 0, \tag{10.3.23}$$

then we have the results

$$F_n = 0 \text{ for } n > 1, \tag{10.3.24}$$

and

$$\mathcal{M}(t) = \exp\left[\int_{t^i}^{t} dt' : -H(t') :\right]. \tag{10.3.25}$$

Note that (3.25) is consistent with (2.6) because if (3.23) holds, then time ordering makes no difference. Finally, if the Hamiltonian is autonomous, then the integral in (3.25) can be done to give the result (7.4.7).

We close this section with the reminder that the discussion we have just given for the solution of (1.8) for $\mathcal{M}$ applies equally well to the solution of (1.19) for $\mathcal{M}_1$. In the case of $\mathcal{M}_1$, the associated exponential quantity F can be developed as an expansion in powers of H_1^{int}. And, since H_1^{int} may be small, it could be the case that $\mathcal{M}_1$ can be written in single exponent form.

Exercises

10.3.1. Verify (3.11) through (3.15).

10.3.2. Verify (3.17) through (3.20).

10.3.3. Verify (3.21) through (3.22). Hint: To do so, use the identity (5.3.14) and the Jacobi identity; and verify and employ integral identities of the form

$$\int_{t^i}^{t} dt_1 \int_{t^i}^{t} dt_2 ** = \int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2 ** + \int_{t^i}^{t} dt_2 \int_{t^i}^{t_2} dt_1 **, \tag{10.3.26}$$

where $**$ denotes some common integrand.

10.4 Factored Product Solution: Powers of H Expansion

As before, let us replace H by ϵH so that the the equation of motion (1.8) becomes

$$\dot{\mathcal{M}} = \mathcal{M} : -\epsilon H : . \tag{10.4.1}$$

Suppose also that we factor $\mathcal{M}$ in the form

$$\mathcal{M} = \mathcal{M}_1\mathcal{M}_2\mathcal{M}_3\mathcal{M}_4\cdots \tag{10.4.2}$$

where

$$\mathcal{M}_m = \exp(:\epsilon^m G_m:) \tag{10.4.3}$$

and the functions G_m are to be determined. [Note that (4.2) is a forward factorization.] Next differentiate the Ansatz (4.2) to find the result

$$\begin{aligned}\dot{\mathcal{M}} &= \dot{\mathcal{M}}_1\mathcal{M}_2\mathcal{M}_3\mathcal{M}_4\cdots + \mathcal{M}_1\dot{\mathcal{M}}_2\mathcal{M}_3\mathcal{M}_4\cdots \\ &\quad +\mathcal{M}_1\mathcal{M}_2\dot{\mathcal{M}}_3\mathcal{M}_4\cdots + \mathcal{M}_1\mathcal{M}_2\mathcal{M}_3\dot{\mathcal{M}}_4\cdots \\ &\quad +\cdots,\end{aligned} \tag{10.4.4}$$

from which it follows using (4.1) and (4.2) that

$$\begin{aligned}\mathcal{M}^{-1}\dot{\mathcal{M}} &= \cdots\mathcal{M}_4^{-1}\mathcal{M}_3^{-1}\mathcal{M}_2^{-1}\mathcal{M}_1^{-1}\dot{\mathcal{M}}_1\mathcal{M}_2\mathcal{M}_3\mathcal{M}_4\cdots \\ &\quad +\cdots\mathcal{M}_4^{-1}\mathcal{M}_3^{-1}\mathcal{M}_2^{-1}\dot{\mathcal{M}}_2\mathcal{M}_3\mathcal{M}_4\cdots \\ &\quad +\cdots\mathcal{M}_4^{-1}\mathcal{M}_3^{-1}\dot{\mathcal{M}}_3\mathcal{M}_4\cdots \\ &\quad +\cdots\mathcal{M}_4^{-1}\dot{\mathcal{M}}_4\cdots \\ &\quad +\cdots = :-\epsilon H:.\end{aligned} \tag{10.4.5}$$

The various terms in (4.5) can be simplified by the use of adjoint operators. For example, we have the result

$$\begin{aligned}&\cdots\mathcal{M}_4^{-1}\mathcal{M}_3^{-1}\mathcal{M}_2^{-1}\mathcal{M}_1^{-1}\dot{\mathcal{M}}_1\mathcal{M}_2\mathcal{M}_3\mathcal{M}_4\cdots = \\ &\quad\cdots\exp(-\#\epsilon^4G_4\#)\exp(-\#\epsilon^3G_3\#)\exp(-\#\epsilon^2G_2\#)\mathcal{M}_1^{-1}\dot{\mathcal{M}}_1.\end{aligned} \tag{10.4.6}$$

Also, there are relations of the form

$$\mathcal{M}_m^{-1}\dot{\mathcal{M}}_m = \text{ iex }(-\#\epsilon^m G_m\#):\epsilon^m\dot{G}_m:. \tag{10.4.7}$$

Upon using (4.7) and relations of the form (4.6) we find that (4.5) can be rewritten in the form

$$\begin{aligned}&\cdots\exp(-\#\epsilon^4G_4\#)\exp(-\#\epsilon^3G_3\#)\exp(-\#\epsilon^2G_2\#)\text{ iex}(-\#\epsilon G_1\#):\epsilon\dot{G}_1: \\ &+\cdots\exp(-\#\epsilon^4G_4\#)\exp(-\#\epsilon^3G_3\#)\text{ iex }(-\#\epsilon^2G_2\#):\epsilon^2\dot{G}_2: \\ &+\cdots\exp(-\#\epsilon^4G_4\#)\text{ iex }(-\#\epsilon^3G_3\#):\epsilon^3\dot{G}_3: \\ &+\cdots\text{ iex}(-\#\epsilon^4G_4\#):\epsilon^4\dot{G}_4:+\cdots = :-\epsilon H:.\end{aligned} \tag{10.4.8}$$

The colons can now be removed from both sides of (4.8) to give the equivalent result

$$\begin{aligned}&\cdots\exp(-:\epsilon^4G_4:)\exp(-:\epsilon^3G_3:)\exp(-:\epsilon^2G_2:)\text{ iex}(-:\epsilon G_1:)\epsilon\dot{G}_1 \\ &+\cdots\exp(-:\epsilon^4G_4:)\exp(-:\epsilon^3G_3:)\text{ iex }(-:\epsilon^2G_2:)\epsilon^2\dot{G}_2 \\ &+\cdots\exp(-:\epsilon^4G_4:)\text{ iex }(-:\epsilon^3G_3:)\epsilon^3\dot{G}_3 \\ &+\cdots\text{ iex}(-:\epsilon^4G_4:)\epsilon^4\dot{G}_4+\cdots = -\epsilon H.\end{aligned} \tag{10.4.9}$$

Recall that the integrated exponential function has the expansion.

$$\begin{aligned} \text{iex}(w) &= \int_0^1 d\tau \exp(\tau w) = (e^w - 1)/w = \sum_{m=0}^{\infty} w^m/(m+1)! \\ &= 1 + w/2 + w^2/3 + w^3/4 + w^4/5 + \cdots . \end{aligned} \tag{10.4.10}$$

Using this expansion we may expand each of the lines in (4.9) as a Taylor series in ϵ. We find, for example, the results

$$\begin{aligned} &\cdots \exp(- : \epsilon^4 G_4 :) \exp(- : \epsilon^3 G_3 :) \exp(- : \epsilon^2 G_2 :) \text{ iex}(- : \epsilon G_1 :) \epsilon \dot{G}_1 = \\ &\epsilon \dot{G}_1 + \epsilon^2[-(1/2) : G_1 : \dot{G}_1] + \epsilon^3[*] + \epsilon^4[*] + \cdots , \end{aligned} \tag{10.4.11}$$

$$\begin{aligned} &\cdots \exp(- : \epsilon^4 G_4 :) \exp(- : \epsilon^3 G_3 :) \text{ iex } (- : \epsilon^2 G_2 :) \epsilon^2 \dot{G}_2 = \\ &\epsilon^2 \dot{G}_2 + \epsilon^3[*] + \epsilon^4[*] + \cdots , \end{aligned} \tag{10.4.12}$$

$$\begin{aligned} &\cdots \exp(- : \epsilon^4 G_4 :) \text{ iex } (- : \epsilon^3 G_3 :) \epsilon^3 \dot{G}_3 = \\ &\epsilon^3 \dot{G}_3 + \epsilon^4[*] + \cdots , \end{aligned} \tag{10.4.13}$$

$$\begin{aligned} &\cdots \text{ iex}(- : \epsilon^4 G_4 :) \epsilon^4 \dot{G}_4 = \\ &\epsilon^4 \dot{G}_4 + \cdots . \end{aligned} \tag{10.4.14}$$

Next equate powers of ϵ on both sides of (4.9). So doing gives, for example through powers of ϵ^4, the results

$$\epsilon \dot{G}_1 = -\epsilon H, \tag{10.4.15}$$

$$\epsilon^2[\dot{G}_2 - (1/2) : G_1 : \dot{G}_1] = 0, \tag{10.4.16}$$

$$\epsilon^3[\dot{G}_3 +] = 0, \tag{10.4.17}$$

$$\epsilon^4[\dot{G}_4 +] = 0. \tag{10.4.18}$$

We conclude that the G_n obey the equations of motion

$$\dot{G}_1 = -H, \tag{10.4.19}$$

$$\dot{G}_2 = (1/2) : G_1 : \dot{G}_1 = (1/2) : G_1 : (-H), \tag{10.4.20}$$

$$\dot{G}_3 = ?(1/12) : G_1 :^2 (-H) + (1/2) : G_2 : (-H), \tag{10.4.21}$$

$$\dot{G}_4 = ?(1/2) : G_3 : (-H) + (1/12)(: G_1 :: G_2 : + : G_2 :: G_1 :)(-H), \cdots \tag{10.4.22}$$

$$\dot{G}_n = \text{ something involving } H \text{ and the } : G_m : \text{ with } m < n. \tag{10.4.23}$$

These equations of motion are to be solved with the initial conditions

$$G_n(t^i) = 0. \tag{10.4.24}$$

The equations for the $\dot{G}_n$ can be integrated numerically or solved by quadrature. Evidently (4.19) with the initial condition (4.24) has the solution

$$G_1(t) = \int_{t^i}^{t} dt_1[-H(t_1)]. \tag{10.4.25}$$

Now substitute (4.25) into (4.20) to get the result

$$\begin{aligned}\dot{G}_2(t_1) &= (1/2)\int_{t^i}^{t_1} dt_2 : -H(t_2) : [-H(t_1)] \\ &= (1/2)\int_{t^i}^{t_1} dt_2[-H(t_2), -H(t_1)].\end{aligned} \tag{10.4.26}$$

This equation can be integrated, again with the initial condition (4.24), to give the result

$$G_2(t) = (1/2)\int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2[-H(t_2), -H(t_1)]. \tag{10.4.27}$$

Let us introduce the short-hand notation $-j$ for the quantity $-H(t_j)$. With this notation, (4.27) can be written in the more compact form

$$G_2(t) = (1/2)\int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2[-2, -1]. \tag{10.4.28}$$

Similarly, again using this notation, we find the result

$$\begin{aligned}&G_3(t) =?(1/6)\int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2 \int_{t^i}^{t_2} dt_3 \times \\ &\{2[-3, [-2, -1]] - [-2, [-3, -1]]\} \\ &= (1/6)\int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2 \int_{t^i}^{t_2} dt_3 \times \\ &\{[-1, [-2, -3]] + [-3, [-2, -1]]\}.\end{aligned} \tag{10.4.29}$$

Here $2[-3, [-2, -1]]$ is short hand for $2[-H(t_3), [-H(t_2), -H(t_1)]]$, etc. And for G_4 we find the result

$$\begin{aligned}G_4(t) =? \quad &(1/12)\int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2 \int_{t^i}^{t_2} dt_3 \int_{t^i}^{t_3} dt_4 \times \\ &\{+3[-4, [-3, [-2, -1]]] - [-4, [-2, [-3, -1]]] \\ &-[-3, [-4, [-2, -1]]] - [-3, [-2, [-4, -1]]] \\ &-[-2, [-4, [-3, -1]]] + [-2, [-3, [-4, -1]]]\}.\end{aligned} \tag{10.4.30}$$

See Exercise 4.1.

At this point at least three comments are in order. First, we recognize from the structure of the equations (4.19) through (4.23) that the quantity ϵ serves only as a "counting" parameter that counts powers of H, and therefore we may now set $\epsilon = 1$. Next we see from (4.23) that the G_n can be determined successively, and consequently a solution by numerical methods or by quadrature is always possible. Third, we see that all the quantities on the right side of the equations (4.19) through (4.23) lie in the Lie algebra generated by the $H(z^i,t)$ for different values of t. (That is, all the quantities consist of H and Poisson brackets involving only factors of H.) In particular, if all the $H(z^i,t)$ are in *involution*,

$$[H(z^i,t),H(z^i,t')]=0, \tag{10.4.31}$$

then we have the results

$$G_n = 0 \text{ for } n>1, \tag{10.4.32}$$

and we again find the result

$$\mathcal{M}(t)=\exp\left[\int_{t^i}^{t} dt' : -H(t') :\right]. \tag{10.4.33}$$

Finally, if the Hamiltonian is autonomous, then the integral in (4.33) can be done to give the result (7.4.7).

We close this section with the reminder that the discussion we have just given for the solution of (1.8) for $\mathcal{M}$ applies equally well to the solution of (1.19) for $\mathcal{M}_1$. In the case of $\mathcal{M}_1$, the associated quantities G_n can be developed as expansions in powers of H_1^{int}. Note that in this context we have put ourselves in a notationally awkward position: The $\mathcal{M}_1$ appearing in (1.12) is different from and should not be confused with the $\mathcal{M}_1$ appearing in (4.2).

Exercises

10.4.1. Suppose that $\mathcal{M}$ obeys the equation of motion (4.1) and that $\mathcal{M}$ is factored in the *reversed* product form

$$\mathcal{M}=\cdots\mathcal{M}_4\mathcal{M}_3\mathcal{M}_2\mathcal{M}_1 \tag{10.4.34}$$

where

$$\mathcal{M}_m=\exp(:\epsilon^m G_m^{\text{rev}}:) \tag{10.4.35}$$

and the functions G_m^{rev} are to be determined. Find equations of motion for these functions.

10.5 Factored Product Solution: Taylor Expansion about Design Orbit

10.5.1 Background

The discussion so far has been quite general in that no particular use has been made of Taylor expansions in the phase-space variables z. In this section we will explore the use of factored product representations such as those described in Sections 7.6 and 7.8.

Let $H(z,t)$ be a general, possibly time-dependent, Hamiltonian. Suppose that $z^d(t)$ is some given trajectory (which is assumed to be known and will be called the *design* trajectory), and that our task is to characterize all trajectories near z^d. Introduce $2n$ new *deviation* variables ζ by the rule

$$z=z^d+\zeta. \tag{10.5.1}$$

The transformation (5.1) is canonical. Consequently, the time evolution of the deviation variables ζ will also be described by some Hamiltonian. Call this Hamiltonian $H^{\text{new}}(\zeta, t)$. Evidently, the problem of studying trajectories near z^d is equivalent to studying the trajectories governed by $H^{\text{new}}(\zeta, t)$ in the case where ζ is small.

What is the relation between $H(z, t)$ and $H^{\text{new}}(\zeta, t)$? Define a function $\bar{H}(\zeta, t)$ by the rule

$$\bar{H}(\zeta, t) = H[z^d(t) + \zeta, t]. \tag{10.5.2}$$

Here the time dependence of $\bar{H}(\zeta, t)$ arises both from the possible time dependence of H and the time dependence of the design orbit $z^d(t)$. Next suppose that the quantity $\bar{H}(\zeta, t)$ is expressed as a power series in ζ by making the expansion

$$\bar{H}(\zeta, t) = \sum_{m=0}^{\infty} \bar{H}_m(\zeta, t). \tag{10.5.3}$$

Here each quantity $\bar{H}_m(\zeta, t)$ is a homogeneous polynomial of degree m in the components of ζ. Then we claim that $H^{\text{new}}(\zeta, t)$ is given by the relation

$$H^{\text{new}}(\zeta, t) = \sum_{m=2}^{\infty} \bar{H}_m(\zeta, t). \tag{10.5.4}$$

There are at least two ways to verify the truth of (5.4). In analogy with (4.8.2), let us write

$$z^d = (\beta_1 \cdots \beta_n, \alpha_1 \cdots \alpha_n), \tag{10.5.5}$$

$$\zeta = (Q_1 \cdots Q_n, P_1 \cdots P_n). \tag{10.5.6}$$

Introduce the mixed-variable generating function $F_2(q, P, t)$ by the rule

$$F_2(q, P, t) = \sum_{\ell=1}^{n} [\alpha_\ell(t) q_\ell - \beta_\ell(t) P_\ell + q_\ell P_\ell]. \tag{10.5.7}$$

Then, following the rules (4.8.4) and (4.8.5), we find the results

$$Q_\ell = \partial F_2 / \partial P_\ell = q_\ell - \beta_\ell(t), \tag{10.5.8}$$

$$p_\ell = \partial F_2 / \partial q_\ell = P_\ell + \alpha_\ell(t), \tag{10.5.9}$$

which are equivalent to the relation (5.1). Also, following the standard rules, the transformed Hamiltonian produced by the symplectic (canonical) transformation associated with (5.7) is given by the relation

$$H^{\text{new}}(Q, P, t) = H^{\text{new}}(\zeta, t) = H^{\text{old}}(z^d + \zeta, t) + \partial F_2/\partial t = \bar{H}(\zeta, t) + \partial F_2/\partial t. \tag{10.5.10}$$

But from (5.7) we find the result

$$\partial F_2/\partial t = \sum_{\ell=1}^{n} [\dot{\alpha}_\ell(t) q_\ell - \dot{\beta}_\ell(t) P_\ell] = \sum_{\ell=1}^{n} [\dot{\alpha}_\ell(t) \beta_\ell(t) + \dot{\alpha}_\ell(t) Q_\ell - \dot{\beta}_\ell P_\ell] \tag{10.5.11}$$

so that

$$H^{\rm new}(Q,P,t) = H^{\rm new}(\zeta,t) = \bar{H}(\zeta,t) + \sum_{\ell=1}^{n}[\dot{\alpha}_\ell(t)\beta_\ell(t) + \dot{\alpha}_\ell(t)Q_\ell - \dot{\beta}_\ell P_\ell]. \qquad (10.5.12)$$

We see that the second term on the far right of (5.12), the $\partial F_2/\partial t$ component of $H^{\rm new}$, consists only of terms independent of ζ and terms linear in ζ. Consequently, consistent with the claim made in (5.2) through (5.4), the quadratic and higher-order terms in ζ that appear in the expansions of $\bar{H}(\zeta,t)$ and $H^{\rm new}(\zeta,t)$ agree. Also, again consistent with (5.4), we know that $H^{\rm new}(\zeta,t)$ cannot contain terms linear in ζ, for otherwise $\zeta = 0$ would not be a possible trajectory for the equations of motion generated by $H^{\rm new}$. Finally, terms in $H^{\rm new}$ independent of ζ make no contribution to the equations of motion and, consistent with (5.4), can simply be dropped.

A second way to verify the truth of (5.4) is simply to examine equations of motion. According to (5.2.3) a general trajectory satisfies the equation of motion

$$\dot{z} = J\partial_z H(z,t), \qquad (10.5.13)$$

and the given trajectory z^d satisfies the equation of motion

$$\dot{z}^d = J\partial_z H(z,t)\big|_{z=z^d}. \qquad (10.5.14)$$

Also, as is easily verified from (5.2) and (5.3), we have the result

$$\partial_z H(z,t)|_{z=z^d} = \partial_\zeta \bar{H}_1(\zeta,t). \qquad (10.5.15)$$

Now let us insert (5.1) through (5.3) into (5.13) to get the relation

$$\dot{z} = \dot{z}^d + \dot{\zeta} = J\partial_z H(z,t) = J\partial_\zeta H(z^d+\zeta,t) = J\partial_\zeta \bar{H}(\zeta,t) = J\partial_\zeta \sum_{m=1}^{\infty} \bar{H}_m(\zeta,t). \qquad (10.5.16)$$

Finally, subtract (5.14) from (5.16) and use (5.15) and (5.4) to find the result

$$\dot{\zeta} = J\partial_\zeta \sum_{m=1}^{\infty} \bar{H}_m(\zeta,t) - J\partial_\zeta \bar{H}_1(\zeta,t) = J\partial_\zeta \sum_{m=2}^{\infty} \bar{H}_m(\zeta,t) = J\partial_\zeta H^{\rm new}(\zeta,t). \qquad (10.5.17)$$

We see that the time evolution of the deviation variables ζ is indeed governed by $H^{\rm new}$, as claimed. We also note that, according to (5.14) and (5.15), the equation of motion for the design trajectory itself is provided by the relation

$$\dot{z}^d = J\partial_\zeta \bar{H}_1(\zeta,t). \qquad (10.5.18)$$

We close this subsection by observing that there is a variant definition of $H^{\rm new}$ that is often convenient. The relation (5.4) can be rewritten in the form

$$H^{\rm new}(\zeta,t) = \sum_{m=2}^{\infty} \bar{H}_m(\zeta,t) = \bar{H}(\zeta,t) - \bar{H}_0(\zeta,t) - \bar{H}_1(\zeta,t). \qquad (10.5.19)$$

Since term $\bar{H}_0(\zeta,t)$ is independent of ζ, it makes no contribution to the equations of motion and therefore may be dropped. Consequently we may make the alternate definition

$$H^{\rm new}(\zeta,t) = \bar{H}(\zeta,t) - \bar{H}_1(\zeta,t). \qquad (10.5.20)$$

Note that all the definitions (5.12), (5.19), and (5.20) are in closed form and do not actually involve the summation of infinite series.

10.5.2 Term by Term Procedure

To continue the general discussion, let us write H^{new} as given by (5.4) in the form

$$H^{\text{new}} = H_2 + H_3 + H_4 + \cdots = H_2 + H_r. \tag{10.5.21}$$

Alternatively, we may expand H^{new} as given by (5.10) or (5.20) in homogeneous polynomials (in the components of ζ) and omit any irrelevant H_0 term to obtain the same result. Let $\mathcal{M}$ be the transfer map associated with H^{new}. In accord with the spirit of (1.9) and (1.12), let us factor $\mathcal{M}$ in the form

$$\mathcal{M} = \mathcal{M}_r\mathcal{M}_2. \tag{10.5.22}$$

Here, as in (5.21), we use the subscript "r" to denote "remaining" terms. [Note that (5.22) is a reverse factorization.] Following the discussion of Section 10.1, we will require that $\mathcal{M}_2$ obey the equation of motion

$$\dot{\mathcal{M}}_2 = \mathcal{M}_2 : -H_2 : . \tag{10.5.23}$$

Correspondingly, $\mathcal{M}_r$ will obey the equation of motion

$$\dot{\mathcal{M}}_r = \mathcal{M}_r : -H_r^{\text{int}} :, \tag{10.5.24}$$

where the interaction Hamiltonian H_r^{int} is given by rule

$$H_r^{\text{int}}(\zeta^i, t) = H_r(\mathcal{M}_2\zeta^i, t). \tag{10.5.25}$$

We will now describe how to compute $\mathcal{M}_2$ and $\mathcal{M}_r$. Let us begin with $\mathcal{M}_2$. Since H_2 is a quadratic Hamiltonian, its associated transfer map $\mathcal{M}_2$ must be linear. Let $\overline{\zeta}(t)$ be the result of $\mathcal{M}_2(t)$ acting on ζ^i. Then there is a symplectic matrix R such that

$$\overline{\zeta}_a(t) = \mathcal{M}_2(t)\zeta_a^i = \sum_b R_{ab}(t)\zeta_b^i, \tag{10.5.26}$$

or, in more compact vector and matrix notation,

$$\overline{\zeta}(t) = R(t)\zeta^i. \tag{10.5.27}$$

Thus, the computation of $\mathcal{M}_2$ is equivalent to finding the matrix R. Since H_2 is quadratic, there is an associated symmetric matrix $S(t)$ such that H_2 is given by the relation

$$H_2(\zeta^i, t) = (1/2)\sum_{a,b} S_{ab}(t)\zeta_a^i\zeta_b^i. \tag{10.5.28}$$

In analogy with (8.3.64) and (8.3.66), we have the result

$$: -H_2 : \zeta^i = JS\zeta^i. \tag{10.5.29}$$

Suppose both sides of (5.23) are applied to the quantity ζ^i. For the left side we find the result

$$\dot{\mathcal{M}}_2\zeta^i = \dot{\overline{\zeta}} = \dot{R}\zeta^i. \tag{10.5.30}$$

For the right side we find the result

$$\mathcal{M}_2 : -H_2 : \zeta^i = \mathcal{M}_2 JS\zeta^i = JS\mathcal{M}_2\zeta^i = JSR\zeta^i. \tag{10.5.31}$$

Now compare the right sides of (5.30) and (5.31). Since ζ^i is an arbitrary vector, we conclude that R must obey the matrix differential equation

$$\dot{R} = JSR. \tag{10.5.32}$$

Also, the requirement that $\mathcal{M}_2$ be the identity operator $\mathcal{I}$ when $t = t^i$ makes R subject to the initial condition

$$R(t^i) = I. \tag{10.5.33}$$

The differential equation (5.32) with the initial condition (5.33) has a unique solution whose computation, in most cases, requires numerical integration. [The system (5.32) is equivalent to $(2n)^2$ first-order coupled and time-dependent linear equations.] In the special case when the matrices $JS(t)$ and $JS(t')$ commute for all times t and t', one has, in analogy to (3.25), the explicit solution

$$R(t) = \exp\left[\int_{t^i}^{t} JS(t')dt'\right]. \tag{10.5.34}$$

[Indeed, it can be shown that the solution to (5.32) depends entirely upon the Lie algebra generated by the matrices $JS(t)$.] In the even more special case that S (and therefore H_2) is time independent, the integration required in (5.34) is immediate, and one obtains the result

$$R = \exp[(t - t^i)JS]. \tag{10.5.35}$$

In either of the cases corresponding to (5.34) and (5.35) one can write

$$\mathcal{M}_2 = \exp(: f_2 :), \tag{10.5.36}$$

with f_2 given by the relation

$$f_2 = -\int_{t^i}^{t} H_2(t')dt'. \tag{10.5.37}$$

In the general case, if desired, one may polar decompose R in analogy to (6.2.2) and, in analogy to (7.2.10), write $\mathcal{M}_2$ in the form

$$\mathcal{M}_2 = \exp(: f_2^c :)\exp(: f_2^a :). \tag{10.5.38}$$

We turn now to the calculation of $\mathcal{M}_r$. We begin by making the computation of H_r^{int} more explicit. By definition, H_r consists of terms of degree 3 and higher,

$$H_r = H_3 + H_4 + \cdots . \tag{10.5.39}$$

Also, in view of (5.25) and the fact that $\mathcal{M}_2$ produces a linear transformation when acting on ζ^i [see (5.26) and (5.27)], it follows that H_r^{int} has the decomposition

$$H_r^{\text{int}} = H_3^{\text{int}} + H_4^{\text{int}} + \cdots , \tag{10.5.40}$$

where each term $H_m^{\rm int}$ is a homogeneous polynomial of degree m given by the relation

$$H_m^{\rm int}(\zeta^i, t) = H_m(\mathcal{M}_2\zeta^i, t). \tag{10.5.41}$$

[We note in passing that the operations involved in computing (5.41) using (5.26) are analogous to those employed in (8.4.23) except that R^{-1} is replaced by R.]

To see how this works out in a specific case, consider the computation of $H_3^{\rm int}$. The terms of still higher degree are handled analogously. Suppose that H_3 is written in the explicit form

$$H_3(\zeta^i, t) = \sum_{abc} T_{abc}(t)\zeta_a^i\zeta_b^i\zeta_c^i, \tag{10.5.42}$$

where T_{abc} is a set of (possibly time-dependent) coefficients. Then use of (5.41) gives the relation

$$H_3^{\rm int}(\zeta^i, t) = \sum_{abc} T_{abc}(\mathcal{M}_2\zeta_a^i)(\mathcal{M}_2\zeta_b^i)(\mathcal{M}_2\zeta_c^i). \tag{10.5.43}$$

However, thanks to (5.26), the terms on the right side of (5.43) may be evaluated explicitly so that $H_3^{\rm int}$ can be expressed in the form

$$H_3^{\rm int}(\zeta^i, t) = \sum_{abc}\sum_{a'b'c'} T_{abc}R_{aa'}R_{bb'}R_{cc'}\zeta_{a'}^i\zeta_{b'}^i\zeta_{c'}^i. \tag{10.5.44}$$

Finally, the sums in (5.44) can be grouped so that $H_3^{\rm int}$ can be written in the final form

$$H_3^{\rm int}(\zeta^i, t) = \sum_{a'b'c'} T_{a'b'c'}^{\rm int}(t)\,\zeta_{a'}^i\zeta_{b'}^i\zeta_{c'}^i, \tag{10.5.45}$$

where $T^{\rm int}$ is defined by the equation

$$T_{a'b'c'}^{\rm int}(t) = \sum_{abc} T_{abc}(t)R_{aa'}(t)R_{bb'}(t)R_{cc'}(t). \tag{10.5.46}$$

As mentioned earlier, because of the time dependence of R, note that $H_3^{\rm int}$ is in general *time dependent* even if H_3 is not.

Let us write $\mathcal{M}_r$ in reversed factorized form. See Section 7.8. Since $H_r^{\rm int}$ consists of terms of degree 3 and higher, $\mathcal{M}_r$ can be written as the product

$$\mathcal{M}_r = \cdots\mathcal{M}_5\mathcal{M}_4\mathcal{M}_3, \tag{10.5.47}$$

where each factor $\mathcal{M}_m$ is generated by the homogeneous polynomial function $f_m(\zeta^i)$,

$$\mathcal{M}_m = \exp(:f_m:). \tag{10.5.48}$$

[Note that (5.47) is also a reversed factorization.] Our goal is to find equations of motion for the f_m.

From the factorization (5.47) and the product rule for differentiation, it follows that $\dot{\mathcal{M}}_r$ can be written in the form

$$\dot{\mathcal{M}}_r = \cdots + \cdots\dot{\mathcal{M}}_5\mathcal{M}_4\mathcal{M}_3 + \cdots\mathcal{M}_5\dot{\mathcal{M}}_4\mathcal{M}_3 + \cdots\mathcal{M}_5\mathcal{M}_4\dot{\mathcal{M}}_3. \tag{10.5.49}$$

Suppose (5.49) is substituted into the equation of motion (5.24) and both sides of the resulting relation are multiplied by $\mathcal{M}_r^{-1}$. So doing gives the result

$$\begin{aligned} \mathcal{M}_r^{-1}\dot{\mathcal{M}}_r &= \cdots + \mathcal{M}_3^{-1}\mathcal{M}_4^{-1}\mathcal{M}_5^{-1}\dot{\mathcal{M}}_5\mathcal{M}_4\mathcal{M}_3 + \mathcal{M}_3^{-1}\mathcal{M}_4^{-1}\dot{\mathcal{M}}_4\mathcal{M}_3 + \mathcal{M}_3^{-1}\dot{\mathcal{M}}_3 \\ &= :-H_r^{\text{int}}: . \end{aligned} \tag{10.5.50}$$

The various terms appearing in (5.50) can be simplified by the use of adjoint operators. For example, we have the result

$$\mathcal{M}_3^{-1}\mathcal{M}_4^{-1}\mathcal{M}_5^{-1}\dot{\mathcal{M}}_5\mathcal{M}_4\mathcal{M}_3 = \exp(-\#f_3\#)\exp(-\#f_4\#)\mathcal{M}_5^{-1}\dot{\mathcal{M}}_5. \tag{10.5.51}$$

See (8.2.23). Also, we have the relation

$$\mathcal{M}_m^{-1}\dot{\mathcal{M}}_m = \text{iex}(-\#f_m\#) : \dot{f}_m : . \tag{10.5.52}$$

See Appendix C. Upon using (5.51) and (5.52) in (5.50), we find that (5.50) can be rewritten in the form

$$\begin{aligned} \cdots &+ \exp(-\#f_3\#)\exp(-\#f_4\#)\,\text{iex}(-\#f_5\#) : \dot{f}_5 : \\ &+ \exp(-\#f_3\#)\,\text{iex}(-\#f_4\#) : \dot{f}_4 : \\ &+ \text{iex}(-\#f_3\#) : \dot{f}_3 :=: -H_r^{\text{int}} : . \end{aligned} \tag{10.5.53}$$

At this stage the colons can be removed from both sides of (5.53) to give the result

$$\begin{aligned} \cdots &+ \exp(-:f_3:)\exp(-:f_4:)\,\text{iex}(-:f_5:)\dot{f}_5 \\ &+ \exp(-:f_3:)\,\text{iex}(-:f_4:)\dot{f}_4 \\ &+ \text{iex}(-:f_3:)\dot{f}_3 = -H_r^{\text{int}}. \end{aligned} \tag{10.5.54}$$

Let us examine both sides of (5.54) with the aim of equating terms of like degree. From the expansion (8.8.9) we find the result

$$\text{iex}(-:f_m:)\dot{f}_m = (1 - :f_m:/2! + :f_m:^2/3! - \cdots)\dot{f}_m. \tag{10.5.55}$$

According to (7.6.16), the terms of the right side of (5.55) have degrees m, $2m-2$, $3m-4$, etc. Consequently, upon using (5.40), and equating terms of like degree in (5.54), we find the result

$$\begin{aligned} P_m[\cdots &+ \exp(-:f_3:)\exp(-:f_4:)\,\text{iex}(-:f_5:)\dot{f}_5 \\ &+ \exp(-:f_3:)\,\text{iex}(-:f_4:)\dot{f}_4 \\ &+ \text{iex}(-:f_3:)\dot{f}_3] = -H_m^{\text{int}}. \end{aligned} \tag{10.5.56}$$

Here P_m denotes a *projection* operator that projects out terms of degree m. For example, we have the results

$$P_3[\cdots + \text{iex}(-:f_3:)\dot{f}_3] = \dot{f}_3, \tag{10.5.57}$$

$$P_4[\cdots + \text{iex}(-:f_3:)\dot{f}_3] = \dot{f}_4 - (:f_3:/2!)\dot{f}_3, \tag{10.5.58}$$

$$P_5[\cdots + \text{iex}(-:f_3:)\dot{f}_3] = \dot{f}_5 - :f_3:\dot{f}_4 + (:f_3:^2/3!)\dot{f}_3. \tag{10.5.59}$$

The relations (5.56) can now be solved for the various $\dot{f}_m$. We find, for example, through $m = 8$, the results

$$\dot{f}_3 = -H_3^{\rm int}, \tag{10.5.60}$$

$$\dot{f}_4 = -H_4^{\rm int} + (:f_3:/2)(-H_3^{\rm int}), \tag{10.5.61}$$

$$\dot{f}_5 = -H_5^{\rm int} + :f_3:(-H_4^{\rm int}) + (1/3):f_3:^2(-H_3^{\rm int}), \tag{10.5.62}$$

$$\begin{aligned}\dot{f}_6 = & - H_6^{\rm int} + :f_3:(-H_5^{\rm int}) + (1/2):f_4:(-H_4^{\rm int}) \\ & + (1/4):f_4::f_3:(-H_3^{\rm int}) + (1/2):f_3:^2(-H_4^{\rm int}) \\ & + (1/8):f_3:^3(-H_3^{\rm int}),\end{aligned} \tag{10.5.63}$$

$$\begin{aligned}\dot{f}_7 = & - H_7^{\rm int} + :f_3:(-H_6^{\rm int}) + :f_4:(-H_5^{\rm int}) + :f_4::f_3:(-H_4^{\rm int}) \\ & + (1/3):f_4::f_3:^2(-H_3^{\rm int}) + (1/2):f_3:^2(-H_5^{\rm int}) \\ & + (1/6):f_3:^3(-H_4^{\rm int}) + (1/30):f_3:^4(-H_3^{\rm int}),\end{aligned} \tag{10.5.64}$$

$$\begin{aligned}\dot{f}_8 = & - H_8^{\rm int} + :f_3:(-H_7^{\rm int}) + :f_4:(-H_6^{\rm int}) + :f_4::f_3:(-H_5^{\rm int}) \\ & + (1/2):f_4::f_3:^2(-H_4^{\rm int}) + (1/8):f_4::f_3:^3(-H_3^{\rm int}) \\ & + (1/2):f_5:(-H_5^{\rm int}) + (1/2):f_5::f_3:(-H_4^{\rm int}) \\ & + (1/6):f_5::f_3:^2(-H_3^{\rm int}) + (1/2):f_3:^2(-H_6^{\rm int}) \\ & + (1/3):f_4:^2(-H_4^{\rm int}) + (1/6):f_4:^2:f_3:(-H_3^{\rm int}) \\ & + (1/6):f_3:^3(-H_5^{\rm int}) + (1/24):f_3:^4(-H_4^{\rm int}) \\ & + (1/144):f_3:^5(-H_3^{\rm int}),\end{aligned} \tag{10.5.65}$$

$$\dot{f}_m = \text{ something involving } H_m \text{ and the } f_\ell \text{ and } H_\ell \text{ with } \ell < m. \tag{10.5.66}$$

What have we accomplished? From (5.22), (5.38), (5.47), and (5.48) we see that $\mathcal{M}$ has been computed in a reverse factorized product form; and we have found how to calculate the $\mathcal{M}_m$. To find $\mathcal{M}_2$ we need to integrate the equations (5.32) with the initial condition (5.33). Here it is assumed that $z^d(t)$ is known so that H_2 and hence S is known. See (5.28). If this is not the case, then we must also integrate the equations (5.14) or (5.18). That is, we must integrate the equations (5.14) [or (5.18)] and (5.32) as a coupled set. To find the f_m that define $\mathcal{M}_r$ according to (5.47) and (5.48), we must also integrate the equations of motion for the f_m as given by (5.60) through (5.66). The requirement that $\mathcal{M}_r$ be the identity operator $\mathcal{I}$ when $t = t^i$ makes the f_m subject to the initial condition

$$f_m(t^i) = 0. \tag{10.5.67}$$

We note from the form of equations (5.60) through (5.66) that the computation of the f_m requires a knowledge of the $H_\ell^{\rm int}$ with $\ell \leq m$. The computation of $H_\ell^{\rm int}(t)$ in turn requires a knowledge of $R(t)$. [See, for example, (5.43) through (5.46).] Consequently, the equations of motion (5.60) through (5.66) for the f_m must be integrated simultaneously with the equations (5.32) for R. [Moreover, in general $z^d(t)$ must also be known to compute $H_\ell^{\rm int}(t)$.

Thus, if $z^d(t)$ is not known explicitly, then the equations (5.14) or (5.18) must in general also be in the set of equations to be integrated.]

We close this section with the observation that the equations of motion (5.60) through (5.66) for the f_m, like equations (3.11) through (3.15) for the F_n, have the property that the f_m can be determined successively. This property has two consequences. First, a solution of the equations of motion for R, f_3, f_4, $\cdots f_m$ by numerical methods is always possible for any m. Moreover, solution by *quadrature* is also possible. For example, assuming that R has already been determined, (5.60) can be integrated immediately to give the result

$$f_3(\zeta^i,t) = \int_{t^i}^{t} dt_1[-H_3^{\rm int}(\zeta^i,t_1)]. \tag{10.5.68}$$

Here we have used (5.67). Next, this result for f_3 can be substituted into (5.61) and the resulting expression for $\dot{f}_4$ can be integrated to give the relation

$$f_4(\zeta^i,t) = \int_{t^i}^{t} dt_1[-H_4^{\rm int}(\zeta^i,t_1)] + (1/2)\int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2[-H_3^{\rm int}(\zeta^i,t_2), -H_3^{\rm int}(\zeta^i,t_1)]. \tag{10.5.69}$$

Similarly, one finds for f_5 the result

$$\begin{aligned}
f_5 &= \int_{t^i}^{t} dt_1[-H_5^{\rm int}(t_1)]\\
&+ \int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2[-H_3^{\rm int}(t_2), -H_4^{\rm int}(t_1)]\\
&+ (1/3)\int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2 \int_{t^i}^{t_2} dt_3[-H_3^{\rm int}(t_3),[-H_3^{\rm int}(t_2), -H_3^{\rm int}(t_1)]]\\
&+ (1/3)\int_{t^i}^{t} dt_1 \int_{t^i}^{t_1} dt_2 \int_{t^i}^{t_2} dt_3[-H_3^{\rm int}(t_2),[-H_3^{\rm int}(t_3), -H_3^{\rm int}(t_1)]].
\end{aligned} \tag{10.5.70}$$

Evidently, one can find explicit integral representations for all the f_m. We note that to determine any f_m it is necessary to know only the $H_\ell^{\rm int}$ with $\ell \le m$. Moreover, if the $H_\ell^{\rm int}$ do not commute (are not in involution) at different times, which is usually the case, there are feed-up effects: lower-order terms in the Hamiltonian can contribute to higher-order Lie generators. Specifically, we see that the f_m all lie in the Lie algebra generated by the $H_\ell^{\rm int}$. Finally, suppose the time dependencies of the various $H_\ell^{\rm int}$ are sufficiently simple that all the integrations occurring in (5.68), (5.69), (5.70), and analogous expressions for the other f_m can be carried out analytically. Then the equations of motion (5.60) through (5.66) can in principle be solved directly by symbolic manipulation to obtain complete analytic expressions for all desired f_m.

We close this subsection by noting that for simplicity we have derived formulas for R and the f_m when the map $\mathcal{M}$ is written in *reversed* factorized product form. That is, we have written $\mathcal{M}$ in the form

$$\mathcal{M} = \cdots \mathcal{M}_5\mathcal{M}_4\mathcal{M}_3\mathcal{M}_2. \tag{10.5.71}$$

See (5.22) and (5.47). Of course, once we have found $\mathcal{M}$ is reversed factorized product form, we can convert it to the *forward* factorized product form

$$\mathcal{M} = \mathcal{M}_2'\mathcal{M}_3'\mathcal{M}_4'\mathcal{M}_5' \cdots \tag{10.5.72}$$

by means of concatenation formulas. Recall Section 8.4. Alternatively, as will be seen in the next section, there are formulas for the associated R' and f_m', analogous to those for R and the f_m, when $\mathcal{M}$ is written in forward factorized product form. Finally we note, from the work of Section 8.5, that there is a standard procedure for passing from forward to reverse factorized forms.

Exercises

10.5.1. The purpose of this exercise is to verify the relation (5.18) from another perspective. Begin by showing that by Taylor's theorem there is the relation

$$\bar{H}_1(\zeta, t) = \sum_a [\partial H(z,t)/\partial z_a]|_{z=z^d(t)}\zeta_a, \tag{10.5.73}$$

and consequently

$$\partial \bar{H}_1(\zeta, t)/\partial \zeta_a = [\partial H(z,t)/\partial z_a]|_{z=z^d(t)}. \tag{10.5.74}$$

Now verify it follows that

$$J\partial_\zeta \bar{H}_1(\zeta, t) = J[\partial_z H(z,t)]|_{z=z^d(t)} = \dot{z}^d. \tag{10.5.75}$$

10.5.2. Consider a general Hamiltonian system with Hamiltonian $H(z,t)$, and suppose $z^d(t)$ is any particular trajectory for (solution to) the equations of motion associated with this Hamiltonian. Form the variational equations about the trajectory $z^d(t)$. Show that the variational equations also arise from a Hamiltonian, and that this Hamiltonian is the quadratic Hamiltonian $\bar{H}_2$ that appears in the sum (5.3). Show that integrating (5.32) with the initial condition (5.33) provides all possible solutions to the variational equations.

10.6 Forward Factorization and Lie Concatenation Revisited

10.6.1 Preliminary Discussion

The previous section derived a reversed factorized product solution to the equation of motion (1.8) for $\mathcal{M}$. For this present section, and other purposes as well, it is useful to also have formulas for a forward factorized product solution. We begin by finding them.

The main purpose of this section, which is really a diversion from the logical presentation of methods for the computation of maps, is to provide another derivation of the Lie concatenation formulas of Section 8.4. Subsequent sections will return to the main subject of this chapter.

10.6.2 Forward Factorization

As before we write

$$\mathcal{M} = \mathcal{M}_r \mathcal{M}_2 \tag{10.6.1}$$

and require that $\mathcal{M}_2$ again satisfy (5.23). Then we find, as before, that $\mathcal{M}_r$ obeys the equation of motion

$$\dot{\mathcal{M}}_r = \mathcal{M}_r : -H_r^{\text{int}} : \tag{10.6.2}$$

with H_r^{int} again given by (5.25). Now, however, we write $\mathcal{M}_r$ in the forward factorized form

$$\mathcal{M}_r = \mathcal{M}_3 \mathcal{M}_4 \mathcal{M}_5 \cdots \tag{10.6.3}$$

with

$$\mathcal{M}_m = \exp(: f_m :). \tag{10.6.4}$$

We will obtain equations of motion for the f_m shortly, but imagine for the moment that they have already been found and solved. Then we may write $\mathcal{M}$ in the form

$$\mathcal{M} = \exp(: f_3 :) \exp(: f_4 :) \exp(: f_5 :) \cdots \mathcal{M}_2. \tag{10.6.5}$$

Algebraic manipulation now gives the equivalent result

$$\begin{aligned} \mathcal{M} &= \mathcal{M}_2 \mathcal{M}_2^{-1} \exp(: f_3 :) \mathcal{M}_2 \mathcal{M}_2^{-1} \exp(: f_4 :) \mathcal{M}_2 \mathcal{M}_2^{-1} \exp(: f_5 :) \cdots \mathcal{M}_2 \\ &= \mathcal{M}_2 \exp(: f_3^{tr} :) \exp(: f_4^{tr} :) \exp(: f_5^{tr} :) \cdots \end{aligned} \tag{10.6.6}$$

where

$$f_m^{tr} = \mathcal{M}_2^{-1} f_m. \tag{10.6.7}$$

We see that $\mathcal{M}$ as given by (6.6) and (6.7) is in the desired forward factorized product form.

It remains to find the f_m. Differentiating (6.3) gives the result

$$\dot{\mathcal{M}}_r = \dot{\mathcal{M}}_3 \mathcal{M}_4 \mathcal{M}_5 \cdots + \mathcal{M}_3 \dot{\mathcal{M}}_4 \mathcal{M}_5 \cdots + \mathcal{M}_3 \mathcal{M}_4 \dot{\mathcal{M}}_5 \cdots + \cdots . \tag{10.6.8}$$

Next, substitution of (6.8) into the equation of motion (6.2) produces the relation

$$\begin{aligned} \mathcal{M}_r^{-1} \dot{\mathcal{M}}_r &= \cdots \mathcal{M}_5^{-1} \mathcal{M}_4^{-1} \mathcal{M}_3^{-1} \dot{\mathcal{M}}_3 \mathcal{M}_4 \mathcal{M}_5 \cdots + \cdots \mathcal{M}_5^{-1} \mathcal{M}_4^{-1} \dot{\mathcal{M}}_4 \mathcal{M}_5 \cdots \\ &+ \cdots \mathcal{M}_5^{-1} \dot{\mathcal{M}}_5 \cdots + \cdots =: -H_r^{\text{int}} : . \end{aligned} \tag{10.6.9}$$

As in the previous section, the various terms in (6.9) can be simplified by the use of adjoint operators. For example, we have the result

$$\cdots \mathcal{M}_5^{-1} \mathcal{M}_4^{-1} \mathcal{M}_3^{-1} \dot{\mathcal{M}}_3 \mathcal{M}_4 \mathcal{M}_5 \cdots = \cdots \exp(-\#f_5\#) \exp(-\#f_4\#) \mathcal{M}_3^{-1} \dot{\mathcal{M}}_3. \tag{10.6.10}$$

Again we recall the relation

$$\mathcal{M}_m^{-1} \dot{\mathcal{M}}_m = \text{ iex } (-\#f_m\#) : \dot{f}_m : . \tag{10.6.11}$$

Upon using (6.10) and (6.11) in (6.9) we find that it can be rewritten in the form

$$\begin{aligned} &\cdots \exp(-\#f_5\#) \exp(-\#f_4\#) \text{ iex}(-\#f_3\#) : \dot{f}_3 : \\ &+ \cdots \exp(-\#f_5\#) \text{ iex } (-\#f_4\#) : \dot{f}_4 : \\ &+ \cdots \text{ iex}(-\#f_5\#) : \dot{f}_5 :=: -H_r^{\text{int}} : . \end{aligned} \tag{10.6.12}$$

The colons can now be removed from both sides of (6.12) to give the equivalent result

$$\begin{aligned}
&\cdots \exp(- : f_5 :) \exp(- : f_4 :) \operatorname{iex}(- : f_3 :) \dot{f}_3 \\
&+ \cdots \exp(- : f_5 :) \operatorname{iex}\, (- : f_4 :) \dot{f}_4 \\
&+ \cdots \ \operatorname{iex}(- : f_5 :) \dot{f}_5 \\
&+ \cdots = -H_r^{\text{int}}.
\end{aligned} \tag{10.6.13}$$

Finally, similar to the procedure in the previous section, we equate terms of like degree in (6.13). So doing gives the equations of motion

$$\dot{f}_3 = -H_3^{\text{int}}, \tag{10.6.14}$$

$$\dot{f}_4 = -H_4^{\text{int}} + (: f_3 : /2)(-H_3^{\text{int}}), \tag{10.6.15}$$

$$\dot{f}_5 = -H_5^{\text{int}} - (1/6) : f_3 :^2 (-H_3^{\text{int}}) + : f_4 : (-H_3^{\text{int}}), \tag{10.6.16}$$

$$\begin{aligned}
\dot{f}_6 \ \ = \ \ & -H_6^{\text{int}} + (1/24) : f_3 :^3 (-H_3^{\text{int}}) + (1/2) : f_4 : (-H_4^{\text{int}}) \\
& -(1/4) : f_4 :: f_3 : (-H_3^{\text{int}}) + : f_5 : (-H_3^{\text{int}}),
\end{aligned} \tag{10.6.17}$$

$$\begin{aligned}
\dot{f}_7 \ \ = \ \ & -H_7^{\text{int}} - (1/120) : f_3 :^4 (-H_3^{\text{int}}) + (1/6) : f_4 :: f_3 :^2 (-H_3^{\text{int}}) \\
& -(1/2) : f_4 :^2 (-H_3^{\text{int}}) + : f_5 : (-H_4^{\text{int}}) + : f_6 : (-H_3^{\text{int}}),
\end{aligned} \tag{10.6.18}$$

$$\begin{aligned}
\dot{f}_8 \ \ = \ \ & -H_8^{\text{int}} + (1/720) : f_3 :^5 (-H_3^{\text{int}}) - (1/24) : f_4 :: f_3 :^3 (-H_3^{\text{int}}) \\
& -(1/6) : f_4 :^2 (-H_4^{\text{int}}) + (1/6) : f_4 :^2 : f_3 : (-H_3^{\text{int}}) + (1/2) : f_5 : (-H_5^{\text{int}}) \\
& +(1/12) : f_5 :: f_3 :^2 (-H_3^{\text{int}}) - (1/2) : f_5 :: f_4 : (-H_3^{\text{int}}) + : f_6 : (-H_4^{\text{int}}) \\
& + : f_7 : (-H_3^{\text{int}}),
\end{aligned} \tag{10.6.19}$$

$$\dot{f}_m = \text{ expression involving } H_m^{\text{int}} \text{ and the } f_\ell \text{ and } H_\ell^{\text{int}} \text{ with } \ell < m. \tag{10.6.20}$$

As before, these equations can be integrated with the initial condition (5.67).

10.6.3 Alternate Derivation of Lie Concatenation Formulas

The tools are now in hand to carry out the main task of this section: an alternate derivation of the Lie concatenation formulas. According to (8.4.26) in Section 8.4, the problem is to find the h_m in the relation

$$\exp(: h_3 :) \exp(: h_4 :) \cdots = \exp(: f_3^{tr} :) \exp(: f_4^{tr} :) \cdots \exp(: g_3 :) \exp(: g_4 :) \cdots . \tag{10.6.21}$$

[Note that here the f_m^{tr} are given by (8.4.23) and not by (6.7).] To do this, we employ a trick. Define a one-parameter family of maps $\mathcal{N}(\lambda)$ by the rule

$$\mathcal{N}(\lambda) = \exp(\lambda : f_3^{tr} :) \exp(\lambda : f_4^{tr} :) \cdots \exp(\lambda : g_3 :) \exp(\lambda : g_4 :) \cdots . \tag{10.6.22}$$

Then, by construction, $\mathcal{N}(\lambda)$ has the properties

$$\mathcal{N}(0) = \mathcal{I}, \tag{10.6.23}$$

$$\mathcal{N}(1) = \exp(:h_3:)\exp(:h_4:)\cdots\ , \tag{10.6.24}$$

$$\mathcal{N}^{-1}(\lambda) = \cdots\exp(-\lambda:g_4:)\exp(-\lambda:g_3:)\cdots\exp(-\lambda:f_4^{tr}:)\exp(-\lambda:f_3^{tr}:). \tag{10.6.25}$$

From Section 6.4 we know that there is an associated Hamiltonian $H(\lambda)$. In this case, in analogy to (1.8), it can be found from the relation

$$:-H:=\mathcal{N}^{-1}\dot{\mathcal{N}} \tag{10.6.26}$$

where a dot denotes differentiation with respect to λ. From (6.22) we find the result

$$\begin{aligned}
\dot{\mathcal{N}} &= \exp(\lambda:f_3^{tr}:):f_3^{tr}:\exp(\lambda:f_4^{tr}:)\cdots\exp(\lambda:g_3:)\exp(\lambda:g_4:)\cdots \\
&+ \exp(\lambda:f_3^{tr}:)\exp(\lambda:f_4^{tr}:):f_4^{tr}:\cdots\exp(\lambda:g_3:)\exp(\lambda:g_4:)\cdots \\
&+ \cdots \\
&+ \exp(\lambda:f_3^{tr}:)\exp(\lambda:f_4^{tr}:)\cdots\exp(\lambda:g_3:):g_3:\exp(\lambda:g_4:)\cdots \\
&+ \exp(\lambda:f_3^{tr}:)\exp(\lambda:f_4^{tr}:)\cdots\exp(\lambda:g_3:)\exp(\lambda:g_4:):g_4:\cdots \\
&+ \cdots\ .
\end{aligned} \tag{10.6.27}$$

Consequently the Hamiltonian Lie operator $:-H:$ is given by

$$\begin{aligned}
:-H: &= \mathcal{N}^{-1}\dot{\mathcal{N}} \\
&= \cdots\exp(-\lambda:g_4:)\exp(-\lambda:g_3:)\cdots\times \\
&\quad \exp(-\lambda:f_4^{tr}:):f_3^{tr}:\exp(\lambda:f_4^{tr}:)\cdots\exp(\lambda:g_3:)\exp(\lambda:g_4:)\cdots \\
&+ \cdots\exp(-\lambda:g_4:)\exp(-\lambda:g_3:)\cdots:f_4^{tr}:\cdots\exp(\lambda:g_3:)\exp(\lambda:g_4:)\cdots \\
&+ \cdots \\
&+ \cdots\exp(-\lambda:g_4:):g_3:\exp(\lambda:g_4:)\cdots \\
&+ \cdots:g_4:\cdots \\
&+ \cdots\ .
\end{aligned} \tag{10.6.28}$$

This result can also be written more compactly with the aid of adjoint operators to take the form

$$\begin{aligned}
:-H: &= \cdots\exp(-\lambda\#g_4\#)\exp(-\lambda\#g_3\#)\cdots\exp(-\lambda\#f_4^{tr}\#):f_3^{tr}: \\
&+ \cdots\exp(-\lambda\#g_4\#)\exp(-\lambda\#g_3\#)\cdots:f_4^{tr}: \\
&+ \cdots \\
&+ \cdots\exp(-\lambda\#g_4\#):g_3: \\
&+ \cdots:g_4: \\
&+ \cdots\ .
\end{aligned} \tag{10.6.29}$$

The colons can now be removed from both sides of (6.29) to give the equivalent result

$$\begin{aligned}
-H &= \cdots\exp(-\lambda:g_4:)\exp(-\lambda:g_3:)\cdots\exp(-\lambda:f_4^{tr}:)f_3^{tr} \\
&+ \cdots\exp(-\lambda:g_4:)\exp(-\lambda:g_3:)\cdots f_4^{tr} \\
&+ \cdots \\
&+ \cdots\exp(-\lambda:g_4:)g_3 \\
&+ \cdots g_4 \\
&+ \cdots\ .
\end{aligned} \tag{10.6.30}$$

Finally, equating terms of like degree on both sides of (6.30) yields the results

$$-H_2 = 0, \tag{10.6.31}$$

$$-H_3 = f_3^{tr} + g_3, \tag{10.6.32}$$

$$-H_4 = -\lambda : g_3 : f_3^{tr} + f_4^{tr} + g_4, \tag{10.6.33}$$

$$\begin{aligned} -H_5 &= f_5^{tr} + g_5 - \lambda : f_4^{tr} : f_3^{tr} - \lambda : g_3 : f_4^{tr} \\ &+ (1/2)\lambda^2 : g_3 :^2 f_3^{tr} - \lambda : g_4 : f_3^{tr} - \lambda : g_4 : g_3, \text{ etc.} \end{aligned} \tag{10.6.34}$$

Evidently the h_m in (6.24) can be regarded as the solutions to differential equations of the form (6.14) through (6.20) with H given by (6.31) through (6.34). Also, as a consequence of (6.31), we have the result

$$H_m^{\text{int}} = H_m. \tag{10.6.35}$$

It follows that the h_m satisfy the differential equations

$$\dot{h}_3 = -H_3 = f_3^{tr} + g_3, \tag{10.6.36}$$

$$\begin{aligned} \dot{h}_4 &= -H_4 + (: h_3 : /2)(-H_3) \\ &= -\lambda : g_3 : f_3^{tr} + f_4^{tr} + g_4 + (: h_3 : /2)(f_3^{tr} + g_3), \end{aligned} \tag{10.6.37}$$

$$\begin{aligned} \dot{h}_5 &= -H_5 - (1/6) : h_3 :^2 (-H_3) + : h_4 : (-H_3) \\ &= f_5^{tr} + g_5 - \lambda : f_4^{tr} : f_3^{tr} - \lambda : g_3 : f_4^{tr} + (1/2)\lambda^2 : g_3 :^2 f_3^{tr} - \lambda : g_4 : f_3^{tr} \\ &- \lambda : g_4 : g_3 - (1/6) : h_3 :^2 (f_3^{tr} + g_3) + : h_4 : (f_3^{tr} + g_3), \text{ etc.} \end{aligned} \tag{10.6.38}$$

These equations can now be solved. Equation (6.36) has the immediate solution

$$h_3(\lambda) = \lambda(f_3^{tr} + g_3). \tag{10.6.39}$$

When this solution is substituted into (6.37), the result is that h_4 satisfies the differential equation

$$\dot{h}_4 = -\lambda : g_3 : f_3^{tr} + f_4^{tr} + g_4 \tag{10.6.40}$$

with the solution

$$h_4(\lambda) = \lambda(f_4^{tr} + g_4) - (\lambda^2/2) : g_3 : f_3^{tr}. \tag{10.6.41}$$

Next, with this knowledge of $h_3(\lambda)$ and $h_4(\lambda)$, the equation of motion for h_5 becomes

$$\dot{h}_5 = f_5^{tr} + g_5 - 2\lambda : g_3 : f_4^{tr} + \lambda^2 : g_3 :^2 f_3^{tr} - (\lambda^2/2) : f_3^{tr} :^2 g_3, \tag{10.6.42}$$

with the solution

$$h_5(\lambda) = \lambda(f_5^{tr} + g_5) - \lambda^2 : g_3 : f_4^{tr} + (\lambda^3/3) : g_3 :^2 f_3^{tr} - (\lambda^3/6) : f_3^{tr} :^2 g_3. \tag{10.6.43}$$

The $h_m(\lambda)$ for $m = 6, 7, \cdots$ can be found in an analogous way. We conclude that all the $h_m(\lambda)$ can be computed recursively for any value of m.

Finally, the quantities $h_m(1)$ yield the desired h_m in (6.21). Compare, for example, (8.4.31) through (8.4.33) with the $h_m(1)$ computed from (6.39), (6.41), and (6.43). The virtue of this method for deriving the Lie concatenation formulas is that no knowledge of the BCH series coefficients is required and the results are immediately obtained in Lie form. Only the Taylor coefficients in the *exp* and *iex* functions are needed, and these are known to all orders. And since the equations to be integrated are polynomial in λ, all calculations can be carried out to any desired order by symbolic manipulation.

Exercises

10.6.1.

10.7 Direct Taylor Summation

We now return to the task of computing maps. Suppose the Hamiltonian H is autonomous (time independent) and is analytic in z about the origin $z = 0$. Then it has an expansion in homogeneous polynomials of the form

$$H = \sum_{\ell=1}^{\infty} H_\ell(z). \tag{10.7.1}$$

(Here we have omitted a possible constant term since it has no dynamic effect.) We know from Section 7.4 that in this case the associated transfer map $\mathcal{M}$ can be written formally as

$$\mathcal{M} = \exp(-\tau : H :) \tag{10.7.2}$$

where we have used the short-hand notation

$$\tau = t - t^i. \tag{10.7.3}$$

The purpose of this section, among other things, is to explore what happens when there are singularities in τ.

Let us write

$$z_a^f = \mathcal{M} z_a^i \tag{10.7.4}$$

and make the Taylor expansion

$$z_a^f = k_a + \sum_b R_{ab} z_b^i + \sum_{bc} T_{abc} z_b^i z_c^i + \sum_{bcd} U_{abcd} z_b^i z_c^i z_d^i + \cdots . \tag{10.7.5}$$

We also have the result

$$z_a^f = \exp(-\tau : H :) z_a^i = \sum_{j=0}^{\infty} [(-\tau)^j / j!] : H :^j z_a^i. \tag{10.7.6}$$

Here H is to be regarded as a function of the z^i,

$$H = H(z^i) = H_1(z^i) + H_2(z^i) + H_3(z^i) + \cdots . \tag{10.7.7}$$

If we substitute (7.7) into (7.6), carry out all the indicated Poisson brackets, and then group terms by degree, we will find the Taylor coefficients k, R, T, U, etc. that appear in (7.5). Once the Taylor coefficients are known, there is (according to the Factorization Theorem of Sections 7.6 through 7.8) a standard procedure for finding the homogeneous polynomials f_1, f_2^c, f_2^a, f_3, f_4, $\cdots$ such that $\mathcal{M}$ can be written in the form

$$\mathcal{M} = \exp(: f_1 :) \exp(: f_2^c :) \exp(: f_2^a :) \exp(: f_3 :) \exp(: f_4 :) \cdots . \tag{10.7.8}$$

What we wish to investigate at this point is the convergence of the series (7.6) *in the sense* that it produces *series* for the Taylor coefficients k, R, T, U, etc. Note that this issue is separate from the convergence of the Taylor series (7.5) with regard to the variables z^i.

To facilitate this investigation, it is convenient to employ the basis monomials G_r introduced in Sections 7.3 and 8.3. With the aid of these monomials and the inner product (8.3.32), introduce the infinite dimensional matrices M and H by the rules

$$M_{sr} = \langle G_s, \mathcal{M} G_r \rangle, \tag{10.7.9}$$

$$H_{sr} = \langle G_s, :H: G_r \rangle. \tag{10.7.10}$$

Then, as a result of (8.3.41), the operator relation (7.2) has the equivalent matrix formulation

$$M = \exp(-\tau H) = \sum_{j=0}^{\infty} (\tau^j / j!) H^j. \tag{10.7.11}$$

[Here, to test the reader's mental agility, the symbol H stands not for the Hamiltonian (7.7) but rather for the matrix (7.10).] Moreover, the various Taylor coefficients k, R, T, U, etc. in (7.5) are just matrix elements of the form $\langle G_s, \mathcal{M} G_a^1 \rangle$ where we have used the notation G_a^1 to denote the *linear* functions z_a.

In any practical calculation with power series it is necessary to truncate them at some stage. Operations performed on and with truncated power series will be referred to as *Truncated Power Series Algebra* (TPSA). In this context it is useful to introduce a *projection* operator $\mathcal{P}^D$. Let $m(r)$ denote the degree of the monomial G_r. It is defined by the relation

$$m(r) = \sum (u_i + v_i). \tag{10.7.12}$$

See (7.3.33) and (7.3.34). We now define $\mathcal{P}^D$ to be the *linear* operator with the property

$$\mathcal{P}^D G_r = G_r \text{ if } m(r) \leq D,$$

$$\mathcal{P}^D G_r = 0 \text{ if } m(r) > D. \tag{10.7.13}$$

That is, $\mathcal{P}^D$ retains monomials having degree D or less, and discards those with degree larger than D. In terms of the earlier notation associated with (8.9.68) and (8.9.69), $\mathcal{P}^D$ is just the truncation operator $\mathcal{T}(> D)$. Let θ be a step function given by the rule

$$\theta(x) = 0 \text{ for } x < 0,$$

$$\theta(x) = 1 \text{ for } x \geq 0. \tag{10.7.14}$$

Then we also have the equivalent definition

$$\mathcal{P}^D G_r = \theta(D - m) G_r. \tag{10.7.15}$$

Evidently $\mathcal{P}^D$ has the property

$$(\mathcal{P}^D)^2 = \mathcal{P}^D. \tag{10.7.16}$$

Now let $\mathcal{A}$ be any linear operator. In TPSA this operator is realized by its *truncated* counterpart ${}^D\!\mathcal{A}$ defined by the rule

$$ {}^D\!\mathcal{A} = \mathcal{P}^D \mathcal{A} \mathcal{P}^D. \tag{10.7.17}$$

Following (7.3.33) through (7.3.35), let us also use the modified notation G_r^m for the basis monomials. Consider the vector space V^D spanned by the monomials G_r^0, G_r^1, $\cdots$ G_r^D. From the definition (7.17) we evidently have the result

$$\langle G_r^m,\ {}^D\!\mathcal{A} G_{r'}^{m'} \rangle = 0 \text{ if either } m > D \text{ or } m' > D. \tag{10.7.18}$$

It follows that ${}^D\!\mathcal{A}$ maps V^D into itself, and consequently has a representation as a *finite* dimensional matrix ${}^D\!A$ acting on V^D. Indeed, the matrix ${}^D\!A$ has the entries

$${}^D\!A_{sr} = \langle G_s,\ {}^D\!\mathcal{A} G_r \rangle \text{ with } m(s), m(r) \leq D. \tag{10.7.19}$$

Supppose $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ are three linear operators related by the equation

$$\mathcal{C} = \mathcal{A}\mathcal{B}. \tag{10.7.20}$$

Then in general one has the inequalities

$${}^D\mathcal{C} \neq\ {}^D\!\mathcal{A}\ {}^D\mathcal{B}, \tag{10.7.21}$$

$${}^D C \neq\ {}^D A\ {}^D B. \tag{10.7.22}$$

However, if either $\mathcal{A}$ or $\mathcal{B}$ commute with $\mathcal{P}^D$, then (7.21) and (7.22) become equalities. Indeed, from (7.16) and the definitions (7.13) we have the result

$${}^D\!\mathcal{A}\ {}^D\mathcal{B} = \mathcal{P}^D \mathcal{A} \mathcal{P}^D \mathcal{P}^D \mathcal{B} P^D = \mathcal{P}^D \mathcal{A} \mathcal{P}^D \mathcal{B} \mathcal{P}^D. \tag{10.7.23}$$

Evidently if one can move the middle factor $\mathcal{P}^D$ either to the left or to the right with impunity, then (7.21) and (7.22) become equalities.

The truncated counterparts of Lie operators are also not derivations. Consider, for example, the case of a 2-dimensional phase space. We have in accord with (5.3.7) the relation

$$: q : p^4 = (: q : p^2)p^2 + p^2 : q : p^2. \tag{10.7.24}$$

Suppose $D = 3$. Then the counterpart of the left side of (7.24) has the value

$$({}^3\!: q :)p^4 = (\mathcal{P}^3 : q : \mathcal{P}^3)p^4 = 0. \tag{10.7.25}$$

On the other hand, the counterpart of the right side of (7.24) is the quantity

$$[({}^3\!: q :)p^2]p^2 + p^2\ ({}^3\!: q :)p^2 = 4p^3. \tag{10.7.26}$$

Thus $({}^3\!: q :)$ is not a derivation. It follows from analogous calculations that no truncated Lie operator of the form $({}^D\!: f_1 :)$ is a derivation.

However, any $({}^D\!: f_\ell :)$ with $\ell \geq 2$ does at least enjoy the property

$$({}^D\!: f_\ell :) = \mathcal{P}^D : f_\ell : \mathcal{P}^D = \mathcal{P}^D : f_\ell : \text{ when } \ell \geq 2. \tag{10.7.27}$$

Let G_r^m be any monomial. From (7.6.16) and (7.15) we find the results

$$\begin{aligned}({}^D\!\!: f_\ell :)G_r^m &= \mathcal{P}^D : f_\ell : \mathcal{P}^D G_r^m = \mathcal{P}^D : f_\ell : \theta(d-m)G_r^m \\ &= \theta(D-m)\mathcal{P}^D \sum_{r'} c_{r'} G_{r'}^{m+\ell-2} \\ &= \theta(D-m)\theta(D+2-\ell-m)\sum_{r'} c_{r'} G_{r'}^{m+\ell-2}, \end{aligned} \tag{10.7.28}$$

$$\mathcal{P}^D : f_\ell : G_r^m = \mathcal{P}^D \sum_{r'} c_{r'} G_{r'}^{m+\ell-2} = \theta(D+2-\ell-m)\sum_{r'} c_{r'} G_{r'}^{m+\ell-2}. \tag{10.7.29}$$

Here the $c_{r'}$ are certain coefficients whose exact values need not concern us. But from (7.14) we conclude that

$$\theta(D-m)\theta(D+2-\ell-m) = \theta(D+2-\ell-m) \text{ for } \ell \geq 2. \tag{10.7.30}$$

Consequently, (7.27) holds when $\ell \geq 2$.

How close does $({}^D\!\!: f_\ell :)$ with $\ell \geq 2$ come to being a derivation? Let g and h be any two polynomials and suppose $\ell \geq 2$. We find from (4.3.7) and (7.7) the result

$$\begin{aligned}({}^D\!\!: f_\ell :)(gh) &= \mathcal{P}^D : f_\ell : (gh) = \mathcal{P}^D[(: f_\ell : g)h + g : f_\ell : h] \\ &= \mathcal{P}^D\{[({}^D\!\!: f_\ell :)g]h + g\ ({}^D\!\!: f_\ell :)h\} \text{ when } \ell \geq 2. \end{aligned} \tag{10.7.31}$$

We see in general that the factor of $\mathcal{P}^D$ cannot be removed from the right side of (7.31), and thus no $({}^D\!\!: f_\ell :)$ is a derivation. Finally, as the counter example (7.24) through (7.26) shows, neither (7.22) nor (7.31) hold when $\ell = 1$.

The stage is now set to study the question of convergence. Consider the operator $\hat{\mathcal{M}}$ defined by the equation

$$\hat{\mathcal{M}} = \exp[-\tau\ ({}^D\!\!: H :)], \tag{10.7.32}$$

which is the truncated analog of (7.2). Let us compute the matrix element $\langle G_s, \hat{\mathcal{M}} G_r\rangle$ with the assumption that

$$m(s), m(r) \leq D. \tag{10.7.33}$$

We find the result

$$\langle G_s, \hat{\mathcal{M}} G_r\rangle = \sum_{j=0}^{\infty} [(-\tau^j/j!]\langle G_s, ({}^D\!\!: H :)^j G_r\rangle. \tag{10.7.34}$$

Let ${}^D H$ be the matrix associated with $({}^D\!\!: H :)$,

$$({}^D H)_{sr} = \langle G_s,\ ({}^D\!\!: H :)G_r\rangle. \tag{10.7.35}$$

With this notation (7.17) takes the form

$$\langle G_s, \hat{\mathcal{M}} G_r\rangle = [\exp(-\tau\ {}^D H)]_{sr}. \tag{10.7.36}$$

Since ${}^D H$ is a finite dimensional matrix, the exponential series for $\exp(-\tau\ {}^D H)$ *converges* for all τ. See Section 3.7. It follows that all the matrix elements $\langle G_s, \hat{\mathcal{M}} G_r\rangle$, and in particular the matrix elements $\langle G_s, \hat{\mathcal{M}} G_a^1 \rangle$, are well defined for all H of the form (7.7) and any τ and any D.

What happens to the matrix elements in the limit $D \to \infty$ when truncation no longer occurs? We need to distinguish the two cases $H_1 = 0$ and $H_1 \neq 0$. Suppose $H_1 = 0$. Then, according to (7.27), $(^D\!: H :)$ has the property

$$(^D\!: H :) = \mathcal{P}^D : H : . \tag{10.7.37}$$

Consequently, from (7.16) and (7.37), $(^D\!: H :)$ also has the property

$$(^D\!: H :)^j = \mathcal{P}^D : H :^j \mathcal{P}^D = \; [^D(: H :^j)]. \tag{10.7.38}$$

Now make use of this property in either (7.32) or its series and matrix element equivalent. Doing so gives the result

$$\hat{\mathcal{M}} = \; ^D\mathcal{M}. \tag{10.7.39}$$

It follows that all matrix elements

$$(^D M)_{rs} = \langle G_r, \; ^D\mathcal{M} G_s \rangle = \langle G_r, \hat{\mathcal{M}} G_s \rangle \tag{10.7.40}$$

are well defined and are independent of D once D is large enough to satisfy (7.33).

In particular, suppose we wish to compute the coefficients in the Taylor series (7.5) through terms of degree D. Then we need the matrix elements $\langle G_s^m, \mathcal{M} G_a^1 \rangle$ for $m \leq D$. We know that all the constant terms k will vanish since we have assumed $H_1 = 0$. Also, all terms $H_\ell(z)$ in (7.1) with $\ell > (D+1)$ may be discarded since, in view of (7.6) and (7.6.16), they make no contribution to the desired terms having degree D and lower. Similarly, if we compute the terms $(: H :^j) z_a^i$ in (7.6) recursively by the relation

$$: H :^{j+1} z_a^i =: H : (: H :^j z_a^i) = [H, : H :^j z_a^i], \tag{10.7.41}$$

then we may discard at each step all those terms that would produce results having degree larger than D. Thus, all operations may be carried out within TPSA. Finally, since we know that the exponential series is convergent, it is only necessary to carry out the Poisson bracket operation (7.41) and the summation in (7.6) for successive values of j until some convergence criterion is met. Of course, all the caveats described in Section 4.1 concerning the use of Taylor series to evaluate the exponential function also apply here. Thus, as described in the next section, we may wish to consider approaches other than direct Taylor summation.

The case $H_1 \neq 0$ is more complicated. In this case there are simple examples for which not even the constant terms k in (7.5) are well defined. Consider the example of two-dimensional phase space and the Hamiltonian

$$\begin{aligned} H &= p^2/2 - (q+q_0)^4/2 + q_0^4/2 = -2qq_0^3 + (p^2/2 - 3q^2q_0^2) - 2q^3q_0 - q^4/2 \\ &= H_1 + H_2 + H_3 + H_4 \end{aligned} \tag{10.7.42}$$

where q_0 is some constant. Assume that

$$q_0 > 0, \tag{10.7.43}$$

and examine the trajectory with the initial conditions

$$t^i = q^i = p^i = 0. \tag{10.7.44}$$

From energy conservation we have the result

$$\dot{q} = [(q + q_0)^4 - q_0^4]^{1/2}, \tag{10.7.45}$$

and hence the time t along the trajectory is given by the integral

$$t(q) = \int_0^q dq'[(q' + q_0)^4 - q_0^4]^{-1/2}. \tag{10.7.46}$$

This integral is well defined and finite for all $q \geq 0$ including $q = +\infty$. That is, the trajectory reaches $q = \infty$ in *finite* time. Moreover, the trajectory also has infinite momentum at this time. Thus both k_1 and k_2 in (7.5) are divergent as τ approaches $t(\infty)$. Put other way, if in this example we set $\tau = t(\infty)$ in (7.36) and then let $D \to \infty$, we will get divergent results for at least some of the matrix elements, including the matrix elements that yield k_1 and k_2. Note that this nonexistence of at least some of the matrix elements of $\mathcal{M}$ is not due to any defect in the method of direct Taylor summation. It is inherent in $\mathcal{M}$, and must occur no matter how $\mathcal{M}$ is computed. We conclude that Hamiltonians for which $H_1 \neq 0$ must be handled with care and on a case by case basis.

Exercises

10.7.1.

10.8 Scaling, Splitting, and Squaring

Let us assume that H is autonomous and that $H_1 = 0$. Then we know that all matrix elements are in principle well defined. However, we also know from Section 4.1 that direct computation of $\exp(-\tau\ {}^D\!H)$ by simply summing the exponential series can be problematic. We therefore wish to explore alternatives.

One approach is to use scaling and squaring. In this section we will try to generalize, for the case of operators, the method used to exponentiate matrices in Section 4.1. The matrix elements corresponding to the Taylor coefficients in (7.5), and in fact all the matrix elements of ${}^D\!M$, can be computed reliably from the exponential series if τ is sufficiently small. Indeed, as described earlier, to compute the Taylor coefficients one simply carries out within TPSA the Poisson brackets indicated in (7.6) for successive values of j until some convergence criterion is met; and if τ is sufficiently small we know that this convergence criterion is met for modest values of j. However, successively squaring the resulting map is not so easy: We might consider computing the full matrix ${}^D\!M$ for a small (scaled) value of τ and then successively squaring the result. The full matrix ${}^D\!M$ has dimension $[S(D, d) + 1] \times [S(D, d) + 1]$, and this number can be very large. (See Section 7.9 and Table 7.2.) Thus, computing and successively squaring it requires considerable effort. Alternatively, one might consider squaring the map using the Taylor form (7.5). In this case one must successively substitute Taylor series into themselves. This process too might be quite time consuming. Yet another approach is to compute the Taylor series for the scaled map, factorize the scaled map, and then successively square the map in the factorized Lie form (7.6.3). This process might be faster.

At this point, and after some thought, one wonders if it might be possible to compute the Lie generators for the scaled map directly without going through the intermediate steps of computing the scaled Taylor map and then factorizing the result. If so, such an approach might both be considerably faster and require less storage. The first part of this section is devoted to describing such a procedure. It is then applied to scaling and squaring.

We know that, as a special case of the previous discussion in Section 10.5, the map (7.2) has the representation

$$\mathcal{M} = \exp(t : -H :) = \cdots \exp(: f_5 :) \exp(: f_4 :) \exp(: f_3 :)\mathcal{R}. \tag{10.8.1}$$

Here we have replaced the symbol τ by t since we will soon need τ for other purposes. We have also assumed $H_1 = 0$. Equation (8.1) may be viewed as a kind of *splitting* formula that writes $\exp(t : -H :)$ as a product of factors having desirable properties. (We will learn more about other splitting formulas in Section 10.10.) As such, it has three advantages: First, (as a consequence of the Factorization Theorem) its form is fixed and potentially exact. See Section 7.6. Second, it can be concatenated easily with other maps of the same form. See Section 8.4. Consequently, it can be squared repeatedly with relative ease. Third, the exact f_ℓ are *entire* (analytic everywhere except at ∞) functions of t, and have rapidly convergent Taylor expansions in t for small t.

We will now describe how to find the Taylor expansions (in t) for the f_ℓ. Let us begin with f_2, which is equivalent to determining $\mathcal{R}$. From (5.18), (5.31), and (5.32) we find the result

$$\mathcal{R} = \mathcal{M}_2 = \exp(: f_2 :) \tag{10.8.2}$$

with

$$f_2(z^i, t) = -tH_2(z^i). \tag{10.8.3}$$

Evidently f_2 is entire in t.

According to the equations of motion (5.60) through (5.66) for the remaining f_ℓ, we need to know the interaction Hamiltonian terms $H_m^{\rm int}$. Following (5.41), they are given by the relations

$$H_m^{\rm int}(z^i, t) = H_m(\mathcal{M}_2 z^i) = \mathcal{M}_2 H_m(z^i) = \exp(t : -H_2 :)H_m(z^i). \tag{10.8.4}$$

We note that because H is assumed to be autonomous, the time dependence of the $H_m^{\rm int}$ comes entirely from the factor $\exp(t : -H_2 :)$. From (8.4) we see that each $H_m^{\rm int}$ has the Taylor expansion

$$H_m^{\rm int} = \sum_{\ell=0}^{\infty} (1/\ell!)(-t)^\ell : H_2 :^\ell H_m. \tag{10.8.5}$$

From (8.4) and (5.25) through (5.35) we know that $H_m^{\rm int}$ can be written as well in the form

$$H_m^{\rm int} = H_m(\mathcal{M}_2 z^i) = H_m(Rz^i) \tag{10.8.6}$$

with

$$R = \exp(tJS). \tag{10.8.7}$$

We know that the matrix exponential function converges and therefore is analytic for all t. Finally, H_m is a polynomial in z^i. It follows that $H_m^{\rm int}$ is entire in t and consequently (8.5) converges for all t.

Let us next compute f_3. From (5.60) and the initial condition $f_3(0) = 0$ we obtain the result

$$f_3 = -\int_0^t dt' H_3^{\rm int}. \tag{10.8.8}$$

This integral can be done using the representation (8.5) to give the series

$$f_3 = \sum_{\ell=1}^{\infty} (1/\ell!)(-t)^\ell : H_2 :^{\ell-1} H_3. \tag{10.8.9}$$

Since the right side of (5.60), namely $H_3^{\rm int}$, is entire in t, it follows from Poincaré's Theorem 1.3.3 that f_3 is entire in t. More simply, we just observe that the integral of an entire function is also an entire function. Either way, we conclude that (8.9) converges for all t.

The computation of f_4 is a bit more involved. From (5.61) and the initial condition $f_4(0) = 0$ we find that f_4 contains two terms:

$$f_4 = -\int_0^t dt' H_4^{\rm int} - (1/2)\int_0^t dt' : f_3 : H_3^{\rm int}. \tag{10.8.10}$$

We will refer to the first term as the *direct* term since it is produced by $H_4^{\rm int}$, which is of the same degree as f_4. The second term will be called a *feed-up* term since it arises from the combined effect of the lower-degree term $H_3^{\rm int}$ and the lower-degree term f_3 (which comes from $H_3^{\rm int}$). Thus we write

$$f_4 = f_4^{\rm d} + f_4^{\rm fu}. \tag{10.8.11}$$

For $f_4^{\rm d}$ we use (8.5) to get a result analogous to that for f_3,

$$f_4^{\rm d} = \sum_{\ell=1}^{\infty} (1/\ell!)(-t)^\ell : H_2 :^{\ell} H_4. \tag{10.8.12}$$

Again we know that $f_4^{\rm d}$ is entire in t and the series (8.12) converges for all t. For the feed-up term we use the series representations (8.5) and (8.9) to get the result

$$\begin{aligned}
f_4^{\rm fu} &= -(1/2)\int_0^t dt' : f_3 : H_3^{\rm int} = -(1/2)\int_0^t dt' [f_3, H_3^{\rm int}] \\
&= -(1/2)\int_0^t dt' \sum_{\ell=1}^{\infty}\sum_{m=1}^{\infty} (1/\ell!)(1/m!)(-t')^{\ell+m} [: H_2^{\ell-1} : H_3, : H_2 :^m H_3] \\
&= (1/2)\sum_{\ell=1}^{\infty}\sum_{m=0}^{\infty} (1/\ell!)(1/m!)[1/(\ell+m+1)](-t)^{\ell+m+1} [: H_2 :^{\ell-1} H_3, : H_2 :^m H_3] \\
&= -(1/12)t^3 [H_3, : H_2 : H_3] + (1/24)t^4 [H_3, : H_2 :^2 H_3] - t^5\{(1/80)[H_3, : H_2 :^3 H_3] \\
&+ (1/120)[: H_2 : H_3, : H_2 :^2 H_3]\} + \cdots .
\end{aligned} \tag{10.8.13}$$

The quantity $f_4^{\rm fu}$ is also entire in t and the series (8.13) converges for all t.

The computation of the remaining f_j is similar. In each case there is a direct term analogous to (8.9) and (8.12). There are also feed-up terms arising from multiple Poisson brackets involving lower degree terms in the Hamiltonian. By contrast, there are no *feed-down* terms. To find a given f_ℓ, it is only necessary to know the $H_{\ell'}$ with $\ell' \leq \ell$. We also observe that all the formulas for the f_ℓ are expressible entirely in terms of Poisson brackets. All formulas involve only operations within the Poisson bracket Lie algebra generated by the H_m. Such results are to be expected in general as a consequence of the BCH theorem. Analogous formulas, but involving instead commutators of vector fields, are to be expected in the non-Hamiltonian case. The coefficients in the various series should be universal. Also, all the series represent entire functions and therefore converge for all values of t. Finally, we note that the rate of convergence of the various series depends only on the properties of $t : H_2 :$ (and hence tH_2), because that is the only term that appears infinitely often in the series.

To fix these ideas more clearly in the mind, we will also consider in some detail the computation of f_5. From (5.62) we find the result

$$f_5 = f_5^{\rm d} + f_5^{\rm fu} \tag{10.8.14}$$

where

$$f_5^{\rm d} = -\int_0^t dt' H_5^{\rm int} \tag{10.8.15}$$

and

$$f_5^{\rm fu} = \int_0^t dt' [: f_3 : (-H_4^{\rm int}) + (1/3) : f_3 :^2 (-H_3^{\rm int})]. \tag{10.8.16}$$

The direct term has the expansion

$$f_5^{\rm d} = \sum_{\ell=1}^{\infty} (1/\ell!)(-t)^\ell : H_2 :^{\ell-1} H_5. \tag{10.8.17}$$

To find the expansion for the feed-up term we insert the previously obtained expressions for $H_3^{\rm int}$, $H_4^{\rm int}$, f_3, and f_4 in (8.16) to obtain the result

$$\begin{aligned}
f_5^{\rm fu} \;&=\; \sum_{\ell=1}^{\infty}\sum_{m=0}^{\infty}(1/\ell!)(1/m!)[1/(\ell+m+1)](-t)^{\ell+m+1}[: H_2 :^{\ell-1} H_3, : H_2 :^m H_4] \\
&-\; (1/3)\sum_{\ell=1}^{\infty}\sum_{m=0}^{\infty}\sum_{n=1}^{\infty}(1/\ell!)(1/m!)(1/n!)[1/(\ell+m+n+1)] \\
&\;\; (-t)^{\ell+m+n+1}[: H_2 :^{\ell-1} H_3, [: H_2 :^m H_3, : H_2 :^{n-1} H_3]] \\
&=\; (1/2)t^2[H_3, H_4] - t^3\{(1/3)[H_3, : H_2 : H_4] + (1/6)[: H_2 : H_3, H_4]\} \\
&+\; t^4\{(-1/24)[H_3, [: H_2 : H_3, H_3]] + (1/8)[H_3, : H_2 :^2 H_4] \\
&+\; (1/8)[: H_2 : H_3, : H_2 : H_4] + (1/24)[: H_2 :^2 H_3, H_4]\} \\
&+\; t^5\{(1/45)[H_3, [: H_2 :^2 H_3, H_3]] - (1/30)[H_3, : H_2 :^3 H_4] \\
&+\; (1/60)[: H_2 : H_3, [: H_2 : H_3, H_3]] - (1/20)[: H_2 : H_3, : H_2 :^2 H_4] \\
&-\; (1/30)[: H_2 :^2 H_3, : H_2, H_4 :] - (1/120)[: H_2 :^3 H_3, H_4]\} + \cdots .
\end{aligned} \tag{10.8.18}$$

At this point we might quote formulas for the f_j for the next few higher values of j and up to some power in t. Instead we observe that, rather than the reverse factorization (8.1) which we write in the form

$$\mathcal{M} = \exp(t : -H :) = \cdots \exp[: f_5(t) :] \exp[: f_4(t) :] \exp[: f_3(t) :]\mathcal{R}(t), \tag{10.8.19}$$

it is often more convenient to have results for the forward factorization

$$\mathcal{M} = \exp(t : -H :) = \mathcal{R}'(t) \exp[: g_3(t) :] \exp[: g_4(t) :] \exp[: g_5(t) :] \cdots . \tag{10.8.20}$$

The relation between the two factorizations is immediate. Suppose we invert both sides of (8.19). Doing so gives the result

$$\exp(t : +H :) = \mathcal{R}^{-1}(t) \exp[- : f_3(t) :] \exp[- : f_4(t) :] \exp[- : f_5(t) :] \cdots . \tag{10.8.21}$$

Next replace t by $-t$ in (8.21) to find the relation

$$\exp(t : -H :) = \mathcal{R}^{-1}(-t) \exp[- : f_3(-t) :] \exp[- : f_4(-t) :] \exp[- : f_5(-t) :] \cdots . \tag{10.8.22}$$

Since the factorization (8.20) is unique, comparison of the first factors on the right sides of (8.20) and (8.22) shows that

$$\mathcal{R}'(t) = \mathcal{R}^{-1}(-t). \tag{10.8.23}$$

But from (8.2) and (8.3) we find that

$$\mathcal{R}^{-1}(-t) = \mathcal{R}(t) \tag{10.8.24}$$

so that

$$\mathcal{R}'(t) = \mathcal{R}(t). \tag{10.8.25}$$

This result is also evident on general grounds. Finally, upon comparing the remaining factors in (8.20) and (8.22), we conclude that

$$g_m(t) = -f_m(-t). \tag{10.8.26}$$

We now quote formulas, obtained by symbolic manipulation, for the first few g_m through terms of order t^5. Again each g_m is the sum of a direct and a feed-up term,

$$g_m = g_m^{\rm d} + g_m^{\rm fu}. \tag{10.8.27}$$

For the direct terms we have in general the formula

$$g_m^{\rm d} = -\sum_{\ell=1}^{\infty} (1/\ell!) t^\ell : H_2 :^{\ell-1} H_m. \tag{10.8.28}$$

For the feed-up terms $g_3^{\rm fu}$ through $g_6^{\rm fu}$ we find, through terms of order t^5, the result

$$g_3^{\rm fu} = 0, \tag{10.8.29}$$

$$\begin{aligned}
g_4^{\rm fu} &= -(1/2)\sum_{\ell=1}^{\infty}\sum_{m=0}^{\infty}(1/\ell!)(1/m!)[1/(\ell+m+1)]t^{\ell+m+1}[:H_2:^{\ell-1}H_3, :H_2:^{m}H_3] \\
&= (1/12)t^3[:H_2:H_3, H_3] + (1/24)t^4[:H_2:^2H_3, H_3] \\
&+ t^5\{(1/80)[:H_2:^3H_3, H_3] + (1/120)[:H_2:^2H_3, :H_2:H_3]\} + \cdots, \qquad (10.8.30)
\end{aligned}$$

$$\begin{aligned}
g_5^{\rm fu} &= -(1/2)t^2[H_3, H_4] - t^3\{(1/3)[H_3, :H_2:H_4] + (1/6)[:H_2:H_3, H_4]\} \\
&- t^4\{(-1/24)[H_3, [:H_2:H_3, H_3]] + (1/8)[H_3, :H_2:^2H_4] \\
&+ (1/8)[:H_2:H_3, :H_2:H_4] + (1/24)[:H_2:^2H_3, H_4]\} \\
&+ t^5\{(1/45)[H_3, [:H_2:^2H_3, H_3]] - (1/30)[H_3, :H_2:^3H_4] \\
&+ (1/60)[:H_2:H_3, [:H_2:H_3, H_3]] - (1/20)[:H_2:H_3, :H_2:^2H_4] \\
&- (1/30)[:H_2:^2H_3, :H_2, H_4:] - (1/120)[:H_2:^3H_3, H_4]\} + \cdots, \qquad (10.8.31)
\end{aligned}$$

$$\begin{aligned}
g_6^{\rm fu} &= -(1/2)t^2[H_3, H_5] \\
&+ t^3\{(-1/6)[H_3, [H_3, H_4]] - (1/3)[H_3, :H_2:H_5] \\
&- (1/3)[H_3, :H_2:H_5] - (1/6)[:H_2:H_3, H_5] \\
&+ (1/12)[:H_2:H_4, H_4]\} \\
&- t^4\{(1/8)[H_3, [H_3, :H_2:H_4]] + (1/6)[H_3, [:H_2:H_3, H_4]] \\
&+ (1/8)[H_3:H_2:^2H_5] - (1/48)[H_4, [:H_2:H_3, H_3]] \\
&+ (1/16)[:H_2:H_3, [H_3, H_4]] + (1/8)[:H_2:H_3, :H_2:H_5] \\
&+ (1/24)[:H_2:^2H_3, H_5] - (1/24)[:H_2:^2H_4, H_4]\} \\
&+ t^5\{(1/80)[H_3, [H_3, [:H_2:H_3, H_3]]] - (1/20)[H_3, [H_3, :H_2:^2H_4]] \\
&- (1/20)[H_3, [:H_2:H_3, :H_2:H_4]] - (1/60)[H_3, [:H_2:^2H_3, H_4]] \\
&- (1/30)[H_3, :H_2:^3H_5] + (1/80)[H_4, [:H_2:^2H_3, H_3]] \\
&- (1/20)[:H_2:H_3, [H_3, :H_2:H_4]] - (1/40)[:H_2:H_3, [:H_2:H_3, H_4]] \\
&- (1/20)[:H_2:H_3, :H_2:^2H_5] + (1/240)[:H_2:H_4, [:H_2:H_3, H_3]] \\
&- (1/60)[:H_2:^2H_3, [H_3, H_4]] - (1/30)[:H_2:^2H_3, :H_2:H_5] \\
&+ (1/20)[:H_2:^2H_4, :H_2:H_4] - (1/120)[:H_2:^3H_3, H_5] \\
&+ (1/180)[:H_2:^3H_4, H_4]\} + \cdots. \qquad (10.8.32)
\end{aligned}$$

These results were obtained using a Mathematica program. See Appendix E.

We have derived the splitting formula (8.1) with the f_m given by (8.9), (8.11) through (8.18), and analogous expressions. Equivalently, we have also derived the splitting formula (8.20) with the g_m given by (8.27) through (8.32) and analogous expressions. How are these formulas to be used? With the aid of scaling and squaring we have the result

$$\begin{aligned}
\mathcal{M} &= \exp(t:-H:) = \{\exp[(t/2^n):-H:]\}^{2^n} \\
&= \{\cdots\{\{\exp[(t/2^n):-H:]\}^2\}^2\cdots\}^2 \ (n \text{ squarings}). \qquad (10.8.33)
\end{aligned}$$

See Section 4.1 and (4.1.6). Now define a quantity τ by writing

$$\tau = t/2^n. \qquad (10.8.34)$$

Next insert, for example, the splitting formula (8.20) into (8.33). Doing so gives the result

$$\begin{aligned}\mathcal{M} &= \exp(t:-H:) = \mathcal{R}(t)\exp[:g_3(t):]\exp[:g_4(t):]\exp[:g_5(t):]\cdots \qquad (10.8.35)\\ &= \{\mathcal{R}(\tau)\exp[:g_3(\tau):]\exp[:g_4(\tau):]\exp[:g_5(\tau):]\cdots\}^{2^n}\\ &= \{\cdots\{\{\mathcal{R}(\tau)\exp[:g_3(\tau):]\exp[:g_4(\tau):]\exp[:g_5(\tau):]\cdots\}^2\}^2\cdots\}^2 \text{ (n squarings).}\end{aligned}$$

Finally, suppose we choose n to be large enough so that τ is sufficiently small that the truncated series (8.28) through (8.32) give accurate results for the $g_m(\tau)$. Then (8.35) gives an accurate result for $\mathcal{M}$.

How large must n be (or, equivalently, how small must τ be) for the truncated series to give accurate results for the $g_m(\tau)$? We have already observed that the convergence of these series depends only on the properties of τH_2. Let us explore what can be said about terms of the form $(\tau:H_2:)^\ell H_m$, which are the common ingredient of all the series. As before we write H_2 in a form analogous to (5.28) except that S is now a constant time-independent matrix. We then find, in analogy to (5.29), the relation

$$:H_2:z_a = -\sum_b (JS)_{ab}z_b. \qquad (10.8.36)$$

Consider the matrix $(-JS)^T$. In general, barring degeneracy, it has $2n$ eigenvectors v^j with eigenvalues λ_j:

$$(-JS)^T v^j = \lambda_j v^j. \qquad (10.8.37)$$

(Here, for the moment, n is the number of degrees of freedom, and *not* the number of squarings.) Define $2n$ first-degree polynomials h_1^j by the rule

$$h_1^j = \sum_a v_a^j z_a. \qquad (10.8.38)$$

In general, again barring degeneracy, the h_1^j will be functionally independent and will span the space of all first-degree polynomials. Let us compute the effect of $:H_2:$ on the h_1^j. From (8.36) through (8.38) we find the result

$$\begin{aligned}:H_2:h_1^j &= \sum_a v_a^j:H_2:z_a = \sum_{a,b} v_a^j(-JS)_{ab}z_b\\ &= \sum_b\{\sum_a[(-JS)^T]_{ba}v_a^j\}z_b = \lambda_j\sum_b v_b^j z_b\\ &= \lambda_j h_1^j. \qquad (10.8.39)\end{aligned}$$

We know from (7.6.14) that the (linear) operator $:H_2:$ maps the space $\mathcal{P}_m$ into itself. (Note that here $\mathcal{P}_m$ denotes a vector space and not a projection operator.) What is its largest eigenvalue when acting on this space? Consider for example, the degree 3 homogeneous polynomials h_3^{ijk} defined by the relation

$$h_3^{ijk} = h_1^i h_1^j h_1^k. \qquad (10.8.40)$$

In general, these polynomials will span the space of third-degree polynomials. Since $: H_2 :$ is a derivation, and in view of (8.39), we find the result

$$\begin{aligned} : H_2 : h_3^{ijk} &= (: H_2 : h_1^i)h_1^j h_1^k + h_1^i(: H_2 : h_1^j)h_1^k + h_1^i h_1^j(: H_2 : h_1^k) \\ &= (\lambda_i + \lambda_j + \lambda_k)h_3^{ijk}. \end{aligned} \quad (10.8.41)$$

Let $\lambda_{\max}$ be the modulus of the eigenvalue with the largest absolute value,

$$\lambda_{\max} = \max_j \; |\lambda_j|. \quad (10.8.42)$$

We conclude from relations of the form (8.41) that the eigenvalues of $: H_2 :$ when acting on $\mathcal{P}_m$ are bounded by the quantity $(m\lambda_{\max})$.

In general, given the matrix $(-JS)^T$, one can compute its eigenvalues. However, this computation requires some work, and often an estimate that requires less computation is sufficient. Let λ_k be the eigenvalue of $(-JS)^T$ having the largest absolute value. Then from (8.39) we have the result

$$\lambda_k v^k = (-JS)^T v^k = (SJ)v^k. \quad (10.8.43)$$

Here we have used the fact that S is symmetric and J is antisymmetric. Now take norms of both sides of (8.43) to get the result

$$\|\lambda_k v^k\| = \|SJv^k\| \leq \|SJ\|\|v^k\|. \quad (10.8.44)$$

The left side of (8.44) can be manipulated using (3.7.8) and (8.42) to give the relation

$$\|\lambda_k v^k\| = |\lambda_k|\|v^k\| = \lambda_{\max}\|v^k\|, \quad (10.8.45)$$

and we conclude upon comparison with (8.44) that $\lambda_{\max}$ has the bound

$$\lambda_{\max} \leq \|SJ\|. \quad (10.8.46)$$

Moreover, we may also write the equation

$$SJ = -JJSJ. \quad (10.8.47)$$

Now take the norm of both sides of (8.47) to get the bound

$$\|SJ\| = \| - JJSJ\| = \|JJSJ\| \leq \|J\|\|JS\|\|J\|. \quad (10.8.48)$$

Suppose the matrix norm $\| \;\|$ to be employed has the property

$$\|J\| = 1, \quad (10.8.49)$$

which is true of the norm (3.7.15). Then (8.46) through (8.49) may be combined to give the bound

$$\lambda_{\max} \leq \|JS\|. \quad (10.8.50)$$

We now have all the tools in hand to examine the convergence of series that involve terms of the form $(\tau : H_2 :)^\ell H_m$. According to the previous discussion the convergence of such series is governed by the size of the terms $(m\tau\lambda_{\rm max})^\ell$ with $\lambda_{\rm max}$ bounded by (8.50). Suppose, for example, we truncate the series (8.28) for $g_m^d(\tau)$ by selecting some N and discarding all terms with $\ell > N$. Let us estimate the error committed in doing so by examining the size of the first neglected term,

$$\text{first neglected term } = -[1/(N+1)!]\tau^{N+1} : H_2 :^N H_m. \tag{10.8.51}$$

We know from our previous discussion that $(: H_2 :^N H_m)$ behaves at worst according to the estimate

$$: H_2 :^N H_m \sim (m\lambda_{\rm max})^N H_m. \tag{10.8.52}$$

Also, we expect from the first term in (8.28) that $g_m^d(\tau)$ itself will be of order (τH_m). Consequently, using (8.51) and (8.52), we should make the comparison

$$\tau H_m \overset{?}{\leftrightarrow} [1/(N+1)!]\tau^{N+1}(m\lambda_{\rm max})^N H_m. \tag{10.8.53}$$

We conclude that the *relative* error in computing $g_m^d(\tau)$, and hence also $\mathcal{M}(t)$, has the estimate

$$\text{relative error } \sim [1/(N+1)!](\tau m\lambda_{\rm max})^N. \tag{10.8.54}$$

Finally, let us define a quantity λ by the rule

$$\lambda = t\lambda_{\rm max}. \tag{10.8.55}$$

By (8.50) it has the estimate

$$\lambda \le \|tJS\|, \tag{10.8.56}$$

and is dimensionless. With the aid of (8.34), (8.54), and (8.55) we see that the relative error can be written in the form

$$\text{relative error } \sim [1/(N+1)!](m\lambda/2^n)^N. \tag{10.8.57}$$

Suppose, for example, we set $N = 5$, limit our attention to the cases $m \le 8$, and select n such that

$$(8\lambda/2^n) < (1/20). \tag{10.8.58}$$

[Note that λ and hence the required n can be computed in advance using (8.58).] Then we find from (8.57) that the relative error has the estimate

$$\text{relative error } \sim (1/6!)(1/20)^5 \simeq 4 \times 10^{-10}. \tag{10.8.59}$$

Although we have only estimated the error due to truncating the series for g_m^d, we expect that the result of truncating the other series at the same N will be comparable as long as (8.58) is satisfied. We conclude from (8.59) that (just as scaling and squaring works well for matrix exponentiation) scaling, splitting, and squaring works well for computing $\mathcal{M}$ in the factorized product forms (8.1) and (8.20). It has high, controllable, and predictable accuracy. Of course, the n required to satisfy a relation of the form (8.58) varies from Hamiltonian to Hamiltonian. However, just as in the matrix case, the n required to achieve some specific accuracy grows only logarithmically with the norm of (tJS), and for any given

Hamiltonian the accuracy increases very rapidly for increasing n. Correspondingly, because the number of required operations is relatively small and no cancellations are required to occur between large terms, problems with round-off error are minimized. Finally, with regard to computational speed, the method of scaling, splitting, and squaring is far faster than numerical integration. (Of course, numerical integration is required in the nonautonomous case.) It is also faster and, as expected, far more reliable than direct use of the Taylor series (7.6). The price that has been paid for this good performance is that one must know the expansions of the form (8.28) through (8.32) as well as concatenation formulas of the form (8.4.31) through (8.4.36). By contrast, the implementation of the Taylor series (7.6) to any order is straightforward.

Exercises

10.8.1. Show that if the matrix norm has the property (8.49), then one has the equation

$$\|JS\| = \|SJ\| = \|S\|. \tag{10.8.60}$$

10.9 Canonical Treatment of Errors

Let $H(z,t)$ be a general, possibly time-dependent, Hamiltonian that is analytic about the origin and consequently has an expansion in homogeneous polynomials in z of the form

$$H(z,t) = \sum_{m=1}^{\infty} H_m(z,t). \tag{10.9.1}$$

(Here we drop a possible z independent term H_0 since it has no effect on the equations of motion.) Moreover, we shall assume that H_1 is *small* so that $z = 0$ is close to being a solution to the equation of motion generated by H. Such Hamiltonians often arise in connection with the description of errors. For example, we will see in Chapter 26 that both mispowered dipole bending magnets and dipole steering magnets are described by Hamiltonians of this form.

Hamiltonians of the form (9.1) with H_1 small can be treated by the method of Section 10.5 and, in the autonomous case, also by the method of Section 10.7. The method of Section 10.5 requires determination of the design trajectory $z^d(t)$, and then provides an expansion about this trajectory. However, since $z = 0$ is nearly a trajectory, we may prefer an expansion of the transfer map $\mathcal{M}$ about $z = 0$. Such an expansion is provided, in the autonomous case, by the method of Section 10.7, but requires the summation of Taylor series for the exponential function. We have seen that the use of such series may be problematic.

The purpose of this section is to develop a method for expanding the transfer map about $z = 0$ under the assumption that H_1 is small. In essence, we will produce a simultaneous expansion both in z and in powers of some parameter that characterizes the smallness of H_1. This method is applicable to both the time dependent and time independent cases.

The method is based on an *enlargement* of $2n$ dimensional phase space to include the extra variables q_{n+1} and p_{n+1}. This is the same enlargement that was used in Section 9.4, and the method based on it will be referred to as a *canonical* treatment of errors. As before, let us use the symbol $\hat{z}$ to denote the coordinates in this enlarged phase space,

$$\hat{z} = (q_1 \cdots q_n, q_{n+1}, p_1 \cdots p_n, p_{n+1}). \tag{10.9.2}$$

Next, modify the Hamiltonian (9.1) to obtain the Hamiltonian $\hat{H}$ defined by the rule

$$\hat{H}(\hat{z}, t) = (q_{n+1})H_1(z, t) + \sum_{m=2}^{\infty} H_m(z, t). \tag{10.9.3}$$

We will see that the transfer map for the Hamiltonian $\hat{H}$, which we will call $\hat{\mathcal{M}}$, can be computed using the methods we have developed previously, and that $\hat{\mathcal{M}}$ contains the information we seek.

We being by noting that $\hat{z} = 0$ is a trajectory for $\hat{H}$, and hence the transfer map $\hat{\mathcal{M}}$ produced by $\hat{H}$ maps the origin of the enlarged phase space into itself. Indeed, with respect to the enlarged phase-space variables, the Hamiltonian $\hat{H}$ has the expansion

$$\hat{H}(\hat{z}, t) = \sum_{m=2}^{\infty} \hat{H}_m(\hat{z}, t) \tag{10.9.4}$$

with

$$\hat{H}_2(\hat{z}, t) = (q_{n+1})H_1(z, t) + H_2(z, t), \tag{10.9.5}$$

$$\hat{H}_m(\hat{z}, t) = H_m(z, t) \text{ for } m > 2. \tag{10.9.6}$$

Observe that this expansion begins with homogeneous polynomials of degree *two*. Correspondingly, the transfer map $\hat{\mathcal{M}}$ can be written in the factored product form

$$\hat{\mathcal{M}} = \hat{\mathcal{R}} \exp(:\hat{f}_3:) \exp(:\hat{f}_4:) \exp(:\hat{f}_5:) \cdots . \tag{10.9.7}$$

Here the $\hat{f}_m$ denote homogeneous polynomials of degree m in the enlarged phase-space variables $\hat{z}$. The map $\hat{\mathcal{M}}$ can be computed in both the time dependent and time independent cases using the method of Section 10.5, and in the time independent case using the methods of Sections 10.7 and 10.8.

By construction $\hat{H}$ is independent of p_{n+1}, and therefore there is the obvious equation of motion

$$\dot{q}_{n+1} = \partial \hat{H}/\partial p_{n+1} = 0. \tag{10.9.8}$$

It follows that $\hat{\mathcal{M}}$ leaves q_{n+1} unchanged,

$$q^f_{n+1} = \hat{\mathcal{M}} q^i_{n+1} = q^i_{n+1}. \tag{10.9.9}$$

But now the reasoning presented in the beginning of Section 9.4 applies. We conclude that $\hat{\mathcal{R}}$ must satisfy the relation

$$\hat{\mathcal{R}} q_{n+1} = q_{n+1}, \tag{10.9.10}$$

and its associated matrix $\hat{R}$ must (for the case $n = 3$) be of the general form (9.4.84). Also, the $\hat{f}_m$ in (9.7) must be independent of p_{n+1},

$$\partial \hat{f}_m/\partial p_{n+1} = 0. \tag{10.9.11}$$

They will depend only on the z^i and q^i_{n+1}. Finally, we see that the relation

$$\hat{z}^f = \hat{\mathcal{M}}\hat{z}^i \tag{10.9.12}$$

provides an expansion of z^f in terms of z^i and q^i_{n+1}, and p^i_{n+1} does not occur in this expansion.

Evidently the size of q^i_{n+1} governs the effect of the term $(q_{n+1})H_1$ in (9.3), and an expansion in powers of q_{n+1} is, in effect, an expansion in terms of powers of H_1. Thus, after some final result for $\hat{\mathcal{M}}$ (or some consequence of $\hat{\mathcal{M}}$) has been obtained as a series in q_{n+1} up to some order, we may set $q_{n+1} = 1$ in this series to obtain a result appropriate to the original Hamiltonian (9.1) when H_1 in this Hamiltonian is treated as a perturbation through the same order.

A simple example helps clarify this approach. Consider the *displaced* one-dimensional harmonic oscillator described by the Hamiltonian

$$H = (p_1^2 + q_1^2)/2 + \delta q_1 \tag{10.9.13}$$

where δ is a small quantity, not to be confused with the δ_a in Section 9.4. It is easily verified that the equations of motion associated with (9.13) have the solution

$$q_1(t) = -\delta + (q_1^i + \delta)\cos t + p_1^i \sin t, \tag{10.9.14}$$

$$p_1(t) = -(q_1^i + \delta)\sin t + p_1^i \cos t, \tag{10.9.15}$$

where the initial time is taken to be $t^i = 0$. According to (9.3) the modified Hamiltonian associated with (9.13) is given by the relation

$$\hat{H} = (p_1^2 + q_1^2)/2 + \delta q_1 q_2. \tag{10.9.16}$$

Since $\hat{H}$ is time independent, and all the $\hat{H}_m$ with $m > 2$ happen to vanish in this simple case, we have the immediate result

$$\hat{\mathcal{M}} = \exp(-t : \hat{H} :) = \exp(-t : \hat{H}_2 :) = \hat{\mathcal{R}}(t). \tag{10.9.17}$$

The map $\hat{\mathcal{R}}(t)$ is in turn described by the matrix $\hat{R}(t)$ given by

$$\hat{R} = \exp(t\hat{J}\hat{S}). \tag{10.9.18}$$

Here we have placed a hat over J to indicate that the 4×4 J is to be used. Also, for convenience, we have used the ordering (9.4.15). Correspondingly, $\hat{J}$ is of the form (3.2.10). With this convention, and according to (5.28), $\hat{S}$ is the matrix

$$\hat{S} = \begin{pmatrix} 1 & 0 & \delta & 0 \\ 0 & 1 & 0 & 0 \\ \delta & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{10.9.19}$$

The exponentiation (9.18) can be carried out to give the result

$$\hat{R} = \begin{pmatrix} \cos t & \sin t & \delta(\cos t - 1) & 0 \\ -\sin t & \cos t & -\delta \sin t & 0 \\ 0 & 0 & 1 & 0 \\ -\delta \sin t & \delta(\cos t - 1) & \delta^2(t - \sin t) & 1 \end{pmatrix}. \tag{10.9.20}$$

Correspondingly, in analogy with (5.27), we find the relations

$$q_1(t) = q_1^i \cos t + p_1^i \sin t + q_2^i \delta(\cos t - 1), \tag{10.9.21}$$

$$p_1(t) = -q_1^i \sin t + p_1^i \cos t - q_2^i \delta \sin t, \tag{10.9.22}$$

$$q_2(t) = q_1^i, \tag{10.9.23}$$

$$p_2(t) = -q_1^i \delta \sin t + p_1^i \delta(\cos t - 1) + q_2^i \delta^2 (t - \sin t) + p_2^i. \tag{10.9.24}$$

We see, as advertised, that (9.21) and (9.22) reduce to (9.14) and (9.15), respectively, when we set $q_2 = 1$. Also, (9.23) is consistent with (9.8) and (9.9). Finally, although of no particular interest for our purposes, (9.24) is consistent with the equation of motion

$$\dot{p}_2 = -\partial \hat{H} / \partial q_2 = -\delta q_1. \tag{10.9.25}$$

In this example we have been able to solve for the map $\hat{\mathcal{M}}$ exactly, and have found that the result for $z(t)$ when q_{n+1} is set to one agrees with the corresponding trajectory, which we were also able to find exactly, for the original problem. For most cases we will only compute $\hat{\mathcal{M}}$ to some order in the phase-space variables including the variable q_{n+1}. In such cases we expect that the correspondingly results for $z(t)$ will be correct through the same order in δ where δ measures the size of H_1.

There is one last observation to be made. In the example just described the computation of $\hat{R}$ was somewhat complicated because it was 4×4, and the information about $p_2(t)$, which was contained in $\hat{R}$, was of no interest. Is there a way of bypassing such complications? In some cases the answer is yes. For the example problem let us rewrite (9.17) in the form

$$\hat{\mathcal{R}}(t) = \exp(: \hat{f}_2 :) \tag{10.9.26}$$

with $\hat{f}_2$ defined by

$$\hat{f}_2 = -t\hat{H}_2 = -t[(p_1^i)^2 + (q_1^i)^2]/2 - t\delta q_1^i q_2^i. \tag{10.9.27}$$

When $\hat{f}_2$ is regarded as a function of q_1^i and p_1^i, which are the dynamical variables of interest, we see that it can be written in the form

$$\hat{f}_2 = f_1^{(2)} + f_2^{(2)} \tag{10.9.28}$$

where $f_1^{(2)}$ is a homogeneous polynomial of degree 1 in the variables of interest,

$$f_1^{(2)} = -t\delta q_2^i q_1^i, \tag{10.9.29}$$

and $f_2^{(2)}$ is a polynomial of degree 2 in these variables,

$$f_2^{(2)} = -t[(p_1^i)^2 + (q_1^i)^2]/2. \tag{10.9.30}$$

Here we have used a subscript on f to indicate degree in the variables of interest, and a superscript to indicate overall degree. This notation is similar to that of Section 9.3 where the subscript served the same purpose and the superscript indicated overall grade. Indeed, at this stage, we may simply view q_2^i as a parameter that plays the same role as ϵ did in Section 9.3.

With the decomposition (9.28) we may rewrite $\hat{\mathcal{M}}$, which we now simply call $\mathcal{M}$ because we are no longer treating q_2 and p_2 as dynamical variables, in the form

$$\mathcal{M} = \exp(: f_1^{(2)} + f_2^{(2)} :). \tag{10.9.31}$$

At this stage we use (9.2.4) to rewrite $\mathcal{M}$ in the form

$$\mathcal{M} = \exp(: f_2 :) \exp(: f_1 :) \tag{10.9.32}$$

where, according to (9.2.7) and (9.2.9),

$$f_2 = f_2^{(2)} = -(t/2)[(p_1^i)^2 + (q_1^i)^2], \tag{10.9.33}$$

$$f_1 = \text{iex}\ (- : f_2^{(2)} :) f_1^{(2)}. \tag{10.9.34}$$

Simple calculation gives the result

$$\begin{aligned} \exp(-\tau : f_2^{(2)} :) f_1^{(2)} &= \exp[(\tau t/2) : (p_1^i)^2 + (q_1^i)^2 :](-t\delta q_2^i q_1^i) \\ &= -t\delta q_2^i \exp[(\tau t/2) : (p_1^i)^2 + (q_1^i)^2 :] q_1^i \\ &= -t\delta q_2^i [q_1^i \cos(t\tau) - p_1^i \sin(t\tau)]. \end{aligned} \tag{10.9.35}$$

Consequently, we find for f_1 the explicit result

$$\begin{aligned} f_1 &= \text{iex}\ (- : f_2^{(2)} :) f_1^{(2)} = \int_0^1 d\tau \exp(-\tau : f_2^{(2)} :) f_1^{(2)} \\ &= -t\delta q_2^i \int_0^1 d\tau [q_1^i \cos(t\tau) - p_1^i \sin(t\tau)] \\ &= -\delta q_2^i [q_1^i \sin t + p_1^i (\cos t - 1)]. \end{aligned} \tag{10.9.36}$$

Here we have used (8.7.9). Let us now find the effect of $\mathcal{M}$, with $\mathcal{M}$ given by (9.32), when it acts on q_1^i and p_1^i. We find for exp(: f_1 :) and exp(: f_2 :) the results

$$\exp(: f_1 :) q_1^i = q_1^i + \delta q_2^i (\cos t - 1), \tag{10.9.37}$$

$$\exp(: f_1 :) p_1^i = p_1^i - \delta q_2^i \sin t, \tag{10.9.38}$$

$$\exp(: f_2 :) q_1^i = q_1^i \cos t + p_1^i \sin t, \tag{10.9.39}$$

$$\exp(: f_2 :) p_1^i = -q_1^i \sin t + p_1^i \cos t. \tag{10.9.40}$$

Consequently we find for $\mathcal{M}$ the net results

$$q_1(t) = \mathcal{M} q_1^i = \delta q_2^i (\cos t - 1) + q_1^i \cos t + p_1^i \sin t, \tag{10.9.41}$$

$$p_1(t) = \mathcal{M} p_1^i = -\delta q_2^i \sin t - q_1^i \sin t + p_1^i \cos t. \tag{10.9.42}$$

Let us compare these results with those of (9.14) and (9.15), which can be written in the form

$$q_1(t) = \delta(\cos t - 1) + q_1^i \cos t + p_1^i \sin t, \tag{10.9.43}$$

$$p_1(t) = -\delta \sin t - q_1^i \sin t + p_1^i \cos t. \tag{10.9.44}$$

We see that (9.41) and (9.43), and (9.42) and (9.44), agree when we set $q_2^i = 1$. Note that, had we wished, we could have set $q_2^i = 1$ at an earlier stage before carrying out the calculations (9.37) through (9.42).

There is a moral to be learned from this last exercise. Quite generally, we see that once $\hat{\mathcal{M}}$ is computed in the form (9.7) with the aid of the auxiliary variable q_{n+1}, then each $\hat{f}_m$ may be decomposed in the form

$$\hat{f}_m = f_m^m(z) + (q_{n+1})f_{m-1}^m(z) + \cdots + (q_{n+1})^m f_0^m. \tag{10.9.45}$$

Here, as in Section 9.4, the superscript m on f_ℓ^m indicates that the quantity is associated with $\hat{f}_m$, and the subscript ℓ indicates that the quantity is homogeneous of degree ℓ in the variables z. After this is done, we may view q_{n+1} as playing the role of ϵ and use the calculus of Section 9.3 to manipulate the $(q_{n+1})^{m-\ell} f_\ell^m$ to produce a map of the form

$$\mathcal{M} = \exp(:f_1:)\mathcal{R}\exp(:f_3:)\exp(:f_4:)\cdots, \tag{10.9.46}$$

and then set $q_{n+1} = 1$ in all the f_m to obtain a final map that involves only the variables z. Better yet, we may simply use the shrinker of Section 9.4 to obtain $\mathcal{M}$ from $\hat{\mathcal{M}}$.

Exercises

10.9.1. Verify that (9.20) is of the form (9.4.46).

10.9.2. The map (9.32), with f_1 and f_2 given by (9.33) and (9.36), is written in reverse factorized form. See Section 7.8. Rewrite the map in forward factorized form. See (9.2.30). Before doing so, set $q_2^i = 1$. Verify that use of the forward factorized map also gives (9.43) and (9.44).

10.10 Wei-Norman and Fer Methods

10.10.1 Wei-Norman Equations

Exercises

10.10.1. Exercise on Wei-Norman equations.

10.10.2 Accelerated Procedure: The Fer Expansion

10.11 Symplectic Integration

Symplectic numerical integration methods, like the numerical integration methods described in Chapter 2, seek to compute trajectories accurately through some order in the time step h or with some prescribed over all accuracy. However, they are special in that they are designed for Hamiltonian systems and seek to satisfy the requirement that the resulting transfer map $\mathcal{M}$ between initial and final conditions be exactly (to machine precision) symplectic.

Many aspects of the construction of symplectic integrators involve map-like methods including Zassenhaus formulas. Moreover, for some symplectic integration methods, in the course of computing a trajectory it is also possible to compute in a natural way the transfer map about this trajectory. Thus, symplectic integration and symplectic maps are closely related, and their discussion could logically form part of this chapter. However, the subject is sufficiently vast to deserve a chapter if its own, and will be treated in Chapter 12.

10.12 Taylor Methods and the Complete Variational Equations

The work of the previous sections dealt for the most part with Hamiltonian systems and the representation of their associated *symplectic* transfer maps $\mathcal{M}$ by Lie transformations. In this section we will work with differential equations that are not necessarily Hamiltonian in form. This may occur if the dynamical system under consideration is intrinsically non-Hamiltonian. Also, there is the possibility that the dynamical system does have a Hamiltonian description, but we do not wish to use it. For example, we may want to consider charged-particle motion but do not wish to be concerned with scalar and vector potentials. Rather, we would prefer to work only in terms of electric and magnetic fields $\boldsymbol{E}$ and $\boldsymbol{B}$. One such option is to employ the first-order set of equations (1.6.68) and (1.6.69). Another is to employ a first-order set of equations obtained from the second-order set (1.6.74) by the usual means. Finally, the equations may be Hamiltonian in form, but we do not wish to exploit their Hamiltonian structure. However the equations arise, we will seek *Taylor* representations for their associated transfer maps

Let us recapitulate some of the contents of Section 1.3. Consider any set of m first-order differential equations of the form

$$\dot{z}_a = f_a(z_1, \cdots, z_m; t; \lambda_1, \cdots, \lambda_n), \quad a = 1, \cdots, m. \tag{10.12.1}$$

Here t is the independent variable and the z_a are dependent variables. Unlike the Hamiltonian case, the z_a need not be canonical variables and they need not be even in number. See (1.3.4). The λ_b are possible parameters.

Let the quantities z_a^0 be initial conditions specified at some initial time $t = t^0$,

$$z_a(t^0) = z_a^0. \tag{10.12.2}$$

Then, under mild conditions imposed on the functions f_a that appear on the right side of (12.1) and thereby define the set of differential equations, there exists a *unique* solution

$$z_a(t) = g_a(z_1^0, \cdots, z_m^0; t^0, t; \lambda_1, \cdots, \lambda_n), \quad a = 1, m \tag{10.12.3}$$

of (12.1) with the property

$$z_a(t^0) = g_a(z_1^0, \cdots, z_m^0; t^0, t^0; \lambda_1, \cdots, \lambda_n) = z_a^0, \quad a = 1, m. \tag{10.12.4}$$

Now assume that the functions f_a are analytic (within some domain) in the quantities z_a, the time t, and the parameters λ_b. Then, according to Poincaré's theorem, the solution given by (12.3) will be analytic (again within some domain) in the initial conditions z_a^0, the times t^0 and t, and the parameters λ_b.

If the solution $z_a(t)$ is analytic in the initial conditions z_a^0 and the parameters λ_b, then it is possible to expand it in the form of a Taylor series, with time-dependent coefficients, in the variables z_a^0 and λ_b. The aim of this section is to describe how these Taylor coefficients can be found as solutions to what we will call the *complete variational* equations.

To aid further discussion, it is useful to also rephrase our goal in the context of maps. Suppose we rewrite the set of first-order differential equations (12.1) in the more compact vector form

$$\dot{\boldsymbol{z}} = \boldsymbol{f}(\boldsymbol{z}; t; \boldsymbol{\lambda}). \tag{10.12.5}$$

Then, again using vector notation, their solution can be written in the form

$$\boldsymbol{z}(t) = \boldsymbol{g}(\boldsymbol{z}^0; t^0, t; \boldsymbol{\lambda}). \tag{10.12.6}$$

That is, the quantities $\boldsymbol{z}(t)$ at any time t are uniquely specified by the initial quantities $\boldsymbol{z}^0$ given at the initial time t^0.

We capitalize on this fact by introducing a slightly different notation. First, use t^i instead of t^0 to denote the *initial* time. Similarly use $\boldsymbol{z}^i$ to denote initial conditions by writing

$$\boldsymbol{z}^i = \boldsymbol{z}^0 = \boldsymbol{z}(t^i). \tag{10.12.7}$$

Next, let t^f be some *final* time, and define final conditions $\boldsymbol{z}^f$ by writing

$$\boldsymbol{z}^f = \boldsymbol{z}(t^f). \tag{10.12.8}$$

Then, with this notation, (12.6) can be rewritten in the form

$$\boldsymbol{z}^f = \boldsymbol{g}(\boldsymbol{z}^i; t^i, t^f; \boldsymbol{\lambda}). \tag{10.12.9}$$

We now view (12.9) as a *map* that sends the initial conditions $\boldsymbol{z}^i$ to the final conditions $\boldsymbol{z}^f$. This map will be called the *transfer map* between the times t^i and t^f, and will be denoted by the symbol $\mathcal{M}$. What we have emphasized is that a set of first-order differential equations of the form (12.5) can be integrated to produce a transfer map $\mathcal{M}$. We express the fact that $\mathcal{M}$ sends $\boldsymbol{z}^i$ to $\boldsymbol{z}^f$ in symbols by writing

$$\boldsymbol{z}^f = \mathcal{M}\boldsymbol{z}^i, \tag{10.12.10}$$

Recall the analogous discussion in Section 1.4 We also note that $\mathcal{M}$ is always invertible: Given $\boldsymbol{z}^f$, t^f, and t^i, we can always integrate backward in time from the moment $t = t^f$ to the moment $t = t^i$ and thereby find the initial conditions $\boldsymbol{z}^i$.

In the context of maps, our goal is to find a Taylor representation for $\mathcal{M}$. If parameters are present, we may wish to have an expansion in them as well. Initially, we will seek Taylor expansions of final conditions in terms of initial conditions. Subsequently, we will seek expansions of final conditions in terms of both initial conditions and parameters.

10.12.1 Case of No or Ignored Parameter Dependence

Suppose the equations (12.1) do not depend on any parameters λ_b or we do not wish to make expansions in them. We may then suppress their appearance to rewrite (12.1) in the form

$$\dot{z}_a = f_a(z,t), \quad a = 1, m. \tag{10.12.11}$$

Suppose, as in Section 10.5, that z^d is some given *design* solution to these equations, and we wish to study solutions in the vicinity of this solution. As before, we introduce deviation variables ζ_a by writing

$$z_a = z_a^d + \zeta_a. \tag{10.12.12}$$

Then the equations of motion (12.11) take the form

$$\dot{z}_a^d + \dot{\zeta}_a = f_a(z^d + \zeta, t). \tag{10.12.13}$$

We assume that the right side of (12.11) is analytic about z^d. See Theorem 3.3 of Section 1.3. Then we may write the relation

$$f_a(z^d + \zeta, t) = f_a(z^d, t) + g_a(z^d, t, \zeta) \tag{10.12.14}$$

where each g_a has a Taylor expansion of the form

$$g_a(z^d, t, \zeta) = \sum_r g_a^r(t) G_r(\zeta). \tag{10.12.15}$$

Here the $G_r(\zeta)$ are the various monomials in the variables ζ_b labeled by an *index* r using some convenient labeling scheme. (See, for examples, Tables 12.1 and 12.3. For more detail about monomial labeling schemes, see Section 39.2 and Appendix S.2.1.) And the g_a^r are (generally) time-dependent coefficients that we will call *forcing terms*.[1] By construction, all the monomials occurring in the right side of (12.15) have degree one or greater. We note that the g_a^r are known once $z^d(t)$ is given. By assumption, z^d is a solution of (12.11) and therefore satisfies the relations

$$\dot{z}_a^d = f_a(z^d, t). \tag{10.12.16}$$

It follows that the deviation variables satisfy the equations of motion

$$\dot{\zeta}_a = g_a(z^d, t, \zeta) = \sum_r g_a^r(t) G_r(\zeta). \tag{10.12.17}$$

These equations are evidently generalizations of the usual variational equations (see Exercise 4.6 of Section 1.4), and will be called the *complete variational* equations.

Consider the solution to the complete variational equations with *initial* conditions ζ_b^i specified at some initial time t^i. Under the conditions of Theorem 3.3 in Section 1.3 we expect that this solution will be an analytic function of the initial conditions ζ_b^i. (Here we have used the notation of Section 1.4.) Also, since the right side of (12.17) vanishes when all $\zeta_b = 0$ [all the monomials G_r appearing in (12.17) have degree one or greater], $\zeta(t) = 0$ is

[1] Here and in what follows the quantities g_a are not to be confused with those appearing in (12.3).

a solution to (12.17). It follows that the solution to the complete variational equations has a Taylor expansion of the form

$$\zeta_a(t) = \sum_r h_a^r(t) G_r(\zeta^i) \tag{10.12.18}$$

where the $h_a^r(t)$ are functions to be determined, and again all the monomials that occur have degree one or greater. When the quantities $h_a^r(t)$ are evaluated at some *final* time t^f, (12.18) provides a representation of the transfer map $\mathcal{M}$ about the design orbit in the Taylor form

$$\zeta_a^f = \zeta_a(t^f) = \sum_r h_a^r(t^f) G_r(\zeta^i). \tag{10.12.19}$$

10.12.2 Inclusion of Parameter Dependence

What can be done if we desire to have an expansion in parameters as well? Suppose that there are n such parameters, or that we wish to have expansions in n of them. The work of the previous section can be extended to handle this case by means of a simple trick: View the n parameters as additional *variables*, and "augment" the set of differential equations by additional differential equations that ensure these additional variables remain constant.

In detail, suppose we label the parameters so that those in which we wish to have an expansion are $\lambda_1 \cdots \lambda_n$. Introduce n additional variables $z_{m+1}, \cdots z_\ell$ where $\ell = m + n$ by making the replacements

$$\lambda_b \rightarrow z_{m+b}, \quad b = 1, n. \tag{10.12.20}$$

Next augment the equations (12.1) by n more of the form

$$\dot{z}_a = 0, \quad a = m + 1, \ell. \tag{10.12.21}$$

By this device we can rewrite the equations (12.1) in the form

$$\dot{z}_a = f_a(z, t), \quad a = 1, \ell \tag{10.12.22}$$

with the understanding that

$$f_a = f_a(z; t; \lambda^{\text{rem}}) \quad a = 1, m, \tag{10.12.23}$$

where λ^{rem} denotes the other *remaining* parameters, if any, and

$$f_a = 0, \quad a = m + 1, \ell. \tag{10.12.24}$$

For the first m equations we impose, as before, the initial conditions

$$z_a(t^i) = z_a^i, \quad a = 1, m. \tag{10.12.25}$$

For the remaining equations we impose the initial conditions

$$z_a(t^i) = \lambda_{a-m}, \quad a = m + 1, \ell. \tag{10.12.26}$$

Note that the relations (12.21) then ensure that the z_a for $a > m$ retain these values for all t.

To continue, let $z^d(t)$ be some design solution. Then, by construction, we have the result

$$z_a^d(t) = \lambda_{a-m}^d = \lambda_{a-m}, \quad a = m+1, \ell. \tag{10.12.27}$$

Again introduce deviation variables by writing

$$z_a = z_a^d + \zeta_a \quad a = 1, \ell. \tag{10.12.28}$$

Then the quantities ζ_a for $a > m$ will describe deviations in the parameter values about the values λ_{a-m}^d. Moreover, because we have assumed analyticity in the parameters as well, relations of the forms (12.14) and (12.15) will continue to hold except that the $G_r(\zeta)$ are now the various monomials in the ℓ variables ζ_b. Relations of the forms (12.16) and (12.17) will also hold with the provisos (12.24) and

$$g_a^r(t) = 0, \quad a = m+1, \ell. \tag{10.12.29}$$

Therefore, we will only need to integrate the equations of the forms (12.16) and (12.17) for $a \leq m$. Finally, relations of the form (12.19) will continue to hold for $a \leq m$ supplemented by the relations

$$\zeta_a^f = \zeta_a^i, \quad a = m+1, \ell. \tag{10.12.30}$$

Since the $G_r(\zeta^i)$ now involve ℓ variables, the relations of the form (12.19) will provide an expansion of the final quantities ζ_a^f (for $a \leq m$) in terms of the initial quantities ζ_a^i (for $a \leq m$) and also the parameter deviations ζ_a^i with $a = m+1, \ell$.

10.12.3 Solution of Complete Variational Equations Using Forward Integration

This subsection and Subsection 12.5 describe two methods for the solution of the complete variational equations. This subsection describes the method that employs integration forward in time, and is the conceptually simpler of the two methods.

To determine the functions h_a^r, let us insert the expansion (12.18) into both sides of (12.17). With r'' as a dummy index, the left side becomes the relation

$$\dot{\zeta}_a = \sum_{r''} \dot{h}_a^{r''}(t) G_{r''}(\zeta^i). \tag{10.12.31}$$

For the right side we find the intermediate result

$$\sum_r g_a^r(t) G_r(\zeta) = \sum_r g_a^r(t)\ G_r\!\left(\sum_{r'} h_1^{r'}(t) G_{r'}(\zeta^i), \cdots \sum_{r'} h_m^{r'}(t) G_{r'}(\zeta^i)\right). \tag{10.12.32}$$

However, since the G_r are monomials, there are relations of the form

$$G_r\!\left(\sum_{r'} h_1^{r'}(t) G_{r'}(\zeta^i), \cdots \sum_{r'} h_m^{r'}(t) G_{r'}(\zeta^i)\right) = \sum_{r''} U_r^{r''}(h_n^s) G_{r''}(\zeta^i), \tag{10.12.33}$$

and therefore the right side of (12.17) can be rewritten in the form

$$\sum_r g_a^r(t) G_r(\zeta) = \sum_{r''} \sum_r g_a^r(t) U_r^{r''}(h_n^s) G_{r''}(\zeta^i). \tag{10.12.34}$$

The notation $U_r^{r''}(h_n^s)$ is employed to indicate that these quantities might (at this stage of the argument) depend on all the h_n^s with n ranging from 1 to ℓ, and s ranging over all possible values.

Now, in accord with (12.17), equate the right sides of (12.31) and (12.34) to obtain the relation

$$\sum_{r''} \dot{h}_a^{r''}(t) G_{r''}(\zeta^i) = \sum_{r''} \sum_r g_a^r(t) U_r^{r''}(h_n^s) G_{r''}(\zeta^i). \tag{10.12.35}$$

Since the monomials $G_{r''}(\zeta^i)$ are linearly independent, we must have the result

$$\dot{h}_a^{r''}(t) = \sum_r g_a^r(t) U_r^{r''}(h_n^s). \tag{10.12.36}$$

We have found a set of differential equations that must be satisfied by the h_a^r. Moreover, from (12.18), there is the relation

$$\zeta_a(t^i) = \sum_r h_a^r(t^i) G_r(\zeta^i) = \zeta_a^i. \tag{10.12.37}$$

Thus, all the functions $h_a^r(t)$ have a known value at the initial time t^i, and indeed are mostly initially zero. When the equations (12.36) are integrated *forward* from $t = t^i$ to $t = t^f$ to obtain the quantities $h_a^r(t^f)$, the result is the transfer map $\mathcal{M}$ about the design orbit in the Taylor form (12.19).

Let us now examine the structure of this set of differential equations. A key observation is that the functions $U_r^{r''}(h_n^s)$ are *universal.* That is, as (12.33) indicates, they describe certain *combinatorial* properties of monomials. They depend only on the dimension ℓ of the system under study, and are the *same* for all such systems. As (12.17) shows, what are peculiar to any given system are the forcing terms $g_a^r(t)$.

10.12.4 Application of Forward Integration to the Two-Variable Case

To see what is going on in more detail, it is instructive to work out the first nontrivial case, that with $\ell = 2$. For two variables, all monomials in (ζ_1, ζ_2) are of the form $(\zeta_1)^{j_1}(\zeta_2)^{j_2}$. Here, to simplify notation, we have dropped the superscript i. Table 12.1 below shows a convenient way of labeling such monomials, and for this labeling we write

$$G_r(\zeta) = (\zeta_1)^{j_1}(\zeta_2)^{j_2} \tag{10.12.38}$$

with

$$j_1 = j_1(r) \;\text{ and }\; j_2 = j_2(r) \tag{10.12.39}$$

and $D(r)$ denotes the *degree* of each monomial.

Table 10.12.1: A labeling scheme for monomials through degree three in two variables.

r	j_1	j_2	D
1	1	0	1
2	0	1	1
3	2	0	2
4	1	1	2
5	0	2	2
6	3	0	3
7	2	1	3
8	1	2	3
9	0	3	3

Thus, for example,

$$G_1 = \zeta_1, \tag{10.12.40}$$

$$G_2 = \zeta_2, \tag{10.12.41}$$

$$G_3 = \zeta_1^2, \tag{10.12.42}$$

$$G_4 = \zeta_1\zeta_2, \tag{10.12.43}$$

$$G_5 = \zeta_2^2, \text{ etc.} \tag{10.12.44}$$

Again, for more detail about monomial labeling schemes, see Section 39.2 and Appendix S.2.1.

Let us now compute the first few $U_r^{r''}(h_n^s)$. From (12.33) and (12.40) we find the relation

$$G_1\left(\sum_{r'} h_1^{r'} G_{r'}(\zeta), \sum_{r'} h_2^{r'} G_{r'}(\zeta)\right) = \sum_{r'} h_1^{r'} G_{r'}(\zeta) = \sum_{r''} U_1^{r''} G_{r''}(\zeta). \tag{10.12.45}$$

It follows that there is the result

$$U_1^{r''} = h_1^{r''}. \tag{10.12.46}$$

Similarly, from (12.33) and (12.41), we find the result

$$U_2^{r''} = h_2^{r''}. \tag{10.12.47}$$

From (12.33) and (12.42) we find the relation

$$\begin{aligned} G_3\left(\sum_{r'} h_1^{r'} G_{r'}(\zeta), \sum_{r'} h_2^{r'} G_{r'}(\zeta)\right) &= \left(\sum_{r'} h_1^{r'} G_{r'}(\zeta)\right)^2 \\ &= \sum_{s,t} h_1^s h_1^t G_s(\zeta) G_t(\zeta) = \sum_{r''} U_3^{r''} G_{r''}(\zeta). \end{aligned} \tag{10.12.48}$$

Use of (12.48) and inspection of (12.40) through (12.44) yield the results

$$U_3^1 = 0, \tag{10.12.49}$$

$$U_3^2 = 0, \tag{10.12.50}$$

$$U_3^3 = (h_1^1)^2, \tag{10.12.51}$$

$$U_3^4 = 2h_1^1 h_1^2, \tag{10.12.52}$$

$$U_3^5 = (h_1^2)^2. \tag{10.12.53}$$

From (12.33) and (12.43) we find the relation

$$\begin{aligned} G_4\Bigg(\sum_{r'} h_1^{r'} G_{r'}(\zeta), \sum_{r'} h_2^{r'} G_{r'}(\zeta)\Bigg) &= \Bigg(\sum_{r'} h_1^{r'} G_{r'}(\zeta)\Bigg)\Bigg(\sum_{r'} h_2^{r'} G_{r'}(\zeta)\Bigg) \\ = \sum_{s,t} h_1^s h_2^t G_s(\zeta) G_t(\zeta) &= \sum_{r''} U_4^{r''} G_{r''}(\zeta). \end{aligned} \tag{10.12.54}$$

It follows that there are the results

$$U_4^1 = 0, \tag{10.12.55}$$

$$U_4^2 = 0, \tag{10.12.56}$$

$$U_4^3 = h_1^1 h_2^1, \tag{10.12.57}$$

$$U_4^4 = h_1^1 h_2^2 + h_1^2 h_2^1, \tag{10.12.58}$$

$$U_4^5 = h_1^2 h_2^2. \tag{10.12.59}$$

Finally, from (12.33) and (12.44), we find the results

$$U_5^1 = 0, \tag{10.12.60}$$

$$U_5^2 = 0, \tag{10.12.61}$$

$$U_5^3 = (h_2^1)^2, \tag{10.12.62}$$

$$U_5^4 = 2h_2^1 h_2^2, \tag{10.12.63}$$

$$U_5^5 = (h_2^2)^2. \tag{10.12.64}$$

Two features now become apparent. As in Table 12.1, let $D(r)$ be the *degree* of the monomial with label r. Then, from the examples worked out, and quite generally from (12.33), we see that there is the relation

$$U_r^{r''} = 0 \text{ when } D(r) > D(r''). \tag{10.12.65}$$

It follows that the sum on the right side of (12.36) always terminates. Second, for the arguments h_n^s possibly appearing in $U_r^{r''}(h_n^s)$, we see that there is the relation

$$D(s) \leq D(r''). \tag{10.12.66}$$

It follows, again see (12.36), that the right side of the differential equation for any $h_a^{r''}$ involves only the h_n^s for which (12.66) holds. Therefore, to determine the coefficients $h_a^r(t^f)$ of the Taylor expansion (12.19) through terms of some degree D, it is only necessary to integrate a finite number of equations, and the right sides of these equations involve only the coefficients for this degree and lower.

For example, to continue our discussion of the case of two variables, the equations (12.36) take the form

$$\dot{h}_1^1(t) = \sum_{r=1}^{2} g_1^r(t) U_r^1 = g_1^1(t)h_1^1(t) + g_1^2(t)h_2^1(t), \tag{10.12.67}$$

$$\dot{h}_2^1(t) = \sum_{r=1}^{2} g_2^r(t) U_r^1 = g_2^1(t)h_1^1(t) + g_2^2(t)h_2^1(t), \tag{10.12.68}$$

$$\dot{h}_1^2(t) = \sum_{r=1}^{2} g_1^r(t) U_r^2 = g_1^1(t)h_1^2(t) + g_1^2(t)h_2^2(t), \tag{10.12.69}$$

$$\dot{h}_2^2(t) = \sum_{r=1}^{2} g_2^r(t) U_r^2 = g_2^1(t)h_1^2(t) + g_2^2(t)h_2^2(t), \tag{10.12.70}$$

$$\begin{aligned} \dot{h}_1^3(t) &= \sum_{r=1}^{5} g_1^r(t) U_r^3 \\ &= g_1^1(t)h_1^3(t) + g_1^2(t)h_2^3(t) + g_1^3(t)[h_1^1(t)]^2 \\ &+ g_1^4(t)h_1^1(t)h_2^1(t) + g_1^5(t)[h_2^1(t)]^2, \end{aligned} \tag{10.12.71}$$

$$\begin{aligned} \dot{h}_2^3(t) &= \sum_{r=1}^{5} g_2^r(t) U_r^3 \\ &= g_2^1(t)h_1^3(t) + g_2^2(t)h_2^3(t) + g_2^3(t)[h_1^1(t)]^2 \\ &+ g_2^4(t)h_1^1(t)h_2^1(t) + g_2^5(t)[h_2^1(t)]^2, \end{aligned} \tag{10.12.72}$$

$$\begin{aligned} \dot{h}_1^4(t) &= \sum_{r=1}^{5} g_1^r(t) U_r^4 \\ &= g_1^1(t)h_1^4(t) + g_1^2(t)h_2^4(t) + 2g_1^3(t)h_1^1(t)h_1^2(t) \\ &+ g_1^4(t)[h_1^1(t)h_2^2(t) + h_1^2(t)h_2^1(t)] + 2g_1^5(t)h_2^1(t)h_2^2(t), \end{aligned} \tag{10.12.73}$$

$$\begin{aligned} \dot{h}_2^4(t) &= \sum_{r=1}^{5} g_2^r(t) U_r^4 \\ &= g_2^1(t)h_1^4(t) + g_2^2(t)h_2^4(t) + 2g_2^3(t)h_1^1(t)h_1^2(t) \\ &+ g_2^4(t)[h_1^1(t)h_2^2(t) + h_1^2(t)h_2^1(t)] + 2g_2^5(t)h_2^1(t)h_2^2(t), \end{aligned} \tag{10.12.74}$$

$$\begin{aligned} \dot{h}_1^5(t) &= \sum_{r=1}^{5} g_1^r(t) U_r^5 \\ &= g_1^1(t)h_1^5(t) + g_1^2(t)h_2^5(t) + g_1^3(t)[h_1^2(t)]^2 \\ &+ g_1^4(t)h_1^2(t)h_2^2(t) + g_1^5(t)[h_2^2(t)]^2, \end{aligned} \tag{10.12.75}$$

$$\begin{aligned}\dot{h}_2^5(t) &= \sum_{r=1}^{5} g_2^r(t)U_r^5 \\ &= g_2^1(t)h_1^5(t) + g_2^2(t)h_2^5(t) + g_2^3(t)[h_1^2(t)]^2 \\ &+ g_2^4(t)h_1^2(t)h_2^2(t) + g_2^5(t)[h_2^2(t)]^2, \text{ etc.}\end{aligned} \tag{10.12.76}$$

From (12.37) we have the initial conditions

$$h_a^r(t^i) = \delta_a^r. \tag{10.12.77}$$

We see that if we desire only the degree one terms in the expansion (12.18), then it is only necessary to integrate the equations (12.67) through (12.70) with the initial conditions (12.77). [A moment's reflection shows that, for the case of two variables and no parameters, these are just the (linear) variational equations for the matrix L in Exercise 4.6 of Section 1.4.] We see that if we desire only the degree one and degree two terms in the expansion (12.18), then it is only necessary to integrate the equations (12.67) through (12.76) with the initial conditions (12.77), etc.

10.12.5 Solution of Complete Variational Equations Using Backward Integration

There is another method of determining the h_a^r that is surprising, ingenious, and in some ways superior to that just described. It involves integrating *backward* in time.[2]

Let us rewrite (12.19) in the slightly more explicit form

$$\zeta_a^f = \sum_r h_a^r(t^i, t^f)G_r(\zeta^i) \tag{10.12.78}$$

to indicate that there are two times involved, t^i and t^f. From this perspective, (12.36) is a set of differential equations for the quantities $(\partial/\partial t)h_a^r(t^i, t)$ that is to be integrated and evaluated at $t = t^f$. An alternate procedure is to seek a set of differential equations for the quantities $(\partial/\partial \bar{t})h_a^r(\bar{t}, t^f)$ that is to be integrated and evaluated at $\bar{t} = t^i$.

As a first step in considering this alternative, rewrite (12.78) in the form

$$\zeta_a^f = \sum_r h_a^r(\bar{t}, t^f)G_r(\zeta(\bar{t})). \tag{10.12.79}$$

Now reason as follows: If $\bar{t}$ is varied and at the same time the quantities $\zeta(\bar{t})$ are varied (evolve) so as to remain on the solution to (12.17) having final conditions ζ^f, then the quantities ζ^f must remain *unchanged.* Consequently, there is the differential equation result

$$0 = d\zeta_a^f/d\bar{t} = \sum_r [(\partial/\partial\bar{t})h_a(\bar{t}, t^f)]G_r(\zeta(\bar{t})) + \sum_r h_a^r(\bar{t}, t^f)(d/d\bar{t})G_r(\zeta(\bar{t})). \tag{10.12.80}$$

Let us introduce the notation $\dot{h}_a^r(\bar{t}, t^f)$ for $(\partial/\partial\bar{t})h_a^r(\bar{t}, t^f)$ so that the first term on the right side of (12.80) can be rewritten in the form

$$\sum_r [(\partial/\partial\bar{t})h_a^r(\bar{t}, t^f)]G_r(\zeta) = \sum_r \dot{h}_a^r G_r(\zeta). \tag{10.12.81}$$

[2] To integrate backward numerically, simply replace h by $-h$ where h is the step size. See Chapter 2.

Next, begin working on the second term on the right side of (12.80) by replacing the summation index r by the dummy index r',

$$\sum_r h_a^r(\bar{t}, t^f)(d/d\bar{t})G_r(\zeta(\bar{t})) = \sum_{r'} h_a^{r'}(\bar{t}, t^f)(d/d\bar{t})G_{r'}(\zeta(\bar{t})). \tag{10.12.82}$$

Now carry out the indicated differentiation using the chain rule and the relation (12.17) which describes how the quantities ζ vary along a solution,

$$(d/d\bar{t})G_{r'}(\zeta(\bar{t})) = \sum_b (\partial G_{r'}/\partial\zeta_b)(d\zeta_b/d\bar{t}) = \sum_{br''} (\partial G_{r'}/\partial\zeta_b)g_b^{r''}(\bar{t})G_{r''}(\zeta(\bar{t})). \tag{10.12.83}$$

Watch closely: Since the G_r are simply standard monomials in the ζ, there must be relations of the form

$$[(\partial/\partial\zeta_b)G_{r'}(\zeta)]G_{r''}(\zeta) = \sum_r C_{br'r''}^r G_r(\zeta) \tag{10.12.84}$$

where the $C_{br'r''}^r$ are *universal constant coefficients* that describe certain combinatorial properties of monomials. As a result, the second term on the right side of (12.80) can be written in the form

$$\sum_{r'} h_a^{r'}(\bar{t}, t^f)(d/d\bar{t})G_{r'}(\zeta(\bar{t})) = \sum_r G_r(\zeta) \sum_{br'r''} C_{br'r''}^r g_b^{r''}(\bar{t})h_a^{r'}(\bar{t}, t^f). \tag{10.12.85}$$

Since the monomials G_r are linearly independent, the relations (12.80) through (12.85) imply the result

$$\dot{h}_a^r(\bar{t}, t^f) = -\sum_{br'r''} C_{br'r''}^r g_b^{r''}(\bar{t})h_a^{r'}(\bar{t}, t^f). \tag{10.12.86}$$

This is a set of differential equations for the h_a^r that are to be integrated from $\bar{t} = t^f$ *back* to $\bar{t} = t^i$. Also, evaluating (12.79) for $\bar{t} = t^f$ gives the results

$$\zeta_a^f = \sum_r h_a^r(t^f, t^f)G_r(\zeta_a^f), \tag{10.12.87}$$

from which it follows that (with the usual polynomial labeling) the h_a^r satisfy the final conditions

$$h_a^r(t^f, t^f) = \delta_a^r. \tag{10.12.88}$$

Therefore the solution to (12.86) is uniquely defined. Finally, it is evident from the definition (12.84) that the coefficients $C_{br'r''}^r$ satisfy the relation

$$C_{br'r''}^r = 0 \text{ unless } [D(r') - 1] + D(r'') = D(r). \tag{10.12.89}$$

Consequently, since $D(r'') \geq 1$ in (12.86), it follows from (12.89) that the only $h_a^{r'}$ that occur on the right side of (12.86) are those that satisfy

$$D(r') \leq D(r). \tag{10.12.90}$$

Similarly, the only $g_b^{r''}$ that occur are those that satisfy

$$D(r'') \leq D(r). \tag{10.12.91}$$

Therefore, as before, to determine the coefficients h_a^r of the Taylor expansion (12.19) through terms of some degree D, it is only necessary to integrate a finite number of equations, and the right sides of these equations involve only the coefficients for this degree and lower.

Comparison of the differential equation sets (12.36) and (12.86) shows that the latter has the remarkable property of being *linear* in the unknown quantities h_a^r. This feature means that the evaluation of the right side of (12.86) involves only the retrieval of certain universal constants $C^r_{br'r''}$ and straight-forward multiplication and addition. By contrast, the use of (12.36) requires evaluation of the fairly complicated *nonlinear* functions $U_r^{r''}(h_n^s)$. Finally, it is easier to insure that a numerical integration procedure is working properly for a set of linear differential equations than it is for a nonlinear set.

The only complication in the use of (12.86) is that the equations must be integrated backwards in $\bar{t}$. Correspondingly the equations (12.16) for the design solution must also be integrated backwards since they supply the required quantities g_a^r through use of (12.14) and (12.15). This is no problem if the final quantities $z^d(t^{\text{fin}})$ are known. However if only the initial quantities $z^d(t^{\text{in}})$ are known, then the equations (12.16) for z^d must first be integrated forward in time to find the final quantities $z^d(t^{\text{fin}})$.

10.12.6 The Two-Variable Case Revisited

For clarity, let us also apply this second method to the two-variable case. Table 12.2 shows the nonzero values of $C^r_{br'r''}$ for $r \in [1,5]$ obtained using (12.40) through (12.44) and (12.84). Note that the rules (12.89) hold. Use of this Table shows that in the two-variable case the equations (12.86) take the form

$$\dot{h}_1^1(\bar{t},t^f) = -\ g_1^1(\bar{t})h_1^1(\bar{t},t^f) - g_2^1(\bar{t})h_1^2(\bar{t},t^f), \tag{10.12.92}$$

$$\dot{h}_2^1(\bar{t},t^f) = -\ g_1^1(\bar{t})h_2^1(\bar{t},t^f) - g_2^1(\bar{t})h_2^2(\bar{t},t^f), \tag{10.12.93}$$

$$\dot{h}_1^2(\bar{t},t^f) = -\ g_1^2(\bar{t})h_1^1(\bar{t},t^f) - g_2^2(\bar{t})h_1^2(\bar{t},t^f), \tag{10.12.94}$$

$$\dot{h}_2^2(\bar{t},t^f) = -\ g_1^2(\bar{t})h_2^1(\bar{t},t^f) - g_2^2(\bar{t})h_2^2(\bar{t},t^f), \tag{10.12.95}$$

$$\dot{h}_1^3(\bar{t},t^f) = -\ 2g_1^1(\bar{t})h_1^3(\bar{t},t^f) - g_1^3(\bar{t})h_1^1(\bar{t},t^f) - g_2^1(\bar{t})h_1^4(\bar{t},t^f) - g_2^3(\bar{t})h_1^2(\bar{t},t^f), \tag{10.12.96}$$

$$\dot{h}_2^3(\bar{t},t^f) = -\ 2g_1^1(\bar{t})h_2^3(\bar{t},t^f) - g_1^3(\bar{t})h_2^1(\bar{t},t^f) - g_2^1(\bar{t})h_2^4(\bar{t},t^f) - g_2^3(\bar{t})h_2^2(\bar{t},t^f), \tag{10.12.97}$$

$$\begin{aligned}\dot{h}_1^4(\bar{t},t^f) &= -\ g_1^1(\bar{t})h_1^4(\bar{t},t^f) - 2g_1^2(\bar{t})h_1^3(\bar{t},t^f) - g_1^4(\bar{t})h_1^1(\bar{t},t^f) \\ &-\ 2g_2^1(\bar{t})h_1^5(\bar{t},t^f) - g_2^2(\bar{t})h_1^4(\bar{t},t^f) - g_2^4(\bar{t})h_1^2(\bar{t},t^f),\end{aligned} \tag{10.12.98}$$

$$\begin{aligned}\dot{h}_2^4(\bar{t},t^f) &= -\ g_1^1(\bar{t})h_2^4(\bar{t},t^f) - 2g_1^2(\bar{t})h_2^3(\bar{t},t^f) - g_1^4(\bar{t})h_2^1(\bar{t},t^f) \\ &-\ 2g_2^1(\bar{t})h_2^5(\bar{t},t^f) - g_2^2(\bar{t})h_2^4(\bar{t},t^f) - g_2^4(\bar{t})h_2^2(\bar{t},t^f),\end{aligned} \tag{10.12.99}$$

$$\dot{h}_1^5(\bar{t},t^f) = -\ g_1^2(\bar{t})h_1^4(\bar{t},t^f) - g_1^5(\bar{t})h_1^1(\bar{t},t^f) - 2g_2^2(\bar{t})h_1^5(\bar{t},t^f) - g_2^5(\bar{t})h_1^2(\bar{t},t^f), \tag{10.12.100}$$

$$\dot{h}_2^5(\bar{t},t^f) = -\ g_1^2(\bar{t})h_2^4(\bar{t},t^f) - g_1^5(\bar{t})h_2^1(\bar{t},t^f) - 2g_2^2(\bar{t})h_2^5(\bar{t},t^f) - g_2^5(\bar{t})h_2^2(\bar{t},t^f),\ \text{etc.} \tag{10.12.101}$$

As advertised, the right sides of (12.92) through (12.101) are indeed simpler than those of (12.67) through (12.76).

Table 10.12.2: Nonzero values of $C^r_{br'r''}$ for $r \in [1, 5]$ in the two-variable case.

r	b	r'	r''	C
1	1	1	1	1
1	2	2	1	1
2	1	1	2	1
2	2	2	2	1
3	1	1	3	1
3	1	3	1	2
3	2	2	3	1
3	2	4	1	1
4	1	1	4	1
4	1	3	2	2
4	1	4	1	1
4	2	2	4	1
4	2	4	2	1
4	2	5	1	2
5	1	1	5	1
5	1	4	2	1
5	2	2	5	1
5	2	5	2	1

10.12.7 Application to Duffing's Equation

As an extension of our discussion of the case of two variables (and no parameters), let us apply the results obtained so far to Duffing's equation (1.4.27) described earlier in Section 1.4. Recall that by a suitable change of variables this equation can be brought to the form

$$q'' + 2\beta q' + q + q^3 = -\epsilon \sin \omega\tau. \tag{10.12.102}$$

Here, for notational convenience, a prime denotes $d/d\tau$. For our purposes, particularly for the parameter expansion soon to be made in the next subsection, it is useful to make the further change of variables

$$q = \omega Q, \tag{10.12.103}$$

$$\omega = 1/\sigma, \tag{10.12.104}$$

$$\omega\tau = t. \tag{10.12.105}$$

When this is done, there are the relations

$$q' = \omega^2 \dot{Q} \tag{10.12.106}$$

and

$$q'' = \omega^3 \ddot{Q} \tag{10.12.107}$$

where now a dot denotes d/dt. [Note that the variable t here is different from that in (1.4.27).] Correspondingly, Duffing's equation takes the form

$$\ddot{Q} + 2\beta\sigma\dot{Q} + \sigma^2 Q + Q^3 = -\epsilon\sigma^3 \sin t. \tag{10.12.108}$$

Finally, this equation can be converted to a first-order pair of the form (12.11) by writing

$$Q = z_1, \tag{10.12.109}$$

$$\dot{Q} = z_2. \tag{10.12.110}$$

Doing so gives the system

$$\dot{z}_1 = z_2, \tag{10.12.111}$$

$$\dot{z}_2 = -2\beta\sigma z_2 - \sigma^2 z_1 - z_1^3 - \epsilon\sigma^3 \sin t, \tag{10.12.112}$$

and we see that there are the relations

$$f_1(z,t) = z_2, \tag{10.12.113}$$

$$f_2(z,t) = -2\beta\sigma z_2 - \sigma^2 z_1 - z_1^3 - \epsilon\sigma^3 \sin t. \tag{10.12.114}$$

Now we are ready to carry out the expansions (12.14) and (12.15). We find the results

$$f_1(z^d + \zeta, t) = z_2^d + \zeta_2, \tag{10.12.115}$$

$$\begin{aligned} f_2(z^d + \zeta, t) &= -2\beta\sigma(z_2^d + \zeta_2) - \sigma^2(z_1^d + \zeta_1) - (z_1^d + \zeta_1)^3 - \epsilon\sigma^3 \sin t \\ &= -[2\beta\sigma z_2^d + \sigma^2 z_1^d + (z_1^d)^3 + \epsilon\sigma^3 \sin] - \{[\sigma^2 + 3(z_1^d)^2]\zeta_1 + 2\beta\sigma\zeta_2\} \\ &\quad -3z_1^d\zeta_1^2 - \zeta_1^3. \end{aligned} \tag{10.12.116}$$

Note that the right sides of (12.115) and (12.116) contain only terms of degree 3 and lower in the deviation variables ζ_a. It follows that for Duffing's equation the only nonzero forcing terms are given by the relations

$$g_1^2 = 1, \tag{10.12.117}$$

$$g_2^1 = -\sigma^2 - 3(z_1^d)^2, \tag{10.12.118}$$

$$g_2^2 = -2\beta\sigma, \tag{10.12.119}$$

$$g_2^3 = -3z_1^d, \tag{10.12.120}$$

$$g_2^6 = -1. \tag{10.12.121}$$

And, according to (12.16), the design solution z^d obeys the equations (12.111) and (12.112) with $z = z^d$.

At this point we pause to look particularly at the lowest-degree (linear) variational equations because they have a special simplicity in the Duffing case. Let L be the matrix defined by the relation

$$L = \begin{pmatrix} h_1^1 & h_1^2 \\ h_2^1 & h_2^2 \end{pmatrix} \tag{10.12.122}$$

so that

$$\zeta(t) = L(t)\zeta^i + O[(\zeta^i)^2]. \qquad (10.12.123)$$

See (12.18). Then equations (12.67) through (12.70) are equivalent to the matrix equation

$$\dot{L} = AL \qquad (10.12.124)$$

where A is the matrix

$$A = \begin{pmatrix} g_1^1 & g_1^2 \\ g_2^1 & g_2^2 \end{pmatrix}. \qquad (10.12.125)$$

From (12.125) and (12.117) through (12.121), we find the result

$$\text{tr } A = g_1^1 + g_2^2 = -2\beta\sigma. \qquad (10.12.126)$$

Also, in the Duffing case, we may set $t^i = 0$ and require the initial condition

$$L(0) = I. \qquad (10.12.127)$$

Based on the results of Exercise 1.4.6, we conclude that there is the relation

$$\det L(t) = \exp(-2\beta\sigma t), \qquad (10.12.128)$$

and, in particular, for $t = t^f = 2\pi$ there is the relation

$$\det L(2\pi) = \exp(-4\pi\beta\sigma) = \exp(-4\pi\beta/\omega). \qquad (10.12.129)$$

Thus, for the Duffing equation, we are able to find the determinant of the linear part of the transfer map *analytically.* Note also the remarkable feature that for the Duffing equation the determinant of the linear part of the transfer map does not depend on z^d. The determinant is the *same* for any trajectory.

Returning to our main discussion, suppose, for example, that we specify the values of β, ϵ, and ω, and then integrate the system (12.111) and (12.112) from $t = 0$ to $t = 2\pi$. So doing produces an example of the stroboscopic map $\mathcal{M}$ described in Subsection 1.4.3. Suppose, further, that we require that the design solution z^d be periodic (with period 2π) thus yielding a fixed point of $\mathcal{M}$,

$$z_a^d(2\pi) = z_a^d(0). \qquad (10.12.130)$$

We will see in Chapter 28 that such fixed points exist. Using this $z^d(t)$, we may integrate from $t = t^i = 0$ to $t = t^f = 2\pi$ the equations (12.67) through (12.76), etc., with the g_a^r given by (12.117) through (12.121) and the initial conditions given by (12.77). Alternatively, we may integrate (12.92) though (12.101), etc. from $\bar{t} = t^f = 2\pi$ back to $\bar{t} = t^i = 0$ with the final conditions (12.88). Carrying out either method determines the quantities $h_a^r(0, 2\pi)$, and we see from (12.19) or (12.78) that we have found a Taylor expansion for $\mathcal{M}$ about the *fixed* point $z^d(0)$.

10.12.8 Application to Duffing's Equation Including some Parameter Dependence

Suppose, as described in Subsection 12.2, we wish to include some parameter dependence. Figures 28.8.5 and 28.8.6 in Chapter 28 show, for example, a portion of the Feigenbaum diagram for Duffing's equation as ω is varied. Evidently ω is a parameter and therefore, according to Theorem 3.3 of Section 1.3, it should be possible to Taylor expand the solution to Duffing's equation with respect to ω as well as with respect to the initial conditions. Equivalently, we will seek an expansion of the solution of (12.108) with respect to the parameter σ. See (12.104).

Following the method of Subsection 12.2, we augment the first-order equation set associated with (12.108) by adding the equation

$$\dot{\sigma} = 0. \tag{10.12.131}$$

Then we may view σ as a variable, and (12.131) guarantees that this variable remains a constant. Taken together, (12.108) and (12.131) may be converted to a first-order triplet of the form (12.22) by writing (12.109), (12.110), and

$$\sigma = z_3. \tag{10.12.132}$$

Doing so gives the system

$$\dot{z}_1 = z_2, \tag{10.12.133}$$

$$\dot{z}_2 = -2\beta z_3 z_2 - z_3^2 z_1 - z_1^3 - \epsilon z_3^3 \sin t, \tag{10.12.134}$$

$$\dot{z}_3 = 0, \tag{10.12.135}$$

and we see that there are the relations

$$f_1(z,t) = z_2, \tag{10.12.136}$$

$$f_2(z,t) = -\ 2\beta z_3 z_2 - z_3^2 z_1 - z_1^3 - \epsilon z_3^3 \sin t, \tag{10.12.137}$$

$$f_3(z,t) = 0. \tag{10.12.138}$$

As before, we introduce deviation variables using (12.12) and carry out the steps (12.13) through (12.19). In particular, we write

$$z_3 = z_3^d + \zeta_3 = \sigma^d + \zeta_3. \tag{10.12.139}$$

This time we are working with monomials in the three variables ζ_1, ζ_2, and ζ_3. [That is, a ranges from 1 to 3 in (12.18).] They are conveniently labeled using the indices r given in Table 12.3 below. We see, for example, that if we desire to work with monomials through degree 2, the index r should range from 1 through 9.

With regard to the expansions (12.14) and (12.15), we find the results

$$f_1(z^d + \zeta, t) = z_2^d + \zeta_2, \tag{10.12.140}$$

Table 10.12.3: A labeling scheme for monomials through degree three in three variables.

r	j_1	j_2	j_3	D
1	1	0	0	1
2	0	1	0	1
3	0	0	1	1
4	2	0	0	2
5	1	1	0	2
6	1	0	1	2
7	0	2	0	2
8	0	1	1	2
9	0	0	2	2
10	3	0	0	3
11	2	1	0	3
12	2	0	1	3
13	1	2	0	3
14	1	1	1	3
15	1	0	2	3
16	0	3	0	3
17	0	2	1	3
18	0	1	2	3
19	0	0	3	3

$$\begin{aligned} f_2(z^d+\zeta,t) &= -\ 2\beta(z_3^d+\zeta_3)(z_2^d+\zeta_2)-(z_3^d+\zeta_3)^2(z_1^d+\zeta_1) \\ &\quad -(z_1^d+\zeta_1)^3-\epsilon(z_3^d+\zeta_3)^3\sin t \\ &= [-2\beta z_2^d z_3^d - z_1^d(z_3^d)^2-(z_1^d)^3-\epsilon(z_3^d)^3\sin t] \\ &\quad -[3(z_1^d)^2+(z_3^d)^2]\zeta_1-2\beta z_3^d\zeta_2-[2\beta z_2^d+2z_1^d z_3^d+3\epsilon(z_3^d)^2\sin t]\zeta_3 \\ &\quad -2\beta\zeta_2\zeta_3-(z_1^d+3\epsilon z_3^d)\zeta_3^2-z_3^d\zeta_1\zeta_2-3z_1^d\zeta_1^2 \\ &\quad -\zeta_1^3-\zeta_1\zeta_3^2-\epsilon(\sin t)\zeta_3^3. \end{aligned} \tag{10.12.141}$$

$$f_3(z^d+\zeta,t)=0. \tag{10.12.142}$$

Note the right sides of (12.140) through (12.142) are at most cubic in the deviation variables ζ_a. Therefore, from Table 12.3, we see that the index r for the g_a^r should range from 1 through 19. It follows that for Duffing's equation (with σ parameter expansion) the *only* nonzero forcing terms are given by the relations

$$g_1^2=1, \tag{10.12.143}$$

$$g_2^1=-3(z_1^d)^2-(z_3^d)^2, \tag{10.12.144}$$

$$g_2^2=-2\beta z_3^d, \tag{10.12.145}$$

$$g_2^3=-2\beta z_2^d-2z_1^d z_3^d-3\epsilon(z_3^d)^2\sin t, \tag{10.12.146}$$

$$g_2^4 = -3z_1^d, \tag{10.12.147}$$

$$g_2^6 = -2z_3^d, \tag{10.12.148}$$

$$g_2^8 = -2\beta, \tag{10.12.149}$$

$$g_2^9 = -z_1^d - 3\epsilon z_3^d \sin t, \tag{10.12.150}$$

$$g_2^{10} = -1, \tag{10.12.151}$$

$$g_2^{15} = -1, \tag{10.12.152}$$

$$g_2^{19} = -\epsilon \sin t. \tag{10.12.153}$$

If we choose to use forward integration, and again restrict our attention to monomials through degree 2, the relevant $U_r^{r''}(h_n^s)$ are given by the relations

$$U_1^{r''} = h_1^{r''}, \tag{10.12.154}$$

$$U_2^{r''} = h_2^{r''}, \tag{10.12.155}$$

$$U_3^{r''} = h_3^{r''}, \tag{10.12.156}$$

$$U_4^r = 0 \text{ for } r \leq 3, \tag{10.12.157}$$

$$U_4^4 = (h_1^1)^2, \tag{10.12.158}$$

$$U_4^5 = 2h_1^1 h_1^2, \tag{10.12.159}$$

$$U_4^6 = 2h_1^1 h_1^3, \tag{10.12.160}$$

$$U_4^7 = (h_1^2)^2, \tag{10.12.161}$$

$$U_4^8 = 2h_1^2 h_1^3, \tag{10.12.162}$$

$$U_4^9 = (h_1^3)^2, \tag{10.12.163}$$

$$U_5^r = 0 \text{ for } r \leq 3, \tag{10.12.164}$$

$$U_5^4 = h_1^1 h_2^1, \tag{10.12.165}$$

$$U_5^5 = h_1^1 h_2^2 + h_1^2 h_2^1, \tag{10.12.166}$$

$$U_5^6 = h_1^3 h_2^1 + h_1^1 h_2^3, \tag{10.12.167}$$

$$U_5^7 = h_1^2 h_2^2, \tag{10.12.168}$$

$$U_5^8 = h_1^3 h_2^2 + h_1^2 h_2^3, \tag{10.12.169}$$

$$U_5^9 = h_1^3 h_2^3, \tag{10.12.170}$$

$$U_6^r = 0 \text{ for } r \leq 3, \tag{10.12.171}$$

$$U_6^4 = h_1^1 h_3^1, \tag{10.12.172}$$

$$U_6^5 = h_1^1 h_3^2 + h_1^2 h_3^1, \tag{10.12.173}$$

$$U_6^6 = h_1^3 h_3^1 + h_1^1 h_3^3, \tag{10.12.174}$$

$$U_6^7 = h_1^2 h_3^2, \tag{10.12.175}$$

$$U_6^8 = h_1^2 h_3^3 + h_1^3 h_3^2, \tag{10.12.176}$$

$$U_6^9 = h_1^3 h_3^3, \tag{10.12.177}$$

$$U_7^r = 0 \text{ for } r \leq 3, \tag{10.12.178}$$

$$U_7^4 = (h_1^2)^2, \tag{10.12.179}$$

$$U_7^5 = 2h_2^1 h_2^2, \tag{10.12.180}$$

$$U_7^6 = 2h_2^1 h_2^3, \tag{10.12.181}$$

$$U_7^7 = (h_2^2)^2, \tag{10.12.182}$$

$$U_7^8 = 2h_2^2 h_2^3, \tag{10.12.183}$$

$$U_7^9 = (h_2^3)^2, \tag{10.12.184}$$

$$U_8^r = 0 \text{ for } r \leq 3, \tag{10.12.185}$$

$$U_8^4 = h_2^1 h_3^1, \tag{10.12.186}$$

$$U_8^5 = h_2^1 h_3^2 + h_2^2 h_3^1, \tag{10.12.187}$$

$$U_8^6 = h_2^3 h_3^1 + h_2^1 h_3^3, \tag{10.12.188}$$

$$U_8^7 = h_2^2 h_3^2, \tag{10.12.189}$$

$$U_8^8 = h_2^2 h_3^3 + h_2^3 h_3^2, \tag{10.12.190}$$

$$U_8^9 = h_2^3 h_3^3, \tag{10.12.191}$$

$$U_9^r = 0 \text{ for } r \leq 3, \tag{10.12.192}$$

$$U_9^4 = (h_3^1)^2, \tag{10.12.193}$$

$$U_9^5 = 2h_3^1 h_3^2, \tag{10.12.194}$$

$$U_9^6 = 2h_3^1 h_3^3, \tag{10.12.195}$$

$$U_9^7 = (h_3^2)^2, \tag{10.12.196}$$

$$U_9^8 = 2h_3^2 h_3^3, \tag{10.12.197}$$

$$U_9^9 = (h_3^3)^2. \tag{10.12.198}$$

As before, the rules (12.65) and (12.66) hold.

The relevant equations for the h_a^r become

$$\dot{h}_a^{r''} = \sum_{r=1}^{3} g_a^r U_r^{r''} \text{ for } r'', a \in [1,3], \tag{10.12.199}$$

$$\dot{h}_a^{r''} = \sum_{r=1}^{9} g_a^r U_r^{r''} \text{ for } r'' \in [4,9] \text{ and } a \in [1,3]. \tag{10.12.200}$$

The initial conditions are

$$h_a^r(t^i) = \delta_a^r \text{ for } a \in [1,3] \text{ and } r \in [1,9]. \tag{10.12.201}$$

Note that because of (12.135) and (12.201) we can actually restrict a in the differential equations (12.199) and (12.200) for the $h_a^{r''}$ to the values $a = 1$ and $a = 2$ since we have the relations

$$h_3^r(t) = \delta_3^r \text{ for all } t. \tag{10.12.202}$$

For the use of backward integration in the three-variable case, Table 12.4 gives the nonzero values of $C_{br'r''}^r$ for $r \in [1, 9]$. In this method the equations (12.86) are to be integrated from $\bar{t} = t^{\text{fin}}$ back to $\bar{t} = t^{\text{in}}$ with the final conditions (12.88). From these equations, and the information about the $g_b^{r''}$ given by (12.143) through (12.153), it is not immediately obvious that there is the relation

$$h_3^r(\bar{t}, t^f) = \delta_3^r \text{ for all } \bar{t}, \tag{10.12.203}$$

which is consistent with (12.202). (And perhaps this is one minor drawback of this method.) However inspection of Table 12.4 for the cases $b = 1$ and $b = 2$ [the possibility $b = 3$ need not be considered because, as can be checked, $g_3^r = 0$ for all r] reveals there is no nonzero coefficient with $r' = 3$. Therefore, $\dot{h}_3^r(\bar{t}, t^f) = 0$ is the solution to (12.86) and (12.88) with $a = 3$ for the $g_b^{r''}$ in question. Correspondingly, as before, only the equations (12.86) with $a = 1$ and $a = 2$ need be integrated.

For further detail including numerical examples, see Section 29.12.

Table 10.12.4: Nonzero values of $C^{r}_{br'r''}$ for $r \in [1, 9]$ in the three-variable case.

r	b	r'	r''	C
1	1	1	1	1
1	2	2	1	1
1	3	3	1	1
2	1	1	2	1
2	2	2	2	1
2	3	3	2	1
3	1	1	3	1
3	2	2	3	1
3	3	3	3	1
4	1	1	4	1
4	1	4	1	2
4	2	2	4	1
4	2	5	1	1
4	3	3	4	1
4	3	6	1	1
5	1	1	5	1
5	1	4	2	2
5	1	5	1	1
5	2	2	5	1
5	2	5	2	1
5	2	7	1	2
5	3	3	5	1
5	3	6	2	1
5	3	8	1	1
6	1	1	6	1
6	1	4	3	2
6	1	6	1	1
6	2	2	6	1
6	2	5	3	1
6	2	8	1	1
6	3	3	6	1
6	3	6	3	1
6	3	3	1	2
7	1	1	7	1
7	1	5	2	1
7	2	2	7	1
7	2	7	2	2
7	3	3	7	1
7	3	8	2	1
8	1	1	8	1
8	1	5	3	1
8	1	6	2	1

r	b	r'	r''	C
8	2	2	8	1
8	2	7	3	2
8	2	8	2	1
8	3	3	8	1
8	3	8	3	1
8	3	9	2	2
9	1	1	9	1
9	1	6	3	1
9	2	2	9	1
9	2	8	3	1
9	3	3	9	1
9	3	9	3	2

10.12.9 Taylor Methods for the Hamiltonian Case

Suppose the equations of motion (12.1) arise from some Hamiltonian H. Then we may introduce deviation variables ζ as in Section 10.5, and employ the Hamiltonian $H^{\text{new}}(\zeta, t)$ of (5.4) to find the evolution of the variables ζ. Expand H^{new} in terms of the monomials $G_{r''}$ by writing

$$H^{\text{new}}(\zeta, t) = \sum_{r''} H^{r''}(t) G_{r''}(\zeta). \tag{10.12.204}$$

According to (5.4), all the $G_{r''}$ have degree two or greater. From Hamilton's equations of motion we have the result

$$\dot{\zeta}_a = [\zeta_a, H^{\text{new}}] = \sum_{r''} H^{r''}(t)[\zeta_a, G_{r''}]. \tag{10.12.205}$$

Since the monomials G_r form a basis, there is an expansion of the form

$$[\zeta_a, G_{r''}] = \sum_r E^r_{ar''} G_r \tag{10.12.206}$$

where the $E^r_{ar''}$ are universal coefficients that again describe certain combinatorial properties of monomials. With the aid of these coefficients (12.205) can be rewritten in the form

$$\dot{\zeta}_a = \sum_r (\sum_{r''} H^{r''} E^r_{ar''}) G_r. \tag{10.12.207}$$

Comparison of (12.17) and (12.207) gives the result

$$g^r_a(t) = \sum_{r''} E^r_{ar''} H^{r''}(t). \tag{10.12.208}$$

This result for the g^r_a may now be used to find the h^r_a in (12.19) by employing either forward or backward integration.

For the method of backward integration there is a further simplification. In its derivation we had to compute the quantities $(d/d\bar{t})G_{r'}(\zeta(\bar{t}))$. See (12.83). In the Hamiltonian case this quantity is given by the rule

$$(d/d\bar{t})G_{r'}(\zeta(\bar{t})) = [G_{r'}, H^{\text{new}}]. \tag{10.12.209}$$

See (1.7.4). Next insert the expansion (12.204) into (12.209) to find the result

$$(d/d\bar{t})G_{r'}(\zeta(\bar{t})) = \sum_{r''} H^{r''}(\bar{t})[G_{r'}, G_{r''}]. \tag{10.12.210}$$

Again, there are universal coefficients $F^{r}_{r'r''}$ describing certain monomial combinatorial properties such that

$$[G_{r'}, G_{r''}] = \sum_{r} F^{r}_{r'r''} G_r. \tag{10.12.211}$$

Note that as a consequence of (7.6.14) there is the restriction

$$F^{r}_{r'r''} = 0 \text{ unless } D(r') + D(r'') - 2 = D(r). \tag{10.12.212}$$

As a result of (12.210) and (12.211) we may write the relation

$$(d/d\bar{t})G_{r'}(\zeta(\bar{t}) = \sum_{r} \left(\sum_{r''} F^{r}_{r'r''} H^{r''}(\bar{t}) \right) G_r. \tag{10.12.213}$$

As a last step combine (12.80) through (12.82) and (12.213) to get the result

$$\dot{h}^{r}_{a}(\bar{t}, t^f) = -\sum_{r'r''} F^{r}_{r'r''} H^{r''}(\bar{t}) h^{r'}_{a}(\bar{t}, t^f). \tag{10.12.214}$$

These equations are the Hamiltonian analog of the general equations (12.86). As before, they are to be integrated from $\bar{t} = t^f$ back to $\bar{t} = t^i$ with the final conditions (12.88).

Let us compare the method just described for Taylor maps in the Hamiltonian case and that for factored product maps given in Section 10.5. Inspection of the equations (5.60) through (5.66) for the Lie generators show that they become ever more complicated with increasing order. By contrast, the equations (12.214) for the Taylor coefficients can be found easily to any desired order since the coefficients $F^{r}_{r'r''}$, which describe the monomial Poisson bracket results (12.211), can be obtained easily using Truncated Power Series Algebra. See Section 39.8. The price that must be paid for this ease of programming and computing to any desired order is the need to integrate many more differential equations. The numbers of equations that must be integrated in the Taylor and Lie methods are the same as those required to specify a map in Taylor or factored product Lie form, respectively. See Table 7.10.2.

Of course, no matter how a Taylor map is computed, if it arises from Hamiltonian equations it will be symplectic. And we know from the factorization Theorem of Section 7.6 that once a symplectic map is known in Taylor form, it can be rewritten in the factored product Lie form (7.6.3).

Exercises

10.12.1. Verify the entries in Table 12.2.

10.12.2. Consider the differential equation

$$\dot{x} = tx^2 \tag{10.12.215}$$

with the initial condition x^i at the initial time t^i. Verify that a particular solution is given by $x^d = 0$, and is the solution with $x(t^i) = 0$. Show that it has the general solution

$$x^f = x^i/\{1 - (x^i/2)[(t^f)^2 - (t^i)^2]\}. \tag{10.12.216}$$

Find the first few terms in the Taylor expansion of x^f in powers of x^i. This expansion is, in essence, an expansion about the solution x^d. Also compute the Taylor expansion of x^f in powers of x^i using both the methods of forward and backward integration. Verify that all your results agree.

10.12.3. Verify the derivation of (12.214).

10.12.4. Let $(\eta_1, \eta_2, \cdots \eta_{2n})$ be $2n$ "dummy" variables. Define functions $Z_a(\bar{t}, t^f; \eta)$ by the rule

$$Z_a(\bar{t}, t^f; \eta) = \sum_r h_a^r(\bar{t}, t^f) G_r(\eta), \tag{10.12.217}$$

which can also be written in the equivalent form

$$Z_a(\bar{t}, t^f; \eta) = \sum_{r'} h_a^{r'}(\bar{t}, t^f) G_{r'}(\eta). \tag{10.12.218}$$

Also define a function $H^{\text{new}}(\bar{t}, \eta)$ by the rule

$$H^{\text{new}}(\bar{t}, \eta) = \sum_{r''} H^{r''}(\bar{t}) G_{r''}(\eta). \tag{10.12.219}$$

Using (12.218), (12.219), (12.211), and (12.214) show that

$$[Z_a, H^{\text{new}}]_\eta = \sum_r \left(\sum_{r'r''} h_a^{r'} H^{r''} F_{r'r''}^r \right) G_r = -\sum_r \dot{h}_a^r(\bar{t}, t^f) G_r(\eta). \tag{10.12.220}$$

Hence, show that (12.114) is equivalent to the differential equation

$$\partial Z_a/\partial \bar{t} = -[Z_a, H^{\text{new}}]_\eta. \tag{10.12.221}$$

10.12.5. Suppose the complete variational equations (12.17) happen to be autonomous. That is, the functions g_a^r do not depend on time. In the Hamiltonian case assume, equivalently, that the H^r do not depend on time. This will be the case for idealized beam-line elements. See Chapters 13 and 14. In the case of equation (12.86), define quantities $K_{rr'}$ by the rule

$$K_{rr'} = \sum_{br''} C_{br'r''}^r g_b^{r''}. \tag{10.12.222}$$

They will be time independent since, by hypothesis, the $g_b^{r''}$ are time independent. In the case of equation (12.214), define quantities $K_{rr'}$ by the rule

$$K_{rr'} = \sum_{r''} F^r_{r'r''} H^{r''}. \tag{10.12.223}$$

They too will be time independent if the $H^{r''}$ are. Show that in either case, equation (12.86) or equation (12.214), there is the common result

$$\dot{h}^r_a = -\sum_{r'} K_{rr'} h^{r'}_a \tag{10.12.224}$$

which, when written in vector form, becomes the vector-matrix equation

$$\dot{h}_a = -K h_a. \tag{10.12.225}$$

Verify that (12.225) has the solution

$$h_a(\bar{t}, t^f) = e^{-(\bar{t}-t^f)K} h_a(t^f, t^f), \tag{10.12.226}$$

and consequently there is the result

$$h_a(t^i, t^f) = e^{-(t^i-t^f)K} h_a(t^f, t^f). \tag{10.12.227}$$

Also, according to (12.88), the vectors $h_a(t^f, t^f)$ are given by the relation

$$h^r_a(t^f, t^f) = \delta^r_a. \tag{10.12.228}$$

Thus, the Taylor map can be found explicitly in the autonomous case in terms of the matrix $\exp[-(t^i - t^f)K]$. This matrix can be computed using the scaling and squaring method of Section 4.1. Compare the results above for the Hamiltonian case with those found in Section 10.7, and in particular that of (7.11). Hint: For simplicity consider the two-variable case and suppose H^{new} consists of only an H_2 and an H_3.

Bibliography

General References

[1] E. Forest, *Beam Dynamics: A New Attitude and Framework*, Harwood Academic Press (1998).

[2] A.J. Dragt and E. Forest, "Computation of nonlinear behavior of Hamiltonian systems using Lie algebraic methods", *J. Math. Phys.* **24**, 2734 (1983).

Magnus Equations, Wei-Norman Equations, Fer Expansions, and General Lie-Algebraic References

[3] W. Magnus, "On the exponential solution of differential equations for a linear operator", *Commun. Pure Appl. Math.* **7**, 649 (1954).

[4] S. Blanes, F. Casas, J. Oteo, and J. Ros, "The Magnus expansion and some of its applications", *Physics Reports* **470**, 151-238 (2009).

[5] F. Fer, "Résolution de l'équation matricielle $\dot{U} = pU$ par produit infini d'exponentielles matricielles", *Bull. Classe Sci. Acad. Roy. Belg.* **44**, 818 (1958).

[6] E. Wichmann, "Note on the Algebraic Aspect of the Integration of a System of Ordinary Linear Differential Equations", *J. Math. Phys.* **6**, 876 (1961).

[7] P. Moan and J. Niesen, "Convergence of the Magnus Series", eprint arXiv:math/0609198 (2006).

[8] J. Wei and E. Norman, "Lie Algebraic Solution of Linear Differential Equations", *J. Math. Phys.* **4**, 575 (1963).

[9] J. Wei and E. Norman, "On global representations of the solutions of linear differential equations as products of exponentials", *Proc. Amer. Math. Soc.* **15**, 327 (1964).

[10] P.-V. Koseleff, "Formal Calculus for Lie Methods in Hamiltonian Mechanics", Ph.D. thesis, École Polytechnique (1993).

[11] M. Torres-Torriti and H. Michalska, "A Software Package for Lie-Algebraic Computations", *SIAM Review* **47**, 722 (2005).

[12] F. Casas, J.A. Oteo, and J. Ros, "Lie algebraic approach to Fer's expansions for classical Hamiltonian systems", *J. Phys. A: Math. Gen.* **24**, 4037 (1991).

[13] S. Blanes, F. Casas, J.A. Oteo, and J. Ros, "Magnus and Fer expansions for matrix differential equations: The convergence problem", *J. Phys. A: Math. Gen.* **31**, 259 (1998).

[14] S. Blanes, F. Casas, and A. Murua, "Splitting and composition methods in the numerical integration of differential equations", arXiv:0812.0377v1 [math.NA] 1 Dec 2008.

[15] A. Iserles and S.P. Nørsett, "On the solution of linear differential equations in Lie groups", *Phil. Trans. R. Soc. Lond. A* **357**, 983 (1999).

Taylor Maps

[16] The method of Backward Integration described in Sections 10.12.5 through 10.12.8 was discovered by F. Neri circa 1986.

Chapter 11

Geometric/Structure-Preserving Integration: Integration on Manifolds

Overview

The subject of *geometric integration*, also sometimes called *structure-preserving integration*, deals both with the construction of numerical integrators on manifolds and with the construction of numerical integrators that respect various group properties. This chapter provides an introduction to the subject of numerical integration on manifolds, mostly by way of three examples: rigid body motion, spin and qubits, and motion of a charged particle in a static magnetic field. Symplectic integration, the main topic of the next chapter, is a special case of geometric integration that respects a particular group property.[1]

Suppose a set of first-order ordinary differential equations is formulated in terms of coordinates in some ambient space and suppose it is known that, when these equations are integrated exactly, some class of trajectories is confined to a lower dimensional manifold in the ambient space, often a manifold associated with some group. If these same equations are integrated numerically with the aid of some integrator that is locally accurate through terms of order h^m one may expect, unless the integrator has special properties, that trajectories will deviate, at each integration step, from this lower dimensional manifold by terms of order h^{m+1}. In this context, a geometric numerical integrator is an integrator that, even though it may make local errors of order h^{m+1}, is guaranteed to produce trajectories that remain on the desired manifold, perhaps exactly to machine precision, or at least to substantially higher order in h. The construction and study of such numerical integrators is the subject of numerical integration on manifolds.[2]

Again suppose a set of first-order ordinary differential equations is formulated in terms of coordinates, call them z, in some ambient space. Let t^i and t^f be initial and final times. Now suppose it is known that the relation/map between $z^i = z(t^i)$ and $z^f = z(t^f)$ for

[1] Although for brevity we will not do so, for precision we should use the terms geometric *numerical* integration and, by association, symplectic *numerical* integration. Strictly speaking, this distinction is necessary because the term *geometric integration* can also refer to the extension of the integration procedures of ordinary calculus to integration over manifolds based on the use of differential forms.

[2] From this perspective numerical integrators that preserve integrals of motion, such as energy and angular momentum, are special instances of geometric integrators.

variable z^i, arising from exactly integrating initial conditions from $t = t^i$ to $t = t^f$, belongs to some group G that acts on the ambient space. What happens if these same equations are integrated numerically with the aid of some integrator that is locally accurate through terms of order h^m? Let t^b be the time at the *beginning* of some integration step. We may expect, unless the integrator has special properties, that the relation/map between $z(t^b)$ and $z(t^b+h)$ will differ from an element of G by terms of order h^{m+1}. In this context, a geometric integrator is an integrator with the property that, even though it may make local errors of order h^{m+1}, the relation/map between $z(t^b)$ and $z(t^b+h)$, for each integration step, is a map that belongs to G, perhaps exactly to machine precision, or at least to substantially higher order in h. Since for such a geometric integrator the maps for all successive integration steps belong to G, their product will also belong to G, and thus the map relating $z(t^i)$ and $z(t^f)$ will also belong to G.[3] In the Hamiltonian case, the ambient space is phase space, the group G is the group of all symplectic maps, and a symplectic numerical integrator is an integrator designed to produce a relation between z^i and z^f that is a symplectic map.

11.1 Numerical Integration on Manifolds: Rigid-Body Motion

As a first example of integration on manifolds, we will consider the problem of determining the motion of a rigid body with one point fixed and subject to an external torque $\boldsymbol{N}$ about the fixed point. We begin with the *kinematics* of rigid-body motion, and then follow with the *dynamics* of rigid-body motion.[4] Subsequent discussion treats various formulations of the equations of motion and various methods for their numerical integration.

11.1.1 Angular Velocity and Rigid-Body Kinematics

To describe rigid-body *kinematics*, suppose $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, $\boldsymbol{e}_3$ is an orthonormal right-hand triad of vectors that is fixed in space, and suppose $\boldsymbol{f}_1(t)$, $\boldsymbol{f}_2(t)$, $\boldsymbol{f}_3(t)$ is an orthonormal right-hand triad of vectors that is fixed in the body.[5] The $\boldsymbol{f}_j(t)$ will generally be time dependent because the body is expected to rotate, and they will be taken to have the initial values

$$\boldsymbol{f}_j(t^0) = \boldsymbol{e}_j. \tag{11.1.1}$$

Since the $\boldsymbol{e}_j$ and the $\boldsymbol{f}_j(t)$ are both orthonormal basis sets, there is a unique *orthogonal* matrix $R(t)$ such that

$$\boldsymbol{f}_j(t) = R(t)\boldsymbol{e}_j. \tag{11.1.2}$$

[3] Note that unlike the case of Chapter 10, we do not seek here, at least primarily, to compute the transfer map $\mathcal{M}$ that relates $z(t^i)$ and $z(t^f)$. Rather, we seek trajectories but require that, were all trajectories to be considered, the relation between z^i and z^f for all possible initial conditions z^i would be a map that belongs to G.

[4] The kinematic equations are the analog of the single-particle equations $\boldsymbol{v} = d\boldsymbol{r}/dt$ or $d\boldsymbol{r}/dt = \boldsymbol{v}$, and the dynamic equations are the analog of the equations $d\boldsymbol{v}/dt = \boldsymbol{F}/m$.

[5] We require that the space-fixed and body-fixed triads have the fixed point of the body as their common origin.

See Section 3.6.3. In view of (1.1) there is the relation

$$R(t^0) = I \tag{11.1.3}$$

and therefore, by continuity, $R(t) \in SO(3, \mathbb{R})$.

Let $(\omega_1^{bf}, \omega_2^{bf}, \omega_3^{bf})$ be the *components* of the *angular velocity* in the *body-fixed* frame. They are *defined* by the rules

$$\omega_j^{bf}(t) = -\boldsymbol{f}_k(t) \cdot \dot{\boldsymbol{f}}_\ell(t) \tag{11.1.4}$$

where j, k, ℓ is any cyclic permutation of $1, 2, 3$. Some implications of this definition of angular velocity are explored in Exercise 1.1. Here we examine how the $\omega_j^{bf}(t)$ are related to $R(t)$. From (1.2) there are the relations

$$\dot{\boldsymbol{f}}_\ell(t) = \dot{R}(t)\boldsymbol{e}_\ell, \tag{11.1.5}$$

and therefore (1.4) can be rewritten in the form

$$\omega_j^{bf}(t) = -[R(t)\boldsymbol{e}_k] \cdot [\dot{R}(t)\boldsymbol{e}_\ell] = -\boldsymbol{e}_k \cdot [R^T(t)\dot{R}(t)\boldsymbol{e}_\ell] = -\boldsymbol{e}_k \cdot [R^{-1}(t)\dot{R}(t)\boldsymbol{e}_\ell]. \tag{11.1.6}$$

Define a matrix $A(t)$ by the rule

$$A(t) = R^{-1}(t)\dot{R}(t) = R^T(t)\dot{R}(t). \tag{11.1.7}$$

With its use, (1.6) can be written in the yet more compact form

$$\omega_j^{bf}(t) = -\boldsymbol{e}_k \cdot A(t)\boldsymbol{e}_\ell. \tag{11.1.8}$$

What can be said about the matrix $A(t)$? From the orthogonality condition

$$R^T(t)R(t) = I \tag{11.1.9}$$

we conclude that

$$\dot{R}^T(t)R(t) + R^T(t)\dot{R}(t) = 0; \tag{11.1.10}$$

and from (1.7) we conclude that

$$A^T(t) = \dot{R}^T(t)R(t). \tag{11.1.11}$$

See Exercise 1.1. It follows from (1.10) that $A(t)$ is antisymmetric,

$$A^T(t) + A(t) = 0, \tag{11.1.12}$$

and therefore can be written in the form

$$A(t) = a_1(t)L^1 + a_2(t)L^2 + a_3(t)L^3 = \boldsymbol{a} \cdot \boldsymbol{L} \tag{11.1.13}$$

where the coefficients $a_j(t)$ are to be determined. Here we have used the notation of Exercise 3.7.30. Now insert (1.13) into (1.8). Computation using the properties of the L^j gives the results

$$a_j(t) = \omega_j^{bf}(t). \tag{11.1.14}$$

For example,

$$\begin{aligned}\omega_1^{bf} &= -\boldsymbol{e}_2 \cdot A\boldsymbol{e}_3 = -\boldsymbol{e}_2 \cdot [(\boldsymbol{a} \cdot \boldsymbol{L})\boldsymbol{e}_3] = -\boldsymbol{e}_2 \cdot (\boldsymbol{a} \times \boldsymbol{e}_3) \\ &= \boldsymbol{e}_2 \cdot (\boldsymbol{e}_3 \times \boldsymbol{a}) = (\boldsymbol{e}_2 \times \boldsymbol{e}_3) \cdot \boldsymbol{a} = \boldsymbol{e}_1 \cdot \boldsymbol{a} = a_1. \end{aligned} \tag{11.1.15}$$

To complete the discussion of rigid-body kinematics, rewrite (1.7) in the form

$$\dot{R} = RA \tag{11.1.16}$$

and introduce the notation

$$A = a_1 L^1 + a_2 L^2 + a_3 L^3 = \omega_1^{bf} L^1 + \omega_2^{bf} L^2 + \omega_3^{bf} L^3 = \boldsymbol{\omega}^{bf} \cdot \boldsymbol{L}. \tag{11.1.17}$$

By substituting (1.17) into (1.16) we find for $R(t)$ the *kinematic* differential equation of motion

$$\dot{R} = R\, \boldsymbol{\omega}^{bf} \cdot \boldsymbol{L}. \tag{11.1.18}$$

Note that, since R is 3×3, the matrix differential equation (1.18) amounts to nine first-order differential equations. Also note, in passing, that $\boldsymbol{\omega}^{bf} \cdot \boldsymbol{L} \in so(3, \mathbb{R})$.

11.1.2 Angular Velocity and Rigid-Body Dynamics

We now turn to the *dynamics* of rigid-body motion. Suppose the body-fixed frame is oriented in the body in such a way that the moment of inertia tensor is diagonal with diagonal entries I_1, I_2, I_3. (Caution! We will also continue to use, as we have already done, the symbol I without a subscript to denote the 3×3 identity matrix.) Then, from the work of Euler, we know that there are the *dynamic* equations of motion

$$\dot{\omega}_1^{bf} = N_1^{bf}/I_1 + \omega_2^{bf}\omega_3^{bf}(I_2 - I_3)/I_1, \tag{11.1.19}$$

$$\dot{\omega}_2^{bf} = N_2^{bf}/I_2 + \omega_3^{bf}\omega_1^{bf}(I_3 - I_1)/I_2, \tag{11.1.20}$$

$$\dot{\omega}_3^{bf} = N_3^{bf}/I_3 + \omega_1^{bf}\omega_2^{bf}(I_1 - I_2)/I_3. \tag{11.1.21}$$

Here the $N_j^{bf}(R, \boldsymbol{\omega^{bf}}, t)$ are the components of $\boldsymbol{N}$ in the body-fixed frame.[6]

11.1.3 Problem of Integrating the Combined Kinematic and Dynamic Equations

Taking both the kinematic and dynamic equations into account, our task is to integrate the nine kinematic differential equations (1.18) and the three dynamic differential equations (1.19) through (1.21) given, at time t^i, some initial orientation $R(t^i)$ and some initial angular velocity $\boldsymbol{\omega}^{bf}(t^i)$. In this context, the ambient space is $9 + 3 = 12$ dimensional.

But now we see that there is a computational problem. We know that R must be an orthogonal matrix. See (1.9). Also, it easily verified that *exact* integration of the matrix

[6] Note that we have allowed the possibility that $\boldsymbol{N}$ is $\boldsymbol{\omega}^{bf}$ dependent, and therefore have gone beyond a Lagrangian/Hamiltonian formulation in allowing for the possibility of dissipative forces.

differential equations (1.18), no matter how $\boldsymbol{\omega}^{bf}$ depends on the time t, maintains the orthogonality condition (1.9) if the initial matrix $R(t^i)$ is orthogonal. See Exercise 1.2. Thus we know that, although R has nine entries, it must lie on the 3-dimensional manifold $SO(3,\mathbb{R})$. However, if some *numerical* integration method is used that is locally accurate only through terms of order h^m, then we expect that the orthogonality condition (1.9) will also be maintained only through terms of order h^m. Thus, in the course of numerical integration, R may be expected to move off the manifold $SO(3,\mathbb{R})$. Moreover, if R is not orthogonal, then the quantities $N_j^{bf}(R,\boldsymbol{\omega}^{bf},t)$ required in (1.19) through (1.21) are not defined since they are only physically specified when R is orthogonal.[7]

11.1.4 Solution by Projection

What to do? One approach is to orthogonalize the R provided after or during the course of each integration step, whenever an orthogonal R is needed to compute the N_j^{bf} or the actual orientation of the body. The orthogonalization can be done, for example, using any of the methods of Section 3.6.4 and Exercise 4.5.7. This process of orthogonalization is an example of a general problem: Given a submanifold embedded in some larger manifold (or ambient space), and given an element in the larger manifold, how does one find a related element in the submanifold, and what is the optional choice for such an element? In the nomenclature of geometric integration, the process for determining such an element is called *projection*.[8]

11.1.5 Solution by Parameterization: Euler Angles

Another approach is to parameterize R in such a way that it is guaranteed to be orthogonal. For example, suppose we employ the Euler parameterization (3.7.207). Then use of (1.18), rewritten in the form

$$\boldsymbol{\omega}^{bf}\cdot\boldsymbol{L} = R^{-1}\dot{R}, \tag{11.1.22}$$

gives the relations

$$\omega_1^{bf} = -\dot{\phi}\sin\theta\cos\psi + \dot{\theta}\sin\psi, \tag{11.1.23}$$

$$\omega_2^{bf} = \dot{\phi}\sin\theta\sin\psi + \dot{\theta}\cos\psi, \tag{11.1.24}$$

$$\omega_3^{bf} = \dot{\phi}\cos\theta + \dot{\psi}, \tag{11.1.25}$$

[7] Note that a similar problem arises in the numerical integration of (10.4.28) with the initial condition (10.4.29). Although the ambient space is $(2n)^2$ dimensional, the solution of (10.4.28) is required to originate and remain in the $n(2n+1)$ dimensional submanifold $Sp(2n,\mathbb{R})$.

[8] For the problem at hand we need some projection $E^9 \rightarrow SO(3,\mathbb{R})$ where E^9 denotes 9-dimensional Euclidean space. Recall also the problem of matrix symplectification, projections $E^{(2n)^2} \rightarrow Sp(2n,\mathbb{R})$, treated in Subsection 3.6.5 and Chapter 4. Observe that all the integration methods of Chapter 2 require the evaluation of the right side $\boldsymbol{f}$ at intermediate points, and generally these points will not be on the desired manifold, but only nearby. Therefore $\boldsymbol{f}$ may not even be defined at these points unless $\boldsymbol{f}$ can be extrapolated off the manifold to nearby points in the ambient space. Alternatively, the intermediate points can be projected from the ambient space onto the manifold, and $\boldsymbol{f}$ is then computed at these projected intermediate points. In either case one has to ensure that extrapolation or projection does not spoil the desired accuracy of the integration method.

from which it follows that

$$\dot{\phi} = (1/\sin\theta)(\omega_2^{bf}\sin\psi - \omega_1^{bf}\cos\psi), \tag{11.1.26}$$

$$\dot{\theta} = \omega_1^{bf}\sin\psi + \omega_2^{bf}\cos\psi, \tag{11.1.27}$$

$$\dot{\psi} = \omega_3^{bf} + (\cot\theta)(\omega_1^{bf}\cos\psi - \omega_2\sin\psi). \tag{11.1.28}$$

See Exercise 1.3. In terms of Euler angles, our task is to integrate the equations (1.26) through (1.28) and (1.19) through (1.21) where now

$$N_j^{bf} = N_j^{bf}(\phi,\theta,\psi,\dot{\phi},\dot{\theta},\dot{\psi},t). \tag{11.1.29}$$

11.1.6 Problem of Kinematic Singularities

Have we achieved our goal? Only in a fashion. It is true that equations (1.26) through (1.28) and (1.19) through (1.21) constitute a set of six first-order equations of motion, which is what we expect for a system with three degrees of freedom. Also, $R(\phi,\theta,\psi)$ is guaranteed to be orthogonal no matter how inaccurately the equations of motion are numerically integrated. However, note that the factor $(1/\sin\theta)$ in (1.26) and the factor $(\cot\theta)$ in (1.28) become *singular* when $\theta = 0$ or $\theta = \pi$. Therefore these equations are unsuitable for numerical integration whenever $\theta \simeq 0$ or $\theta \simeq \pi$. The singularity at $\theta = 0$ is particularly alarming because it means that Euler angles do not provide a good coordinate patch in the vicinity $R \simeq I$.

Now there is nothing *a priori* to prevent $\theta \simeq 0$ or $\theta \simeq \pi$ from happening over the course of a rigid body's motion. (For example, a top is allowed to be vertical or inverted.) These singularities are *kinematic* in the sense that they are an artifact of our choice of coordinate system, and are not intrinsic to the motion of the system being considered. (See Exercise 8.2.11.) However, it can be shown from topological considerations that singularities of this type must arise no mater how R is parameterized if only three parameters are used. A global three-variable and singularity-free parameterization of $SO(3,\mathbb{R})$ is impossible. Consequently, if only three variables are used, it is necessary to change coordinate patches whenever a singularity of the coordinate system in current use is approached.[9] This complication might appear to make it difficult to write a robust three-variable numerical integration procedure that would apply for all possible rigid-body motions.[10] However, we will eventually entertain the possibility of changing the coordinate system frequently, and perhaps at every integration step.

[9] Moreover, even if a trajectory does not pass directly through a kinematic singularity, the presence of a singularity still affects the integration of nearby trajectories because numerical integration is based on the assumption of analyticity and the use of Taylor series, and nearby singularities affect the convergence of Taylor series.

[10] We also remark that the kind of problem we have encountered here is expected to occur quite generally whenever one seeks to numerically integrate trajectories that are known, and hence required, to lie within some group manifold. Because group manifolds do not generally have the topology of Euclidean space, there is generally no global singularity-free coordinate system that can be used.

11.1.7 Quaternions to the Rescue

Our ancestors have discovered that, for rigid-body motion, this troublesome singularity problem can be overcome in an optimum way by the use of (unit) quaternion parameters $w = (w_0, w_1, w_2, w_3)$. Again see Exercise 8.2.11. Note that there are now four parameters, rather then three as in the Euler-angle case, and they are related by the constraint

$$w \cdot w = \sum_{j=0}^{3} w_j^2 = 1. \tag{11.1.30}$$

Quaternion parameters and $SO(3,\mathbb{R})$ matrices are connected by (8.2.73). Quaternion and angle-axis parameters are connected by the relations

$$w_0 = \cos(\theta/2), \tag{11.1.31}$$

$$\boldsymbol{w} = -\boldsymbol{n}\sin(\theta/2). \tag{11.1.32}$$

Quaternion and Euler-angle parameters are connected by the relations

$$w_0 = \cos(\theta/2)\cos[(1/2)(\phi+\psi)], \tag{11.1.33}$$

$$w_1 = -\sin(\theta/2)\sin[(1/2)(-\phi+\psi)], \tag{11.1.34}$$

$$w_2 = -\sin(\theta/2)\cos[(1/2)(-\phi+\psi)], \tag{11.1.35}$$

$$w_3 = -\cos(\theta/2)\sin[(1/2)(\phi+\psi)]. \tag{11.1.36}$$

See Exercise 1.4.

More to the point for our purposes, the angular velocities are given in terms of quaternion parameters by the relations

$$\omega_1^{bf} = 2(w_1\dot{w}_0 - w_0\dot{w}_1 + w_3\dot{w}_2 - w_2\dot{w}_3), \tag{11.1.37}$$

$$\omega_2^{bf} = 2(w_2\dot{w}_0 - w_0\dot{w}_2 + w_1\dot{w}_3 - w_3\dot{w}_1), \tag{11.1.38}$$

$$\omega_3^{bf} = 2(w_3\dot{w}_0 - w_0\dot{w}_3 + w_2\dot{w}_1 - w_1\dot{w}_2). \tag{11.1.39}$$

See Exercise 1.5. To these relations we add the further relation

$$0 = -\sum_{j=0}^{3} w_j\dot{w}_j, \tag{11.1.40}$$

which follows from differentiating (1.30). Together the equations (1.37) through (1.40) can be written in the vector/matrix form

$$\begin{pmatrix} 0 \\ \omega_1^{bf}/2 \\ \omega_2^{bf}/2 \\ \omega_3^{bf}/2 \end{pmatrix} = \begin{pmatrix} -w_0 & -w_1 & -w_2 & -w_3 \\ w_1 & -w_0 & w_3 & -w_2 \\ w_2 & -w_3 & -w_0 & w_1 \\ w_3 & w_2 & -w_1 & -w_0 \end{pmatrix} \begin{pmatrix} \dot{w}_0 \\ \dot{w}_1 \\ \dot{w}_2 \\ \dot{w}_3 \end{pmatrix}. \tag{11.1.41}$$

Remarkably, the 4 × 4 matrix on the right side of (1.41) is orthogonal! See Exercise 1.6. Consequently (1.41) can be inverted easily to give the relation

$$\begin{pmatrix} \dot{w}_0 \\ \dot{w}_1 \\ \dot{w}_2 \\ \dot{w}_3 \end{pmatrix} = \begin{pmatrix} -w_0 & w_1 & w_2 & w_3 \\ -w_1 & -w_0 & -w_3 & w_2 \\ -w_2 & w_3 & -w_0 & -w_1 \\ -w_3 & -w_2 & w_1 & -w_0 \end{pmatrix} \begin{pmatrix} 0 \\ \omega_1^{bf}/2 \\ \omega_2^{bf}/2 \\ \omega_3^{bf}/2 \end{pmatrix}. \tag{11.1.42}$$

Taking components of (1.42) yields the four kinematic equations of motion

$$\dot{w}_0 = (1/2)(\omega_1^{bf} w_1 + \omega_2^{bf} w_2 + \omega_3^{bf} w_3), \tag{11.1.43}$$

$$\dot{w}_1 = (1/2)(-\omega_1^{bf} w_0 - \omega_2^{bf} w_3 + \omega_3^{bf} w_2), \tag{11.1.44}$$

$$\dot{w}_2 = (1/2)(\omega_1^{bf} w_3 - \omega_2^{bf} w_0 - \omega_3^{bf} w_1), \tag{11.1.45}$$

$$\dot{w}_3 = (1/2)(-\omega_1^{bf} w_2 + \omega_2^{bf} w_1 - \omega_3^{bf} w_0). \tag{11.1.46}$$

It is these four kinematic equations, along with the three dynamic equations (1.19) through (1.21), that are to be integrated. Note that equations (1.43) through (1.46) are singularity free. Indeed, they are *linear* in the w_j. They are therefore ideal for numerical integration. For further elaboration on this point, see the discussion at the end of Exercise 1.13.

It is easily verified that exact integration of the differential equations (1.43) through (1.46) preserves the unit sphere condition (1.30) if this condition is satisfied at some initial time t^i. See Exercise 1.8. However, if we are integrating the equations of motion numerically by some routine that is only exact through order h^m, we may expect that the condition (1.30) will be violated by terms of order h^{m+1} at each integration step. That is, instead of remaining on the unit sphere S^3, w will become a general point in the ambient four-dimensional space E^4. One simple procedure to overcome this problem is to repeatedly project, by simple scaling, $w \in E^4$ back onto the unit sphere S^3 any time a unit w is required to compute $R(w)$. So doing requires little computational overhead.

11.1.8 Modification of the Quaternion Kinematic Equations of Motion

Another procedure is to modify the kinematic equations of motion. Define an *error* quantity ϵ that measures the departure of w from S^3 by the rule

$$\epsilon = 1 - w \cdot w. \tag{11.1.47}$$

Replace equations (1.43) through (1.46) by the modified equations

$$\dot{w}_0 = (1/2)(\omega_1^{bf} w_1 + \omega_2^{bf} w_2 + \omega_3^{bf} w_3) + k\epsilon w_0, \tag{11.1.48}$$

$$\dot{w}_1 = (1/2)(-\omega_1^{bf} w_0 - \omega_2^{bf} w_3 + \omega_3^{bf} w_2) + k\epsilon w_1, \tag{11.1.49}$$

$$\dot{w}_2 = (1/2)(\omega_1^{bf} w_3 - \omega_2^{bf} w_0 - \omega_3^{bf} w_1) + k\epsilon w_2, \tag{11.1.50}$$

$$\dot{w}_3 = (1/2)(-\omega_1^{bf} w_2 + \omega_2^{bf} w_1 - \omega_3^{bf} w_0) + k\epsilon w_3, \tag{11.1.51}$$

where k is some constant satisfying $k > 0$. Evidently the modified equations have the same solutions as the original equations as long as $\epsilon = 0$. But, if for some reason $\epsilon \neq 0$, the modified equations *drive* ϵ to 0 if k is positive! Again see Exercise 1.8. Thus, even when integrated numerically, the modified equations (1.48) through (1.51) along with (1.19) through (1.21) provide a satisfactory description of rigid body motion, and are commonly used for this purpose in applications that range from inertial guidance (including space craft orientation/control) through robotics to virtual reality.

11.1.9 Local Coordinate Patches

So far, in the case of rigid-body motion, we have been able to finesse the problem of maintaining the manifold condition (1.9) either by some projection $E^9 \rightarrow SO(3, \mathbb{R})$ or the use of a remarkable coordinate system, namely quaternions, and the simpler projection $E^3 \rightarrow S^3$ by scaling, or by modification of the quaternion kinematic equations of motion. Are there other approaches, and in particular are there approaches that are also applicable in a more general context?

One procedure is to introduce a local coordinate patch at each integration step. Suppose, for a given integration step, we write

$$R(t) = R^b R^v(t). \tag{11.1.52}$$

Here

$$R^b = R(t^b) \tag{11.1.53}$$

where t^b is the time at the *beginning* of the integration step, and R^v is a *variable* rotation matrix near the identity with the property

$$R^v(t^b) = I. \tag{11.1.54}$$

From (1.52) we have the relation

$$\dot{R} = R^b \dot{R}^v, \tag{11.1.55}$$

and substituting this result and (1.52) into (1.18) yields the relation

$$\dot{R}^v = R^v\, \boldsymbol{\omega}^{bf} \cdot \boldsymbol{L}. \tag{11.1.56}$$

We wish to integrate (1.56) to find $R^v(t^b + h)$ starting with the initial condition (1.54). Then, having done so, we have from (1.52) the result

$$R(t^b + h) = R^b R^v(t^b + h). \tag{11.1.57}$$

Note that in view of (1.54) what is needed in any given instance, if parameters are to be employed, is a local coordinate patch in the vicinity of the identity.

11.1.10 Canonical Coordinates of the Second Kind: Tait-Bryan Angles

In the case of $SO(3,\mathbb{R})$, how can we parameterize R^v near the identity? One possibility is to employ canonical coordinates of the second kind (see Section 7.9) to write, for example, the Ansatz

$$R^v(\lambda) = \exp(\lambda_1 L^1)\exp(\lambda_2 L^2)\exp(\lambda_3 L^3) \tag{11.1.58}$$

with the quantities λ_j being parameters. [In the context of rigid-body motion, the quantities λ_j in (1.58) are called *Tait-Bryan* or *Cardan* angles.] Unlike the Euler-angle parameterization, application of the BCH formula shows that such parameterizations are well defined for R^v near I and the λ_j near 0 since all three L^j are employed.[11]

For these coordinates use of (1.22) yields the relations

$$\omega_1^{bf} = \dot\lambda_1\cos(\lambda_2)\cos(\lambda_3) + \dot\lambda_2\sin(\lambda_3), \tag{11.1.59}$$

$$\omega_2^{bf} = -\dot\lambda_1\cos(\lambda_2)\sin(\lambda_3) + \dot\lambda_2\cos(\lambda_3), \tag{11.1.60}$$

$$\omega_3^{bf} = \dot\lambda_1\sin(\lambda_2) + \dot\lambda_3. \tag{11.1.61}$$

And inverting the relations (1.59) through (1.61) yields the kinematic equations of motion

$$\dot\lambda_1 = [1/\cos(\lambda_2)][\omega_1^{bf}\cos(\lambda_3) - \omega_2^{bf}\sin(\lambda_3)], \tag{11.1.62}$$

$$\dot\lambda_2 = \omega_1^{bf}\sin(\lambda_3) + \omega_2^{bf}\cos(\lambda_3), \tag{11.1.63}$$

$$\dot\lambda_3 = \omega_3^{bf} - \tan(\lambda_2)[\omega_1^{bf}\cos(\lambda_3) - \omega_2^{bf}\sin(\lambda_3)]. \tag{11.1.64}$$

See Exercise 1.9. In terms of these coordinates our task is to integrate the equations (1.62) through (1.64) and (1.19) through (1.21) where now

$$N_j^{bf} = N_j^{bf}(\lambda_1,\lambda_2,\lambda_3,\dot\lambda_1,\dot\lambda_2,\dot\lambda_3,t). \tag{11.1.65}$$

Observe, as anticipated, that the equations of motion are nonsingular for small λ_j. However they are singular when $\lambda_2 = \pm\pi/2$. (For the source of these singularities, again see Exercise 1.9.) Thus, like the case for Euler angles, these coordinates cannot be used globally.

11.1.11 Canonical Coordinates of the First Kind: Angle-Axis Parameters

Alternatively, inspired by the angle-axis parameterization (3.7.199), another possibility is to introduce parameters $\lambda_1, \lambda_2, \lambda_3$ and make the Ansatz

$$R^v(\lambda) = \exp(\lambda_1 L^1 + \lambda_2 L^2 + \lambda_3 L^3) = \exp(\boldsymbol{\lambda}\cdot\boldsymbol{L}). \tag{11.1.66}$$

[11] Canonical coordinates of the first and second kind are both well defined near the origin for any finite-dimensional Lie group. However, we know that in some cases canonical coordinates of the first kind cannot be set up globally. See Section 3.8. The global status of canonical coordinates of the second kind is less clear. In general there is a global polar decomposition, which amounts to a hybrid of canonical coordinates of the first and second kinds.

That is, we have expressed $R^v \in SO(3,\mathbb{R})$ and near the identity in terms of elements in the Lie algebra $so(3,\mathbb{R})$; and we have parameterized the Lie algebra. Since the map between group elements near the identity and Lie elements near the origin is well defined, we know that this parameterization is well defined for R^v near I and the λ_j near 0. As described in Section 7.9, this parameterization amounts to using canonical coordinates of the first kind.

This parameterization yields the remarkably symmetric looking kinematic equations of motion

$$\dot{\boldsymbol{\lambda}} = \boldsymbol{\omega}^{bf} + (1/2)(\boldsymbol{\lambda} \times \boldsymbol{\omega}^{bf}) + [(\boldsymbol{\lambda} \cdot \boldsymbol{\omega}^{bf})\boldsymbol{\lambda} - (\boldsymbol{\lambda} \cdot \boldsymbol{\lambda})\boldsymbol{\omega}^{bf}]\{1/|\boldsymbol{\lambda}|^2 - [1/(2|\boldsymbol{\lambda}|)]\cot(|\boldsymbol{\lambda}|/2)\}. \tag{11.1.67}$$

See Exercise 1.10. For the inverse relation that expresses $\boldsymbol{\omega}^{bf}$ in terms of $\boldsymbol{\lambda}$ and $\dot{\boldsymbol{\lambda}}$, see Exercise 1.11.

Observe, again as anticipated, that the equations of motion are nonsingular for small λ_j, namely when $|\boldsymbol{\lambda}| < 2\pi$. But they are singular when $|\boldsymbol{\lambda}| = 2\pi$. We expect this singularity to occur because we see from (3.7.188) and (3.7.202) that the individual components of $\boldsymbol{n}$ are not defined in terms of v or R when $\theta = |\boldsymbol{\lambda}| = 2\pi$.

We reiterate that even when coordinates can be set up globally, as in the case of $SO(3,\mathbb{R})$ using either Euler angles or angle-axis parameters, there may still be singularities in the equations of motion because at some points the inverse map from the group to the Lie algebra may not be well defined. We have seen an example of this problem in the case of Euler angles. Every element of $SO(3,\mathbb{R})$ can be written in Euler form, but at $\theta = 0$ and $\theta = \pi$ the inverse map is not well defined. The same is true for angle-axis parameters. Every element of $SO(3,\mathbb{R})$ can be written in angle-axis form, but at $\theta = 2\pi$ the inverse map is not well defined.

11.1.12 Cayley Parameters

Quadratic groups, including $SO(3,\mathbb{R})$ and $SU(2)$, can also be parameterized near the identity in terms of Cayley parameters. When Cayley parameters are used for $SO(3,\mathbb{R})$ or $SU(2)$, call them μ_j, the task is to find differential equations that specify the $\dot{\mu}_j$ in terms of the μ_j and ω^{bf}.

We first consider the case of $SO(3,\mathbb{R})$ and employ the Cayley parameterization

$$R^v(\boldsymbol{\mu}) = (I + \boldsymbol{\mu} \cdot \boldsymbol{L})(I - \boldsymbol{\mu} \cdot \boldsymbol{L})^{-1}. \tag{11.1.68}$$

For this parameterization it can be shown that there are the kinematic equations of motion

$$\dot{\boldsymbol{\mu}} = (1/2)[\boldsymbol{\omega}^{bf} + (\boldsymbol{\mu} \times \boldsymbol{\omega}^{bf}) + (\boldsymbol{\mu} \cdot \boldsymbol{\omega}^{bf})\boldsymbol{\mu}]. \tag{11.1.69}$$

See Exercise 1.13.

Next consider the case of $SU(2)$ and employ the Cayley parameterization

$$u^v(\boldsymbol{\mu}) = (I + \boldsymbol{\mu} \cdot \boldsymbol{K})(I - \boldsymbol{\mu} \cdot \boldsymbol{K})^{-1}. \tag{11.1.70}$$

Here we have again used the notation of Exercise 3.7.30. In this case there are the kinematic equations of motion

$$\dot{\boldsymbol{\mu}} = (1/2)[\boldsymbol{\omega}^{bf} + (\boldsymbol{\mu} \times \boldsymbol{\omega}^{bf}) + (1/2)(\boldsymbol{\mu} \cdot \boldsymbol{\omega}^{bf})\boldsymbol{\mu} - (1/4)(\boldsymbol{\mu} \cdot \boldsymbol{\mu})\boldsymbol{\omega}^{bf}]. \tag{11.1.71}$$

Again see Exercise 1.13.

What can be said about the singularity structure of the kinematic equations of motion (1.69) and (1.71)? Strangely enough, both sets appear to be singularity free! However, this appearance is deceptive because both sets of kinematic equations of motion are singular in $\boldsymbol{\mu}$ at infinity, and there is the possibility that this singularity can be encountered in *finite* time. Yet again see Exercise 1.13.

11.1.13 Summary of Integration Using Local Coordinates

Upon employing (1.58) or (1.66) in (1.56) we can find first-order differential equations that, analogous to the relations (1.26) through (1.28), specify $\dot{\lambda}_1, \dot{\lambda}_2, \dot{\lambda}_3$ in terms of the λ_j and ω_j^{bf}. Moreover, these equations will be singularity free in the neighborhood the origin in λ space. What we have done is to convert the group-space differential equation (1.56) into a set of differential equations for the parameters λ_j. These equations, with the initial conditions

$$\lambda_j(t^b) = 0, \tag{11.1.72}$$

see (1.54), can be integrated numerically for one integration step using any convenient method to find the quantities $\lambda_j(t^b + h)$. If h is sufficiently small, the $\lambda(t)$ for $t \in [t^b, t^b + h]$ will remain small, and the differential equations specifying the $\dot{\lambda}_j$ will remain singularity free over the course of integration. Once the quantities $\lambda_j(t^b + h)$ have been found, $R(t^b + h)$ is given, depending on what type of canonical parameterization has been employed, by either (1.58) or (1.66).

As a modification of this procedure, one may integrate for k steps to find $\lambda_j(t^b + kh)$ and subsequently $R(t^b + kh)$. What is required then is continual checking that the λ_j have not come too close to singular values.

Evidently generalizations of the methods just described can in principle be employed for any Lie group. The difficulties encountered lie only in determining the equations that specify the $\dot{\lambda}_j$, see for example Exercises 1.9 and 1.10, and in evaluating the exponentials that occur with the use of (1.58) or (1.66). Assuming these exponentials can be evaluated accurately, the result for $R^v(t^b + h)$ will lie on the group manifold to high accuracy even though the $\lambda_j(t^b + h)$ are only exact through terms of order h^m. Finally, we must assume that the group multiplications involved in (1.57) can also be carried out with high accuracy.

Similarly, with the use of Cayley parameters, the difficulties lie only in determining the equations that specify the $\dot{\mu}_j$ [see for example (1.69) and (1.71)], and in carrying out the matrix inversions and multiplications that occur with the use of a Cayley representation [see (1.68) and (1.70)]. Assuming these inversions and multiplications can be evaluated accurately, the results for $R^v(t^b + h)$ or $u^v(t^b + h)$ will lie on the associated group manifold to high accuracy even though the $\mu_j(t^b + h)$ are only exact through terms of order h^m. Finally, we must again assume that the group multiplications involved in (1.57), for example, can also be carried out with high accuracy.

11.1.14 Integration in the Lie Algebra: Exponential Representation

With a suitable translation of the origin in time, the differential equation (1.56) with the initial condition (1.54) is a special case of the general differential equation of the form

$$\dot{M}(t) = M(t)A(t) \tag{11.1.73}$$

with the initial condition

$$M(0) = I, \tag{11.1.74}$$

and our goal is to find $M(h)$. Here $M(t)$ is expected to belong to some Lie group G and is near the identity for small t; and $A(t)$ belongs to the Lie algebra of G, which we denote by $\mathcal{L}(G)$. See Appendix C. In particular, we wish to obtain $M(h)$ by numerical means with a possible local error of order h^{m+1} and, despite this possible error, we want to guarantee that $M(h)$ is in G.

Since $M(t)$ is near the identity, we may write

$$M(t) = \exp[B(t)] \tag{11.1.75}$$

where $B(t)$ is in $\mathcal{L}(G)$ and is small (near the origin). If we can find $B(h)$ with a local error of order h^{m+1} and if, despite this possible error, we can assure that $B(h)$ is in $\mathcal{L}(G)$, then we know that $M(h) = \exp[B(h)]$ will have the desired local accuracy and is guaranteed to be in G. We will now see that a suitable $B(h)$ can be found by converting the differential equation (1.73), a *group* differential equation for the group elements $M(t)$ in terms of the group elements $M(t)$ and the Lie elements $A(t)$, into a *Lie* differential equation for the Lie elements $B(t)$ in terms of the Lie elements $B(t)$ and the Lie elements $A(t)$.

Before proceeding further, here is a chance to learn some terminology: Thinking geometrically, we may view $M(t)$ as a path in G, and we may view $B(t)$ as a path in $\mathcal{L}(G)$. In this context, the path $B(t)$ is said to be a *lift* of the path $M(t)$. That is, we may view $\mathcal{L}(G)$, the Lie algebra of G, as lying "above" the group G. Correspondingly, we may say that the path $B(t)$ is obtained by "lifting" the path $M(t)$ in the group G up to a path in $\mathcal{L}(G)$. Upon solving (1.75) for B in terms of M, we have the relation

$$B(t) = \log[M(t)], \tag{11.1.76}$$

and we see that the lift in question is accomplished by the logarithmic function. Conversely, in view of (1.75), the exponential function "lowers" the path $B(t)$ in $\mathcal{L}(G)$ down to the path $M(t)$ in the group G.

Let us continue. Given the differential equation (1.73) for the path $M(t)$, what is the associated differential equation for the path $B(t)$? The relation (1.75) can be differentiated to yield the result

$$\dot{M}(t) = M(t)\,\mathrm{iex}[-\#B(t)\#]\dot{B}(t). \tag{11.1.77}$$

See (*) in Appendix C. Also, in view of (1.73), the relation (1.77) can be rewritten in the form

$$\mathrm{iex}[-\#B(t)\#]\dot{B}(t) = A(t). \tag{11.1.78}$$

Finally, as in Section 10.3, the relation (1.78) can be inverted to become

$$\dot{B}(t) = \{\mathrm{iex}[-\#B(t)\#]\}^{-1}A(t). \tag{11.1.79}$$

We have obtained a (somewhat fearsome looking) differential equation for B in terms of B and A. Our task is to integrate this equation from $t = 0$ to $t = h$ with, in view of (1.74), the initial condition

$$B(0) = 0. \tag{11.1.80}$$

What can we make of (1.79)? The function $[\mathrm{iex}(-w)]^{-1}$ has an expansion of the form

$$[\mathrm{iex}(-w)]^{-1} = \sum_{\ell=0}^{\infty} b_\ell w^\ell \tag{11.1.81}$$

where the coefficients b_ℓ are known. Again see Appendix C. Correspondingly, the differential equation (1.79) is equivalent to the equation

$$\begin{aligned}\dot{B}(t) &= \sum_{\ell=0}^{\infty} b_\ell [\#B(t)\#]^\ell A(t)\\ &= \{b_0 + b_1[\#B(t)\#] + b_2[\#B(t)\#]^2 + \cdots\}A(t)\\ &= b_0A(t) + b_1\{B(t), A(t)\} + b_2\{B(t), \{B(t), A(t)\}\} + \cdots . \end{aligned}\tag{11.1.82}$$

In general, all terms in the expansion (1.82) need to be retained. That is, in general, we need to sum the series (1.82) which, in the general case, can be a formidable task.[12]

However, suppose we only wish to obtain $B(h)$ through some order in h. From the Magnus expansion, see Section 10.3, we know that $B(t)$ is of order h for $t \in [0, h]$. It follows that the term $b_\ell[\#B(t)\#]^\ell A(t)$ is of order h^ℓ and therefore contributes a term of order $h^{\ell+1}$ to $B(h)$. Suppose we truncate the series (1.82) beyond $\ell = n$ with n even. Then the size of the first omitted term will be of order h^{n+3}.[13] The result is the replacement of the differential equation (1.82) with the truncated equation

$$\begin{aligned}\dot{B}(t) &= \sum_{\ell=0}^{\ell=n} b_\ell [\#B(t)\#]^\ell A(t)\\ &= b_0A(t) + b_1\{B(t), A(t)\} + b_2\{B(t), \{B(t), A(t)\}\} + \cdots\\ &\quad + b_n\{B(t), \{B(t), \{\cdots\{B(t), A(t)\}\cdots\}\}\}, \end{aligned}\tag{11.1.83}$$

and the understanding is that this truncated equation is to be integrated only over the interval $t \in [0, h]$. Then the $B(h)$ so obtained will be correct through order h^m with $m = n + 2$. We have obtained a tractable problem. We have also been introduced to a new idea: the equation of motion to be integrated may be *modified* depending on the desired local accuracy for the integrated result.

[12] Exercise 1.10 shows, in effect, how this summation can be done for the cases of the Lie algebras $su(2)$ and $so(3, \mathbb{R})$.

[13] Examination of the b_ℓ, see Appendix C, reveals (save for $b_1 = 1/2$) that they all vanish for odd ℓ.

Still it might appear that, even with all our efforts, not much has been accomplished. We will soon see that progress has indeed been made. First, let us perform a sanity check on the results obtained so far. Suppose that $B(t) \in \mathcal{L}(G)$. From the definition

$$\dot{B}(t) = \lim_{\epsilon \to 0} [B(t+\epsilon) - B(t)]/\epsilon \tag{11.1.84}$$

we see that only vector space operations are involved in the calculation of $\dot{B}(t)$, and therefore $\dot{B}(t)$, which appears on the left side of (1.83), must also be in $\mathcal{L}(G)$. But is the right side of (1.83) in $\mathcal{L}(G)$? We know that $A(t)$ is in $\mathcal{L}(G)$. Also, all the rest of the right side of (1.83) involves sums of commutators of $B(t)$ with $A(t)$. By definition, $\mathcal{L}(G)$ is closed under addition and commutation. Therefore the right side of (1.83) is also in $\mathcal{L}(G)$. We conclude (1.83) is sane at least to the extent that both sides are in $\mathcal{L}(G)$.

But now we make a key observation: Suppose (1.83) is integrated by some numerical integrator to find $B(t)$. Examination of the various numerical integration schemes (for example all those discussed in Chapter 2) reveals that they all involve just *linear* combinations of the right side of the differential equation in question evaluated at various times and coordinate values. We know, by definition, that $\mathcal{L}(G)$ is closed under addition. Therefore, if the right side of the differential equation is known to be in $\mathcal{L}(G)$ at all evaluation points, then the result of numerically integrating such an equation is guaranteed to be in $\mathcal{L}(G)$. Since we have verified that the right side of (1.83) is in $\mathcal{L}(G)$, it follows that the $B(t)$ obtained by numerical integration, whatever the accuracy of the method, is guaranteed to be in $\mathcal{L}(G)$. Finally, if we wish to obtain $B(h)$ with an accuracy of some desired order in h, we may truncate (1.82) to obtain (1.83) with a known accuracy, and then integrate (1.83) using any integrator whose order equals or exceeds the accuracy of (1.83).

In summary, where conveniently feasible, integration in the Lie algebra is advantageous compared to integration in the group because numerical integration schemes, no matter their accuracy, generally preserve Lie algebraic structure.[14] Put another way, suppose the matrices A and B are $k \times k$. Then (1.83) is a differential equation in an ambient k^2 dimensional space. Numerical integration of (1.83) by any of the standard methods produces a sequence of points in this ambient space corresponding to the times t^n. We have seen that if the initial point is in the subspace $\mathcal{L}(G)$, then all subsequent points will also be in $\mathcal{L}(G)$. In our case, the initial point is given by (1.80), and is obviously in $\mathcal{L}(G)$. Therefore all subsequent points will also be in $\mathcal{L}(G)$.[15]

11.1.15 Integration in the Lie Algebra: Cayley Representation

We have used the exponential and logarithmic functions to relate G and $\mathcal{L}(G)$. For quadratic groups, which are often of interest, one may also use a Cayley representation to provide a

[14] Note that the work of Subsection 10.4.2 essentially employs a hybrid of canonical coordinates of the first and second kinds for the the nonlinear part of a symplectic map and formulates differential equations in the Lie algebra. The work of Section 10.3 and Subsection 10.4.3 also formulates differential equations in the Lie algebra.

[15] In the context of the exponential representation and in the case of finite-dimensional groups, the strategy of integrating in the Lie algebra and making the truncation (1.83) was pioneered by *Munthe-Kaas*; and the use of Runge-Kutta in this setting is often referred to as *RKMK* integration.

map between group elements near the identity and Lie algebra elements near the origin. That is, lowering and lifting can be done using the Cayley transformation and its inverse. See Section 3.12. Where possible, employing this approach yields a Lie algebraic differential equation analogous to (1.82), but with the advantage that only three terms appear on the right side. Therefore, *no* truncation is required. See Exercise 1.12.

We again seek to integrate equations of the form (1.73) with the initial condition (1.74) where now $M(t)$ is expected to belong to some quadratic Lie group G, and $A(t)$ belongs to its associated Lie algebra. What we again desire is a way of numerically integrating (1.73) that guarantees $M(t)$ is in G even though the numerical solution may be locally exact only through terms of order h^m.

When M belongs to a quadratic group G, we may employ the Cayley parameterization

$$M = (I + V)(I - V)^{-1}, \tag{11.1.85}$$

where V is in the Lie algebra of G. The challenge now is is to find the equation of motion for V. In Exercise 1.12 you will show that the desired result is the equation of motion

$$\dot{V} = (1/2)(A + \{V, A\} - VAV). \tag{11.1.86}$$

Note that, in contrast to (1.82) whose right side contains an infinite number of terms, the right side of (1.86) contains only three terms. Therefore no truncation is necessary and, unlike the case of $B(t)$, the accuracy to which $V(t)$ is calculated depends only on the method of integration.

Is this result sane? For a quadratic group G we know that V is in the Lie algebra. By an argument identical to that made in Subsection 1.14 for the case of $\dot{B}(t)$, it follows that $\dot{V}(t)$ must also be in the Lie algebra of G. But is the right side of (1.86) in the Lie algebra of G? Evidently, since A is in the Lie algebra of G, the first two terms on the right side of (1.86) are in the Lie algebra of G. What about the third term VAV? It can be shown that for a quadratic group the quantity VAV is also in the Lie algebra of G. Again see Exercise 1.12. Therefore (1.86) is sane at least to the extent that both its sides are in the Lie algebra of G.

Suppose (1.86) is integrated by some numerical integrator to find $V(t)$ under the assumption that V is initially in the Lie algebra of G. We repeat the key observation of Subsection 1.14: Examination of the usual numerical integration schemes, see Chapter 2, reveals that they all involve just linear combinations of the right side of the differential equation in question evaluated at various times and coordinate values. Therefore, if the right side is known to be in the Lie algebra of G for all evaluation points, then the result of numerically integrating such an equation is guaranteed to be in the Lie algebra of G, no matter what the local accuracy of the integrator or the step size employed. Since we have verified that the right side of (1.86) is in the Lie algebra of G, it follows that $V(t)$ will be in the Lie algebra of G if it is initially in the Lie algebra of G. Finally, since $V(t)$ is in the Lie algebra of G, it follows that $M(t)$ given by (1.85) is in G.

We have achieved our goal of, in effect, numerically integrating (1.73) in such a way that $M(t)$ is guaranteed to be in G. All that is required is accurate matrix multiplication in the calculation of the right side of (1.86) and accurate matrix inversion and multiplication in the evaluation of (1.85). Note, however, that this procedure cannot be carried out globally since the Cayley parameterization (1.85) cannot be made globally. It may therefore be necessary to change coordinate systems (by left or right group translation) from time to time during the course of a numerical integration in order to stay clear of the singularities associated with any given Cayley parametrization.[16]

11.1.16 Parameterization of G and $\mathcal{L}(G)$

To reiterate a point, if the matrices M or B are $k \times k$, then equations of the kind (1.73) or (1.82) involve k^2 variables. By contrast, the group G, and correspondingly its Lie algebra $\mathcal{L}(G)$, generally have much smaller dimension. Therefore it may be advantageous to parameterize the group or Lie algebra and to convert the differential equations for the group or Lie algebra into (usually) far fewer differential equations for the parameters. This is what was done for the case of $SO(3,\mathbb{R})$ by the use of quaternion parameters, the use of Tait-Bryan angles, and the use of angle-axis parameters.

By the introduction of a basis it is also possible to parameterize the Lie elements that occur in a Cayley formulation. That is, we parameterize the A and V appearing in (1.86). Again, as illustrated in Subsection 1.12, the result is (usually) far fewer equations that need to be integrated. See also Exercise 1.13.

11.1.17 Quaternions Revisited

We close this subsection by remarking that, in the case of $SO(3,\mathbb{R})$ and in the context of local coordinates, there is a still better approach, which again uses quaternions: Namely, suppose we again write (1.56) but now parameterize R^v in terms of quaternions. In this case the relations(1.43) through (1.46) will continue to hold and, in view of (1.54), there will be the initial conditions

$$w_0(t^b) = 1, \tag{11.1.87}$$

$$w_j(t^b) = 0 \text{ for } j = 1,2,3. \tag{11.1.88}$$

Also, if h is sufficiently small, the w_j for $j = 1,2,3$ will remain small over the course of a single integration step. We may therefore enforce the condition (1.30) by writing

$$w_0 = [1 - (w_1^2 + w_2^2 + w_3^2)]^{1/2} \tag{11.1.89}$$

and insert this result into (1.44) through (1.46). So doing gives the modified equations of motion

$$\dot{w}_1 = (1/2)\{-\omega_1^{bf}[1 - (w_1^2 + w_2^2 + w_3^2)]^{1/2} - \omega_2^{bf} w_3 + \omega_3^{bf} w_2\}, \tag{11.1.90}$$

$$\dot{w}_2 = (1/2)\{\omega_1^{bf} w_3 - \omega_2^{bf}[1 - (w_1^2 + w_2^2 + w_3^2)]^{1/2} - \omega_3^{bf} w_1\}, \tag{11.1.91}$$

$$\dot{w}_3 = (1/2)\{-\omega_1^{bf} w_2 + \omega_2^{bf} w_1 - \omega_3^{bf}[1 - (w_1^2 + w_2^2 + w_3^2)]^{1/2}\}. \tag{11.1.92}$$

These three kinematic equations, whose right sides depend only on (w_1, w_2, w_3) and on $(\omega_1^{bf}, \omega_2^{bf}, \omega_3^{bf})$, along with the three dynamic equations (1.19) through (1.21), are to be

[16] It might appear the the right side of (1.86) is singularity free. However, it is singular at infinity, and this singularity can be reached in finite time. See the discussion in Exercise 1.13.

integrated over the interval $t \in [t^b, t^b + h]$. (And any kind of integrator can be used.) Whenever w_0 is needed, it is to be computed from (1.89). Note that in view of (8.2.73), no exponentials need be computed to find $R^v(t)$. Also, the kinematic equations are mostly linear and involve, at worst, the calculation of a square root. Moreover, if desired, the equivalent of matrix multiplication in $SO(3, \mathbb{R})$ can be carried out at the quaternion level where the operations are computationally simpler. See (5.10.24) and Exercises 8.2.10 and 8.2.11. For all these reasons, relatively few floating-point operations are needed to carry out an integration step. Finally, although the quaternion parameters (w_1, w_2, w_3) may only be computed with a local error of order h^{m+1}, the resulting matrix $R_v(t^b+h)$ will be orthogonal to a very high accuracy limited only by roundoff error.

Exercises

11.1.1. The purpose of this exercise is to explore some of the consequences of the definition of angular velocity given in Subsection 1.1. The first task is a bit of housekeeping. In deriving (1.10) through (1.12) it was tacitly assumed that

$$[(d/dt)R]^T = (d/dt)(R^T). \tag{11.1.93}$$

That is, the operations of differentiating and transposing commute. Verify that this is so for any matrix.

Now move on to the main task. The *components* $\omega_j^{bf}(t)$ of the angular velocity in the *body-fixed* frame are defined by the rule (1.4). Accordingly, define the angular velocity *vector* $\boldsymbol{\omega}(t)$ by the rule

$$\boldsymbol{\omega}(t) = \sum_j \omega_j^{bf}(t)\boldsymbol{f}_j(t). \tag{11.1.94}$$

Let us compute the vector $\boldsymbol{\omega} \times \boldsymbol{f}_1$. Verify the chain of relations

$$\begin{aligned}
\boldsymbol{\omega} \times \boldsymbol{f}_1 &= \omega_2^{bf}\boldsymbol{f}_2 \times \boldsymbol{f}_1 + \omega_3^{bf}\boldsymbol{f}_3 \times \boldsymbol{f}_1 = -\omega_2^{bf}\boldsymbol{f}_3 + \omega_3^{bf}\boldsymbol{f}_2 \\
&= (\boldsymbol{f}_3 \cdot \dot{\boldsymbol{f}}_1)\boldsymbol{f}_3 - (\boldsymbol{f}_1 \cdot \dot{\boldsymbol{f}}_2)\boldsymbol{f}_2 \\
&= (\boldsymbol{f}_3 \cdot \dot{\boldsymbol{f}}_1)\boldsymbol{f}_3 + (\boldsymbol{f}_2 \cdot \dot{\boldsymbol{f}}_1)\boldsymbol{f}_2 + (\boldsymbol{f}_1 \cdot \dot{\boldsymbol{f}}_1)\boldsymbol{f}_1 \\
&= \dot{\boldsymbol{f}}_1.
\end{aligned} \tag{11.1.95}$$

Here we have used the fact

$$\boldsymbol{f}_1 \cdot \boldsymbol{f}_2 = 0 \tag{11.1.96}$$

from which it follows that

$$\dot{\boldsymbol{f}}_1 \cdot \boldsymbol{f}_2 + \boldsymbol{f}_1 \cdot \dot{\boldsymbol{f}}_2 = 0, \tag{11.1.97}$$

and the fact

$$\boldsymbol{f}_1 \cdot \boldsymbol{f}_1 = 1 \tag{11.1.98}$$

from which it follows that

$$\boldsymbol{f}_1 \cdot \dot{\boldsymbol{f}}_1 = 0. \tag{11.1.99}$$

It is easily checked that there are two more relations like (1.95), and that they together give the general equations of motion

$$\dot{\boldsymbol{f}}_j(t) = \boldsymbol{\omega}(t) \times \boldsymbol{f}_j(t). \tag{11.1.100}$$

Verify from (1.100) that

$$\boldsymbol{f}_j \times \dot{\boldsymbol{f}}_j = \boldsymbol{f}_j \times (\boldsymbol{\omega} \times \boldsymbol{f}_j) = \boldsymbol{\omega} - (\boldsymbol{\omega} \cdot \boldsymbol{f}_j)\boldsymbol{f}_j, \tag{11.1.101}$$

and therefore

$$\sum_j \boldsymbol{f}_j \times \dot{\boldsymbol{f}}_j = 3\boldsymbol{\omega} - \sum_j (\boldsymbol{\omega} \cdot \boldsymbol{f}_j)\boldsymbol{f}_j = 3\boldsymbol{\omega} - \boldsymbol{\omega} = 2\boldsymbol{\omega}. \tag{11.1.102}$$

Conclude that $\boldsymbol{\omega}$ is also given by the relation

$$\boldsymbol{\omega} = (1/2) \sum_j \boldsymbol{f}_j \times \dot{\boldsymbol{f}}_j. \tag{11.1.103}$$

Moreover, suppose $\boldsymbol{v}(t)$ is any vector that is “fixed” in the body. That is, suppose $\boldsymbol{v}(t)$ has an expansion of the form

$$\boldsymbol{v}(t) = \sum_j v_j \boldsymbol{f}_j(t) \tag{11.1.104}$$

where the components v_j are *constant* numbers. Show that it follows from (1.100) that $\boldsymbol{v}$ obeys the equation of motion

$$\dot{\boldsymbol{v}}(t) = \boldsymbol{\omega}(t) \times \boldsymbol{v}(t). \tag{11.1.105}$$

Let us compute the $\omega_j^{bf}(t)$ for a special case. Suppose $R(t)$ is of the form

$$R(t) = \exp[\theta(t)\boldsymbol{n} \cdot \boldsymbol{L}] \tag{11.1.106}$$

where $\boldsymbol{n}$ is a constant vector. Verify that

$$\dot{R}(t) = \exp[\theta(t)\boldsymbol{n} \cdot \boldsymbol{L}]\ \dot{\theta}(t)\boldsymbol{n} \cdot \boldsymbol{L}. \tag{11.1.107}$$

Show, from (1.22), that in this case

$$\omega_j^{bf}(t) = \dot{\theta}(t)n_j. \tag{11.1.108}$$

This special case illustrates why the name *angular velocity* is appropriate.

What can be said more generally? Suppose $R(t)$ is a time-dependent matrix in $SO(3,\mathbb{R})$. Show that

$$\begin{aligned} R(t+dt) &= R(t) + \dot{R}(t)dt + O[(dt)^2] = R(t)[I + R^{-1}(t)\dot{R}(t)dt] + O[(dt)^2] \\ &= R(t)[I + (dt)\boldsymbol{\omega}^{bf}(t) \cdot \boldsymbol{L}] + O[(dt)^2] \\ &= R(t)\exp[(dt)\boldsymbol{\omega}^{bf}(t) \cdot \boldsymbol{L}] + O[(dt)^2]. \end{aligned} \tag{11.1.109}$$

You have shown that, through terms of order dt, $R(t+dt)$ is gotten from $R(t)$ by translating $R(t)$ on the right with the near-identity element $\exp[(dt)\boldsymbol{\omega}^{bf}(t) \cdot \boldsymbol{L}]$.

Can $R(t+dt)$ and $R(t)$ instead be related by translating $R(t)$ on the left with a near-identity element? Show that the answer is *yes*. Verify the manipulations

$$\begin{aligned} R(t+dt) &= R(t)+\dot{R}(t)dt+O[(dt)^2]=[I+dt\dot{R}(t)R^{-1}(t)]R(t)+O[(dt)^2] \\ &= \{I+dtR(t)[R^{-1}(t)\dot{R}(t)]R^{-1}(t)\}R(t)+O[(dt)^2] \\ &= \{I+dtR(t)[\boldsymbol{\omega}^{bf}(t)\cdot\boldsymbol{L}]R^{-1}(t)\}R(t)+O[(dt)^2]. \end{aligned} \tag{11.1.110}$$

Next verify that

$$R(t)[\boldsymbol{\omega}^{bf}(t)\cdot\boldsymbol{L}]R^{-1}(t)=\boldsymbol{\omega}^{sf}(t)\cdot\boldsymbol{L} \tag{11.1.111}$$

where $\boldsymbol{\omega}^{sf}(t)$ is defined by the relation

$$\boldsymbol{\omega}^{sf}(t)=R(t)\boldsymbol{\omega}^{bf}(t). \tag{11.1.112}$$

See (8.2.59). Combine (1.110) through (1.112) to get the result

$$\begin{aligned} R(t+dt) &= \{I+dtR(t)[\boldsymbol{\omega}^{bf}(t)\cdot\boldsymbol{L}]R^{-1}(t)\}R(t)+O[(dt)^2] \\ &= \{I+dt\boldsymbol{\omega}^{sf}(t)\cdot\boldsymbol{L}\}R(t)+O[(dt)^2] \\ &= \exp[(dt)\boldsymbol{\omega}^{sf}(t)\cdot\boldsymbol{L}]R(t)+O[(dt)^2]. \end{aligned} \tag{11.1.113}$$

You have shown that, through terms of order dt, $R(t+dt)$ is gotten from $R(t)$ by translating $R(t)$ on the left with the near-identity element $\exp[(dt)\boldsymbol{\omega}^{sf}(t)\cdot\boldsymbol{L}]$. Verify also that (1.113) implies the relation

$$\boldsymbol{\omega}^{sf}(t)\cdot\boldsymbol{L}=\dot{R}(t)R^{-1}(t), \tag{11.1.114}$$

which is to be compared with (1.22).

According to (1.112), the quantities we have called ω_j^{sf} are related to the ω_j^{bf} by the rule

$$\omega_j^{sf}=\sum_k R_{jk}\omega_k^{bf}. \tag{11.1.115}$$

What is their significance? Recall the angular velocity vector $\boldsymbol{\omega}(t)$ defined by (1.94). Rewrite this definition using the dummy index k to obtain the equivalent definition

$$\boldsymbol{\omega}(t)=\sum_k \omega_k^{bf}(t)\boldsymbol{f}_k(t). \tag{11.1.116}$$

Since the *space-fixed* vectors $\boldsymbol{e}_j$ form a basis, we may make the expansion

$$\boldsymbol{\omega}(t)=\sum_j[\boldsymbol{e}_j\cdot\boldsymbol{\omega}(t)]\boldsymbol{e}_j. \tag{11.1.117}$$

Use the representation (1.116) to compute the expansion coefficients $[\boldsymbol{e}_j\cdot\boldsymbol{\omega}(t)]$. Verify that so doing gives the result

$$\begin{aligned} \boldsymbol{e}_j\cdot\boldsymbol{\omega}(t) &= \sum_k \omega_k^{bf}(t)[\boldsymbol{e}_j\cdot\boldsymbol{f}_k(t)]=\sum_k \omega_k^{bf}(t)\{\boldsymbol{e}_j\cdot[R(t)\boldsymbol{e}_k(t)]\} \\ &= \sum_k \omega_k^{bf}(t)R_{jk}(t)=\sum_k R_{jk}(t)\omega_k^{bf}(t)=\omega_j^{sf}(t), \end{aligned} \tag{11.1.118}$$

where use has been made of (1.2) and (1.115). By putting everything together, show that (1.117) can be rewritten in the form

$$\boldsymbol{\omega}(t) = \sum_j \omega_j^{sf}(t)\boldsymbol{e}_j. \tag{11.1.119}$$

We conclude that the quantities $\omega_j^{sf}(t)$ are the *space-fixed* components of the angular velocity vector $\boldsymbol{\omega}(t)$.

Finally show, for the special case (1.106), that

$$\omega_j^{sf}(t) = \omega_j^{bf}(t). \tag{11.1.120}$$

11.1.2. The purpose of this exercise is to show that the matrix differential equation (1.18) preserves the orthogonality condition (1.9), and conversely.

Begin with the converse. We already know from the orthogonality assumption (1.9) that the matrix $(R^{-1}\dot{R})$ must be antisymmetric. See (1.12). Therefore for orthogonal 3×3 matrices a relation of the form (1.18) must hold.

Now prove the main assertion. Show, by taking transposes, that (1.18) implies the relation

$$(d/dt)R^T = -\boldsymbol{\omega}^{bf} \cdot \boldsymbol{L}\ R^T. \tag{11.1.121}$$

Next show that there is the relation

$$(d/dt)(RR^T) = [(d/dt)R]R^T + R(d/dt)R^T = R(\boldsymbol{\omega}^{bf} \cdot \boldsymbol{L})R^T - R(\boldsymbol{\omega}^{bf} \cdot \boldsymbol{L})R^T = 0. \tag{11.1.122}$$

Here use has been made of (1.18) and (1.121). Assume that R is orthogonal at the initial time t^i,

$$R(t^i)R(t^i)^T = I = R^T(t^i)R(t^i). \tag{11.1.123}$$

Verify that the solution to the differential equation (1.122) with the initial condition (1.123) is

$$R(t)R(t)^T = I = R^T(t)R(t). \tag{11.1.124}$$

You have shown that (1.18) preserves orthogonality.

11.1.3. The purpose of this exercise is to derive equations (1.23) through (1.28), the expressions for the ω_j^{bf} in terms of Euler angles. Recall the Euler-angle parameterization (3.7.207),

$$R(t) = \exp[\phi(t)L^3]\exp[\theta(t)L^2]\exp[\psi(t)L^3]. \tag{11.1.125}$$

According to (1.22), we must compute the quantities ω_j^{bf} defined by the relation

$$\boldsymbol{\omega}^{bf} \cdot \boldsymbol{L} = R^{-1}\dot{R}. \tag{11.1.126}$$

Verify that R^{-1} is given by the relation

$$R^{-1} = \exp(-\psi L^3)\exp(-\theta L^2)\exp(-\phi L^3). \tag{11.1.127}$$

Show that $\dot{R}$ is given by the relation

$$\begin{aligned}\dot{R} &= \dot{\phi}L^3\exp(\phi L^3)\exp(\theta L^2)\exp(\psi L^3)\\ &+ \exp(\phi L^3)\dot{\theta}L^2\exp(\theta L^2)\exp(\psi L^3)\\ &+ \exp(\phi L^3)\exp(\theta L^2)\exp(\psi L^3)\dot{\psi}L^3.\end{aligned} \tag{11.1.128}$$

Now begins the fun. Verify that $R^{-1}\dot{R}$ is given by the seemingly hopeless expression

$$\begin{aligned}R^{-1}\dot{R} &= \exp(-\psi L^3)\exp(-\theta L^2)\exp(-\phi L^3)\dot{\phi}L^3\exp(\phi L^3)\exp(\theta L^2)\exp(\psi L^3)\\ &+ \exp(-\psi L^3)\exp(-\theta L^2)\exp(-\phi L^3)\exp(\phi L^3)\dot{\theta}L^2\exp(\theta L^2)\exp(\psi L^3)\\ &+ \exp(-\psi L^3)\exp(-\theta L^2)\exp(-\phi L^3)\exp(\phi L^3)\exp(\theta L^2)\exp(\psi L^3)\dot{\psi}L^3.\end{aligned} \tag{11.1.129}$$

Simplify each of the three lines in (1.129) so that they become

$$\begin{aligned}&\exp(-\psi L^3)\exp(-\theta L^2)\exp(-\phi L^3)\dot{\phi}L^3\exp(\phi L^3)\exp(\theta L^2)\exp(\psi L^3)\\ &= \dot{\phi}\exp(-\psi L^3)\exp(-\theta L^2)L^3\exp(\theta L^2)\exp(\psi L^3),\end{aligned} \tag{11.1.130}$$

$$\begin{aligned}&\exp(-\psi L^3)\exp(-\theta L^2)\exp(-\phi L^3)\exp(\phi L^3)\dot{\theta}L^2\exp(\theta L^2)\exp(\psi L^3)\\ &= \dot{\theta}\exp(-\psi L^3)L^2\exp(\psi L^3),\end{aligned} \tag{11.1.131}$$

$$\begin{aligned}&\exp(-\psi L^3)\exp(-\theta L^2)\exp(-\phi L^3)\exp(\phi L^3)\exp(\theta L^2)\exp(\psi L^3)\dot{\psi}L^3\\ &= \dot{\psi}L^3.\end{aligned} \tag{11.1.132}$$

Line (1.132) is as simple as we could desire. The next more complicated line is (1.131). Show, using the machinery of Exercise 8.2.10, that

$$\exp(-\psi L^3)L^2\exp(\psi L^3) = \cos(\psi)L^2 + \sin(\psi)L^1. \tag{11.1.133}$$

Thus, the right side of (1.131) becomes

$$\dot{\theta}[\cos(\psi)L^2 + \sin(\psi)L^1]. \tag{11.1.134}$$

Finally, work on line (1.130). Show that

$$\exp(-\theta L^2)L^3\exp(\theta L^2) = \cos(\theta)L^3 - \sin(\theta)L^1. \tag{11.1.135}$$

Next show that

$$\begin{aligned}&\exp(-\psi L^3)[\cos(\theta)L^3 - \sin(\theta)L^1]\exp(\psi L^3)\\ &= \cos(\theta)L^3 - \sin(\theta)\exp(-\psi L^3)L^1\exp(\psi L^3)\\ &= \cos(\theta)L^3 - \sin(\theta)[\cos(\psi)L^1 - \sin(\psi)L^2].\end{aligned} \tag{11.1.136}$$

By combining (1.135) and (1.136), verify that

$$\begin{aligned}&\exp(-\psi L^3)\exp(-\theta L^2)L^3\exp(\theta L^2)\exp(\psi L^3)\\&=\cos(\theta)L^3-\sin(\theta)[\cos(\psi)L^1-\sin(\psi)L^2].\end{aligned}\tag{11.1.137}$$

Thus, the right side of (1.130) becomes

$$\dot{\phi}\{\cos(\theta)L^3-\sin(\theta)[\cos(\psi)L^1-\sin(\psi)L^2]\}.\tag{11.1.138}$$

All the necessary ingredients are at hand. By combining (1.129) through (1.138), show that

$$\begin{aligned}R^{-1}\dot{R} &= \dot{\phi}\{\cos(\theta)L^3-\sin(\theta)[\cos(\psi)L^1-\sin(\psi)L^2]\}\\&+ \dot{\theta}[\cos(\psi)L^2+\sin(\psi)L^1]\\&+ \dot{\psi}L^3\\&= (-\dot{\phi}\sin\theta\cos\psi+\dot{\theta}\sin\psi)L^1\\&+ (\dot{\phi}\sin\theta\sin\psi+\dot{\theta}\cos\psi)L^2\\&+ (\dot{\phi}\cos\theta+\dot{\psi})L^3.\end{aligned}\tag{11.1.139}$$

Verify, upon equating coefficients of the L^j in (1.126) and (1.139), that the relations (1.23) through (1.25) follow. Finally, verify that inverting the relations (1.23) through (1.25) yields the relations (1.26) through (1.28).

11.1.4. The purpose of this exercise is to derive equations (1.31) and (1.32), the relation between quaternion and angle-axis parameters, and equations (1.33) through (1.36), the relation between quaternion and Euler-angle parameters. For this purpose, it is convenient to exploit the homomorphism between $SO(3,\mathbb{R})$ and $SU(2)$. Review Exercises 3.7.30, 5.10.13, 8.2.10, and 8.2.11.

Suppose $u \in SU(2)$ is parameterized in terms of angle-axis parameters by writing

$$u=\exp(\theta\boldsymbol{n}\cdot\boldsymbol{K})=I\cos(\theta/2)+2(\boldsymbol{n}\cdot\boldsymbol{K})\sin(\theta/2).\tag{11.1.140}$$

See (3.7.186) and (3.7.188). Suppose that u is also parameterized in terms of unit quaternion matrices by writing

$$u(w)=w_0\sigma^0+i\boldsymbol{w}\cdot\boldsymbol{\sigma}.\tag{11.1.141}$$

Equate (1.140) and (1.141) and employ the relations

$$I=\sigma^0,\tag{11.1.142}$$

$$\boldsymbol{K}=(-i/2)\boldsymbol{\sigma}.\tag{11.1.143}$$

Use the linear independence of the matrices σ^0 through σ^3 to verify (1.31) and (1.32).

Next suppose $u \in SU(2)$ is parameterized in terms of Euler angles as in (3.7.195) in Exercise 3.7.30. Verify that there is the decomposition

$$u(\phi,\theta,\psi) = \begin{pmatrix} \cos(\theta/2)\exp[-(i/2)(\phi+\psi)] & -\sin(\theta/2)\exp[(i/2)(-\phi+\psi)] \\ \sin(\theta/2)\exp[-(i/2)(-\phi+\psi)] & \cos(\theta/2)\exp[(i/2)(\phi+\psi)] \end{pmatrix}$$
$$= \cos(\theta/2)\cos[(1/2)(\phi+\psi)]\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - i\sin(\theta/2)\sin[(1/2)(-\phi+\psi)]\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$
$$-i\sin(\theta/2)\cos[(1/2)(-\phi+\psi)]\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} - i\cos(\theta/2)\sin[(1/2)(\phi+\psi)]\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{11.1.144}$$

Thus, using the definition of the Pauli matrices, verify that

$$\begin{aligned} u &= \sigma^0\{\cos(\theta/2)\cos[(1/2)(\phi+\psi)]\} \\ &- i\sigma^1\{\sin(\theta/2)\sin[(1/2)(-\phi+\psi)]\} \\ &- i\sigma^2\{\sin(\theta/2)\cos[(1/2)(-\phi+\psi)]\} \\ &- i\sigma^3\{\cos(\theta/2)\sin[(1/2)(\phi+\psi)]\}. \end{aligned} \tag{11.1.145}$$

Finally, verify that comparison of (1.141) and (1.145), and use of the linear independence of the matrices σ^0 through σ^3, yields (1.33) through (1.36).

11.1.5. The purpose of this exercise is to derive equations (1.37) through (1.39), the expressions for the angular velocities ω_j^{bf} in terms of quaternion parameters. Surprisingly, this task turns out to be computationally simpler than the Euler-angle case of Exercise 1.3. Review Exercise 1.4 above for notation.

The relation (1.22) specifies the ω_j^{bf} in terms of $SO(3,\mathbb{R})$ quantities. Based on the homomorphism between $SO(3,\mathbb{R})$ and $SU(2)$ (see Exercises 3.7.30, 5.10.13, 8.2.10, and 8.2.11), verify that there is an associated $SU(2)$ relation given by the formula

$$u^{-1}\dot{u} = \omega_1^{bf}K^1 + \omega_2^{bf}K^2 + \omega_3^{bf}K^3 = \boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}. \tag{11.1.146}$$

Indeed, suppose u is parameterized in terms of Euler angles by writing

$$u(t) = \exp[\phi(t)K^3]\exp[\theta(t)K^2]\exp[\psi(t)K^3]. \tag{11.1.147}$$

See (3.7.194). Show that then use of (1.147) in (1.146) reproduces the results of Exercise 1.3.

Instead, we will parameterize u in terms of unit quaternion matrices by writing (1.141). Then we have the results

$$u^{-1} = u^\dagger = w_0\sigma^0 - i\boldsymbol{w}\cdot\boldsymbol{\sigma} \tag{11.1.148}$$

and

$$\dot{u}(w) = \dot{w}_0\sigma^0 + i\dot{\boldsymbol{w}}\cdot\boldsymbol{\sigma}. \tag{11.1.149}$$

Verify that carrying out the required multiplication $u^{-1}\dot{u}$ yields the intermediate result

$$\begin{aligned} u^{-1}\dot{u} &= [w_0\sigma^0 - i\boldsymbol{w}\cdot\boldsymbol{\sigma}][\dot{w}_0\sigma^0 + i\dot{\boldsymbol{w}}\cdot\boldsymbol{\sigma}] \\ &= w_0\dot{w}_0\sigma^0 + iw_0\dot{\boldsymbol{w}}\cdot\boldsymbol{\sigma} - i\dot{w}_0\boldsymbol{w}\cdot\boldsymbol{\sigma} + (\boldsymbol{w}\cdot\boldsymbol{\sigma})(\dot{\boldsymbol{w}}\cdot\boldsymbol{\sigma}). \end{aligned} \tag{11.1.150}$$

Now use (5.7.44) to write

$$(\boldsymbol{w}\cdot\boldsymbol{\sigma})(\dot{\boldsymbol{w}}\cdot\boldsymbol{\sigma}) = (\boldsymbol{w}\cdot\dot{\boldsymbol{w}})\sigma^0 + i(\boldsymbol{w}\times\dot{\boldsymbol{w}})\cdot\boldsymbol{\sigma}. \tag{11.1.151}$$

Substitute (1.151) into (1.150) to yield the next intermediate result

$$u^{-1}\dot{u} = (w_0\dot{w}_0 + \boldsymbol{w}\cdot\dot{\boldsymbol{w}})\sigma^0 + i(w_0\dot{\boldsymbol{w}} - \dot{w}_0\boldsymbol{w} + \boldsymbol{w}\times\dot{\boldsymbol{w}})\cdot\boldsymbol{\sigma}. \tag{11.1.152}$$

But in view of (1.40), which follows from the requirement that u be a unit quaternion, the first term on the right of (1.152) vanishes, and (1.152) therefore becomes

$$u^{-1}\dot{u} = i(w_0\dot{\boldsymbol{w}} - \dot{w}_0\boldsymbol{w} + \boldsymbol{w}\times\dot{\boldsymbol{w}})\cdot\boldsymbol{\sigma}. \tag{11.1.153}$$

There is also the result (1.143), and therefore (1.153) can be written as

$$u^{-1}\dot{u} = -2(w_0\dot{\boldsymbol{w}} - \dot{w}_0\boldsymbol{w} + \boldsymbol{w}\times\dot{\boldsymbol{w}})\cdot\boldsymbol{K}. \tag{11.1.154}$$

Upon comparing (1.146) and (1.154), show that

$$\omega_j^{bf} = -2(w_0\dot{\boldsymbol{w}} - \dot{w}_0\boldsymbol{w} + \boldsymbol{w}\times\dot{\boldsymbol{w}})\cdot\boldsymbol{e}_j. \tag{11.1.155}$$

Verify that (1.37) through (1.39) are equivalent to (1.155).

11.1.6. The purpose of this exercise is to verify that the 4×4 matrices appearing in (1.41) and (1.42) are orthogonal. Suppose $(w_0, w_1, w_2, w_3)^T$ is a unit four vector. That is, suppose (1.30) is satisfied. Consider the mapping of this four vector into the space of 4×4 matrices given by the rule

$$\begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{pmatrix} \rightarrow M(w) = \begin{pmatrix} w_0 & -w_1 & -w_2 & -w_3 \\ w_1 & w_0 & w_3 & -w_2 \\ w_2 & -w_3 & w_0 & w_1 \\ w_3 & w_2 & -w_1 & w_0 \end{pmatrix}. \tag{11.1.156}$$

Verify that all the columns of M are unit vectors. Verify that all the columns of M are mutually orthogonal. It follows that M is an orthogonal matrix,

$$M^T M = I. \tag{11.1.157}$$

Verify that the matrices appearing in (1.41) and (1.42) are $-M^T$ and $-M$, and hence both are also orthogonal.

11.1.7. Intrigued by the remarkable mapping (1.156), this exercise is devoted to a further exploration of what is going on. We will learn that what is involved is a mapping of a pair of unit quaternions into $SO(4,\mathbb{R})$.

Show that since M is orthogonal, it must satisfy

$$\det M = \pm 1. \tag{11.1.158}$$

Let e_0 be the unit vector

$$e_0 = (1,0,0,0)^T. \tag{11.1.159}$$

Verify that

$$M(e_0) = I, \tag{11.1.160}$$

and therefore

$$\det M(e_0) = +1. \tag{11.1.161}$$

Show that any unit vector $w \in S^3$ is connected to e_0 by a continuous path in S^3. [Hint: Show, for example, that $SO(4,\mathbb{R})$ acts transitively on S^3.] Verify that the mapping (1.156) is continuous. Show that it follows, by continuity, that

$$\det M(w) = +1, \tag{11.1.162}$$

and therefore $M(w) \in SO(4,\mathbb{R})$. That is, (1.156) produces a map

$$S^3 \to SO(4,\mathbb{R}). \tag{11.1.163}$$

Can all elements of $SO(4,\mathbb{R})$ be written in the form $M(w)$? No, because $w \in S^3$ involves three parameters, and we know that $SO(4,\mathbb{R})$ is six dimensional. What elements in $SO(4,\mathbb{R})$ can be written in the form $M(w)$? Let's see. Define three matrices D^1, D^2, D^3 by the rules

$$D^1 = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}, \tag{11.1.164}$$

$$D^2 = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \tag{11.1.165}$$

$$D^3 = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \tag{11.1.166}$$

Verify the relation

$$M(w) = w_0 I + w_1 D^1 + w_2 D^2 + w_3 D^3. \tag{11.1.167}$$

Show that the matrices D^j satisfy the relations

$$(D^j)^2 = -I, \tag{11.1.168}$$

$$D^1D^2 = -D^2D^1 = -D^3, \tag{11.1.169}$$

$$D^2D^3 = -D^3D^2 = -D^1, \tag{11.1.170}$$

$$D^3D^1 = -D^1D^3 = -D^2. \tag{11.1.171}$$

By employing these relations, show that

$$M(w)M(w') = M(w'') \tag{11.1.172}$$

with

$$\begin{aligned} w_0'' &= w_0w_0' - w_1w_1' - w_2w_2' - w_3w_3', \\ w_1'' &= w_0w_1' + w_0'w_1 - w_2w_3' + w_3w_2', \\ w_2'' &= w_0w_2' + w_0'w_2 - w_3w_1' + w_1w_3', \\ w_3'' &= w_0w_3' + w_0'w_3 - w_1w_2' + w_2w_1', \end{aligned} \tag{11.1.173}$$

which, with the notation $\boldsymbol{w} = (w_1, w_2, w_3)$, can be written more compactly in the form

$$w_0'' = w_0w_0' - \boldsymbol{w} \cdot \boldsymbol{w}', \tag{11.1.174}$$

$$\boldsymbol{w}'' = w_0\boldsymbol{w}' + w_0'\boldsymbol{w} - \boldsymbol{w} \times \boldsymbol{w}'. \tag{11.1.175}$$

Show that if $w \in S^3$ and $w' \in S^3$, then $w'' \in S^3$. Show also that

$$M(w + w') = M(w) + M(w'). \tag{11.1.176}$$

Observe that the relations (1.174) and (1.175) are exactly those for quaternion matrix multiplication. See Section 5.10.4 and Exercise 5.10.15. Verify that there is the correspondence

$$\begin{aligned} I &\leftrightarrow e, \\ -D^1 &\leftrightarrow j, \\ -D^2 &\leftrightarrow, k \\ -D^3 &\leftrightarrow \ell. \end{aligned} \tag{11.1.177}$$

That is, the 4×4 matrices $I, -D^1, -D^2, -D^3$ provide a representation for quaternion algebra. Observe that, unlike the 2×2 representation given by the complex matrices in (5.10.64), these 4×4 matrices are all real.

Verify that the D^j are antisymmetric,

$$(D^j)^T = -D^j. \tag{11.1.178}$$

Verify that

$$[M(w)]^T = M(w^*) \tag{11.1.179}$$

with the definitions

$$w_0^* = w_0, \quad w_1^* = -w_1, \quad w_2^* = -w_2, \quad w_3^* = -w_3. \tag{11.1.180}$$

Verify that

$$(w^*)^* = w. \tag{11.1.181}$$

Verify that $w^* \in S^3$ if $w \in S^3$, and verify the relation

$$M(w^*) = [M(w)]^{-1}. \tag{11.1.182}$$

Thus, verify that the $M(w)$ for $w \in S^3$ form a group.

Finally, what elements in $SO(4,\mathbb{R})$ can be written in the form $M(w)$? We know from Exercise 5.10.13 that unit quaternions form a group that is isomorphic to $SU(2)$. We also know from Exercises 4.4.19 and 4.3.20 that $so(4,\mathbb{R})$ is the direct sum of two commuting $su(2)$ Lie algebras. From these same exercises we know that the H^j generate the $SU(2)$ associated with one of these $su(2)$ Lie algebras. Verify from (4.3.152) and (1.161) through (1.163) that there is the relation

$$D^j = -2H^j. \tag{11.1.183}$$

Show that it follows that the elements in $SO(4,\mathbb{R})$ that can be written in the form $M(w)$ are those that belong to the $SU(2)$ subgroup generated by the H^j. Verify, in the terminology of Exercises 4.3.19 and 4.3.20, that these are all elements of the form $\exp(\boldsymbol{t}\cdot\boldsymbol{H})$. Find the relation between $w \in S^3$ and $\boldsymbol{t}$. Lastly verify that, as expected from the work of Exercise 4.3.19, that all $M(w)$ with $w \in S^3$ are also symplectic with respect to the J of (4.3.65).

But wait, as they say in infomercials, there's more! As just stated, we know from Exercises 4.3.19 and 4.3.20 that $SO(4,\mathbb{R})$ is the direct product of two commuting $SU(2)$ subgroups, and we know from Exercise 5.10.13 that any $SU(2)$ has an associated unit quaternion equivalent. Therefore, there should be *two* unit quaternions associated with $SO(4,\mathbb{R})$. How can the second unit quaternion be found/employed?

Motivated by (1.183), we might define matrices E^j by the rule

$$E^j \overset{?}{=} -2G^j \tag{11.1.184}$$

with the G^j given by (4.3.151). This is a workable possibility. However, a choice that yields more aesthetic results is to make the definitions

$$\begin{aligned} E^1 &= 2G^3, \\ E^2 &= -2G^2, \\ E^3 &= -2G^1. \end{aligned} \tag{11.1.185}$$

The E^j have the explicit form

$$E^1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}, \tag{11.1.186}$$

$$E^2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \tag{11.1.187}$$

$$E^3 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}. \tag{11.1.188}$$

Form the matrix $N(w)$ by the rule

$$N(w) = w_0 I + w_1 E^1 + w_2 E^2 + w_3 E^3. \tag{11.1.189}$$

Verify that $N(w)$ has the explicit form

$$N(w) = \begin{pmatrix} w_0 & w_1 & w_2 & w_3 \\ -w_1 & w_0 & w_3 & -w_2 \\ -w_2 & -w_3 & w_0 & w_1 \\ -w_3 & w_2 & -w_1 & w_0 \end{pmatrix}. \tag{11.1.190}$$

Check that the columns of $N(w)$ are mutually orthogonal and, if $w \in S^3$, are also unit vectors. It follows that $N(w)$ is an orthogonal matrix if $w \in S^3$. Check that

$$N(e_0) = I, \tag{11.1.191}$$

and therefore show, as was done for $M(w)$, that $N(w) \in SO(4, \mathbb{R})$. Thus the correspondence

$$\begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{pmatrix} \rightarrow N(w) = \begin{pmatrix} w_0 & w_1 & w_2 & w_3 \\ -w_1 & w_0 & w_3 & -w_2 \\ -w_2 & -w_3 & w_0 & w_1 \\ -w_3 & w_2 & -w_1 & w_0 \end{pmatrix} \tag{11.1.192}$$

also provides a mapping

$$S^3 \rightarrow SO(4, \mathbb{R}). \tag{11.1.193}$$

Review the relations (1.168) through (1.182). Show that there are completely analogous relations for the E^j and $N(w)$. Among other things, the 4×4 matrices $I, -E^1, -E^2, -E^3$ also provide a representation for quaternion algebra. Also verify that $M(w')$ and $N(w)$ commute.

Finally, in analogy with (4.3.166) and (4.3.174), define a matrix $O(w, w')$ by the rule

$$O(w, w') = N(w)M(w'). \tag{11.1.194}$$

Call the pair $(w, w') = S^3 \times S^3$ a double three-sphere. The relation (1.194) provides a two-to-one mapping of the double three-sphere into $SO(4, \mathbb{R})$,

$$S^3 \times S^3 \rightarrow SO(4, \mathbb{R}). \tag{11.1.195}$$

11.1.8. The purpose of this exercise is to determine the behavior of $w \cdot w$ when w evolves as in equations (1.42) or in equations (1.48) through (1.51), and to explore the nature of these equations. Let Ω be the vector

$$\Omega = \begin{pmatrix} 0 \\ \omega_1^{bf}/2 \\ \omega_2^{bf}/2 \\ \omega_3^{bf}/2 \end{pmatrix}. \tag{11.1.196}$$

Verify that the equations (1.48) through (1.51) can be written in the compact form

$$\dot{w} = -M(w)\Omega + k\epsilon w \tag{11.1.197}$$

with $M(w)$ defined by (1.156). It follows that

$$(w, \dot{w}) = -(w, M(w)\Omega) + k\epsilon(w, w). \tag{11.1.198}$$

Verify that

$$M^T(w)w = e_0 \tag{11.1.199}$$

with e_0 given by (1.159). Next verify that

$$(w, M(w)\Omega) = (M^T(w)w, \Omega) = (e_0, \Omega) = 0. \tag{11.1.200}$$

It follows that

$$(1/2)(d/dt)(w, w) = (w, \dot{w}) = k\epsilon(w, w), \tag{11.1.201}$$

and therefore $w \cdot w$ is conserved if $k = 0$, which is the case for the equations of motion (1.42). In particular, if $w \cdot w = 1$ initially, it will remain so.

What about the evolution of $w \cdot w$ when $k \neq 0$? Verify that

$$\dot{\epsilon} = -2(w, \dot{w}). \tag{11.1.202}$$

Show that, consequently, (1.201) and (1.202) together yield the relation

$$\dot{\epsilon} = -2k\epsilon(1 - \epsilon). \tag{11.1.203}$$

Verify that (1.203) has the implicit solution

$$\epsilon(t)/[1 - \epsilon(t)] = [\epsilon_0/(1 - \epsilon_0)] \exp[-2k(t - t_0)] \tag{11.1.204}$$

where

$$\epsilon_0 = \epsilon(t_0). \tag{11.1.205}$$

To solve (1.204) explicitly for $\epsilon(t)$, let

$$r(t) = [\epsilon_0/(1 - \epsilon_0)] \exp[-2k(t - t_0)]. \tag{11.1.206}$$

Show that

$$\epsilon(t) = r(t)/[1 + r(t)]. \tag{11.1.207}$$

Evidently, for $k > 0$, $\epsilon(t) \to 0$ as $t \to +\infty$, essentially exponentially. Correspondingly, $w \cdot w \to 1$ as $t \to +\infty$.

There is an alternate way of looking at the equations (1.43) through (1.46) or (1.47) through (1.51) that emphasizes linearity in w. Let $A(\boldsymbol{\omega}^{bf})$ be the matrix defined by the rule

$$A(\boldsymbol{\omega}^{bf}) = (1/2) \begin{pmatrix} 0 & \omega_1^{bf} & \omega_2^{bf} & \omega_3^{bf} \\ -\omega_1^{bf} & 0 & \omega_3^{bf} & -\omega_2^{bf} \\ -\omega_2^{bf} & -\omega_3^{bf} & 0 & \omega_1^{bf} \\ -\omega_3^{bf} & \omega_2^{bf} & -\omega_1^{bf} & 0 \end{pmatrix}. \tag{11.1.208}$$

Show that (1.43) through (1.46) can be written in the vector/matrix form

$$\dot{w} = A(\boldsymbol{\omega}^{bf})w. \tag{11.1.209}$$

Verify that A is antisymmetric, and therefore it immediately follows from (1.209) that

$$(w, \dot{w}) = (w, Aw) = 0. \tag{11.1.210}$$

Similarly, show that (1.48) through (1.51) can be written in the form

$$\dot{w} = A(\boldsymbol{\omega}^{bf})w + k\epsilon w. \tag{11.1.211}$$

In this case, it follows directly from (1.211) and the antisymmetry of A that

$$(w, \dot{w}) = (w, Aw) + k\epsilon(w, w) = k\epsilon(w, w), \tag{11.1.212}$$

as before.

11.1.9. The aim of the this exercise is to find the ω_j^{bf} corresponding to the Tait-Bryan angle parameterization (1.58). The tools for this purpose will be very similar to those employed in Exercise 1.3, which you should review. Our discussion begins with the key relation

$$(R^v)^{-1}\dot{R}^v = \boldsymbol{\omega}^{bf} \cdot \boldsymbol{L}. \tag{11.1.213}$$

Show that this relation follows from (1.56).

For R^v given by (1.58), verify that

$$(R^v)^{-1} = \exp(-\lambda_3 L^3)\exp(-\lambda_2 L^2)\exp(-\lambda_1 L^1) \tag{11.1.214}$$

and

$$\begin{aligned}\dot{R}^v &= \dot{\lambda}_1 L^1 \exp(\lambda_1 L^1)\exp(\lambda_2 L^2)\exp(\lambda_3 L^3)\\ &+ \exp(\lambda_1 L^1)\dot{\lambda}_2 L^2 \exp(\lambda_2 L^2)\exp(\lambda_3 L^3)\\ &+ \exp(\lambda_1 L^1)\exp(\lambda_2 L^2)\exp(\lambda_3 L^3)\dot{\lambda}_3 L^3.\end{aligned} \tag{11.1.215}$$

Next verify that $(R^v)^{-1}\dot{R}^v$ is given by the expression

$$\begin{aligned}(R^v)^{-1}\dot{R}^v &= \exp(-\lambda_3 L^3)\exp(-\lambda_2 L^2)\exp(-\lambda_1 L^1)\dot{\lambda}_1 L^1 \exp(\lambda_1 L^1)\exp(\lambda_2 L^2)\exp(\lambda_3 L^3)\\ &+ \exp(-\lambda_3 L^3)\exp(-\lambda_2 L^2)\exp(-\lambda_1 L^1)\exp(\lambda_1 L^1)\dot{\lambda}_2 L^2 \exp(\lambda_2 L^2)\exp(\lambda_3 L^3)\\ &+ \exp(-\lambda_3 L^3)\exp(-\lambda_2 L^2)\exp(-\lambda_1 L^1)\exp(\lambda_1 L^1)\exp(\lambda_2 L^2)\exp(\lambda_3 L^3)\dot{\lambda}_3 L^3.\end{aligned} \tag{11.1.216}$$

Simplify each of the three lines in (1.216) so that they become

$$\begin{aligned}&\exp(-\lambda_3 L^3)\exp(-\lambda_2 L^2)\exp(-\lambda_1 L^1)\dot{\lambda}_1 L^1 \exp(\lambda_1 L^1)\exp(\lambda_2 L^2)\exp(\lambda_3 L^3)\\ &\qquad = \dot{\lambda}_1 \exp(-\lambda_3 L^3)\exp(-\lambda_2 L^2) L^1 \exp(\lambda_2 L^2)\exp(\lambda_3 L^3),\end{aligned} \tag{11.1.217}$$

$$\begin{aligned}&\exp(-\lambda_3 L^3)\exp(-\lambda_2 L^2)\exp(-\lambda_1 L^1)\exp(\lambda_1 L^1)\dot{\lambda}_2 L^2\exp(\lambda_2 L^2)\exp(\lambda_3 L^3)\\&\qquad = \dot{\lambda}_2\exp(-\lambda_3 L^3)L^2\exp(\lambda_3 L^3),\end{aligned} \tag{11.1.218}$$

$$\begin{aligned}&\exp(-\lambda_3 L^3)\exp(-\lambda_2 L^2)\exp(-\lambda_1 L^1)\exp(\lambda_1 L^1)\exp(\lambda_2 L^2)\exp(\lambda_3 L^3)\dot{\lambda}_3 L^3\\&\qquad = \dot{\lambda}_3 L^3.\end{aligned} \tag{11.1.219}$$

Line (1.219) is as simple as we could desire. The next more complicated line is (1.218). Show, using the machinery of Exercise 8.2.10, that

$$\exp(-\lambda_3 L^3)L^2\exp(\lambda_3 L^3) = \cos(\lambda_3)L^2 + \sin(\lambda_3)L^1. \tag{11.1.220}$$

Thus, the right side of (1.218) becomes

$$\dot{\lambda}_2[\cos(\lambda_3)L^2 + \sin(\lambda_3)L^1]. \tag{11.1.221}$$

Finally, work on line (1.217). Show that

$$\exp(-\lambda_2 L^2)L^1\exp(\lambda_2 L^2) = \cos(\lambda_2)L^1 + \sin(\lambda_2)L^3. \tag{11.1.222}$$

Next show that

$$\begin{aligned}&\exp(-\lambda_3 L^3)[\cos(\lambda_2)L^1 + \sin(\lambda_2)L^3]\exp(\lambda_3 L^3)\\&= \sin(\lambda_2)L^3 + \cos(\lambda_2)\exp(-\lambda_3 L^3)L^1\exp(\lambda_3 L^3)\\&= \sin(\lambda_2)L^3 + \cos(\lambda_2)[\cos(\lambda_3)L^1 - \sin(\lambda_3)L^2].\end{aligned} \tag{11.1.223}$$

By combining (1.222) and (1.223), verify that

$$\begin{aligned}&\exp(-\lambda_3 L^3)\exp(-\lambda_2 L^2)L^1\exp(\lambda_2 L^2)\exp(\lambda_3 L^3)\\&= \sin(\lambda_2)L^3 + \cos(\lambda_2)[\cos(\lambda_3)L^1 - \sin(\lambda_3)L^2].\end{aligned} \tag{11.1.224}$$

Thus, the right side of (1.217) becomes

$$\dot{\lambda}_1\{\sin(\lambda_2)L^3 + \cos(\lambda_2)[\cos(\lambda_3)L^1 - \sin(\lambda_3)L^2]\}. \tag{11.1.225}$$

Now we have everything we need. By combining (1.216) through (1.225), show that

$$\begin{aligned}(R^v)^{-1}\dot{R}^v &= \dot{\lambda}_1\{\sin(\lambda_2)L^3 + \cos(\lambda_2)[\cos(\lambda_3)L^1 - \sin(\lambda_3)L^2]\}\\&+ \dot{\lambda}_2[\cos(\lambda_3)L^2 + \sin(\lambda_3)L^1]\\&+ \dot{\lambda}_3 L^3\\&= [\dot{\lambda}_1\cos(\lambda_2)\cos(\lambda_3) + \dot{\lambda}_2\sin(\lambda_3)]L^1\\&+ [-\dot{\lambda}_1\cos(\lambda_2)\sin(\lambda_3) + \dot{\lambda}_2\cos(\lambda_3)]L^2\\&+ [\dot{\lambda}_1\sin(\lambda_2) + \dot{\lambda}_3]L^3.\end{aligned} \tag{11.1.226}$$

Verify, upon equating coefficients of the L^j in (1.213) and (1.226), that there are the relations

$$\omega_1^{bf} = \dot{\lambda}_1 \cos(\lambda_2)\cos(\lambda_3) + \dot{\lambda}_2 \sin(\lambda_3), \tag{11.1.227}$$

$$\omega_2^{bf} = -\dot{\lambda}_1 \cos(\lambda_2)\sin(\lambda_3) + \dot{\lambda}_2 \cos(\lambda_3), \tag{11.1.228}$$

$$\omega_3^{bf} = \dot{\lambda}_1 \sin(\lambda_2) + \dot{\lambda}_3. \tag{11.1.229}$$

Finally, verify that inverting the relations (1.227) through (1.229) yields the equations of motion

$$\dot{\lambda}_1 = [1/\cos(\lambda_2)][\omega_1^{bf}\cos(\lambda_3) - \omega_2^{bf}\sin(\lambda_3)], \tag{11.1.230}$$

$$\dot{\lambda}_2 = \omega_1^{bf}\sin(\lambda_3) + \omega_2^{bf}\cos(\lambda_3), \tag{11.1.231}$$

$$\dot{\lambda}_3 = \omega_3^{bf} - \tan(\lambda_2)[\omega_1^{bf}\cos(\lambda_3) - \omega_2^{bf}\sin(\lambda_3)]. \tag{11.1.232}$$

Verify that these equations of motion are nonsingular for small λ_j. Note, however, that there are singularities when $\lambda_2 = \pm\pi/2$. What causes these singularities? Show that

$$\begin{aligned}
&\exp(\lambda_1 L^1)\exp[(\pi/2)L^2]\exp(\lambda_3 L^3)\\
&= \exp(\lambda_1 L^1)\exp[(\pi/2)L^2]\exp(\lambda_3 L^3)\exp[-(\pi/2)L^2]\exp[(\pi/2)L^2]\\
&= \exp(\lambda_1 L^1)\exp(\lambda_3 L^1)\exp[(\pi/2)L^2]\\
&= \exp[(\lambda_1 + \lambda_3)L^1]\exp[(\pi/2)L^2],
\end{aligned} \tag{11.1.233}$$

$$\begin{aligned}
&\exp(\lambda_1 L^1)\exp[-(\pi/2)L^2]\exp(\lambda_3 L^3)\\
&= \exp(\lambda_1 L^1)\exp[(-\pi/2)L^2]\exp(\lambda_3 L^3)\exp[(\pi/2)L^2]\exp[-(\pi/2)L^2]\\
&= \exp(\lambda_1 L^1)\exp(-\lambda_3 L^1)\exp[-(\pi/2)L^2]\\
&= \exp[(\lambda_1 - \lambda_3)L^1]\exp[-(\pi/2)L^2].
\end{aligned} \tag{11.1.234}$$

Verify that only the combination $(\lambda_1 + \lambda_3)$ is well defined when $\lambda_2 = \pi/2$, and only the combination $(\lambda_1 - \lambda_3)$ is well defined when $\lambda_2 = -\pi/2$. Therefore the parameterization (1.58) fails when $\lambda_2 = \pm\pi/2$.

11.1.10. The aim of this exercise, which is not easy, is to find the $\dot{\lambda}_j$ in terms of the ω_j^{bf} for the angle-axis parameterization (1.66). Although not essential, it is convenient for this purpose to use the $SU(2)$ version (1.146) of the $SO(3,\mathbb{R})$ relation (1.22). In analogy to (1.66), begin by writing

$$u(\boldsymbol{\lambda}) = \exp(\boldsymbol{\lambda}\cdot\boldsymbol{K}). \tag{11.1.235}$$

From the rule for differentiating the exponential function there is the result

$$(d/dt)u = u\ \mathrm{iex}(-\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)(d/dt)(\boldsymbol{\lambda}\cdot\boldsymbol{K}), \tag{11.1.236}$$

which can be rewritten in the form

$$\dot{u} = u\ \mathrm{iex}(-\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)(\dot{\boldsymbol{\lambda}}\cdot\boldsymbol{K}). \tag{11.1.237}$$

See Appendix C. Combine (1.146) and (1.237) to show that

$$\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K} = \mathrm{iex}(-\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)(\dot{\boldsymbol{\lambda}}\cdot\boldsymbol{K}). \tag{11.1.238}$$

Now for some daring steps. Show that, as in Section 10.3, the relation (1.238) can be inverted to become

$$\begin{aligned}\dot{\boldsymbol{\lambda}}\cdot\boldsymbol{K} &= [\mathrm{iex}(-\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)]^{-1}(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})\\ &= [I+(1/2)\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#+(1/12)(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^2+\cdots](\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})\\ &= (\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})+(1/2)\{(\boldsymbol{\lambda}\cdot\boldsymbol{K}),(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})\}\\ &\quad +(1/12)\{(\boldsymbol{\lambda}\cdot\boldsymbol{K}),\{(\boldsymbol{\lambda}\cdot\boldsymbol{K}),(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})\}\}+\cdots. \qquad (11.1.239)\end{aligned}$$

The result (1.239) is quite general in the sense that analogous results hold for any Lie group. For the specific case of $SU(2)$ we will seek to explicitly sum the series (1.239).

From Appendix C we know that there is the expansion

$$[\mathrm{iex}(-\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)]^{-1}=\sum_{m=0}^{\infty}b_m(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^m. \qquad (11.1.240)$$

Verify that insertion of (1.240) into (1.239) gives the result

$$\begin{aligned}\dot{\boldsymbol{\lambda}}\cdot\boldsymbol{K} &= [\sum_{m=0}^{\infty}b_m(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^m](\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})\\ &= b_0(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})+b_1(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})\\ &\quad +b_2(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^2(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})+b_3(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^3(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})\\ &\quad +b_4(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^4(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})+b_5(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^5(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})+\cdots. \qquad (11.1.241)\end{aligned}$$

Now evaluate the terms appearing in (1.241). Show that there are the results

$$(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})=\{\boldsymbol{\lambda}\cdot\boldsymbol{K},\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}\}=(\boldsymbol{\lambda}\times\boldsymbol{\omega}^{bf})\cdot\boldsymbol{K}, \qquad (11.1.242)$$

$$\begin{aligned}(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^2(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}) &= \{\boldsymbol{\lambda}\cdot\boldsymbol{K},(\boldsymbol{\lambda}\times\boldsymbol{\omega}^{bf})\cdot\boldsymbol{K}\}=[\boldsymbol{\lambda}\times(\boldsymbol{\lambda}\times\boldsymbol{\omega}^{bf})]\cdot\boldsymbol{K}\\ &= (\boldsymbol{\lambda}\cdot\boldsymbol{\omega}^{bf})(\boldsymbol{\lambda}\cdot\boldsymbol{K})-(\boldsymbol{\lambda}\cdot\boldsymbol{\lambda})(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}), \qquad (11.1.243)\end{aligned}$$

$$\begin{aligned}(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^3(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}) &= (\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^2(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})\\ &= -(\boldsymbol{\lambda}\cdot\boldsymbol{\lambda})\{\boldsymbol{\lambda}\cdot\boldsymbol{K},\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}\}\\ &= -(\boldsymbol{\lambda}\cdot\boldsymbol{\lambda})(\boldsymbol{\lambda}\times\boldsymbol{\omega}^{bf})\cdot\boldsymbol{K}\\ &= (i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda})(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}), \qquad (11.1.244)\end{aligned}$$

$$\begin{aligned}(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^4(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}) &= (\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^3(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})\\ &= (\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)(-1)(\boldsymbol{\lambda}\cdot\boldsymbol{\lambda})(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})\\ &= (i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda})(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^2(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}),\ \text{etc.} \qquad (11.1.245)\end{aligned}$$

Note the similarity of (3.7.201) and (1.244).

Verify that putting everything together gives the net result

$$\begin{aligned}
&[\sum_{m=0}^{\infty} b_m(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^m](\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}) = b_0(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})\\
&+b_1(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}) + b_3(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^3(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}) + \cdots\\
&+b_2(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^2(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}) + b_4(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^4(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}) + \cdots\\
&= b_0(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})\\
&+[(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})][b_1 + b_3(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda}) + b_5(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda})^2\cdots]\\
&+[(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^2(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})][b_2 + b_4(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda}) + b_6(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda})^2\cdots]\\
&= b_0(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})\\
&+[(\boldsymbol{\lambda}\times\boldsymbol{\omega}^{bf})\cdot\boldsymbol{K}][b_1 + b_3(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda}) + b_5(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda})^2\cdots]\\
&+[(\boldsymbol{\lambda}\cdot\boldsymbol{\omega}^{bf})(\boldsymbol{\lambda}\cdot\boldsymbol{K}) - (\boldsymbol{\lambda}\cdot\boldsymbol{\lambda})(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})][b_2 + b_4(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda}) + b_6(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda})^2\cdots].
\end{aligned} \tag{11.1.246}$$

Also verify that combining (1.241) and the last result in (1.246) gives the relation

$$\begin{aligned}
&\dot{\boldsymbol{\lambda}} = b_0\boldsymbol{\omega}^{bf}\\
&+(\boldsymbol{\lambda}\times\boldsymbol{\omega}^{bf})[b_1 + b_3(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda}) + b_5(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda})^2\cdots]\\
&+[(\boldsymbol{\lambda}\cdot\boldsymbol{\omega}^{bf})\boldsymbol{\lambda} - (\boldsymbol{\lambda}\cdot\boldsymbol{\lambda})\boldsymbol{\omega}^{bf}][b_2 + b_4(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda}) + b_6(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda})^2\cdots].
\end{aligned} \tag{11.1.247}$$

What remains to be done is to sum the series in (1.247). Begin by defining a quantity w by the rule

$$w = \sqrt{(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda})} = i|\boldsymbol{\lambda}|. \tag{11.1.248}$$

With this definition we have the relations

$$[b_1 + b_3(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda}) + b_5(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda})^2\cdots] = (1/w)\sum_{\text{odd } m} b_m w^m, \tag{11.1.249}$$

$$[b_2 + b_4(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda}) + b_6(i\boldsymbol{\lambda}\cdot i\boldsymbol{\lambda})^2\cdots] = (1/w^2)\sum_{\text{even } m>0} b_m w^m. \tag{11.1.250}$$

We also know, see Appendix C, that

$$\sum_{m=0}^{\infty} b_m w^m = w/[1-\exp(-w)], \tag{11.1.251}$$

from which it follows that

$$\sum_{m>0} b_m w^m = w/[1-\exp(-w)] - 1. \tag{11.1.252}$$

(Here we have used the result $b_0 = 1$, which you should check.) Verify the identity

$$\begin{aligned}
w/[1-\exp(-w)] &= w\exp(w/2)/[\exp(w/2)-\exp(-w/2)]\\
&= (w/2)[\cosh(w/2)+\sinh(w/2)]/\sinh(w/2)\\
&= w/2 + (w/2)\coth(w/2).
\end{aligned} \tag{11.1.253}$$

Use this identity to show that

$$\sum_{m>0} b_m w^m = [w/2] + [(w/2)\coth(w/2) - 1]. \tag{11.1.254}$$

From (1.254), by equating odd and even parts, show that

$$\sum_{\text{odd } m} b_m w^m = w/2, \tag{11.1.255}$$

$$\sum_{\text{even } m>0} b_m w^m = (w/2)\coth(w/2) - 1. \tag{11.1.256}$$

Use these results to show that

$$(1/w) \sum_{\text{odd } m} b_m w^m = 1/2, \tag{11.1.257}$$

$$(1/w^2) \sum_{\text{even } m>0} b_m w^m = [1/(2w)]\coth(w/2) - 1/w^2. \tag{11.1.258}$$

Verify that employing (1.248) in (1.258) yields the result

$$(1/w^2) \sum_{\text{even } m>0} b_m w^m = 1/|\boldsymbol{\lambda}|^2 - [1/(2|\boldsymbol{\lambda}|)]\cot(|\boldsymbol{\lambda}|/2). \tag{11.1.259}$$

At last, verify the final (and amazing) result

$$\dot{\boldsymbol{\lambda}} = \boldsymbol{\omega}^{bf} + (1/2)(\boldsymbol{\lambda} \times \boldsymbol{\omega}^{bf}) + [(\boldsymbol{\lambda} \cdot \boldsymbol{\omega}^{bf})\boldsymbol{\lambda} - (\boldsymbol{\lambda} \cdot \boldsymbol{\lambda})\boldsymbol{\omega}^{bf}]\{1/|\boldsymbol{\lambda}|^2 - [1/(2|\boldsymbol{\lambda}|)]\cot(|\boldsymbol{\lambda}|/2)\}. \tag{11.1.260}$$

Check that wherever $|\boldsymbol{\lambda}|$ appears in (1.260), it appears as an *even* power. Therefore, there is no overall ambiguity in (1.260) despite the sign ambiguity present in the definition (1.248). Show, moreover, that the right side of (1.260) is *analytic* in the components of $\boldsymbol{\lambda}$ for $|\boldsymbol{\lambda}| < 2\pi$, but is singular when $|\boldsymbol{\lambda}| = 2\pi$. As stated earlier, we expect this singularity to occur because we see from (3.7.188) and (3.7.202) that the individual components of $\boldsymbol{n}$ are not defined in terms of v or R when $\theta = |\boldsymbol{\lambda}| = 2\pi$.

There is another way of writing (1.260) that is of interest. Define a function $f(\boldsymbol{\lambda})$ by the rule

$$f(\boldsymbol{\lambda}) = 1/|\boldsymbol{\lambda}|^2 - [1/(2|\boldsymbol{\lambda}|)]\cot(|\boldsymbol{\lambda}|/2). \tag{11.1.261}$$

Verify that $f(\boldsymbol{\lambda})$ is even in $|\boldsymbol{\lambda}|$, is analytic in the components of $\boldsymbol{\lambda}$ for $|\boldsymbol{\lambda}| < 2\pi$, and is singular when $|\boldsymbol{\lambda}| = 2\pi$. Define a 3×3 matrix $M(\boldsymbol{\lambda})$ by the rule

$$M(\boldsymbol{\lambda}) = I + (1/2)(\boldsymbol{\lambda} \cdot \boldsymbol{L}) + f(\boldsymbol{\lambda})(\boldsymbol{\lambda} \cdot \boldsymbol{L})^2. \tag{11.1.262}$$

Verify that (1.260) can also be written in the form

$$\dot{\boldsymbol{\lambda}} = M(\boldsymbol{\lambda})\boldsymbol{\omega}^{bf}, \tag{11.1.263}$$

which highlights linearity in $\boldsymbol{\omega}^{bf}$ and the role of the matrix $\boldsymbol{\lambda} \cdot \boldsymbol{L}$.

11.1.11. Review Exercise 1.10. It found the $\dot{\lambda}_j$ in terms of the ω_j^{bf} for the parameterization (1.66). For some purposes it is useful to also find the ω_j^{bf} in terms of the $\dot{\lambda}_j$. That is the aim of this exercise.

There are at least two ways to proceed. The first begins with the relation (1.238) and makes an expansion of the form

$$\text{iex}(-\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)=\sum_{m=0}^{\infty}d_m(\#\boldsymbol{\lambda}\cdot\boldsymbol{K}\#)^m. \tag{11.1.264}$$

This expansion is then manipulated, in the spirit of Exercise 1.10, to find and sum expansions that specify the ω_j^{bf} in terms of the $\dot{\lambda}_j$.

A second way exploits more of what we already know from Exercise 1.10. Suppose we could invert the matrix $M(\boldsymbol{\lambda})$ given in (1.262). Then we could rewrite (1.263) in the form

$$\boldsymbol{\omega}^{bf}=[M(\boldsymbol{\lambda})]^{-1}\dot{\boldsymbol{\lambda}}, \tag{11.1.265}$$

and we would have found the ω_j^{bf} in terms of the $\dot{\lambda}_j$.

We now proceed to construct $[M(\boldsymbol{\lambda})]^{-1}$. Let $N(\boldsymbol{\lambda})$ be a 3×3 matrix of the form

$$N(\boldsymbol{\lambda})=I+a(\boldsymbol{\lambda})(\boldsymbol{\lambda}\cdot\boldsymbol{L})+b(\boldsymbol{\lambda})(\boldsymbol{\lambda}\cdot\boldsymbol{L})^2 \tag{11.1.266}$$

where $a(\boldsymbol{\lambda})$ and $(\boldsymbol{\lambda})$ are coefficients to be determined. Next form the product of $M(\boldsymbol{\lambda})$ and $N(\boldsymbol{\lambda})$. Verify that so doing yields the result

$$\begin{aligned} MN &= I+(1/2)(\boldsymbol{\lambda}\cdot\boldsymbol{L})+f(\boldsymbol{\lambda}\cdot\boldsymbol{L})^2\\ &+ a(\boldsymbol{\lambda}\cdot\boldsymbol{L})+(1/2)a(\boldsymbol{\lambda}\cdot\boldsymbol{L})^2+af(\boldsymbol{\lambda}\cdot\boldsymbol{L})^3\\ &+ b(\boldsymbol{\lambda}\cdot\boldsymbol{L})^2+(1/2)b(\boldsymbol{\lambda}\cdot\boldsymbol{L})^3+bf(\boldsymbol{\lambda}\cdot\boldsymbol{L})^4. \end{aligned} \tag{11.1.267}$$

Verify the property

$$(\boldsymbol{\lambda}\cdot\boldsymbol{L})^3=-|\boldsymbol{\lambda}|^2\boldsymbol{\lambda}\cdot\boldsymbol{L}, \tag{11.1.268}$$

and employ it in (1.267) to achieve the net result

$$\begin{aligned} & MN=I+(1/2)(\boldsymbol{\lambda}\cdot\boldsymbol{L})+f(\boldsymbol{\lambda}\cdot\boldsymbol{L})^2\\ + \quad & a(\boldsymbol{\lambda}\cdot\boldsymbol{L})+(1/2)a(\boldsymbol{\lambda}\cdot\boldsymbol{L})^2-|\boldsymbol{\lambda}|^2af(\boldsymbol{\lambda}\cdot\boldsymbol{L})\\ + \quad & b(\boldsymbol{\lambda}\cdot\boldsymbol{L})^2-|\boldsymbol{\lambda}|^2(1/2)b(\boldsymbol{\lambda}\cdot\boldsymbol{L})-|\boldsymbol{\lambda}|^2bf(\boldsymbol{\lambda}\cdot\boldsymbol{L})^2\\ = \quad & I+[(1/2)+a-|\boldsymbol{\lambda}|^2af-|\boldsymbol{\lambda}|^2(1/2)b](\boldsymbol{\lambda}\cdot\boldsymbol{L})\\ + \quad & [f+(1/2)a+b-|\boldsymbol{\lambda}|^2bf](\boldsymbol{\lambda}\cdot\boldsymbol{L})^2. \end{aligned} \tag{11.1.269}$$

Suppose we can arrange that the coefficients of $(\boldsymbol{\lambda}\cdot\boldsymbol{L})$ and $(\boldsymbol{\lambda}\cdot\boldsymbol{L})^2$ in the net result (1.269) vanish. So doing requires the conditions

$$(1/2)+a-|\boldsymbol{\lambda}|^2af-|\boldsymbol{\lambda}|^2(1/2)b=0 \tag{11.1.270}$$

and

$$f + (1/2)a + b - |\boldsymbol{\lambda}|^2 bf = 0. \tag{11.1.271}$$

If these conditions are met, it follows that $MN = I$, and therefore

$$N(\boldsymbol{\lambda}) = [M(\boldsymbol{\lambda})]^{-1}. \tag{11.1.272}$$

Verify that the linear equations (1.270)and (1.271) can be written in the standard form

$$(1 - |\boldsymbol{\lambda}|^2 f)a - (|\boldsymbol{\lambda}|^2/2)b = -1/2, \tag{11.1.273}$$

$$(1/2)a + (1 - |\boldsymbol{\lambda}|^2 f)b = -f. \tag{11.1.274}$$

Show from the definition (1.261) of f that

$$(1 - |\boldsymbol{\lambda}|^2 f) = (|\boldsymbol{\lambda}|/2)\cot(|\boldsymbol{\lambda}|/2). \tag{11.1.275}$$

Show that (1.273) and (1.274) have the solution

$$a = -(2/|\boldsymbol{\lambda}|^2)\sin^2(|\boldsymbol{\lambda}|/2), \tag{11.1.276}$$

$$b = (1/|\boldsymbol{\lambda}|^2)[1 - (1/|\boldsymbol{\lambda}|)\sin(|\boldsymbol{\lambda}|)]. \tag{11.1.277}$$

Evaluate $a(0)$ and $b(0)$ and verify that both $a(\boldsymbol{\lambda})$ and $b(\boldsymbol{\lambda})$ are analytic in the components of $\boldsymbol{\lambda}$ everywhere except at ∞.

11.1.12. The aim of this exercise is to explore the use of Cayley transforms and parameterizations for the purpose of integration on manifolds. Review Section 3.12 and Exercises 3.12.1, 3.12.5, and 3.12.6 to recall the subject of Cayley transforms for quadratic groups G. We will begin with the general case. Then in a following exercise, we will study in more detail the cases of $SO(3, \mathbb{R})$ and $SU(2)$, which are more tractable. Specifically, in this exercise we will study equations of motion of the form

$$\dot{M}(t) = M(t)A(t) \tag{11.1.278}$$

where $M(t)$ is expected to belong to some quadratic Lie group G and $A(t)$ belongs to its associated Lie algebra. What we seek is a way of numerically integrating (1.278) that guarantees $M(t)$ is in G even though the numerical solution may be locally exact only through terms of order h^m.

We could also study the related equations of motion of the form

$$\dot{N}(t) = \bar{A}(t)N(t) \tag{11.1.279}$$

but, as will be seen from some of the work of Exercise 2.7, this case is equivalent to the case (1.278) under the substitutions $N = M^{-1}$ and $\bar{A} = -A$.

From Appendix C we know in general that exact integration of (1.278) assures that $M(t)$ is in G. Here, before going further, your first task is to provide a simple proof of this fact in the case of quadratic groups.

Begin with the converse claim. Let $M(t)$ be some path in matrix space. By Taylor's theorem there is the result

$$M(t+dt)=M(t)+dt\dot{M}(t)+O[(dt)^2]. \tag{11.1.280}$$

Define a matrix $A(t)$ by the rule

$$A(t)=M^{-1}(t)\dot{M}(t). \tag{11.1.281}$$

Verify that (1.278) follows from (1.281). What remains is to determine the properties of $A(t)$.

For $M(t)$ to belong to a quadratic group G it must satisfy a relation of the form

$$M^T(t)LM(t)=L. \tag{11.1.282}$$

See Exercise 3.12.5. From (1.282) verify that

$$M^T(t+dt)LM(t+dt)=L. \tag{11.1.283}$$

Since $M(t)$ and $M(t+dt)$ are nearby matrices, it follows that the product $M^{-1}(t)M(t+dt)$ must be near the identity. Indeed, verify that we may write

$$\begin{aligned} M^{-1}(t)M(t+dt) &= M^{-1}(t)[M(t)+dt\dot{M}(t)]+O[(dt)^2] \\ &= I+dtM^{-1}(t)\dot{M}(t)+O[(dt)^2] \\ &= I+dtA(t)+O[(dt)^2] \\ &= \exp[dtA(t)]+O[(dt)^2]. \end{aligned} \tag{11.1.284}$$

Rewrite (1.284) in the form

$$M(t+dt)=M(t)\exp[dtA(t)]+O[(dt)^2]. \tag{11.1.285}$$

Show that employing this relation in (1.283) and equating powers of dt yields the condition

$$A^T(t)L+LA(t)=0, \tag{11.1.286}$$

which demonstrates that $A(t)$ belongs to the Lie algebra of G.

Next consider the direct claim. Suppose that $M(t)$ satisfies (1.278) and that $A(t)$ satisfies (1.286). Assume also that at some time t^0 there is the relation

$$M^T(t^0)LM(t^0)=L. \tag{11.1.287}$$

Such will be the case, in particular, if $M(t^0)$=I. Your task is to show that then (1.282) must hold for all t.

Begin by showing that (1.287) can be rewritten in the form

$$[M^{-1}(t^0)]^TL[M^{-1}(t^0)]=L. \tag{11.1.288}$$

Next, from the identity

$$M^{-1}(t)M(t)=I \tag{11.1.289}$$

and (1.278), show that

$$(d/dt)[M^{-1}(t)] = -A(t)M^{-1}(t). \tag{11.1.290}$$

As a further step show that

$$\begin{aligned}(d/dt)[(M^{-1})^T LM^{-1}] &= \{(d/dt)[(M^{-1})^T]\}LM^{-1} + (M^{-1})^T L(d/dt)(M^{-1}) \\ &= [(d/dt)(M^{-1})]^T LM^{-1} + (M^{-1})^T L(d/dt)(M^{-1}) \\ &= [-AM^{-1}]^T LM^{-1} + (M^{-1})^T L[-AM^{-1}] \\ &= -(M^{-1})^T[A^T L + LA]M^{-1} = 0. \end{aligned} \tag{11.1.291}$$

Verify that the unique solution to the differential equation (1.291) with the initial condition (1.288) is the relation

$$[M^{-1}(t)]^T L[M^{-1}(t)] = L. \tag{11.1.292}$$

Finally, show that (1.282) follows from (1.292).

With this background work out of the way, the main task of this exercise is to consider use of the Cayley parameterization. Specifically, for M we employ the Cayley parameterization

$$M = (I+V)(I-V)^{-1}, \tag{11.1.293}$$

see (3.12.36), and your task is to find the equation of motion for V. In particular, you are to show how this may be done starting from (1.278) rewritten as(1.281).

Begin by writing M in the form

$$M = BC \tag{11.1.294}$$

where

$$B = I + V \tag{11.1.295}$$

and

$$C = D^{-1} \tag{11.1.296}$$

with

$$D = I - V. \tag{11.1.297}$$

Show that the product differentiation rule yields the result

$$\dot{M} = \dot{B}C + B\dot{C}. \tag{11.1.298}$$

Simple calculation with (1.295) yields the result

$$\dot{B} = \dot{V}. \tag{11.1.299}$$

The calculation of $\dot{C}$ is more involved. Show from (1.296) and the product differentiation rule that

$$CD = I, \tag{11.1.300}$$

$$\dot{C}D + C\dot{D} = 0, \tag{11.1.301}$$

and therefore

$$\dot{C} = -C\dot{D}C. \tag{11.1.302}$$

Show that use of (1.297) gives the result

$$\dot{D} = -\dot{V}, \tag{11.1.303}$$

and therefore

$$\dot{C} = C\dot{V}C. \tag{11.1.304}$$

Verify that combining the results obtained so far gives the next conclusion

$$\dot{M} = \dot{V}C + BC\dot{V}C. \tag{11.1.305}$$

Manipulate some more. Verify the steps

$$\begin{aligned}\dot{M} &= \dot{V}C + BC\dot{V}C = C^{-1}C\dot{V}C + BC\dot{V}C = (C^{-1} + B)C\dot{V}C \\ &= (D + B)C\dot{V}C = [(I - V) + (I + V)]C\dot{V}C = 2C\dot{V}C.\end{aligned} \tag{11.1.306}$$

According to (1.281) what is needed is the quantity $M^{-1}\dot{M}$. Show that

$$M^{-1} = C^{-1}B^{-1} \tag{11.1.307}$$

and therefore

$$M^{-1}\dot{M} = 2C^{-1}B^{-1}C\dot{V}C. \tag{11.1.308}$$

Verify that B and C commute and therefore B^{-1} and C^{-1} commute. It follows that

$$C^{-1}B^{-1}C = B^{-1}C^{-1}C = B^{-1} \tag{11.1.309}$$

so that

$$M^{-1}\dot{M} = 2C^{-1}B^{-1}C\dot{V}C = 2B^{-1}\dot{V}C, \tag{11.1.310}$$

from which we conclude, with the aid of (1.281), that

$$A = 2B^{-1}\dot{V}C. \tag{11.1.311}$$

Solve (1.311) for $\dot{V}$ to yield the result

$$\dot{V} = (1/2)BAC^{-1} = (1/2)(I + V)A(I - V) = (1/2)(A + VA - AV - VAV), \tag{11.1.312}$$

which can be written as

$$\dot{V} = (1/2)(A + \{V, A\} - VAV). \tag{11.1.313}$$

You have found $\dot{V}$ in terms of A. Note that, in contrast to (1.82) whose right side contains an infinite number of terms, the right side of (1.313) contains only three terms.

Is this result sane? For a quadratic group G we know that V is in the Lie algebra. From the definition

$$\dot{V}(t) = \lim_{\epsilon \to 0}[V(t + \epsilon) - V(t)]/\epsilon \tag{11.1.314}$$

we see that only vector space operations are involved in the calculation of $\dot{V}(t)$, and therefore $\dot{V}(t)$ must also be in the Lie algebra of G. But is the right side of (1.313) in the Lie algebra of G? Evidently, since A is in the Lie algebra of G, the first two terms on the right side of (1.313) are in the Lie algebra of G. What about the third term VAV? According to Exercise 3.12.4 the condition for A and V to be in the Lie algebra of G is that

$$L^{-1}A^T L = -A, \tag{11.1.315}$$

$$L^{-1}V^T L = -V. \tag{11.1.316}$$

See (3.12.34). Verify it follows by matrix manipulation that

$$L^{-1}(VAV)^T L = -VAV, \tag{11.1.317}$$

and therefore VAV is also in the Lie algebra of G. Therefore (1.313) is sane at least to the extent that both its sides are in the Lie algebra of G.

To return to the main discussion, suppose (1.313) is integrated by some numerical integrator to find $V(t)$ under the assumption that V is initially in the Lie algebra of G. We repeat the key observation of Subsection 1.14: Examination of the usual numerical integration schemes, see Chapter 2, reveals that they all involve just linear combinations of the right side of the differential equation in question evaluated at various times and coordinate values. Therefore, if the right side is known to be in the Lie algebra of G for all evaluation points, then the result of numerically integrating such an equation is guaranteed to be in the Lie algebra of G, no matter what the local accuracy of the integrator or the step size employed. Since we have verified that the right side of (1.313) is in the Lie algebra of G, it follows that $V(t)$ will be in the Lie algebra of G if it is initially in the Lie algebra of G. Finally, since $V(t)$ is in the Lie algebra of G, it follows that $M(t)$ given by (1.293) is in G.

We have achieved our goal of, in effect, numerically integrating (1.278) in such a way that $M(t)$ is guaranteed to be in G. Note, however, that this procedure cannot be carried out globally since the Cayley parameterization (1.293) cannot be made globally. It may therefore be necessary to change coordinate systems (by left or right group translation) from time to time during the course of a numerical integration in order to stay clear of the singularities associated with any given Cayley parametrization.

Although we have achieved our goal, there is still one undesirable feature of our procedure. Namely, if A and V are $k \times k$ matrices, then the integration of (1.313) involves the integration of k^2 equations. Generally the group G has dimension considerably less than k^2. What we would like is a way of parameterizing the Lie algebra of G and a procedure that only involves the integration of differential equations for these parameters. In the next exercise we will illustrate such a procedure for the cases of $SO(3,\mathbb{R})$ and $SU(2)$ where the necessary operations can be carried out relatively easily.

We close this exercise with a small variation. Suppose the relations (1.293) through (1.307) remain in force, but the task is to find

$$\bar{A} = \dot{M}M^{-1} \tag{11.1.318}$$

rather than (1.281). Verify, using (1.306) and (1.307), that

$$\dot{M}M^{-1} = 2C\dot{V}CC^{-1}B^{-1} = 2C\dot{V}B^{-1}. \tag{11.1.319}$$

Show that solving (1.319) for $\dot{V}$ with the aid of (1.318) yields the result

$$\begin{aligned} \dot{V} &= (1/2)C^{-1}\bar{A}B = (1/2)D\bar{A}B = (1/2)I - V)\bar{A}(I+V) \\ &= (1/2)(\bar{A} - \{V, \bar{A}\} - V\bar{A}V). \end{aligned} \tag{11.1.320}$$

This result will be of use in Subsection 2.9.

11.1.13. The aim of this exercise is to apply the methods of the previous exercise to the cases of $SO(3, \mathbb{R})$ and $SU(2)$ including parameterization of the associated Lie algebras. In the case of $SO(3, \mathbb{R})$ we seek to integrate the equation

$$\dot{R} = R\, \boldsymbol{\omega}^{bf} \cdot \boldsymbol{L}. \tag{11.1.321}$$

Recall (1.18). And in the case of $SU(2)$ we seek to integrate (1.146) rewritten in the form

$$\dot{u} = u\, \boldsymbol{\omega}^{bf} \cdot \boldsymbol{K}. \tag{11.1.322}$$

In these two cases we write the Cayley parameterizations

$$R(\boldsymbol{\mu}) = (I + \boldsymbol{\mu} \cdot \boldsymbol{L})(I - \boldsymbol{\mu} \cdot \boldsymbol{L})^{-1} \tag{11.1.323}$$

and

$$u(\boldsymbol{\mu}) = (I + \boldsymbol{\mu} \cdot \boldsymbol{K})(I - \boldsymbol{\mu} \cdot \boldsymbol{K})^{-1}. \tag{11.1.324}$$

What you are to find is the relation between $\dot{\boldsymbol{\mu}}$ and $\boldsymbol{\omega}^{bf}$ for these two cases. At this point you should review Exercise 1.12 if you have not previously studied it.

For the case of $SO(3, \mathbb{R})$ compare (1.321) with (1.278) and compare (1.323) with (1.293). Show that in this case there are the correspondences

$$A = \boldsymbol{\omega}^{bf} \cdot \boldsymbol{L}, \tag{11.1.325}$$

$$V = \boldsymbol{\mu} \cdot \boldsymbol{L},\ \dot{V} = \dot{\boldsymbol{\mu}} \cdot \boldsymbol{L}. \tag{11.1.326}$$

For the case of $SU(2)$ compare (1.322) with (1.278) and compare (1.324) with (1.293). Show that in this case there are the correspondences

$$A = \boldsymbol{\omega}^{bf} \cdot \boldsymbol{K}, \tag{11.1.327}$$

$$V = \boldsymbol{\mu} \cdot \boldsymbol{K},\ \dot{V} = \dot{\boldsymbol{\mu}} \cdot \boldsymbol{K}. \tag{11.1.328}$$

What remains is to insert these results into (1.313) and to work out the consequences. Begin with the case of $SO(3, \mathbb{R})$. Show that use of (1.325) and (1.326) in (1.313) yields the result

$$\dot{\boldsymbol{\mu}} \cdot \boldsymbol{L} = (1/2)[\boldsymbol{\omega}^{bf} \cdot \boldsymbol{L} + \{\boldsymbol{\mu} \cdot \boldsymbol{L}, \boldsymbol{\omega}^{bf} \cdot \boldsymbol{L}\} - (\boldsymbol{\mu} \cdot \boldsymbol{L})(\boldsymbol{\omega}^{bf} \cdot \boldsymbol{L})(\boldsymbol{\mu} \cdot \boldsymbol{L})]. \tag{11.1.329}$$

Next manipulate the terms appearing on the right side of this equation. The commutator term is easy. It has the value

$$\{\boldsymbol{\mu} \cdot \boldsymbol{L}, \boldsymbol{\omega}^{bf} \cdot \boldsymbol{L}\} = (\boldsymbol{\mu} \times \boldsymbol{\omega}^{bf}) \cdot \boldsymbol{L}. \tag{11.1.330}$$

See (3.7.183). The evaluation of

$$(\boldsymbol{\mu}\cdot\boldsymbol{L})(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{L})(\boldsymbol{\mu}\cdot\boldsymbol{L}) = \ ? \tag{11.1.331}$$

is more tedious. By using the 3×3 matrix form for each of the factors in (1.331), multiplying out the matrices, and rewriting the result in the form $\boldsymbol{c}\cdot\boldsymbol{L}$, show that

$$(\boldsymbol{\mu}\cdot\boldsymbol{L})(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{L})(\boldsymbol{\mu}\cdot\boldsymbol{L}) = -(\boldsymbol{\mu}\cdot\boldsymbol{\omega}^{bf})(\boldsymbol{\mu}\cdot\boldsymbol{L}). \tag{11.1.332}$$

Verify that combining (1.329) through (1.332) gives the result

$$\dot{\boldsymbol{\mu}}\cdot\boldsymbol{L} = (1/2)[\boldsymbol{\omega}^{bf} + (\boldsymbol{\mu}\times\boldsymbol{\omega}^{bf}) + (\boldsymbol{\mu}\cdot\boldsymbol{\omega}^{bf})\boldsymbol{\mu}]\cdot\boldsymbol{L}. \tag{11.1.333}$$

From this result it follows that there are the equations of motion

$$\dot{\boldsymbol{\mu}} = (1/2)[\boldsymbol{\omega}^{bf} + (\boldsymbol{\mu}\times\boldsymbol{\omega}^{bf}) + (\boldsymbol{\mu}\cdot\boldsymbol{\omega}^{bf})\boldsymbol{\mu}]. \tag{11.1.334}$$

We have achieved the desired goal for the case of $SO(3,\mathbb{R})$. Note that, since R is 3×3 real, the integration of (1.321), and its Cayley counterpart (1.313), involves 9 real differential equations. By contrast, the integration of (1.334) involves only 3 real differential equations. Although (1.313) and its integration preserves lie algebraic structure, it does not exploit this structure. By contrast, based on the introduction of a basis, (1.334) exploits Lie algebraic structure. And, of course, if local errors of order h^{m+1} arise in the numerical integration of (1.334), the resulting $R(\boldsymbol{\mu})$ is still guaranteed to be in $SO(3,\mathbb{R})$ because of the Ansatz (1.323).

The case of $SU(2)$ requires more calculations, but these calculations involve only results already known. Verify that inserting (1.327) and (1.328) into (1.313) yields the relation

$$\dot{\boldsymbol{\mu}}\cdot\boldsymbol{K} = (1/2)[\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K} + \{\boldsymbol{\mu}\cdot\boldsymbol{K},\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}\} - (\boldsymbol{\mu}\cdot\boldsymbol{K})(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})(\boldsymbol{\mu}\cdot\boldsymbol{K})]. \tag{11.1.335}$$

Now manipulate the terms in (1.335) using known results. Recall that

$$\{\boldsymbol{\mu}\cdot\boldsymbol{K},\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}\} = (\boldsymbol{\mu}\times\boldsymbol{\omega}^{bf})\cdot\boldsymbol{K}. \tag{11.1.336}$$

See (3.7.182). Next recall that

$$(\boldsymbol{\mu}\cdot\boldsymbol{K})(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K}) = -(1/4)(\boldsymbol{\mu}\cdot\boldsymbol{\omega}^{bf})I + (1/2)(\boldsymbol{\mu}\times\boldsymbol{\omega}^{bf})\cdot\boldsymbol{K}. \tag{11.1.337}$$

See (3.7.176). It follows that

$$(\boldsymbol{\mu}\cdot\boldsymbol{K})(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})(\boldsymbol{\mu}\cdot\boldsymbol{K}) = -(1/4)(\boldsymbol{\mu}\cdot\boldsymbol{\omega}^{bf})(\boldsymbol{\mu}\cdot\boldsymbol{K}) + (1/2)[(\boldsymbol{\mu}\times\boldsymbol{\omega}^{bf})\cdot\boldsymbol{K}](\boldsymbol{\mu}\cdot\boldsymbol{K}). \tag{11.1.338}$$

Verify that

$$\begin{aligned}[(\boldsymbol{\mu}\times\boldsymbol{\omega}^{bf})\cdot\boldsymbol{K}](\boldsymbol{\mu}\cdot\boldsymbol{K}) &= -(1/4)[(\boldsymbol{\mu}\times\boldsymbol{\omega}^{bf})\cdot\boldsymbol{\mu}]I + (1/2)[(\boldsymbol{\mu}\times\boldsymbol{\omega}^{bf})\times\boldsymbol{\mu}]\cdot\boldsymbol{K}\\ &= 0 - (1/2)[\boldsymbol{\mu}\times(\boldsymbol{\mu}\times\boldsymbol{\omega}^{bf})]\cdot\boldsymbol{K} = -(1/2)[(\boldsymbol{\mu}\cdot\boldsymbol{\omega}^{bf})\boldsymbol{\mu} - (\boldsymbol{\mu}\cdot\boldsymbol{\mu})\boldsymbol{\omega}^{bf}]\cdot\boldsymbol{K},\end{aligned} \tag{11.1.339}$$

and that, consequently,

$$\begin{aligned}&(\boldsymbol{\mu}\cdot\boldsymbol{K})(\boldsymbol{\omega}^{bf}\cdot\boldsymbol{K})(\boldsymbol{\mu}\cdot\boldsymbol{K})\\&=-(1/4)(\boldsymbol{\mu}\cdot\boldsymbol{\omega}^{bf})(\boldsymbol{\mu}\cdot\boldsymbol{K})-(1/4)[(\boldsymbol{\mu}\cdot\boldsymbol{\omega}^{bf})\boldsymbol{\mu}-(\boldsymbol{\mu}\cdot\boldsymbol{\mu})\boldsymbol{\omega}^{bf}]\cdot\boldsymbol{K}\\&=[-(1/2)(\boldsymbol{\mu}\cdot\boldsymbol{\omega}^{bf})\boldsymbol{\mu}+(1/4)(\boldsymbol{\mu}\cdot\boldsymbol{\mu})\boldsymbol{\omega}^{bf}]\cdot\boldsymbol{K}.\end{aligned}\tag{11.1.340}$$

Verify that combining (1.335) through (1.340) gives the result

$$\dot{\boldsymbol{\mu}}\cdot\boldsymbol{K}=(1/2)[\boldsymbol{\omega}^{bf}+(\boldsymbol{\mu}\times\boldsymbol{\omega}^{bf})+(1/2)(\boldsymbol{\mu}\cdot\boldsymbol{\omega}^{bf})\boldsymbol{\mu}-(1/4)(\boldsymbol{\mu}\cdot\boldsymbol{\mu})\boldsymbol{\omega}^{bf}]\cdot\boldsymbol{K}.\tag{11.1.341}$$

From this result it follows that there are the equations of motion

$$\dot{\boldsymbol{\mu}}=(1/2)[\boldsymbol{\omega}^{bf}+(\boldsymbol{\mu}\times\boldsymbol{\omega}^{bf})+(1/2)(\boldsymbol{\mu}\cdot\boldsymbol{\omega}^{bf})\boldsymbol{\mu}-(1/4)(\boldsymbol{\mu}\cdot\boldsymbol{\mu})\boldsymbol{\omega}^{bf}].\tag{11.1.342}$$

We have achieved our goal for the case of $SU(2)$. Note that, since u is 2×2 complex, the integration of (1.322), and its Cayley counterpart, involves 8 real differential equations. By contrast, the integration of (1.342) again involves only 3 real differential equations. And, again, if local errors of order h^{m+1} arise in the numerical integration of (1.342), the resulting $u(\boldsymbol{\mu})$ is still guaranteed to be in $SU(2)$ because of the Ansatz (1.324).

What can be said about the singularity structure of Cayley parameterization? Evidently (1.293) is singular when

$$\det(I-V)=0,\tag{11.1.343}$$

that is, when V has +1 as an eigenvalue. Also, (1.311) can be rewritten in the form

$$A=2(I+V)^{-1}\dot{V}(I-V)^{-1}\tag{11.1.344}$$

so that the relation between A and $\dot{V}$ is singular when

$$\det(I+V)=0\ \text{ and }\ \det(I-V)=0.\tag{11.1.345}$$

Strangely enough, the general equation of motion (1.313), and the specific $SO(3,\mathbb{R})$ and $SU(2)$ equations of motion (1.334) and (1.342), appear to be singularity free. This appearance is deceptive, because, for example, (1.334) and (1.342) *are* singular in $\boldsymbol{\mu}$ at infinity, and there is the possibility that this singularity can be encountered in *finite* time.

To realize this possible singularity in the case of $SO(3,\mathbb{R})$, suppose that

$$\omega_j^{bf}=\Omega\delta_{j3},\tag{11.1.346}$$

where Ω is a constant, and make the Ansatz

$$\mu_j=f(t)\delta_{j3}\tag{11.1.347}$$

where f satisfies the initial condition

$$f(0)=0,\tag{11.1.348}$$

but is otherwise to be determined. Show that putting this Ansatz into the equation of motion (1.334) yields the result

$$\dot{f} = (1/2)\Omega(1 + f^2). \tag{11.1.349}$$

Show that the solution to (1.349) with the initial condition (1.348) is

$$f = \tan(\Omega t/2). \tag{11.1.350}$$

Review Exercise 3.12.6. Show that the corresponding $\boldsymbol{\lambda}$ is given by

$$\lambda_j = \Omega t \delta_{j3} \tag{11.1.351}$$

and therefore R is given by

$$R = \exp(\Omega t L^3). \tag{11.1.352}$$

See (3.12.61). Observe that (1.350), and hence (1.347), are singular when $\Omega t = \pi$ and therefore when $|\boldsymbol{\lambda}| = \pi$ which, according to Exercise 3.12.6, is the expected condition for singularity in the case of $SO(3, \mathbb{R})$.

To realize this possible singularity in the case of $SU(2)$, suppose that (1.346) through (1.348) continue to hold. Show that putting this Ansatz into (1.342) yields the differential equation

$$\dot{f} = (\Omega/2)(1 + f^2/4), \tag{11.1.353}$$

and that this equation has the solution

$$f = 2\tan(\Omega t/4). \tag{11.1.354}$$

Show that the corresponding $\boldsymbol{\lambda}$ is again given by

$$\lambda_j = \Omega t \delta_{j3} \tag{11.1.355}$$

and therefore u is given by

$$u = \exp(\Omega t K^3). \tag{11.1.356}$$

See (3.12.73). Observe that (1.354), and hence (1.347), are singular when $\Omega t = 2\pi$ and therefore when $|\boldsymbol{\lambda}| = 2\pi$ which, according to Exercise 3.12.6, is the expected condition for singularity in the case of $SU(2)$.

At this point you, the observant reader, might object that the quaternion parameter equations of motion (1.43) through (1.46), which were lauded as being wonderful, are also singular in the components of w at infinity. They are indeed singular at infinity, but this singularity cannot be reached in *real* time because these equations preserve the condition $w \cdot w = 1$. This singularity can be reached in *complex* time, but because the equations of motion are *linear* in w, the time must be *infinite* complex. Therefore, for quaternion parameterization, there are no singularities in finite time, real or complex.

11.2 Numerical Integration on Manifolds: Spin and Qubits

As a second example of integration on manifolds, we consider an equation that occurs in several contexts. Let $\boldsymbol{s}(t)$ be a time-dependent 3-dimensional vector that evolves according to the rule

$$d\boldsymbol{s}/dt = \bar{\boldsymbol{\omega}}(t) \times \boldsymbol{s} \tag{11.2.1}$$

where $\bar{\boldsymbol{\omega}}(t)$ is some other specified, possibly time dependent, 3-dimensional vector. This equation is called the *Bloch* equation in the context of nuclear magnetic resonance (NMR or MRI) and electron spin resonance (ESR), and the *Bargmann-Michel-Telegdi* (BMT or *Thomas*-BMT or *Thomas-Frenkel*-BMT) equation in the context of determining the evolution of a particle's spin polarization vector as it traverses some accelerator or beam line. It also occurs in the context of rigid-body motion. See (1.105) in Exercise 1.1. Finally, it is relevant to the general quantum mechanical treatment of two level systems (qubits), and therefore plays a prominent role in quantum information theory and quantum computation. See Exercises 2.15 and 2.16. Note also that the equation of motion (1.6.112) can be written as

$$d\boldsymbol{v}/dt = [-(q/m^*)\boldsymbol{B}] \times \boldsymbol{v}, \tag{11.2.2}$$

which also appears to be of the form (2.1). See Section 3 for further discussion of this observation.

Suppose our task is to find $\boldsymbol{s}(t)$ given $\bar{\boldsymbol{\omega}}(t)$ and the initial vector

$$\boldsymbol{s}^0 = \boldsymbol{s}(t^0) \tag{11.2.3}$$

at time t^0.[17] It is easily verified that the equations of motion (2.1) preserve $\boldsymbol{s} \cdot \boldsymbol{s}$. Define a quantity s^* by the rule

$$s^* = \sqrt{\boldsymbol{s}^0 \cdot \boldsymbol{s}^0}, \tag{11.2.4}$$

and let S^{2*} denote the two-sphere of radius s^* imbedded in the ambient space E^3. With this notation, we may say that the equations of motion (2.1) preserve S^{2*}, and are equations of motion on the manifold S^{2*} embedded in the ambient space E^3. Note also that the equations of motion (2.1) are *linear*. Therefore, for many applications, there is no loss in generality in taking $\boldsymbol{s}^0$ to be a unit vector: $\boldsymbol{s}^0 \in S^2$ where, as usual, S^2 denotes the unit two-sphere ($s^* = 1$). Solutions corresponding to initial conditions that are not unit vectors can be obtained by simple scaling of the solutions corresponding to unit-vector initial conditions. Therefore, unless specifically specified otherwise, we will work with the S^2 case. However, where useful, we will treat explicitly the general S^{2*} case.

Our task is to integrate (2.1) numerically in such a way that, even if local errors of order h^{m+1} are made, the solution is guaranteed to be in S^2. One procedure for so doing is to employ any of the standard methods of Chapter 2 for one step at a time (thereby making local errors of order h^{m+1}) and to then project after each step the resulting $\boldsymbol{s}$ back onto S^2 by simple scaling. Alternatively, we may consider other approaches that parameterize S^2 or exploit other features of the problem. The purpose of this section is to describe several such approaches.

[17] For a discussion of the inverse problem, namely given $\boldsymbol{s}(t)$ find $\bar{\boldsymbol{\omega}}(t)$, see Exercise 2.1.

11.2.1 Constrained Cartesian Coordinates Are Not Global

In Cartesian coordinates (2.1) yields the three coupled linear equations

$$\dot{s}_1 = \bar{\omega}_2 s_3 - \bar{\omega}_3 s_2, \tag{11.2.5}$$

$$\dot{s}_2 = \bar{\omega}_3 s_1 - \bar{\omega}_1 s_3, \tag{11.2.6}$$

$$\dot{s}_3 = \bar{\omega}_1 s_2 - \bar{\omega}_2 s_1. \tag{11.2.7}$$

To recapitulate, if they are integrated numerically by a method that makes local errors of order h^{m+1}, the quantity $\boldsymbol{s} \cdot \boldsymbol{s}$ will generally be locally preserved only through terms of order h^m. If we wish to preserve the condition $\boldsymbol{s} \in S^2$ to machine precision, one simple procedure is to project $\boldsymbol{s} \in E^3$ back onto S^2 by simple scaling after each integration step.

Another procedure, assuming $\boldsymbol{s}^0 \in S^2$, is to enforce the condition $\boldsymbol{s} \in S^2$ by making the definition

$$s_1 = +(1 - s_2^2 - s_3^2)^{1/2} \tag{11.2.8}$$

and inserting this definition/constraint into the equations (2.6) and (2.7) to yield the equations of motion

$$\dot{s}_2 = \bar{\omega}_3 (1 - s_2^2 - s_3^2)^{1/2} - \bar{\omega}_1 s_3, \tag{11.2.9}$$

$$\dot{s}_3 = \bar{\omega}_1 s_2 - \bar{\omega}_2 (1 - s_2^2 - s_3^2)^{1/2}. \tag{11.2.10}$$

In essence, we have parameterized S^2 by the coordinates s_2 and s_3. If $\boldsymbol{s}$ is initially in the front hemisphere, $s_1 > 0$, these equations for s_2 and s_3 can be integrated as long as $(s_2^2 + s_3^2) < 1$. However, they become singular if $\boldsymbol{s}$ crosses the plane $s_1 = 0$ (which divides the front and rear hemispheres), as is certainly mathematically/physically possible, and they therefore cannot be generally used to produce a global solution.

11.2.2 Polar-Angle Coordinates Are Not Global

Yet another possibility is to parameterize $\boldsymbol{s} \in S^2$ by the use of polar-angle coordinates θ and ϕ. Make the Ansatz

$$s_1 = \sin(\theta)\cos(\phi), \tag{11.2.11}$$

$$s_2 = \sin(\theta)\sin(\phi), \tag{11.2.12}$$

$$s_3 = \cos(\theta). \tag{11.2.13}$$

This Ansatz guarantees $\boldsymbol{s} \in S^2$. From (2.11) through (2.13) we find the relations

$$\dot{s}_1 = \dot{\theta}\cos(\theta)\cos(\phi) - \dot{\phi}\sin(\theta)\sin(\phi), \tag{11.2.14}$$

$$\dot{s}_2 = \dot{\theta}\cos(\theta)\sin(\phi) + \dot{\phi}\sin(\theta)\cos(\phi), \tag{11.2.15}$$

$$\dot{s}_3 = -\dot{\theta}\sin(\theta). \tag{11.2.16}$$

Solving these relations for $\dot{\theta}$ and $\dot{\phi}$ yields the results

$$\dot{\theta} = -[1/\sin(\theta)]\dot{s}_3, \tag{11.2.17}$$

$$\dot{\phi} = -[1/\sin(\theta)][\dot{s}_1 \sin(\phi) - \dot{s}_2 \cos(\phi)]. \tag{11.2.18}$$

Finally, combining (2.5) through (2.7) and (2.11) through (2.13) with (2.17) and (2.18) yields the equations of motion

$$\dot{\theta} = -\bar{\omega}_1 \sin(\phi) + \bar{\omega}_2 \cos(\phi), \tag{11.2.19}$$

$$\dot{\phi} = \bar{\omega}_3 - [\cos(\theta)/\sin(\theta)][\bar{\omega}_1 \cos(\phi) + \bar{\omega}_2 \sin(\phi)]. \tag{11.2.20}$$

Observe that (2.20) is singular at the poles $\theta = 0$ and $\theta = \pi$. (Note that ϕ is ill defined at the poles, and consequently these singularities are to be expected). Therefore these equations are also not suitable for finding global solutions.

11.2.3 Local Tangent-Space Coordinates

One way to insure that a numerical trajectory will remain on an invariant manifold is to introduce local coordinates in the ambient space at some point on the manifold, locally parameterize the manifold, formulate differential equations for these parameters, and finally numerically integrate the differential equations for the parameters. By so doing, even if single-step errors of order h^{m+1} occur in the parameters over the process of integration, the resulting trajectory is guaranteed to remain on the manifold. We will illustrate this process for the equation of motion (2.1) and, for future use, we will explicitly treat the general case S^{2*}.

Let $\boldsymbol{s}^b$ be some point on the the manifold S^{2*} at the *beginning* of an integration step to be initiated at time t^b. Parameterize points in the ambient space and in the vicinity of $\boldsymbol{s}^b$ by writing

$$\boldsymbol{s}(t) = \boldsymbol{s}^b + \boldsymbol{s}^v(t) \tag{11.2.21}$$

where $\boldsymbol{s}^v(t)$ is a *variable* vector with the property

$$\boldsymbol{s}^v(t^b) = 0. \tag{11.2.22}$$

Next work to insure that $\boldsymbol{s}(t)$ remains on S^{2*} as t varies. Let $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, $\boldsymbol{e}_3$ be a fixed right-hand triad of orthonormal vectors. Construct a second right-hand triad of orthonormal vectors $\boldsymbol{f}_1$, $\boldsymbol{f}_2$, $\boldsymbol{f}_3$ associated with $\boldsymbol{s}^b$ as follows: Begin by defining

$$\boldsymbol{f}_1 = \boldsymbol{s}^b/s^*. \tag{11.2.23}$$

Next examine the quantities $\boldsymbol{e}_j \cdot \boldsymbol{f}_1$ and select the $\boldsymbol{e}_j$ for which $\boldsymbol{e}_j \cdot \boldsymbol{f}_1$ has the least magnitude. This will the $\boldsymbol{e}_j$ that is most nearly perpendicular to $\boldsymbol{s}^b$. Use this $\boldsymbol{e}_j$, call it $\boldsymbol{e}_k$, to define the unit vector

$$\boldsymbol{f}_2 = (\boldsymbol{e}_k \times \boldsymbol{f}_1)/|\boldsymbol{e}_k \times \boldsymbol{f}_1|. \tag{11.2.24}$$

By construction $\boldsymbol{f}_2$ is perpendicular to $\boldsymbol{f}_1$,

$$\boldsymbol{f}_1 \cdot \boldsymbol{f}_2 = 0. \tag{11.2.25}$$

Finally, define the unit vector $\boldsymbol{f}_3$ by the rule

$$\boldsymbol{f}_3 = \boldsymbol{f}_1 \times \boldsymbol{f}_2. \tag{11.2.26}$$

Continue on by expressing $\boldsymbol{s}^v(t)$ in the form

$$\boldsymbol{s}^v(t) = s_1^{vf}(t)\boldsymbol{f}_1 + s_2^{vf}(t)\boldsymbol{f}_2 + s_3^{vf}(t)\boldsymbol{f}_3. \tag{11.2.27}$$

Here the superscript f reminds us that the $\boldsymbol{f}_j$ have been employed as a basis. That is, the s_j^{vf} are the components of $\boldsymbol{s}^v$ with respect to the $\boldsymbol{f}_j$ basis. By construction, vectors of the form

$$\boldsymbol{s}^{tan} = \boldsymbol{s}^b + s_2^{vf}\boldsymbol{f}_2 + s_3^{vf}\boldsymbol{f}_3 \tag{11.2.28}$$

comprise the *tangent* space to S^{2*} at $\boldsymbol{s}^b$; and we may view s_2^{vf} and s_3^{vf} as being tangent-space coordinates. Combining (2.21) and (2.27) gives the result

$$\boldsymbol{s}(t) = [s^* + s_1^{vf}(t)]\boldsymbol{f}_1 + s_2^{vf}(t)\boldsymbol{f}_2 + s_3^{vf}(t)\boldsymbol{f}_3. \tag{11.2.29}$$

Now enforce the condition that $\boldsymbol{s}(t)$ lie in S^{2^*}. So doing gives the relation

$$[s^* + s_1^{vf}(t)]^2 + [s_2^{vf}(t)]^2 + [s_3^{vf}(t)]^2 = (s^*)^2, \tag{11.2.30}$$

from which it follows that

$$s_1^{vf}(t) = \{(s^*)^2 - [s_2^{vf}(t)]^2 - [s_3^{vf}(t)]^2\}^{1/2} - s^*. \tag{11.2.31}$$

We see that in the vicinity of $\boldsymbol{s}^b$ the manifold S^{2*} can be parameterized by s_2^{vf} and s_3^{vf} providing (2.31) is used to specify s_1^{vf}.

What remains is to find equations of motion for s_2^{vf} and s_3^{vf}. The first step is to expand $\bar{\boldsymbol{\omega}}(t)$ in terms of the $\boldsymbol{f}_j$ by writing

$$\bar{\boldsymbol{\omega}}(t) = \sum_j [\bar{\boldsymbol{\omega}}(t)\cdot\boldsymbol{f}_j]\boldsymbol{f}_j = \sum_j \bar{\omega}_j^f(t)\boldsymbol{f}_j \tag{11.2.32}$$

where we have made the definition

$$\bar{\omega}_j^f(t) = \bar{\boldsymbol{\omega}}(t)\cdot\boldsymbol{f}_j \tag{11.2.33}$$

and again the superscript f reminds us that the $\boldsymbol{f}_j$ has been employed as a basis. Use of (2.29) gives for the left side of (2.1) the result

$$d\boldsymbol{s}(t)/dt = \dot{s}_1^{vf}(t)\boldsymbol{f}_1 + \dot{s}_2^{vf}(t)\boldsymbol{f}_2 + \dot{s}_3^{vf}(t)\boldsymbol{f}_3. \tag{11.2.34}$$

Use of (2.29) and (2.32) gives for the right side side of (2.1) the result

$$\begin{aligned}\bar{\boldsymbol{\omega}}(t)\times\boldsymbol{s} &= [\bar{\omega}_2^f(t)s_3^{vf}(t) - \bar{\omega}_3^f(t)s_2^{vf}(t)]\boldsymbol{f}_1\\ &+ \{\bar{\omega}_3^f(t)[s^* + s_1^{vf}(t)] - \bar{\omega}_1^f(t)s_3^{vf}(t)\}\boldsymbol{f}_2\\ &+ \{\bar{\omega}_1^f(t)s_2^{vf}(t) - \bar{\omega}_2^f(t)[s^* + s_1^{vf}(t)]\}\boldsymbol{f}_3.\end{aligned} \tag{11.2.35}$$

Now equate the second and third components of (2.34) and (2.35) to find the relations

$$\dot{s}_2^{vf}(t) = \bar{\omega}_3^f(t)[s^* + s_1^{vf}(t)] - \bar{\omega}_1^f(t)s_3^{vf}(t), \tag{11.2.36}$$

$$\dot{s}_3^{vf}(t) = \bar{\omega}_1^f(t)s_2^{vf}(t) - \bar{\omega}_2^f(t)[s^* + s_1^{vf}(t)]. \tag{11.2.37}$$

Finally, employing (2.31) in (2.36) and (2.37) yields the equations of motion

$$\dot{s}_2^{vf}(t) = \bar{\omega}_3^f(t)\{(s^*)^2 - [s_2^{vf}(t)]^2 - [s_3^{vf}(t)]^2\}^{1/2} - \bar{\omega}_1^f(t)s_3^{vf}(t), \tag{11.2.38}$$

$$\dot{s}_3^{vf}(t) = \bar{\omega}_1^f(t)s_2^{vf}(t) - \bar{\omega}_2^f(t)\{(s^*)^2 - [s_2^{vf}(t)]^2 - [s_3^{vf}(t)]^2\}^{1/2}. \tag{11.2.39}$$

It is these equations that are to be numerically integrated from the time t^b to the time $t^b + h$ (or perhaps $t^b + kh$) starting with the initial conditions $s_2^{vf}(t^b) = s_3^{vf}(t^b) = 0$.[18] Then, once $\boldsymbol{s}^v(t^b + h)$ [or perhaps $\boldsymbol{s}^v(t^b + kh)$] has been obtained, $\boldsymbol{s}(t^b + h)$ [or perhaps $\boldsymbol{s}(t^b + kh)$] is given by (2.21).[19] At this point, the whole process just described is repeated as often as desired. That is, the vectors $\boldsymbol{f}_j$ are reconstructed based on the most recently obtained $\boldsymbol{s}$, etc.

We close this subsection by noting that, as was the case with constrained Cartesian coordinates and polar-angle coordinates, there are only *two* equations to be integrated, namely (2.38) and (2.39), whereas working in the ambient space E^3 as in (2.5) through (2.7) required the integration of *three* equations.

11.2.4 Exploiting Connection with Rigid-Body Kinematics

We next consider approaches related to those used in the rigid-body case. Suppose we seek the general solution of (2.1). Observe, with the aid of the matrices L^j, that (2.1) can be written in the form

$$d\boldsymbol{s}/dt = (\bar{\boldsymbol{\omega}} \cdot \boldsymbol{L})\boldsymbol{s}. \tag{11.2.40}$$

Recall (3.7.200). Also, since (2.1) is linear, we may make the general Ansatz

$$\boldsymbol{s}(t) = S(t)\boldsymbol{s}^0 \tag{11.2.41}$$

where S is a 3×3 matrix to be determined. Now insert (2.41) into (2.40) to find the relation

$$\dot{S}(t)\boldsymbol{s}^0 = (\bar{\boldsymbol{\omega}} \cdot \boldsymbol{L})S(t)\boldsymbol{s}^0. \tag{11.2.42}$$

Since we wish $\boldsymbol{s}^0$ to be an arbitrary unit vector and (2.42) is linear, the relation (2.42) is equivalent to the matrix differential equation

$$\dot{S}(t) = (\bar{\boldsymbol{\omega}} \cdot \boldsymbol{L})S(t), \tag{11.2.43}$$

and from (2.3) and (2.41) we find the initial condition

$$S(t^0) = I. \tag{11.2.44}$$

In summary, solving (2.43) with the initial condition (2.44) provides the general solution to (2.1). This approach has the advantage that once $S(t)$ has been found, $\boldsymbol{s}(t)$ can be found

[18] Observe that the equations of motion (2.38) and (2.39) agree with the equations of motion (2.9) and (2.10) in the case where the $\boldsymbol{f}_j$ agree with the $\boldsymbol{e}_j$ and $s^* = 1$, which is a nice check of our work.

[19] If $k > 1$ is attempted, one must monitor $[(s_2^v)^2 + (s_3^v)^2]$ to ensure that the square root singularity in (2.31) is not approached too closely.

for *all* initial conditions $\boldsymbol{s}^0$ by the easy computation (2.41). In essence, $S(t)$ is the transfer map associated with the differential equation (2.1). By contrast, if (2.5) through (2.7), or (2.9) and (2.10), or (2.19) and (2.20), or (2.38) and (2.39) are employed, these differential equations must be integrated afresh for each different initial $\boldsymbol{s}^0$.

At this point we observe that (2.43) and (1.18) are quite similar. Indeed, suppose we pass back and forth between matrices R and S by the rule

$$R = S^{-1} \text{ or } S = R^{-1}, \tag{11.2.45}$$

from which it follows that

$$R(t^0) = I. \tag{11.2.46}$$

Then it can be shown from (2.43) that there is the relation

$$\dot{R} = R(\boldsymbol{\omega}^{bf} \cdot \boldsymbol{L}) \tag{11.2.47}$$

with

$$\boldsymbol{\omega}^{bf} = -\bar{\boldsymbol{\omega}}. \tag{11.2.48}$$

See Exercise 2.7.

We know that R is orthogonal and therefore, from (2.45), we conclude that S is also orthogonal. We also observe that in the orthogonal case the relation (2.45) is equivalent to the computationally simpler relation

$$R = S^T \text{ or } S = R^T. \tag{11.2.49}$$

We see that all the machinery developed for and the conclusions drawn about rigid body motion in Section 1 are also applicable here.

11.2.5 What Just Happened? Generalizations

In the last subsection we saw that the problem of determining the path $\boldsymbol{s}(t)$ in the manifold S^2 and satisfying the *manifold* differential equation (2.1) [or, equivalently (2.40)] with the initial condition (2.3) was converted into finding a path $S(t)$ in the group $SO(3,\mathbb{R})$ that satisfied the *group* differential equation (2.43) with the initial condition (2.44). We also observe that the group $SO(3,\mathbb{R})$ acts *transitively* on the manifold S^2. (Evidently any point in S^2 can be rotated into any other point in S^2.) Therefore S^2 is a homogeneous space with respect to the group $SO(3,\mathbb{R})$, and is in fact isomorphic to the coset space $SO(3,\mathbb{R})/SO(2,\mathbb{R})$. Recall the discussion of homogeneous spaces in Subsections 5.12.3 through 5.12.5.

Observe moreover that (2.41) is a relation that sends the path $S(t)$ in the group $SO(3,\mathbb{R})$ to the path $\boldsymbol{s}(t)$ in the manifold S^2. The path $S(t)$ begins at the identity, see (2.44), and the path $\boldsymbol{s}(t)$ begins at $\boldsymbol{s}^0$. In the language of manifold theory, the relation (2.41) is said to *push forward* the path $S(t) \in SO(3,\mathbb{R})$ to produce the path $\boldsymbol{s}(t) \in S^2$.[20] Correspondingly, the group differential equation (2.43) is the *pullback* to $SO(3,\mathbb{R})$ of the S^2 manifold differential equation (2.1). Finally, since $SO(3,\mathbb{R})$ acts on and *preserves* S^2, we are guaranteed that $\boldsymbol{s}(t)$ will be in S^2 if $S(t)$ is in $SO(3,\mathbb{R})$. Therefore a numerical integrator that preserves

[20] Put another way, $\boldsymbol{s}(t)$ is the orbit of $\boldsymbol{s}^0$ under the action of the $SO(3,\mathbb{R})$ group elements $S(t)$.

the $SO(3,\mathbb{R})$ group manifold for any group differential equation automatically produces a pushed-forward path that is guaranteed to lie in the S^2 manifold.

It should now be evident that this strategy has some general applicability. Suppose we are given a manifold, call it Γ, perhaps embedded in some larger ambient space, and a first-order differential equation for a path, call it $\gamma(t)$, that from the differential equation and the initial condition γ^0 can be shown to lie in Γ. Suppose we can find a group, call it $\mathcal{G}$, such that $\mathcal{G}$ acts transitively on Γ so that Γ is a homogeneous space with respect to $\mathcal{G}$. Introduce the notation $G(t)$ to denote a path in $\mathcal{G}$ with the beginning point

$$G(t^0) = I. \tag{11.2.50}$$

Then this path in $\mathcal{G}$ may be pushed forward to produce a path $\gamma(t)$ in Γ. In a notation analogous to that of Subsection 5.12.4, we write

$$\gamma(t) = T_{G(t)}(\gamma^0) \tag{11.2.51}$$

where T_G describes the action of $\mathcal{G}$ on Γ. With some suitable parameterization of $\mathcal{G}$, perhaps involving an embedding in an ambient space of its own, the differential equation for $\gamma(t)$ can be pulled back to produce a group differential equation for $G(t)$. And if this group differential equation can be integrated numerically in such a way that the group manifold is preserved, say either by parameterizing $\mathcal{G}$ or its Lie algebra or by integrating in its Lie algebra, then the $\gamma(t)$ given by (2.51) will be (locally) accurate through terms of order h^m and is guaranteed to lie in Γ. Thus, whatever means can be found to integrate group differential equations numerically in such a way that the group manifold is preserved, by the same means one has found a procedure to numerically integrate in a manifold preserving way all differential equations defined on the homogeneous spaces associated with $\mathcal{G}$.

11.2.6 Exploiting an Important Simplification: Lie Taylor Factorization and Lie Taylor Runge Kutta

We observe that there is one way in which the context for (2.43) is simpler than that for (1.18). Namely, in (2.43) $\bar{\boldsymbol{\omega}}$ is assumed to be a *given* function of t *independent* of S whereas in (1.18) $\boldsymbol{\omega}^{bf}$ must be determined dynamically from the Euler equations (1.19) through (1.21) which themselves may depend on R. This simplification can be used to good advantage in integrating (2.43) numerically. Of course, since $S(t)$ is orthogonal when computed exactly, we will want a numerical integrator that guarantees this property for the numerical solution. In this subsection we will see how this simplification can be exploited to perform what we call *Lie Taylor factorization*, and in so doing we will produce in effect a special kind of Runge Kutta that we will call *Lie Taylor Runge Kutta*.

Subsequently, in the next two subsections, we will explore how this requirement that $\bar{\boldsymbol{\omega}}(t)$ be known in advance can be relaxed in the context of two other special forms of Runge Kutta that we will call *factored Lie Runge Kutta* and *Magnus Lie Runge Kutta*. A final subsection revisits the use of integration in the Lie algebra.

For the purposes of this subsection it is convenient to capitalize on the homomorphism between $SO(3,\mathbb{R})$ and $SU(2)$. Let u be the 2×2 matrix that satisfies the differential equation

$$\dot{u}(t) = (\bar{\boldsymbol{\omega}} \cdot \boldsymbol{K})u(t) \tag{11.2.52}$$

with the initial condition

$$u(t^0) = I. \tag{11.2.53}$$

By the results of Section 1.3, $u(t)$ is uniquely defined. Also, it can be shown that $u(t)$ is in $SU(2)$ for all t. See Exercise 2.4. Next, use the homomorphism between $SO(3,\mathbb{R})$ and $SU(2)$ to define $S(t)$ by the rule

$$S_{\alpha\beta}(t) = S_{\alpha\beta}[u(t)] = (1/2)\mathrm{tr}[u^\dagger(t)\sigma^\alpha u(t)\sigma^\beta]. \tag{11.2.54}$$

See (8.2.54). Then it can be shown that $S(t)$ satisfies (2.43) with the initial condition (2.44). See Exercise 2.5.[21]

Since u is a 2×2 complex matrix, it lives in the ambient space $\mathbb{C}^4 = E^8$. However, we know that u is also in $SU(2)$ for all t. Therefore (2.52) is an equation of motion on the three-dimensional manifold $SU(2)$ imbedded in the ambient space E^8. Based on the assumption that $\bar{\boldsymbol{\omega}}$ is a given function of t, what we seek is a method, beyond those already described, for numerically integrating (2.52) in such a way that u is guaranteed to remain in $SU(2)$.[22] Such a method is presented below.

In the language of Section 2.6, let H be the duration of a meso integration step, and suppose H is divided into M micro steps each of duration $h = H/M$. Let t^b be the time at which a meso step is to be initiated so that we wish to integrate from t^b to $t^b + H$. We also suppose that

$$u^b = u(t^b) \tag{11.2.55}$$

is known and is an element of $SU(2)$.

Introduce a relative time τ by the rule

$$t = t^b + \tau \tag{11.2.56}$$

so that, in terms of τ, we wish to integrate from $\tau = \tau^0 = 0$ to $\tau = \tau^M = H$. Also, define a quantity $\hat{\boldsymbol{\omega}}(\tau)$ by the rule

$$\hat{\boldsymbol{\omega}}(\tau) = \bar{\boldsymbol{\omega}}(t^b + \tau). \tag{11.2.57}$$

In the spirit of (1.52), write

$$u(t) = u^v(t)u^b \tag{11.2.58}$$

with u^v being a variable matrix near the identity satisfying

$$u^v(t^b) = I. \tag{11.2.59}$$

[21] Why *can* we move from $SO(3,\mathbb{R})$ matrices to $SU(2)$ matrices? It can be shown that the solution to any matrix differential equation of the forms $dM/dt = AM$ or $dM/dt = MA$ is governed by the Lie algebra generated by the $A(t)$ at different times. See the paper by *E. Wichmann* cited in the Bibliography for Chapter 10. See also Section 10.3 and Appendix C. We may therefore use any set of matrices that obey the Lie algebra in question. In our case this Lie algebra is $so(3,\mathbb{R})$ or, equivalently, $su(2)$. Why *should* we move from $SO(3,\mathbb{R})$ to $SU(2)$? It is computationally advantageous to use the matrices that have the lowest dimension. In this case, the matrices with lowest dimension that satisfy $su(2)$ commutation rules are the K^j.

[22] By methods "already described" we mean that u, like R, can be parameterized in terms of Euler or Tait-Bryan angles or angle-axis or quaternion or Cayley parameters, or can be integrated in its Lie algebra; and so doing produces equations of motion analogous to those for rigid-body motion.

Finally, define a variable matrix $\hat{u}^v(\tau)$ by the rule

$$\hat{u}^v(\tau) = u^v(t^b + \tau). \tag{11.2.60}$$

It then follows from (2.52), (2.53), and (2.55) through (2.60) that $\hat{u}^v(\tau)$ obeys the equation of motion

$$d\hat{u}^v(\tau)/d\tau = [\hat{\boldsymbol{\omega}}(\tau) \cdot \boldsymbol{K}]\hat{u}^v(\tau) \tag{11.2.61}$$

with the initial condition

$$\hat{u}^v(0) = I. \tag{11.2.62}$$

Suppose the $M+1$ vectors $\hat{\boldsymbol{\omega}}^n$ are given with

$$\hat{\boldsymbol{\omega}}^n = \hat{\boldsymbol{\omega}}(\tau^{(n)}) \tag{11.2.63}$$

and

$$\tau^{(n)} = nh \;\text{ for }\; n = 0, 1, \cdots M. \tag{11.2.64}$$

We want to use this information to find a numerical approximation to $\hat{u}^v(H)$ that is both accurate and exactly in $SU(2)$. How to proceed? First use the $M+1$ vectors $\hat{\boldsymbol{\omega}}^n$ to produce a polynomial *fit* to $\hat{\boldsymbol{\omega}}(\tau)$ of the form

$$\hat{\boldsymbol{\omega}}^{\text{fit}}(\tau) = \sum_{m=0}^{M} \boldsymbol{c}_m \tau^m. \tag{11.2.65}$$

Here we have introduced the notation $\tau^{(n)}$ to denote the n^{th} sampling point and the notation τ^m to denote the m^{th} power of τ. We also remark that some sampling procedure other than equal spacing could be used to obtain the expansion (2.65). All we need for present purposes are the expansion coefficients $\boldsymbol{c}_m$ in (2.65), and need not specify how they are to be obtained.

Next, assume that $\hat{u}^v(\tau)$ has a *factorized Taylor* approximation of the form

$$\hat{u}^v(\tau) \simeq \hat{u}^{v\text{fac}}(\tau) = \exp(\tau^{M+1}\boldsymbol{d}_M \cdot \boldsymbol{K})\exp(\tau^M \boldsymbol{d}_{M-1} \cdot \boldsymbol{K}) \cdots \exp(\tau \boldsymbol{d}_0 \cdot \boldsymbol{K}). \tag{11.2.66}$$

Note that this Ansatz satisfies the relation

$$\hat{u}^{v\text{fac}}(0) = I, \tag{11.2.67}$$

as is desirable in view of (2.62).[23] Finally, insert (2.65) and (2.66) into (2.61) to yield the approximate relation

$$d\hat{u}^{v\text{fac}}(\tau)/d\tau = [\hat{\boldsymbol{\omega}}^{\text{fit}}(\tau) \cdot \boldsymbol{K}]\hat{u}^{v\text{fac}}(\tau), \tag{11.2.68}$$

which can also be written in the form

$$[d\hat{u}^{v\text{fac}}(\tau)/d\tau][\hat{u}^{v\text{fac}}(\tau)]^{-1} = \hat{\boldsymbol{\omega}}^{\text{fit}}(\tau) \cdot \boldsymbol{K}. \tag{11.2.69}$$

[23] The justification for the Ansatz (2.66) is as follows: Under the assumption that $\hat{\boldsymbol{\omega}}(\tau)$ is analytic in τ, the solution $\hat{u}^v(\tau)$ to (2.61) will be analytic in τ. See Section 1.3. It follows that the logarithm of $\hat{u}^v(\tau)$, the $su(2)$ element corresponding to $\hat{u}^v(\tau)$, can be expanded as a power series in τ assuming $\hat{u}^v(\tau)$ is near the origin, which it is for small τ. See Subsection 3.7.1. Finally, we may pass from a power series in the exponent to a product of exponentials with the aid of the BCH series, thereby yielding the factorization (2.66).

The strategy now is to equate powers of τ on both sides of (2.69) to determine the vectors $\boldsymbol{d}_n$ in terms of the vectors $\boldsymbol{c}_m$.

As an example, let us see how this procedure plays out for the case $M = 3$. Then there are the results

$$\hat{\boldsymbol{\omega}}^{\text{fit}}(\tau) = \boldsymbol{c}_0 + \boldsymbol{c}_1\tau + \boldsymbol{c}_2\tau^2 + \boldsymbol{c}_3\tau^3, \tag{11.2.70}$$

$$\hat{u}^{v\text{fac}}(\tau) = \exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(\tau\boldsymbol{d}_0\cdot\boldsymbol{K}), \tag{11.2.71}$$

$$[\hat{u}^{v\text{fac}}(\tau)]^{-1} = \exp(-\tau\boldsymbol{d}_0\cdot\boldsymbol{K})\exp(-\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(-\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(-\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K}), \tag{11.2.72}$$

$$\begin{aligned}
d\hat{u}^{v\text{fac}}(\tau)/d\tau \;=\;& \exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})(\boldsymbol{d}_0\cdot\boldsymbol{K})\exp(\tau\boldsymbol{d}_0\cdot\boldsymbol{K})\\
+\;& \exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})(2\tau\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(\tau\boldsymbol{d}_0\cdot\boldsymbol{K})\\
+\;& \exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})(3\tau^2\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(\tau\boldsymbol{d}_0\cdot\boldsymbol{K}).\\
+\;& (4\tau^3\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(\tau\boldsymbol{d}_0\cdot\boldsymbol{K}).
\end{aligned} \tag{11.2.73}$$

Next, in view of (2.69), combine (2.72) and (2.73) to yield the result

$$\begin{aligned}
&[d\hat{u}^{v\text{fac}}(\tau)/d\tau][\hat{u}^{v\text{fac}}(\tau)]^{-1} =\\
&\exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})(\boldsymbol{d}_0\cdot\boldsymbol{K})\exp(\tau\boldsymbol{d}_0\cdot\boldsymbol{K})\times\\
&\exp(-\tau\boldsymbol{d}_0\cdot\boldsymbol{K})\exp(-\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(-\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(-\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\\
&+\exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})(2\tau\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(\tau\boldsymbol{d}_0\cdot\boldsymbol{K})\times\\
&\exp(-\tau\boldsymbol{d}_0\cdot\boldsymbol{K})\exp(-\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(-\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(-\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\\
&+\exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})(3\tau^2\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(\tau\boldsymbol{d}_0\cdot\boldsymbol{K})\times\\
&\exp(-\tau\boldsymbol{d}_0\cdot\boldsymbol{K})\exp(-\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(-\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(-\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\\
&+(4\tau^3\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(\tau\boldsymbol{d}_0\cdot\boldsymbol{K})\times\\
&\exp(-\tau\boldsymbol{d}_0\cdot\boldsymbol{K})\exp(-\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(-\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(-\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K}).
\end{aligned} \tag{11.2.74}$$

After cancellations of various factors in (2.74) against their inverses, (2.74) simplifies to become

$$\begin{aligned}
&[d\hat{u}^{v\text{fac}}(\tau)/d\tau][\hat{u}^{v\text{fac}}(\tau)]^{-1} =\\
&\exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})(\boldsymbol{d}_0\cdot\boldsymbol{K})\times\\
&\exp(-\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(-\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(-\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\\
&+\exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})(2\tau\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(-\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(-\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\\
&+\exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})(3\tau^2\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(-\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\\
&+(4\tau^3\boldsymbol{d}_3\cdot\boldsymbol{K}).
\end{aligned} \tag{11.2.75}$$

Now expand the various terms on the right side of (2.75) as power series in τ through terms of order τ^3. The first term becomes

$$\begin{aligned}
&\exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})(\boldsymbol{d}_0\cdot\boldsymbol{K})\times\\
&\exp(-\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(-\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(-\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K}) =\\
&\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})(\boldsymbol{d}_0\cdot\boldsymbol{K})\exp(-\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(-\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K}) + O(\tau^4).
\end{aligned} \tag{11.2.76}$$

Observe that

$$\begin{aligned}\exp(\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})(\boldsymbol{d}_0\cdot\boldsymbol{K})\exp(-\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K}) &= \boldsymbol{d}_0\cdot\boldsymbol{K}+\tau^2\{\boldsymbol{d}_1\cdot\boldsymbol{K}),\boldsymbol{d}_0\cdot\boldsymbol{K}\}+O(\tau^4)\\ &= \boldsymbol{d}_0\cdot\boldsymbol{K}+\tau^2(\boldsymbol{d}_1\times\boldsymbol{d}_0)\cdot\boldsymbol{K}+O(\tau^4).\end{aligned}\tag{11.2.77}$$

See (3.7.182) and (8.2.5). It follows that

$$\begin{aligned}&\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})(\boldsymbol{d}_0\cdot\boldsymbol{K})\exp(-\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(-\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})=\\ &\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})(\boldsymbol{d}_0\cdot\boldsymbol{K})\exp(-\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})+\tau^2(\boldsymbol{d}_1\times\boldsymbol{d}_0)\cdot\boldsymbol{K}+O(\tau^4)=\\ &\boldsymbol{d}_0\cdot\boldsymbol{K}+\tau^3\{\boldsymbol{d}_2\cdot\boldsymbol{K}),\boldsymbol{d}_0\cdot\boldsymbol{K}\}+\tau^2(\boldsymbol{d}_1\times\boldsymbol{d}_0)\cdot\boldsymbol{K}+O(\tau^4)=\\ &\boldsymbol{d}_0\cdot\boldsymbol{K}+\tau^3(\boldsymbol{d}_2\times\boldsymbol{d}_0)\cdot\boldsymbol{K}+\tau^2(\boldsymbol{d}_1\times\boldsymbol{d}_0)\cdot\boldsymbol{K}+O(\tau^4).\end{aligned}\tag{11.2.78}$$

The net result is that the first term has the expansion

$$\begin{aligned}&\exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})(\boldsymbol{d}_0\cdot\boldsymbol{K})\times\\ &\exp(-\tau^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(-\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(-\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})=\\ &\boldsymbol{d}_0\cdot\boldsymbol{K}+\tau^2(\boldsymbol{d}_1\times\boldsymbol{d}_0)\cdot\boldsymbol{K}+\tau^3(\boldsymbol{d}_2\times\boldsymbol{d}_0)\cdot\boldsymbol{K}+O(\tau^4).\end{aligned}\tag{11.2.79}$$

What remains is to expand the second and third terms. The second term has the expansion

$$\begin{aligned}&\exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})(2\tau\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(-\tau^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(-\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})=\\ &2\tau\boldsymbol{d}_1\cdot\boldsymbol{K}+O(\tau^4).\end{aligned}\tag{11.2.80}$$

The third term has the expansion

$$\exp(\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})(3\tau^2\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(-\tau^4\boldsymbol{d}_3\cdot\boldsymbol{K})=3\tau^2\boldsymbol{d}_2\cdot\boldsymbol{K}+O(\tau^4).\tag{11.2.81}$$

The fourth term, $4\tau^3\boldsymbol{d}_3\cdot\boldsymbol{K}$, is already as simple as possible.

Now gather all the terms together. The result, as a power series in τ, is that (2.75) becomes

$$\begin{aligned}&[d\hat{u}^{v\mathrm{fac}}(\tau)/d\tau][\hat{u}^{v\mathrm{fac}}(\tau)]^{-1}=\\ &\boldsymbol{d}_0\cdot\boldsymbol{K}+2\tau\boldsymbol{d}_1\cdot\boldsymbol{K}+\tau^2[3\boldsymbol{d}_2+(\boldsymbol{d}_1\times\boldsymbol{d}_0)]\cdot\boldsymbol{K}+\tau^3[4\boldsymbol{d}_3+(\boldsymbol{d}_2\times\boldsymbol{d}_0)]\cdot\boldsymbol{K}+O(\tau^4).\end{aligned}\tag{11.2.82}$$

We are ready to equate powers of τ on both sides of (2.69). So doing yields the relations

$$\boldsymbol{d}_0=\boldsymbol{c}_0,\tag{11.2.83}$$

$$2\boldsymbol{d}_1=\boldsymbol{c}_1,\tag{11.2.84}$$

$$3\boldsymbol{d}_2 + (\boldsymbol{d}_1 \times \boldsymbol{d}_0) = \boldsymbol{c}_2, \tag{11.2.85}$$

$$4\boldsymbol{d}_3 + (\boldsymbol{d}_2 \times \boldsymbol{d}_0) = \boldsymbol{c}_3; \tag{11.2.86}$$

and these relations have the solution

$$\boldsymbol{d}_0 = \boldsymbol{c}_0, \tag{11.2.87}$$

$$\boldsymbol{d}_1 = \boldsymbol{c}_1/2, \tag{11.2.88}$$

$$\boldsymbol{d}_2 = (1/3)\boldsymbol{c}_2 + (1/6)(\boldsymbol{c_0} \times \boldsymbol{c}_1), \tag{11.2.89}$$

$$\boldsymbol{d}_3 = (1/4)\boldsymbol{c}_3 + (1/12)(\boldsymbol{c_0} \times \boldsymbol{c}_2) + (1/24)[\boldsymbol{c_0} \times (\boldsymbol{c_0} \times \boldsymbol{c}_1)]. \tag{11.2.90}$$

In the case $M = 3$ we have found the approximation

$$\hat{u}^v(H) \simeq \hat{u}^{v\text{fac}}(H) = \exp(H^4\boldsymbol{d}_3 \cdot \boldsymbol{K})\exp(H^3\boldsymbol{d}_2 \cdot \boldsymbol{K})\exp(H^2\boldsymbol{d}_1 \cdot \boldsymbol{K})\exp(H\boldsymbol{d}_0 \cdot \boldsymbol{K}) \tag{11.2.91}$$

with the coefficients $\boldsymbol{d}_0$ through $\boldsymbol{d}_3$ given by (2.87) through (2.90). And in general we have the result

$$\hat{u}^v(H) \simeq \hat{u}^{v\text{fac}}(H) = \cdots \exp(H^{n+1}\boldsymbol{d}_n \cdot \boldsymbol{K})\exp(H^n\boldsymbol{d}_{n-1} \cdot \boldsymbol{K}) \cdots \exp(H\boldsymbol{d}_0 \cdot \boldsymbol{K}). \tag{11.2.92}$$

What can be said about the error in this approximation? We begin by observing that any given $\boldsymbol{d}_n$ depends only on the $\boldsymbol{c}_m$ with $m \leq n$, and is independent of M as long as $M \geq n$. Also if we use $(M + 1)$ sampling points to find $(M + 1)$ values of $\hat{\boldsymbol{\omega}}$, and use these values to compute $\boldsymbol{c}_0$ through $\boldsymbol{c}_M$, then we can find $\boldsymbol{d}_0$ through $\boldsymbol{d}_M$ using relations of the kind (2.87) through (2.90). With this information we can compute $\hat{u}^v(H)$ given by

$$\hat{u}^v(H) \simeq \hat{u}^{v\text{fac}}(H) = \exp(H^{M+1}\boldsymbol{d}_M \cdot \boldsymbol{K})\exp(H^M\boldsymbol{d}_{M-1} \cdot \boldsymbol{K}) \cdots \exp(H\boldsymbol{d}_0 \cdot \boldsymbol{K}), \tag{11.2.93}$$

which is locally accurate through terms of order H^{M+1}, and exactly in $SU(2)$. In effect, we have produced a special kind of Runge Kutta that we will call *Lie Taylor Runge Kutta*. Indeed, in the terminology of Runge-Kutta integration, we may think of $(M + 1)$ as being the number of stages. Thus, the local accuracy of $\hat{u}^v(H)$ equals the number of stages. Comparison of this result with the entries of Table 2.3.1 shows that this performance of Lie Taylor Runge Kutta equals or exceeds that of ordinary explicit Runge-Kutta; and reference to (T.166) and (T.169) shows that Lie Taylor Runge Kutta has one order lower accuracy than Newton Cotes for odd values of $M + 1$, and equal accuracy for even values of $M + 1$.[24]

There is one other observation that is worth consideration. In determining the $\boldsymbol{c}_0$ through $\boldsymbol{c}_M$ there is no need for the sampling points to lie within the interval of meso-step integration as was done in (2.64). Suppose, for example, that we wish to integrate from the time $t = t^0$ to the time $t = t^0 + T$ using N meso steps each of duration

$$H = T/N. \tag{11.2.94}$$

[24] Note that here we are demanding more than the integration of an equation of the form (T.157) since in this instance the right side of (2.52) depends on u as well as t.

Also suppose that over the *full* interval $t \in [t^0, t^0 + T]$ the vector function $\bar{\boldsymbol{\omega}}(t)$ can be well fit by a polynomial of degree M in t with vector coefficients. Under this assumption, the coefficients of this polynomial can be obtained by evaluating $\bar{\boldsymbol{\omega}}(t)$ at $(M+1)$ sampling points. Use this global polynomial to form the local expansion (2.65) for each meso-step integration, and use a relation of the form (2.93) to find the result for each meso-step integration. Then the local error for each meso step is of order H^{M+2}, and the global error in integrating from the time $t = t^0$ to the time $t = t^0 + T$ is given by

$$\text{global error} \approx NH^{M+2} = N(T/N)^{M+2} = T(T/N)^{M+1}. \tag{11.2.95}$$

We see that the global *integration* error can be made arbitrarily small by increasing N (and, correspondingly, decreasing H) *without* changing the number of sampling points $(M + 1)$. Put another way, for sufficiently large N, the full global error is *only* the error associated with making the global polynomial fit to $\bar{\boldsymbol{\omega}}(t)$ in the full interval $t \in [t^0, t^0 + T]$. The goodness of this polynomial fit in turn depends only on the analytic properties of $\bar{\boldsymbol{\omega}}(t)$. In particular, a good fit is easiest to achieve when $\bar{\boldsymbol{\omega}}(t)$ does not vary too rapidly over the interval $[t^0, t^0 + T]$.[25] Finally, we may truncate (2.93) at M' with $M' < M$ and still achieve convergence, but at the slower rate of

$$\text{global error} \approx NH^{M'+2} = N(T/N)^{M'+2} = T(T/N)^{M'+1}. \tag{11.2.96}$$

In this case we only need to work out formulas for the $\boldsymbol{d}_n$ with $n \leq M'$

Where applicable, employing the ideas just described should result in a substantial savings in computer time because only $(M + 1)$ evaluations of $\bar{\boldsymbol{\omega}}(t)$ are required for the full integration run.

11.2.7 Factored Lie Runge Kutta

Purpose, Motivation, and Plan

The purpose of this subsection is to describe a special form of Runge Kutta designed to preserve group properties. We will see that for integrating equations of the form (2.43), unlike Lie Taylor Runge Kutta, it does not require the values of $\bar{\boldsymbol{\omega}}(t)$ in advance.

By way of motivation, suppose we seek to integrate (2.43) by the simplest Runge-Kutta method, namely the crude Euler method of Section 2.2. Doing so yields the stepping formula

$$S^{n+1} = S^n + h\dot{S}^n = S^n + h[\bar{\boldsymbol{\omega}}(t^n) \cdot \boldsymbol{L}]S^n = \{I + h[\bar{\boldsymbol{\omega}}(t^n) \cdot \boldsymbol{L}]\}S^n. \tag{11.2.97}$$

Observe that $\{I + h[\bar{\boldsymbol{\omega}}(t^n) \cdot \boldsymbol{L}]\}$ is generally not an orthogonal matrix. Consequently, S^{n+1} will generally not be orthogonal even if S^n is. Consider, instead, the modified stepping formula

$$S^{n+1} = \exp[h\bar{\boldsymbol{\omega}}(t^n) \cdot \boldsymbol{L}]S^n. \tag{11.2.98}$$

As can be seen by expanding $\exp[h\bar{\boldsymbol{\omega}}(t^n) \cdot \boldsymbol{L}]$ in powers of h, (2.98) and (2.97) agree through terms of order h. Therefore, use of (2.98) provides an integration algorithm that is of the

[25] More precisely, $\bar{\boldsymbol{\omega}}(t)$ needs to be analytic in the complex t plane in a disk of radius $T/2$ and centered on $t = t^0 + T/2$.

same order as the Euler method (2.97). However the algorithm (2.98), even though (like Euler) it makes local errors of order h^2, preserves $SO(3,\mathbb{R})$ exactly because $\exp[h\bar{\boldsymbol{\omega}}(t^n)\cdot\boldsymbol{L}]$ is orthogonal.

The relation (2.98) provides an example of what we call a *factored Lie Runge-Kutta* algorithm designed to preserve some group, in this case $SO(3,\mathbb{R})$.[26] We now discuss the possibility of finding such algorithms that are of order h^m with $m > 1$.

The equation (2.43) is a special case of a more general equation, and some results are known about factored Lie Runge Kutta for this more general equation. Our plan is to discuss this more general equation, and then apply the known results for the more general equation to the special case (2.43).

Factored Lie Runge Kutta

Let G be some Lie group of $n\times n$ matrices, and let Y denote matrices in G. Next assume that there is some $n\times n$ matrix function $A(Y,t)$ such that $A(Y,t)$ is in the Lie algebra of G for all $Y\in G$ and all t. Let t^0 be some initial time and let Y^0 be some initial matrix in G. Consider the matrix differential equation

$$\dot{Y}(t) = A(Y,t)Y(t). \tag{11.2.99}$$

Then it can be shown that the solution to (2.99) lies in G for all time.[27] See Exercise 2.8. Comparison of (2.43) and (2.99) shows that (2.43) is a special case of (2.99) with S playing the role of Y and $[\bar{\boldsymbol{\omega}}(t)\cdot\boldsymbol{L}]$ playing the role of $A(Y,t)$:

$$S \leftrightarrow Y, \tag{11.2.100}$$

$$[\bar{\boldsymbol{\omega}}(t)\cdot\boldsymbol{L}] \leftrightarrow A(Y,t). \tag{11.2.101}$$

Note that for the special case (2.43) the matrices $A(Y,t)$ are in fact *independent* of Y.

Crouch, *Grossman*, and others have developed factored Lie Runge-Kutta algorithms for the numerical integration of (2.99). These algorithms are constructed in such a way that, although they may make local errors of order h^{m+1}, $Y(t)$ is guaranteed to lie in G to machine precision and evaluations of $A(Y,t)$ are required only for matrices Y in G.

Applying crude Euler to (2.99) produces the stepping rule

$$Y^{n+1} = Y^n + h\dot{Y}^n = [I + hA(Y^n,t^n)]Y^n \tag{11.2.102}$$

which, to the same order in h, can be rewritten in the exponential form

$$Y^{n+1} = \{\exp[hA(Y^n,t^n)]\}Y^n. \tag{11.2.103}$$

Suppose $Y^n\in G$. Then, by assumption, $A(Y^n,t^n)$ is in the Lie algebra of G, from which it follows that $\{\exp[hA(Y^n,t^n)]\}\in G$, and therefore $Y^{n+1}\in G$. The stepping rule (2.103) preserves G.

[26] The significance of the adjective *factored* will become apparent subsequently. See (2.113).

[27] Note that there is a consistency consideration here. If the solution $Y(t)$ were to leave G even though the initial matrix Y^0 is in G, then $A(Y,t)$ could become undefined.

Consider a single-stage Butcher tableau of the form

$$\begin{array}{c|c} c_1 & a_{11} \\ \hline & b_1 \end{array} \ . \tag{11.2.104}$$

It describes crude Euler when $c_1 = a_{11} = 0$ and $b_1 = 1$. See Exercise 2.3.3. Define quantities Y_1^n and K_1 by the rules

$$Y_1^n = Y^n, \tag{11.2.105}$$

$$K_1 = A(Y_1^n, t^n + hc_1) = A(Y_1^n, t^n) = A(Y^n, t^n), \tag{11.2.106}$$

Then we see that (2.103) can be written in the form

$$Y^{n+1} = \exp(hb_1 K_1)Y^n. \tag{11.2.107}$$

Thus, there is a correspondence between (2.103) and the Butcher tableau for crude Euler.

Next consider a two-stage Butcher tableau of the form

$$\begin{array}{c|cc} c_1 & 0 & 0 \\ c_2 & a_{21} & 0 \\ \hline & b_1 & b_2 \end{array} \ . \tag{11.2.108}$$

Note that the matrix a is strictly lower triangular, and therefore the associated Runge-Kutta method is explicit, as is the single-stage method specified by (2.104) when $a_{11} = 0$. Also, we continue to enforce the consistency condition (2.3.16) so that, in fact, $c_1 = 0$. Corresponding to the Butcher tableau (2.108), consider the following rule for stepping from Y^n to Y^{n+1}:

$$Y_1^n = Y^n, \tag{11.2.109}$$

$$K_1 = A(Y_1^n, t^n + hc_1) = A(Y_1^n, t^n) = A(Y^n, t^n), \tag{11.2.110}$$

$$Y_2^n = [\exp(ha_{21}K_1)]Y^n, \tag{11.2.111}$$

$$K_2 = A(Y_2^n, t^n + hc_2), \tag{11.2.112}$$

$$Y^{n+1} = [\exp(hb_2K_2)\exp(hb_1K_1)]Y^n. \tag{11.2.113}$$

Observe that the term appearing in square brackets on the right side of (2.113) is *factored* into a product of group elements, each in exponential form.

Let us examine the ingredients in this rule. Assuming $Y^n \in G$, we see from (2.109) that $Y_1^n \in G$. Next, we see from (2.110) that K_1 is in the Lie algebra of G. Now look at (2.111). Since K_1 is in the Lie algebra, $[\exp(ha_{21}K_1)]$ is in G, and therefore $Y_2^n \in G$. With regard to (2.112), we see that the arguments of A are in its domain of definition, and therefore K_2 is well defined and in the Lie algebra of G. Finally, examination of (2.113) shows that $Y^{n+1} \in G$. Our goal has at least been partially achieved: Starting from $Y^0 \in G$, we have produced a sequence of matrices Y^1, Y^2, $\cdots$, all of which are in G to machine precision.

What about the error associated with this rule? It can be shown that this algorithm is locally correct through terms of order h^2 (local error of order h^3) if the coefficients a, b, and c satisfy the consistency condition (2.3.16) and the order conditions (2.3.42), and (2.3.43). Thus, in the case of explicit one and two-stage methods, the order conditions on the Butcher tableau and the local accuracy for factored Lie Runge-Kutta are the same as those for the *ordinary* Runge-Kutta methods of Chapter 2.

To continue our discussion, consider a three-stage explicit Butcher tableau of the form

$$\begin{array}{c|ccc} c_1 & 0 & 0 & 0 \\ c_2 & a_{21} & 0 & 0 \\ c_3 & a_{31} & a_{32} & 0 \\ \hline & b_1 & b_2 & b_3 \end{array} \tag{11.2.114}$$

Corresponding to the Butcher tableau (2.114), we make the following rule for stepping from Y^n to Y^{n+1}:

$$Y_1^n = Y^n, \tag{11.2.115}$$

$$K_1 = A(Y_1^n, t^n + hc_1) = A(Y_1^n, t^n) = A(Y^n, t^n), \tag{11.2.116}$$

$$Y_2^n = [\exp(ha_{21}K_1)]Y^n, \tag{11.2.117}$$

$$K_2 = A(Y_2^n, t^n + hc_2), \tag{11.2.118}$$

$$Y_3^n = [\exp(ha_{32}K_2)][\exp(ha_{31}K_1)]Y^n, \tag{11.2.119}$$

$$K_3 = A(Y_3^n, t^n + hc_3), \tag{11.2.120}$$

$$Y^{n+1} = \exp(hb_3K_3)\exp(hb_2K_2)\exp(hb_1K_1)Y^n. \tag{11.2.121}$$

Again we see that the Y_j^n are in G, the arguments of A are in its domain of definition, and consequently the K_j are well defined and in the Lie algebra of G. And examination of (2.121) shows that therefore $Y^{n+1} \in G$. We conclude that our goal has again at least been partially achieved.

What about the error associated with this rule? It can be shown that this algorithm is locally correct through terms of order h^3 (local error of order h^4) if the Butcher tableau satisfies the consistency condition (2.3.16), the order conditions (2.3.42) through (2.3.45), and the *additional* order condition

Additional Order 3:

$$\sum_i b_i^2 c_i + 2\sum_{i<j} b_i c_i b_j = 1/3. \tag{11.2.122}$$

We know that the consistency condition (2.3.16) and the order conditions (2.3.42) through (2.3.45) are necessary and sufficient for the three-stage explicit ordinary Runge-Kutta methods of Chapter 2 to be locally accurate through terms of order h^3. Because of the additional order condition (2.122), more is required for the case of factored Lie Runge Kutta to achieve a local accuracy through terms of order h^3. Fortunately, there are three-stage explicit Butcher tableaux that meet the requirements (2.3.16), (2.3.42) through (2.3.45), *and* (2.122). Two such Butcher tableaux, found by Crouch and Grossman, are given below:

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ -1/24 & -1/24 & 0 & 0 \\ 17/24 & 161/24 & -6 & 0 \\ \hline & 1 & -2/3 & 2/3 \end{array}\;, \tag{11.2.123}$$

$$\begin{array}{c|ccc}
0 & 0 & 0 & 0 \\
3/4 & 3/4 & 0 & 0 \\
17/24 & 119/216 & 17/108 & 0 \\
\hline
 & 13/51 & -2/3 & 24/17
\end{array}\;. \tag{11.2.124}$$

Note the curious feature that use of Butcher tableau (2.123) entails the evaluation of A *outside* the temporal interval $[t^n, t^n + h]$, the interval over which integration is being performed, because for this tableau $c_2 < 0$.

At this point we are prepared to state the general recipe for factored Lie Runge-Kutta methods. We have already discussed the cases of $s = 1$ or $s = 2$ or $s = 3$ stages. Consider a Butcher tableau with s stages with $s > 3$ and again suppose that the matrix a is strictly lower triangular. Use this tableau to make the following stepping rule: For $1 \leq j \leq 3$ define, as before, Y_j^n and K_j by the rules (2.115) through (2.120). And for $4 \leq j \leq s$ make the definitions

$$Y_j^n = \exp(ha_{j,j-1}K_{j-1})\exp(ha_{j,j-2}K_{j-2}) \cdot \ldots \cdot \exp(ha_{j,1}K_1)Y^n, \tag{11.2.125}$$

$$K_j = A(Y_j^n, t^n + hc_j). \tag{11.2.126}$$

Finally, step from Y^n to Y^{n+1} using the rule

$$Y^{n+1} = \exp(hb_sK_s)\exp(hb_{s-1}K_{s-1}) \cdot \ldots \cdot \exp(hb_1K_1)Y^n. \tag{11.2.127}$$

What can be said about the error in this case? For an optimum choice of coefficients, how many stages are required to achieve order m? That is, what is the analog of Table 2.3.1 for the case of factored Lie Runge-Kutta methods? This is a difficult question. For factored Lie Runge-Kutta and for $m \geq 3$, compared to the ordinary Runge-Kutta methods of Chapter 2, there are many more conditions that the entries in the Butcher tableau must meet to achieve order m. For example, to achieve $m = 4$ there are 5 more order conditions for factored Lie Runge Kutta compared to ordinary Runge-Kutta. And, unlike ordinary Runge Kutta, it is impossible with only 4 stages to satisfy all the factored Lie Runge-Kutta order conditions required to achieve $m = 4$. At least $s = 5$ stages are required for factored Lie Runge Kutta to achieve order $m = 4$, some 5-stage Butcher tableaux with this property have been obtained by *Owren and Marthinsen*, and they have published one of them. Finally, It is believed that the minimum number of stages required for factored Lie Runge Kutta to achieve an order m with $m > 4$ grows rapidly with increasing m.

Application of Factored Lie Runge Kutta

As described at the beginning of this subsection, our goal is to find higher-order versions of (2.98). This is now easily done based on what we have learned of factored Lie Runge Kutta. From (2.101) and (2.126) we see for the case of (2.43) that we may write

$$K_j \leftrightarrow \bar{\boldsymbol{\omega}}(t^n + hc_j) \cdot \boldsymbol{L}. \tag{11.2.128}$$

And, using (2.127), we see that there is the stepping rule

$$\begin{aligned} S^{n+1} \;\; = \;\; & \exp[hb_s\bar{\boldsymbol{\omega}}(t^n + hc_s) \cdot \boldsymbol{L}]\exp[hb_{s-1}\bar{\boldsymbol{\omega}}(t^n + hc_{s-1}) \cdot \boldsymbol{L}] \times \\ & \ldots \times \exp[hb_1\bar{\boldsymbol{\omega}}(t^n + hc_1) \cdot \boldsymbol{L}]S^n. \end{aligned} \tag{11.2.129}$$

We have achieved our objective. Given an s-stage factored Lie Runge-Kutta method of order m, (2.129) provides a stepping rule that may make local errors of order h^{m+1}, but is guaranteed to preserve $SO(3,\mathbb{R})$ to machine precision.

For example, based on the Butcher tableau (2.124), there is the third-order (but exactly orthogonality-preserving) stepping rule

$$\begin{aligned} S^{n+1} &= \exp\{(24/17)h\bar{\boldsymbol{\omega}}[t^n+(17/24)h]\cdot\boldsymbol{L}\}\exp\{-(2/3)h\bar{\boldsymbol{\omega}}[t^n+(3/4)h]\cdot\boldsymbol{L}\}\times \\ &\quad \exp\{(13/51)h\bar{\boldsymbol{\omega}}[t^n]\cdot\boldsymbol{L}\}S^n. \end{aligned} \tag{11.2.130}$$

We close this subsection with two comments. The first is based on the observation that the Butcher tableau for a factored Lie Runge-Kutta method can also be used as a Butcher tableau for an ordinary Runge Kutta method. This result follows because the order conditions for factored Lie Runge Kutta contain as a subset all the order conditions for ordinary Runge Kutta. Therefore if a factored Lie Runge-Kutta method is used to track particle spin, which amounts to use of the rule (2.129) to track particle spin, then the same algorithm (the same Butcher tableau) could be used for an ordinary Runge-Kutta routine to compute the particle trajectory. In this way, the same times (t^n+hc_j) would occur in both the spin and particle trajectory routines, thus facilitating the computation of the required quantities $\bar{\boldsymbol{\omega}}(t^n+hc_j)$.

There is a corollary to this observation. In the previous subsection $\bar{\boldsymbol{\omega}}(t)$ was assumed to be a *given* function of t. With the use of factored Lie Runge Kutta this assumption is no longer necessary since there is no need with this method for an explicit fit of the form (2.65) with known coefficients. All that is required at each step are the sampling-point values $\bar{\boldsymbol{\omega}}(t^n+hc_j)$. This is true even if the $\bar{\boldsymbol{\omega}}(t^n+hc_j)$ need be determined dynamically.

At this point it is tempting to imagine that this approach could be applied to the case of spin if there were spin-orbit coupling (Stern-Gerlach effect) so that the equations of motion for the particle trajectory could be visualized as depending on S as well as the particle's position and momentum. However, there are quantum-mechanical reasons why this approach is not applicable.

The Stern-Gerlach effect is an example of *quantum entanglement*. Conceptually, and in a fully quantum treatment, when entering a Stern-Gerlach apparatus a single particle (assumed to have spin 1/2) is described by an initial state vector that is the tensor product of a spin state eigenvector (an eigenvector with eigenvalue $+1$ for $\boldsymbol{n}\cdot\boldsymbol{\sigma}$ for some specified unit vector $\boldsymbol{n}$) and an orbital state vector for a wave packet well localized (consistent within the uncertainty principle) both in position and momentum about some initial point z^i in phase space. After passing through the Stern-Gerlach apparatus the particle is no longer described by a product state vector. That is, the outgoing final quantum state vector *cannot* be written in tensor product form. Rather, it is described by a superposition of two vectors, each expressible in tensor product form. The first vector has a spin state that is an eigenvector with eigenvalue $+1$ for $\boldsymbol{m}\cdot\boldsymbol{\sigma}$ for some specified unit vector $\boldsymbol{m}$ that is determined by the orientation of the Stern-Gerlach apparatus, and an orbital part consisting of a well localized packet about some final point z^{f+}. The second vector has a spin state that is an eigenvector of $\boldsymbol{m}\cdot\boldsymbol{\sigma}$ with eigenvalue -1 and an orbital part consisting of a well localized packet about some final point z^{f-}. Moreover, if the Stern-Gerlach experiment is a success, the points z^{f+} and z^{f-} are sufficiently separated and the associated wave packets sufficiently localized so that there is little overlap between them. Finally, there is a definite phase relation between the two vectors. This state of affairs has no classical analog, and therefore cannot be treated classically.

We also remark that the Stern-Gerlach experiment, which is intended to separate a (single) beam of particles (with different particles characterized by different initial vectors $\boldsymbol{n}$) into two well-separated beams with the first consisting of particles in spin eigenstates characterized by $\boldsymbol{m}$ and the second consisting of particles in spin eigenstates characterized by $-\boldsymbol{m}$, may not necessarily be possible for all kinds of particles. The original Stern-Gerlach experiment was performed with silver atoms that are both *heavy* and were *neutral* (not ionized). There is an argument, due to Bohr and generalized by others, to the effect that it is not possible to magnetically separate a beam, consisting of particles with various spin orientations, into two beams with specified polarizations if the particles are *charged* and too light, or have too small a magnetic moment. For example, it is argued that it is not possible to achieve a Stern-Gerlach effect with electrons.

The essence of the argument, which is semiclassical, is as follows: Particles that are both charged and have a magnetic moment experience two forces when they move in a magnetic field, a Lorentz force due to the field itself and a Stern-Gerlach force due to field inhomogeneities. Since the magnetic field must be divergence free, desirable inhomogeneity (inhomogeneity transverse to the beam direction and in the direction of the main field) employed to produce a Stern-Gerlach force must lead to undesirable inhomogeneity in the other transverse direction. This undesirable field inhomogeneity leads to an undesirable Lorentz force that tends to spread the beam. Therefore the beam must be made small so that only small field variations are actually encountered by particles in the beam. But then, by the quantum uncertainty principle, there will be a corresponding spread in velocity space. This spread in velocity space again leads to an uncertainty in the Lorentz force. It may happen, if the particle mass is too small (thereby leading to a very large spread in velocity space and a corresponding large uncertainty in the Lorentz force) or if the magnetic moment is too small, that the uncertainty in the Lorentz force exceeds the Stern-Gerlach force. The net effect then (if the argument is to be believed) is that only a single final beam is produced whose spread due to the quantum-related uncertainty in the Lorentz force exceeds the splitting expected from the Stern-Gerlach force, thereby washing out any anticipated Stern-Gerlach effect.

The moral to be drawn from these considerations is that a full quantum treatment is required to find reliably the complete effect of an inhomogeneous magnetic field on a beam of particles that are charged and possibly too light or have too small a magnetic moment. See the references to the Stern-Gerlach effect at the end of this chapter.[28]

Although the Stern-Gerlach effect cannot be treated classically, there is a classical problem that is somewhat analogous, namely that of rigid-body motion, for which factored Lie Runge Kutta can be used to good advantage. As described earlier, in the case of rigid-body motion there is the complication that the $\boldsymbol{\omega}^{bf}$ must be determined dynamically from the Euler equations (1.19) through (1.21), which themselves may depend on R. This complication

[28] In fact, the semiclassical arguments that are used to describe the Stern-Gerlach effect even in the case of a neutral beam are suspect because they do not include beam spreading due to the uncertainty principle, the Stern-Gerlach force in the other transverse direction due to the undesirable field inhomogeneity, and the precession of the magnetic moment in the main field. See Exercise 2.17.

causes no problem if both the kinematic equations (1.18) and the dynamic equations (1.19) through (1.21) are integrated using the same factored Lie Runge-Kutta Butcher tableau. In fact, all that is really required is that both the kinematic equations (1.18) and the dynamic equations (1.19) through (1.21) be integrated using Runge-Kutta Butcher tableaux that have the same sampling times, i.e. the same vector c. However, it is important to recognize that the integration method(s) must be capable, as factored Lie Runge Kutta is, of producing the intermediate values Y_j^n so that values of R are available when needed in the various Runge-Kutta stages.

The second comment is that the $s = 5$ and $m = 4$ factored Lie Runge-Kutta Butcher tableau published by Owren and Marthinsen has for its b and c entries the values

$$b_1 = (1 + \kappa + \kappa^2)/(2\kappa + 2\kappa^2), \tag{11.2.131}$$

$$b_2 = 0, \tag{11.2.132}$$

$$b_3 = -(1 + 2\kappa + \kappa^2)/(12 + 6\kappa + 6\kappa^2), \tag{11.2.133}$$

$$b_4 = -1/(2\kappa + 2\kappa^2), \tag{11.2.134}$$

$$b_5 = b_1 = (1 + \kappa + \kappa^2)/(2\kappa + 2\kappa^2), \tag{11.2.135}$$

$$c_1 = 0, \tag{11.2.136}$$

$$c_2 = 3/2, \tag{11.2.137}$$

$$c_3 = 2/3 + \kappa/3 + \kappa^2/6 \simeq * \cdots, \tag{11.2.138}$$

$$c_4 = 1/3 - \kappa/3 - \kappa^2/6 \simeq - * \cdots, \tag{11.2.139}$$

$$c_5 = 1, \tag{11.2.140}$$

where

$$\kappa = 2^{1/3}. \tag{11.2.141}$$

Note that $b_2 = 0$. Thus, when (2.129) is used in this case as a stand-alone formula, it is effectively a *four* stage ($s = 4$) $m = 4$ formula. Observe, however, that $c_3 > 1$. Therefore use of this Butcher tableau involves an evaluation of $\bar{\boldsymbol{\omega}}(t)$ outside the interval $[t^n, t^n + h]$. One might also worry that $c_2 > 1$, which also leads to t values outside $[t^n, t^n + h]$. But, since $b_2 = 0$, this is not a concern. Finally, observe that $c_4 < 0$, so a second evaluation of $\bar{\boldsymbol{\omega}}(t)$ outside the interval $[t^n, t^n + h]$ is also required.

11.2.8 Magnus Lie Runge Kutta

The use of factored Lie Runge Kutta is not particularly attractive for our purposes because only relatively low-order results are available and because sometimes some evaluation points lie outside the interval $[t^n, t^{n+1}]$. But factored Lie Runge Kutta is designed to handle the case (2.99) for which A is allowed to depend on Y. What happens if we relax this requirement, and consider *only* equations of the simpler form

$$\dot{Y}(t) = A(t)Y(t)? \tag{11.2.142}$$

Note that, in view of (2.101), our problem of particular interest, namely (2.61), is of this simpler form.[29] We will learn that for the case (2.142) there are far more attractive results.

Our approach will be to *combine* all the exponents in the Lie Taylor factorization (2.93) to write $\hat{u}^{v\mathrm{fac}}(H)$ in *single-exponent* form. For example, in the case $M = 3$, suppose we write

$$\begin{aligned} \hat{u}^{v}(H) &\simeq \hat{u}^{v\mathrm{fac}}(H) = \exp(H^4\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(H^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(H^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(H\boldsymbol{d}_0\cdot\boldsymbol{K}) \\ &\simeq \exp[G(H)] \end{aligned} \tag{11.2.143}$$

where

$$G(H) = H^4\boldsymbol{e}_3\cdot\boldsymbol{K} + H^3\boldsymbol{e}_2\cdot\boldsymbol{K} + H^2\boldsymbol{e}_1\cdot\boldsymbol{K} + H\boldsymbol{e}_0\cdot\boldsymbol{K}. \tag{11.2.144}$$

Since the use of a single-exponent representation is in the spirit of the Magnus equations and our results will eventually be cast in Runge-Kutta form, we will refer to this procedure as *Magnus Lie Runge Kutta.* Indeed, the quantities $\boldsymbol{e}_n$ could be found in terms of the $\boldsymbol{c}_m$ by integrating the Magnus equations. See Section 10.3. Equivalently, they could be found by making temporal Taylor expansions of equations of the forms (1.67) or (1.83) and equating like powers of t.

Alternatively, since we already know the $\boldsymbol{d}_n$, as a first step we can convert the left side of (2.143) to the right side of (2.143) using the BCH formula (3.7.41). So doing will provide the $\boldsymbol{e}_n$ in terms of the $\boldsymbol{d}_m$. Also, according to (2.87) through (2.90), the $\boldsymbol{d}_m$ are already known in terms of the $\boldsymbol{c}_\ell$. Therefore, in a second step, we can find the $\boldsymbol{e}_n$ in terms of the $\boldsymbol{c}_m$ by simple algebraic substitution. We will now carry out this task.

Begin by observing that

$$\exp(H^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(H\boldsymbol{d}_0\cdot\boldsymbol{K}) \simeq \exp(E) \tag{11.2.145}$$

with

$$\begin{aligned} E &= H^2\boldsymbol{d}_1\cdot\boldsymbol{K} + H\boldsymbol{d}_0\cdot\boldsymbol{K} + (1/2)H^3\{\boldsymbol{d}_1\cdot\boldsymbol{K}, \boldsymbol{d}_0\cdot\boldsymbol{K}\} \\ &\quad +(1/12)H^4\{\boldsymbol{d}_0\cdot\boldsymbol{K}, \{\boldsymbol{d}_0\cdot\boldsymbol{K}, \boldsymbol{d}_1\cdot\boldsymbol{K}\}\} \\ &= H\boldsymbol{d}_0\cdot\boldsymbol{K} + H^2\boldsymbol{d}_1\cdot\boldsymbol{K} + (1/2)H^3\{\boldsymbol{d}_1\cdot\boldsymbol{K}, \boldsymbol{d}_0\cdot\boldsymbol{K}\} \\ &\quad +(1/12)H^4\{\boldsymbol{d}_0\cdot\boldsymbol{K}, \{\boldsymbol{d}_0\cdot\boldsymbol{K}, \boldsymbol{d}_1\cdot\boldsymbol{K}\}\} \\ &= H\boldsymbol{d}_0\cdot\boldsymbol{K} + H^2\boldsymbol{d}_1\cdot\boldsymbol{K} + (1/2)H^3(\boldsymbol{d}_1\times\boldsymbol{d}_0)\cdot\boldsymbol{K} \\ &\quad +(1/12)H^4[\boldsymbol{d}_0\times(\boldsymbol{d}_0\times\boldsymbol{d}_1)]\cdot\boldsymbol{K}. \end{aligned} \tag{11.2.146}$$

Next we find that

$$\begin{aligned} &\exp(H^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(H^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(H\boldsymbol{d}_0\cdot\boldsymbol{K}) \simeq \\ &\qquad \exp(H^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(E) = \exp(F) \end{aligned} \tag{11.2.147}$$

[29] Note also that (10.4.28) is of this form.

with

$$\begin{aligned} F \simeq H^3\boldsymbol{d}_2\cdot\boldsymbol{K} + E + (1/2)\{H^3\boldsymbol{d}_2\cdot\boldsymbol{K}, E\} \simeq \\ H^3\boldsymbol{d}_2\cdot\boldsymbol{K} + E + (1/2)H^4\{\boldsymbol{d}_2\cdot\boldsymbol{K}, \boldsymbol{d}_0\cdot\boldsymbol{K}\} = \\ H^3\boldsymbol{d}_2\cdot\boldsymbol{K} + E + (1/2)H^4(\boldsymbol{d}_2\times\boldsymbol{d}_0)\cdot\boldsymbol{K} = \\ H\boldsymbol{d}_0\cdot\boldsymbol{K} + H^2\boldsymbol{d}_1\cdot\boldsymbol{K} + H^3[\boldsymbol{d}_2\cdot\boldsymbol{K} + (1/2)(\boldsymbol{d}_1\times\boldsymbol{d}_0)\cdot\boldsymbol{K}] \\ +H^4\{(1/2)(\boldsymbol{d}_2\times\boldsymbol{d}_0)\cdot\boldsymbol{K} + (1/12)[\boldsymbol{d}_0\times(\boldsymbol{d}_0\times\boldsymbol{d}_1)]\cdot\boldsymbol{K}\}. \end{aligned} \tag{11.2.148}$$

Finally, we see that

$$\begin{aligned} \exp(H^4\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(H^3\boldsymbol{d}_2\cdot\boldsymbol{K})\exp(H^2\boldsymbol{d}_1\cdot\boldsymbol{K})\exp(H\boldsymbol{d}_0\cdot\boldsymbol{K}) \simeq \\ \exp(H^4\boldsymbol{d}_3\cdot\boldsymbol{K})\exp(F) = \exp(G) \end{aligned} \tag{11.2.149}$$

with

$$G \simeq H^4\boldsymbol{d}_3\cdot\boldsymbol{K} + F. \tag{11.2.150}$$

The net result, through terms of order H^4, is that

$$\begin{aligned} G = H\boldsymbol{d}_0\cdot\boldsymbol{K} + H^2\boldsymbol{d}_1\cdot\boldsymbol{K} + H^3[\boldsymbol{d}_2\cdot\boldsymbol{K} + (1/2)(\boldsymbol{d}_1\times\boldsymbol{d}_0)\cdot\boldsymbol{K}] \\ +H^4\{\boldsymbol{d}_3\cdot\boldsymbol{K} + (1/2)(\boldsymbol{d}_2\times\boldsymbol{d}_0)\cdot\boldsymbol{K} + (1/12)[\boldsymbol{d}_0\times(\boldsymbol{d}_0\times\boldsymbol{d}_1)]\cdot\boldsymbol{K}\}. \end{aligned} \tag{11.2.151}$$

Upon comparing (2.144) and (2.151), we conclude that there are the relations

$$\boldsymbol{e}_0 = \boldsymbol{d}_0, \tag{11.2.152}$$

$$\boldsymbol{e}_1 = \boldsymbol{d}_1, \tag{11.2.153}$$

$$\boldsymbol{e}_2 = \boldsymbol{d}_2 + (1/2)(\boldsymbol{d}_1\times\boldsymbol{d}_0), \tag{11.2.154}$$

$$\boldsymbol{e}_3 = \boldsymbol{d}_3 + (1/2)(\boldsymbol{d}_2\times\boldsymbol{d}_0) + (1/12)[\boldsymbol{d}_0\times(\boldsymbol{d}_0\times\boldsymbol{d}_1)]. \tag{11.2.155}$$

The first step is complete. To finish our task, we employ the relations (2.87) through (2.90) in the relations (2.152) through (2.155) to find the results

$$\boldsymbol{e}_0 = \boldsymbol{c}_0, \tag{11.2.156}$$

$$\boldsymbol{e}_1 = \boldsymbol{c}_1/2, \tag{11.2.157}$$

$$\boldsymbol{e}_2 = (1/3)\boldsymbol{c}_2 - (1/12)(\boldsymbol{c}_0\times\boldsymbol{c}_1), \tag{11.2.158}$$

$$\boldsymbol{e}_3 = (1/4)\boldsymbol{c}_3 - (1/12)(\boldsymbol{c}_0\times\boldsymbol{c}_2). \tag{11.2.159}$$

Remarkably, although there are double cross products in the intermediate results (2.90) and (2.155), there is (due to cancelations) no double cross product in the final results (2.156) through (2.159).

And there is a further remarkable result. Define a vector $\boldsymbol{\Omega}$ by the rule

$$\boldsymbol{\Omega}(H) = H\boldsymbol{e}_0 + H^2\boldsymbol{e}_1 + H^3\boldsymbol{e}_2 + H^4\boldsymbol{e}_3 \tag{11.2.160}$$

so that G can be written the form

$$G(H) = \boldsymbol{\Omega}(H)\cdot\boldsymbol{K}. \tag{11.2.161}$$

By (2.156) through (2.159) we may also write

$$\begin{aligned}\boldsymbol{\Omega}(H) \;&=\; H\boldsymbol{c}_0 + (1/2)H^2\boldsymbol{c}_1 + (1/3)H^3\boldsymbol{c}_2 + (1/4)H^4\boldsymbol{c}_3 \\ &\quad -(1/12)H^3(\boldsymbol{c}_0 \times \boldsymbol{c}_1) - (1/12)H^4(\boldsymbol{c}_0 \times \boldsymbol{c}_2).\end{aligned} \tag{11.2.162}$$

Next, in accord with the stipulation that $M = 3$, define a vector $\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau)$ by the rule

$$\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau) = \sum_{m=0}^{3} \boldsymbol{c}_m \tau^m = \boldsymbol{c}_0 + \boldsymbol{c}_1\tau + \boldsymbol{c}_2\tau^2 + \boldsymbol{c}_3\tau^3, \tag{11.2.163}$$

which is the $M = 3$ version of (2.65). Observe that

$$\int_0^H \hat{\boldsymbol{\omega}}^{\rm fit3}(\tau)\, d\tau = H\boldsymbol{c}_0 + (1/2)H^2\boldsymbol{c}_1 + (1/3)H^3\boldsymbol{c}_2 + (1/4)H^4\boldsymbol{c}_3, \tag{11.2.164}$$

and that the right side of (2.164) also appears on the right side of right (2.162). Therefore we many rewrite (2.162) in the form

$$\boldsymbol{\Omega}(H) \;=\; \int_0^H \hat{\boldsymbol{\omega}}^{\rm fit3}(\tau)\, d\tau - (1/12)H^3(\boldsymbol{c}_0 \times \boldsymbol{c}_1) - (1/12)H^4(\boldsymbol{c}_0 \times \boldsymbol{c}_2). \tag{11.2.165}$$

Enter Legendre Gauss

The occurrence of the integral (2.164) suggests the application of quadrature formulas. Suppose we define two sampling times τ_i by the rule

$$(\tau_1, \tau_2) = (Hx_1, Hx_2) \tag{11.2.166}$$

where x_1 and x_2 are the $k = 2$ Legendre-Gauss sampling points. See (T.1.29). Then, since $k = 2$ Legendre Gauss has $\ell_{\max} = 3$, see (T.1.11), there is the relation

$$\int_0^H \hat{\boldsymbol{\omega}}^{\rm fit3}(\tau)\, d\tau = (H/2)[\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_1) + \hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_2)]. \tag{11.2.167}$$

See (T.1.72).

Also, it can be verified that there is the result

$$\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_1) \times \hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_2) = (1/\sqrt{3})[H(\boldsymbol{c}_0 \times \boldsymbol{c}_1) + H^2(\boldsymbol{c}_0 \times \boldsymbol{c}_2)] + O(H^3). \tag{11.2.168}$$

See Exercise 2.12. It follows that

$$\begin{aligned}-(\sqrt{3}/12)H^2[\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_1) \times \hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_2)] = \\ -(1/12)[H^3(\boldsymbol{c}_0 \times \boldsymbol{c}_1) + H^4(\boldsymbol{c}_0 \times \boldsymbol{c}_2)] + O(H^5).\end{aligned} \tag{11.2.169}$$

Comparison of (2.165) with (2.167) and (2.169) now reveals that, through terms of order H^4, there is the even more remarkable result

$$\boldsymbol{\Omega}(H) = (H/2)[\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_1) + \hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_2)] - (\sqrt{3}/12)H^2[\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_1) \times \hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_2)]. \tag{11.2.170}$$

The ingredients for computing $\boldsymbol{\Omega}(H)$ through terms of order H^4 can be obtained by computing the value of $\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau)$ at just *two* sampling points! By contrast, the utilization of $M = 3$ Lie Taylor factorization requires the evaluation of $\hat{\boldsymbol{\omega}}(\tau)$ at $M + 1 = 4$ sampling points.

Let us summarize what has been accomplished. From (2.143) and (2.161) we have the result

$$\hat{u}^v(H) = \exp[G(H)] = \exp[\boldsymbol{\Omega}(H) \cdot \boldsymbol{K}] \tag{11.2.171}$$

where, through terms of order H^4, $\boldsymbol{\Omega}(H)$ is given by (2.170). Upon making in (2.170) the substitution

$$\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_i) \simeq \hat{\boldsymbol{\omega}}(\tau_i) \tag{11.2.172}$$

we obtain, in effect, a *two*-stage *explicit fourth*-order Runge Kutta method for computing $\hat{u}^v(H)$ with the *guarantee* that $\hat{u}^v(H)$ is *exactly* in $SU(2)$. Note that order 4 is the highest order that can be obtained even for the simplest problem of $k = 2$ quadrature of ordinary functions. Again see (T.1.72).

At this point we make two remarks. The first is that (2.164), the first term in (2.165), is what might be expected if the quantities $\hat{\boldsymbol{\omega}}(\tau) \cdot \boldsymbol{K}$ at various times τ all commuted. It is the analog of the term (10.3.17) in the Magnus expansion. Correspondingly (2.169), the remaining terms in (2.165), takes into account the possibility that the quantities $\hat{\boldsymbol{\omega}}(\tau) \cdot \boldsymbol{K}$ at various times may not all commute. It is the analog of the term (10.3.19) in the Magnus expansion.

The second remark is that if (2.170) through (2.172) are to be used to track particle spin, then Gauss4 could be used to compute the particle trajectory. See the Butcher tableau (2.3.19). This is possible because both algorithms would then use the common sampling times $t^n + \tau_1$ and $t^n + \tau_2$, and Gauss4 is a *collocation* method so that the results at each stage produce accurate values for the orbit and hence accurate values for $\hat{\boldsymbol{\omega}}$ at the sampling times. See Exercise 2.3.12. Moreover Gauss4, when applied to Hamiltonian systems, has the further important property of being symplectic. See Section *.

Enter Newton Cotes

With the idea of quadrature still in mind, we recall from Table T.1.1 that $k = 3$ Newton Cotes also has $\ell_{\rm max} = 3$. Suppose we now define three sampling times τ_i by the rule

$$(\tau_1, \tau_2, \tau_3) = (Hx_1, Hx_2, Hx_3) \tag{11.2.173}$$

where x_1, x_2, and x_3 are the $k = 3$ Newton-Cotes sampling points. See (T.1.15). Then we have the relation

$$\int_0^H \hat{\boldsymbol{\omega}}^{\rm fit3}(\tau)\, d\tau = (H/6)[\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_1) + 4\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_2) + \hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_3)]. \tag{11.2.174}$$

See (T.1.16) and (T.1.66). Also, there is the result

$$\begin{aligned} \hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_1) \times \hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_3) &= \hat{\boldsymbol{\omega}}^{\rm fit3}(0) \times \hat{\boldsymbol{\omega}}^{\rm fit3}(H) \\ &= \boldsymbol{c}_0 \times [\boldsymbol{c}_0 + \boldsymbol{c}_1 H + \boldsymbol{c}_2 H^2 + \boldsymbol{c}_3 H^3] \\ &= H(\boldsymbol{c}_0 \times \boldsymbol{c}_1) + H^2(\boldsymbol{c}_0 \times \boldsymbol{c}_2) + O(H^3). \end{aligned} \tag{11.2.175}$$

It follows that

$$-(1/12)H^2[\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_1)\times\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_3)] = -(1/12)[H^3(\boldsymbol{c}_0\times\boldsymbol{c}_1)+H^4(\boldsymbol{c}_0\times\boldsymbol{c}_2)]+O(H^5). \tag{11.2.176}$$

Comparison of (2.165) with (2.174) and (2.176) reveals that, through terms of order H^4, there is also the result

$$\begin{aligned}\boldsymbol{\Omega}(H) &= (H/6)[\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_1)+4\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_2)+\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_3)]\\ &- (1/12)H^2[\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_1)\times\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_3)].\end{aligned} \tag{11.2.177}$$

The ingredients for computing $\boldsymbol{\Omega}(H)$ through terms of order H^4 can be obtained by computing the value of $\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau)$ at the three sampling times (2.173). Moreover, we observe that these sampling times are the same as those for classic RK4. See the Butcher tableau (2.3.14). It follows that use of (2.171), (2.172), and (2.177) are ideal for tracking particle spin when the particle trajectory is computed using classic RK4 equipped with *dense output*. See Section 2.3.4. The dense output feature would be used to provide values for the coordinates and hence the $\hat{\boldsymbol{\omega}}$ at the times τ_i, and since classic RK4 (although not a collocation method) employs the same sampling times, these interpolated values are expected to be especially accurate.

One can also evaluate the integral on the left side of (2.174) using $k = 4$ Newton Cotes, which also has $\ell_{\rm max} = 3$. See (T.1.20) through (T.1.22) and Table T.1.1. So doing produces a $k = 4$ formula for $\boldsymbol{\Omega}(H)$ that involves four sampling times τ_i given by the rule

$$(\tau_1,\tau_2,\tau_3,\tau_4) = (Hx_1,Hx_2,Hx_3,Hx_4) \tag{11.2.178}$$

where the x_i are the $k = 4$ Newton-Cotes sampling points. See (T.1.20). Use of these sampling points gives, through terms of order H^4, the relation

$$\begin{aligned}\boldsymbol{\Omega}(H) &= (H/8)[\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_1)+3\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_2)+3\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_3)+\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_4)]\\ &- (1/12)H^2[\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_1)\times\hat{\boldsymbol{\omega}}^{\rm fit3}(\tau_4)].\end{aligned} \tag{11.2.179}$$

And, in this case, one can use for trajectory integration the explicit RK4 version given by the Butcher tableau (2.3.15), again equipped with dense output, since this RK4 (although again not a collocation method) has the same sampling times τ_1 through τ_4.

We close this subsection with two final remarks. The first remark is that the results we have just found for spin readily generalize to all equations of the form (2.142). Suppose we seek to integrate (2.142) using a stepping rule of the form

$$Y^{n+1} = Y(t^n+H) = \exp(G_n)Y^n. \tag{11.2.180}$$

Then it can be shown, for example and in analogy with (2.170), that through terms of order H^4 there is the rule

$$G_n = (H/2)[A(t^n+\tau_1)+A(t^n+\tau_2)] - (\sqrt{3}/12)H^2\{A(t^n+\tau_1),A(t^n+\tau_2)\}. \tag{11.2.181}$$

There are also rules analogous to (2.177) and (2.179).

The second remark is that there is a systematic procedure for finding the coefficients of the expansion of $G(H)$ in powers of H in terms of the Taylor expansion of $A(t)$ in powers of t, which we have just done for $M \leq 3$, and some specific results are known for the next few orders including the cases $M \leq 6$. See the paper of *Blanes*, *Casas*, and *Ros* listed under the references to Magnus Lie Runge Kutta in the Bibliography at the end of this chapter. When translated to the context of spin tracking, they provide higher-order generalizations of (2.170), (2.177), and (2.179). For example, a three-stage sixth-order generalization of (2.170) could be constructed that could be used in conjunction with Gauss6.

11.2.9 Integration in the Lie Algebra Revisited

In the previous Subsection 2.8 we studied the integration of the equation of motion

$$\dot{Y}(t) = A(t)Y(t). \tag{11.2.182}$$

We also mentioned earlier, and you will prove in Exercise 2.7, that one can work equally well with an equation of motion having the form

$$\dot{M}(t) = M(t)A(t). \tag{11.2.183}$$

Moreover, from the work of Subsections 1.14 and 1.15 , we know how to integrate (2.183) in its Lie algebra. It follows that we also know how to integrate (2.182) in its Lie algebra. The purpose of this subsection is to study how a generalization of (2.182), namely an equation of motion of the form

$$\dot{Y}(t) = A(Y,t)Y(t), \tag{11.2.184}$$

can be integrated in its Lie algebra.[30] In so doing we will describe an alternative to factored Lie Runge Kutta. Recall Subsection 2.7 and (2.99).

For the purposes of Runge Kutta it is sufficient to describe how to take one step. In the spirit of Subsection 2.6, let t^b be the time at which an integration step is to be initiated so that we wish to integrate from t^b to $t^b + h$. We also suppose that

$$Y^b = Y(t^b) \tag{11.2.185}$$

is known and is an element of the group in question. Write

$$Y(t) = Y^v(t)Y^b \tag{11.2.186}$$

with Y^v being a variable matrix near the identity satisfying

$$Y^v(t^b) = I. \tag{11.2.187}$$

Then it follows from (2.184) and (2.186) that $Y^v(t)$ obeys the equation of motion

$$\dot{Y}^v(t) = A(Y^vY^b, t)Y^v(t) \tag{11.2.188}$$

with the initial condition (2.187).

[30] See Exercise 2.14 for the treatment of a related problem equivalent to the generalization of (2.183).

Next introduce a relative time τ by the rule

$$t = t^b + \tau. \tag{11.2.189}$$

Also, define quantities $\hat{Y}^v$ and $\hat{A}$ by the rules

$$\hat{Y}^v(\tau) = Y^v(t^b + \tau), \tag{11.2.190}$$

$$\hat{A}(\hat{Y}^v, \tau) = A[Y^v(t^b + \tau)Y^b, t^b + \tau] = A[\hat{Y}^v(\tau)Y^b, t^b + \tau]. \tag{11.2.191}$$

It follows from these definitions and (2.188) that $\hat{Y}^v(\tau)$ obeys the equation of motion

$$d\hat{Y}^v(\tau)/d\tau = \hat{A}(\hat{Y}^v, \tau)\hat{Y}^v(\tau) \tag{11.2.192}$$

with the initial condition

$$\hat{Y}^v(0) = I. \tag{11.2.193}$$

Our task is to integrate (2.192) from $\tau = 0$ to $\tau = h$ in such a way that $\hat{Y}^v(h)$ is accurate through terms of order h^m, has possible errors of order h^{m+1}, but is still exactly in the group in question, or at least is in the group through terms of substantially higher order than h^m.

Integration in the Lie Algebra: Exponential Representation

Begin by making the exponential Ansatz

$$\hat{Y}^v(\tau) = \exp[\Omega(\tau)]. \tag{11.2.194}$$

The relation (2.194) can be differentiated and manipulated to yield the result

$$[d\hat{Y}^v(\tau)/d\tau][\hat{Y}^v(\tau)]^{-1} = \text{iex}(\#\Omega\#)(d\Omega/d\tau). \tag{11.2.195}$$

See (*) in Appendix C. Also, the equation of motion (2.192) can be rewritten in the form

$$[d\hat{Y}^v(\tau)/d\tau][\hat{Y}^v(\tau)]^{-1} = \hat{A}(\hat{Y}^v, \tau). \tag{11.2.196}$$

Comparison of (2.195) and (2.196) and use of (2.194) give the result

$$\text{iex}(\#\Omega\#)(d\Omega/d\tau) = \hat{A}[\exp(\Omega), \tau]. \tag{11.2.197}$$

Finally, solving (2.197) for $d\Omega/d\tau$ yields the equation of motion

$$d\Omega/d\tau = [\text{iex}(\#\Omega\#)]^{-1}\hat{A}[\exp(\Omega), \tau]; \tag{11.2.198}$$

and use of (2.193) and (2.194) gives the initial condition

$$\Omega(0) = 0. \tag{11.2.199}$$

The function $[\text{iex}(w)]^{-1}$ has an expansion of the form

$$[\text{iex}(w)]^{-1} = \sum_{\ell=0}^{\infty} c_\ell w^\ell \tag{11.2.200}$$

with $c_0 = 1$, $c_1 = -1/2$, $c_2 = 1/12$, and $c_\ell = b_\ell$ for $\ell > 1$. Again see Appendix C. With the aid of this expansion, (2.198) becomes

$$d\Omega/d\tau = \sum_{\ell=0}^{\infty} c_\ell [\#\Omega(\tau)\#]^\ell \hat{A}[\exp(\Omega), \tau]. \tag{11.2.201}$$

As in the work of Subsection 1.14, the sum in (2.201) can be truncated beyond an even number n to obtain a result that is correct through order h^m with $m = n + 2$. Thus, to this accuracy, the equation to be solved is

$$\begin{aligned} d\Omega/d\tau &= \sum_{\ell=0}^{\ell=n} c_\ell [\#\Omega(\tau)\#]^\ell \hat{A}[\exp(\Omega), \tau] \\ &= c_0 \hat{A}[\exp(\Omega), \tau] + c_1\{\Omega(\tau), \hat{A}[\exp(\Omega), \tau]\} \\ &\quad + c_2\{\Omega(\tau), \{\Omega(\tau), \hat{A}[\exp(\Omega), \tau]\}\} + \cdots \\ &\quad + c_n\{\Omega(\tau), \{\Omega(\tau), \{\cdots\{\Omega(\tau), \hat{A}[\exp(\Omega), \tau]\}\cdots\}\}\}, \end{aligned} \tag{11.2.202}$$

and the understanding is that this truncated equation is to be integrated only over the interval $\tau \in [0, h]$.

For the sake of pedagogy, let us work out a specific case in detail. Suppose we take $n = 2$, in which case we expect a local accuracy through terms of order h^4. Then (2.202) becomes

$$\begin{aligned} d\Omega/d\tau &= \hat{A}[\exp(\Omega), \tau] - (1/2)\{\Omega(\tau), \hat{A}[\exp(\Omega), \tau]\} \\ &\quad + (1/12)\{\Omega(\tau), \{\Omega(\tau), \hat{A}[\exp(\Omega), \tau]\}\}. \end{aligned} \tag{11.2.203}$$

Correspondingly, we will use classic RK4 for integration. To do so, and to conform to previous notation, it is convenient to write (2.203) in the form

$$d\Omega/d\tau = f(\Omega, \tau) \tag{11.2.204}$$

where

$$\begin{aligned} f(\Omega, \tau) &= \hat{A}[\exp(\Omega), \tau] - (1/2)\{\Omega, \hat{A}[\exp(\Omega), \tau]\} \\ &\quad + (1/12)\{\Omega, \{\Omega, \hat{A}[\exp(\Omega), \tau]\}\}. \end{aligned} \tag{11.2.205}$$

For the first stage of classic RK4, in this instance, and adopting a notation analogous to that of (2.3.8) through (2.3.10), we have the results

$$\tau_1 = 0, \tag{11.2.206}$$

$$\Omega_1 = 0, \tag{11.2.207}$$

$$K_1 = f(\Omega_1, \tau_1) = f(0, 0) = \hat{A}[I, 0]. \tag{11.2.208}$$

See (2.201) and (2.3.9), and note that according to (2.3.14) the first row in the Butcher tableau for classic RK4 vanishes. For the second stage there are the results

$$\tau_2 = h/2, \tag{11.2.209}$$

$$\Omega_2 = (h/2)K_1, \tag{11.2.210}$$

$$\begin{aligned} K_2 &= f(\Omega_2, \tau_2) \\ &= \hat{A}[\exp(\Omega_2), h/2] - (1/2)\{\Omega_2, \hat{A}[\exp(\Omega_2), h/2]\} \\ &\quad +(1/12)\{\Omega_2, \{\Omega_2, \hat{A}[\exp(\Omega_2), h/2]\}\}. \end{aligned} \tag{11.2.211}$$

For the third stage there are the results

$$\tau_3 = h/2, \tag{11.2.212}$$

$$\Omega_3 = (h/2)K_2, \tag{11.2.213}$$

$$\begin{aligned} K_3 &= f(\Omega_3, \tau_3) \\ &= \hat{A}[\exp(\Omega_3), h/2] - (1/2)\{\Omega_3, \hat{A}[\exp(\Omega_3), h/2]\} \\ &\quad +(1/12)\{\Omega_3, \{\Omega_3, \hat{A}[\exp(\Omega_3), h/2]\}\}. \end{aligned} \tag{11.2.214}$$

For the fourth stage there are the results

$$\tau_4 = h, \tag{11.2.215}$$

$$\Omega_4 = hK_3, \tag{11.2.216}$$

$$\begin{aligned} K_4 &= f(\Omega_4, \tau_4) \\ &= \hat{A}[\exp(\Omega_4), h] - (1/2)\{\Omega_4, \hat{A}[\exp(\Omega_4), h]\} \\ &\quad +(1/12)\{\Omega_4, \{\Omega_4, \hat{A}[\exp(\Omega_4), h]\}\}. \end{aligned} \tag{11.2.217}$$

The net result of the full RK4 step is that

$$\Omega(h) = (h/6)K_1 + (h/3)K_2 + (h/3)K_3 + (h/6)K_4. \tag{11.2.218}$$

See (2.3.11).

Finally, from (2.194), we find that, through terms of order h^4,

$$\hat{Y}^v(h) = \exp[\Omega(h)]. \tag{11.2.219}$$

Also, from the definitions (2.186), (2.189), and (2.190), we have the relation

$$Y(t^b + h) = Y^v(t^b + h)Y^b = \hat{Y}^v(h)Y^b. \tag{11.2.220}$$

Therefore, upon setting $t^b = t^n$ and making the identification

$$Y^b = Y^n, \tag{11.2.221}$$

we obtain the stepping rule

$$Y^{n+1} = \exp[\Omega(h)]Y^n. \tag{11.2.222}$$

Let us reflect on what has been accomplished. Let G be the group in question and let $\mathcal{L}(G)$ be its Lie algebra. From (2.207) we see that $\Omega_1 \in \mathcal{L}(G)$. Also, since $I \in G$, we see from (2.208) that $\hat{A}$ is evaluated in its domain, and therefore $K_1 \in \mathcal{L}(G)$. Next look at the second stage results. By (2.210), $\Omega_2 \in \mathcal{L}(G)$ since $K_1 \in \mathcal{L}(G)$. Also, since $\Omega_2 \in \mathcal{L}(G)$, there is the result that $\exp(\Omega_2) \in G$ and therefore again $\hat{A}$ is evaluated in its domain. See the right side of (2.211). It follows that all the ingredients in the right side of (2.211) are in $\mathcal{L}(G)$. Moreover they are combined there in such a way that the full right side of (2.211) is in $\mathcal{L}(G)$, and therefore $K_2 \in \mathcal{L}(G)$. Evidently analogous results hold for all the stages so that $\hat{A}$ is always evaluated in its domain and all the K_j are in $\mathcal{L}(G)$. Therefore, by (2.218), $\Omega(h) \in \mathcal{L}(G)$ from which it follows that $\exp[\Omega(h)] \in G$. And from (2.222) we conclude that $Y^{n+1} \in G$ if $Y^n \in G$. Despite generally having made local errors of order h^5, arising both from the truncation of (2.201) and the use of RK4, we have preserved G to machine precision.

Finally we note that, contrary to appearances, the relation (2.222) between Y^{n+1} and Y^n is generally *not* linear. This is the case because generally $\Omega(h)$ depends on Y^n. See the far right side of (2.191) and recall (2.221).

Integration in the Lie Algebra: Cayley Representation

Suppose G is a quadratic group. The results (2.185) through (2.193), since they are general, continue to hold. But if G is a quadratic group, then we may also make, in place of (2.194), the Ansatz

$$\hat{Y}^v(\tau) = \mathrm{cay}[V(\tau)], \tag{11.2.223}$$

where

$$\mathrm{cay}(V) = (I+V)(I-V)^{-1}. \tag{11.2.224}$$

Note that here the definition of $\mathrm{cay}(V)$ given by (2.224) differs by a sign from that given in (3.12.45). Again V will be in $\mathcal{L}(G)$ if $\hat{Y}^v$ is in G, and conversely.

From (2.223) and (2.224) it follows that

$$[d\hat{Y}^v(\tau)/d\tau][\hat{Y}^v(\tau)]^{-1} = 2C(dV/d\tau)B^{-1}, \tag{11.2.225}$$

where B and C are given in terms of V by the relations (1.295) through (1.297). See the calculations at the end of Exercise 1.12. Recall also that (2.192) can be rewritten in the form (2.196). Combining (2.196) and (2.225) gives the relation

$$2C(dV/d\tau)B^{-1} = \hat{A}[\mathrm{cay}(V), \tau]. \tag{11.2.226}$$

Finally, solving (2.226) for $dV/d\tau$ yields the equation of motion

$$dV/d\tau = (1/2)\hat{A}[\mathrm{cay}(V), \tau] - (1/2)\{V, \hat{A}[\mathrm{cay}(V), \tau]\} - (1/2)V\hat{A}[\mathrm{cay}(V), \tau]V. \tag{11.2.227}$$

See (1.320) at the end of Exercise 1.12. And use of (2.193) and (2.223) gives the initial condition

$$V(0) = 0. \tag{11.2.228}$$

Note that the right side of (2.227), unlike (2.201), involves only a finite number of terms, namely three, and therefore in this case no truncation is required.

Because no truncation of the equation of motion has occurred, the only source of error in this case is that associated with numerical integration. Suppose we write (2.227) in the form

$$dV/d\tau = f(V,\tau) \tag{11.2.229}$$

where

$$f(V,\tau) = (1/2)\hat{A}[\mathrm{cay}(V),\tau] - (1/2)\{V,\hat{A}[\mathrm{cay}(V),\tau]\} - (1/2)V\hat{A}[\mathrm{cay}(V),\tau]V. \tag{11.2.230}$$

And suppose for purposes of illustration that we again use classic RK4 for integration. For the first stage of classic RK4, and again adopting a notation analogous to that of (2.3.8) through (2.3.10), we now have the results

$$\tau_1 = 0, \tag{11.2.231}$$

$$V_1 = 0, \tag{11.2.232}$$

$$K_1 = f(V_1,\tau_1) = f(0,0) = (1/2)\hat{A}[I,0]. \tag{11.2.233}$$

See (2.228). For the second stage there are the results

$$\tau_2 = h/2, \tag{11.2.234}$$

$$V_2 = (h/2)K_1, \tag{11.2.235}$$

$$\begin{aligned} K_2 &= f(V_2,\tau_2) \\ &= (1/2)\hat{A}[\mathrm{cay}(V_2),h/2] - (1/2)\{V_2,\hat{A}[\mathrm{cay}(V_2),h/2]\} \\ &\quad -(1/2)V_2\hat{A}[\mathrm{cay}(V_2),h/2]V_2. \end{aligned} \tag{11.2.236}$$

For the third stage there are the results

$$\tau_3 = h/2, \tag{11.2.237}$$

$$V_3 = (h/2)K_2, \tag{11.2.238}$$

$$\begin{aligned} K_3 &= f(V_3,\tau_3) \\ &= (1/2)\hat{A}[\mathrm{cay}(V_3),h/2] - (1/2)\{V_3,\hat{A}[\mathrm{cay}(V_3),h/2]\} \\ &\quad -(1/2)V_3\hat{A}[\mathrm{cay}(V_3),h/2]V_3. \end{aligned} \tag{11.2.239}$$

For the fourth stage there are the results

$$\tau_4 = h, \tag{11.2.240}$$

$$V_4 = hK_3, \tag{11.2.241}$$

$$\begin{aligned} K_4 &= f(V_4,\tau_4) \\ &= (1/2)\hat{A}[\mathrm{cay}(V_4),h] - (1/2)\{V_4,\hat{A}[\mathrm{cay}(V_4),h]\} \\ &\quad -(1/2)V_4\hat{A}[\mathrm{cay}(V_4),h]V_4. \end{aligned} \tag{11.2.242}$$

The net result of the full RK4 step is that

$$V(h) = (h/6)K_1 + (h/3)K_2 + (h/3)K_3 + (h/6)K_4. \tag{11.2.243}$$

Finally, from (2.223), we find that, through terms of order h^4,

$$\hat{Y}^v(h) = \text{cay}[V(h)]. \tag{11.2.244}$$

Look again at the relations (2.220) and (2.221). Combining them with (2.244) yields the stepping rule

$$Y^{n+1} = \text{cay}[V(h)]Y^n. \tag{11.2.245}$$

Let us again reflect on what has been accomplished. Let G be the group in question and let $\mathcal{L}(G)$ be its Lie algebra. From (2.232) we see that $V_1 \in \mathcal{L}(G)$. Also, since $I \in G$, we see from (2.233) that $\hat{A}$ is evaluated in its domain, and therefore $K_1 \in \mathcal{L}(G)$. Next look at the second stage results. By (2.235), $V_2 \in \mathcal{L}(G)$ since $K_1 \in \mathcal{L}(G)$. Also, since $V_2 \in \mathcal{L}(G)$, there is the result that $\text{cay}(V_2) \in G$ and therefore again $\hat{A}$ is evaluated in its domain. See the right side of (2.236). It follows that all the ingredients in the right side of (2.236) are in $\mathcal{L}(G)$. Moreover they are combined there in such a way that the full right side of (2.236) is in $\mathcal{L}(G)$, and therefore $K_2 \in \mathcal{L}(G)$. Recall the discussion in Exercise 1.12. Evidently analogous results hold for all the stages so that $\hat{A}$ is always evaluated in its domain and all the K_j are in $\mathcal{L}(G)$. Therefore, by (2.243), $V(h) \in \mathcal{L}(G)$ from which it follows that $\text{cay}[V(h)] \in G$. And from (2.245) we conclude that $Y^{n+1} \in G$ if $Y^n \in G$. Despite generally having made local errors of order h^5 arising from the use of RK4, we have preserved G to machine precision.

Finally we note again that, contrary to appearances, the relation (2.245) between Y^{n+1} and Y^n is generally *not* linear. This is the case because generally $V(h)$ depends on Y^n. Again see the far right side of (2.191) and recall (2.221).

Exercises

11.2.1. Much of Section 2 discussed the problem of finding $\boldsymbol{s}(t)$ given $\bar{\boldsymbol{\omega}}(t)$, the equation of motion (2.1), and the initial condition (2.3). The aim of this exercise is to treat the inverse *control theory* problem: given $\boldsymbol{s}(t)$, can one find an $\bar{\boldsymbol{\omega}}(t)$ such that solving (2.1) yields $\boldsymbol{s}(t)$? Stated geometrically, given a path in S^2, is there an $\bar{\boldsymbol{\omega}}(t)$ such that solving (2.1) yields this path? You are to show that the answer to this question is *yes*, and that there are in fact many such $\bar{\boldsymbol{\omega}}(t)$. Then you are to show that a knowledge of *two* paths suffices to define $\bar{\boldsymbol{\omega}}(t)$ uniquely.[31] Treatment of the two-path case employs some of the machinery of Exercise 1.1, which you should review.

Begin by showing that, since by assumption $\boldsymbol{s}(t) \in S^2$, there is the condition

$$\boldsymbol{s}(t) \cdot \dot{\boldsymbol{s}}(t) = 0. \tag{11.2.246}$$

Next show that (2.1) can be rewritten in the forms

$$d\boldsymbol{s}/dt = \bar{\boldsymbol{\omega}}(t) \times \boldsymbol{s} = -\boldsymbol{s} \times \bar{\boldsymbol{\omega}}(t) = -[\boldsymbol{s}(t) \cdot \boldsymbol{L}]\bar{\boldsymbol{\omega}}(t), \tag{11.2.247}$$

[31] The idea of considering two paths, as well as the relation (1.103), were suggested by Sateesh Mane.

$$[\boldsymbol{s}(t)\cdot\boldsymbol{L}]\bar{\boldsymbol{\omega}}(t) = -\dot{\boldsymbol{s}}(t). \tag{11.2.248}$$

According to (2.248), given $\boldsymbol{s}(t)$ and $\dot{\boldsymbol{s}}(t)$, we seek to find a vector $\bar{\boldsymbol{\omega}}(t)$ such that the vector/matrix equation (2.248) is satisfied.

Evidently, at any moment t', there are three possibilities:

$$\boldsymbol{s}(t') \in S^2 \text{ and } \boldsymbol{s}(t') = +\boldsymbol{e}_3, \tag{11.2.249}$$

$$\boldsymbol{s}(t') \in S^2 \text{ and } \boldsymbol{s}(t') = -\boldsymbol{e}_3, \tag{11.2.250}$$

$$\boldsymbol{s}(t') \in S^2 \text{ and } \boldsymbol{s}(t') \neq \pm\boldsymbol{e}_3. \tag{11.2.251}$$

First consider the nongeneric possibilities (2.249) and (2.250). Verify that then the relation (2.248) becomes

$$\pm\, L^3\bar{\boldsymbol{\omega}}(t') = -\dot{\boldsymbol{s}}(t'). \tag{11.2.252}$$

Write out $\bar{\boldsymbol{\omega}}(t')$ and $\dot{\boldsymbol{s}}(t')$ in component form,

$$\bar{\boldsymbol{\omega}}(t') = \begin{pmatrix} \bar{\omega}_1(t') \\ \bar{\omega}_2(t') \\ \bar{\omega}_3(t') \end{pmatrix}, \tag{11.2.253}$$

$$\dot{\boldsymbol{s}}(t') = \begin{pmatrix} \dot{s}_1(t') \\ \dot{s}_2(t') \\ 0 \end{pmatrix}. \tag{11.2.254}$$

Here, in writing (2.254), we have used the fact $\dot{s}_3(t') = 0$ which follows from (2.249), (2.250), and the condition (2.246). Show that

$$L^3\bar{\boldsymbol{\omega}}(t') = \begin{pmatrix} -\bar{\omega}_2(t') \\ \bar{\omega}_1(t') \\ 0 \end{pmatrix}. \tag{11.2.255}$$

Therefore verify (2.252) is satisfied in the possibilities (2.249) and (2.250) providing that (respectively)

$$\bar{\omega}_1(t') = \mp\dot{s}_2(t'), \tag{11.2.256}$$

$$\bar{\omega}_2(t') = \pm\dot{s}_1(t'), \tag{11.2.257}$$

$$\bar{\omega}_3(t') = \text{anything}. \tag{11.2.258}$$

What remains is the generic possibility (2.251). For this possibility parameterize $\boldsymbol{s}(t') \in S^2$ in terms of polar angle coordinates as in (2.11) through (2.13), and verify that both $\theta(t')$ and $\phi(t')$ are well defined. Show that, in terms of the Euler angle parameterization (3.7.207), the rotation $R(\phi,\theta,0)$ relates $\boldsymbol{s}(t') \in S^2$ and $\boldsymbol{e}_3$ by the equation

$$\boldsymbol{s}(t') = R[\phi(t'),\theta(t'),0]\boldsymbol{e}_3. \tag{11.2.259}$$

See (3.7.208). Next verify that

$$RL^3R^{-1} = R(\boldsymbol{e}_3\cdot\boldsymbol{L})R^{-1} = [(R\boldsymbol{e}_3)\cdot\boldsymbol{L}] = \boldsymbol{s}(t')\cdot\boldsymbol{L}. \tag{11.2.260}$$

Recall (8.2.59). Show that inserting this result into (2.248) yields, at the generic moment $t = t'$, the relation

$$[RL^3R^{-1}]\bar{\boldsymbol{\omega}}(t') = -\dot{\boldsymbol{s}}(t') \tag{11.2.261}$$

from which it follows that

$$L^3[R^{-1}\bar{\boldsymbol{\omega}}(t')] = -R^{-1}\dot{\boldsymbol{s}}(t'). \tag{11.2.262}$$

Introduce the notation

$$\boldsymbol{u} = R^{-1}\bar{\boldsymbol{\omega}}(t'), \tag{11.2.263}$$

$$\boldsymbol{v} = R^{-1}\dot{\boldsymbol{s}}(t'), \tag{11.2.264}$$

so that (2.262) becomes

$$L^3\boldsymbol{u} = -\boldsymbol{v}. \tag{11.2.265}$$

Verify also that

$$\boldsymbol{e}_3 \cdot \boldsymbol{v} = [R^{-1}\boldsymbol{s}] \cdot [R^{-1}\dot{\boldsymbol{s}}] = \boldsymbol{s} \cdot \dot{\boldsymbol{s}} = 0. \tag{11.2.266}$$

Here we have used the orthogonality of R. Consequently, verify that we may write

$$\boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}, \tag{11.2.267}$$

$$\boldsymbol{v} = \begin{pmatrix} v_1 \\ v_2 \\ 0 \end{pmatrix}. \tag{11.2.268}$$

We are back to the first case we considered, and conclude that

$$u_1 = -v_2, \tag{11.2.269}$$

$$u_2 = v_1, \tag{11.2.270}$$

$$u_3 = \text{anything}. \tag{11.2.271}$$

Verify that (2.269) through (2.271) can be written in the vector/matrix form

$$\boldsymbol{u}(t') = L^3\boldsymbol{v}(t') + g(t')\boldsymbol{e}_3 \tag{11.2.272}$$

where g is any function. We are almost done. Show that inserting (2.263) and (2.264) into (2.272) yields the result

$$R^{-1}\bar{\boldsymbol{\omega}}(t') = L^3R^{-1}\dot{\boldsymbol{s}}(t') + g(t')\boldsymbol{e}_3, \tag{11.2.273}$$

from which it follows, for any generic time $t = t'$, that

$$\begin{aligned} \bar{\boldsymbol{\omega}}(t) &= RL^3R^{-1}\dot{\boldsymbol{s}}(t) + g(t)R\boldsymbol{e}_3 \\ &= [\boldsymbol{s}(t) \cdot \boldsymbol{L}]\dot{\boldsymbol{s}}(t) + g(t)\boldsymbol{s}(t) \\ &= \boldsymbol{s}(t) \times \dot{\boldsymbol{s}}(t) + g(t)\boldsymbol{s}(t). \end{aligned} \tag{11.2.274}$$

If all has been done correctly, (2.274) provides the general solution to (2.248). Verify that this general solution also covers the nongeneric possibilities (2.249) and (2.250) for which the polar angles were not well defined. Verify, by direct computation, that the general solution satisfies (2.248). Hint: Use (3.7.200). In practice we might require that $g(t)$ be continuous, and for convenience we might set $g(t) = 0$.

You have shown that there is a choice of $\bar{\boldsymbol{\omega}}(t)$ that will produce any desired path $\boldsymbol{s}(t) \in S^2$, and that this choice is not unique. What if there are two different paths $\boldsymbol{r}(t) \in S^2$ and $\boldsymbol{s}(t) \in S^2$ that are supposed to be produced by the *same* $\bar{\boldsymbol{\omega}}(t)$. That is, $\boldsymbol{s}(t)$ satisfies (2.1) and $\boldsymbol{r}(t)$ satisfies the analogous relation

$$d\boldsymbol{r}/dt = \bar{\boldsymbol{\omega}}(t) \times \boldsymbol{r}. \tag{11.2.275}$$

Is $\bar{\boldsymbol{\omega}}(t)$ then uniquely determined? You are to show that the answer is *yes*.

Can $\boldsymbol{r}(t) \in S^2$ and $\boldsymbol{s}(t) \in S^2$ be specified arbitrarily and independently? The answer is *no*. There are some mutual restrictions on $\boldsymbol{r}(t) \in S^2$ and $\boldsymbol{s}(t) \in S^2$ since they are supposed to be produced by the same differential equation, namely (2.1) and (2.275). We will call a path on S^2 that satisfies this differential equation an *allowable path*.

First, we must assume that $\boldsymbol{r}^0 \neq \pm\boldsymbol{s}^0$. In the first case the initial conditions would be the same and therefore $\boldsymbol{r}(t)$ and $\boldsymbol{s}(t)$ would be identical. In the second case, since by linearity $-\boldsymbol{s}(t)$ is an allowable path if $\boldsymbol{s}(t)$ is, we have again gained no new information. Therefore we may assume that $\boldsymbol{r}^0$ and $\boldsymbol{s}^0$ are linearly independent. Show that it follows that

$$-1 < \boldsymbol{r}^0 \cdot \boldsymbol{s}^0 < 1 \tag{11.2.276}$$

and

$$|\boldsymbol{r}^0 \times \boldsymbol{s}^0| \neq 0. \tag{11.2.277}$$

Consider the quantity $\boldsymbol{r}(t) \cdot \boldsymbol{s}(t)$. Verify that

$$\begin{aligned}(d/dt)[\boldsymbol{r}(t) \cdot \boldsymbol{s}(t)] &= \dot{\boldsymbol{r}}(t) \cdot \boldsymbol{s}(t) + \boldsymbol{r}(t) \cdot \dot{\boldsymbol{s}}(t) \\ &= [\bar{\boldsymbol{\omega}}(t) \times \boldsymbol{r}(t)] \cdot \boldsymbol{s}(t) + \boldsymbol{r}(t) \cdot [\bar{\boldsymbol{\omega}}(t) \times \boldsymbol{s}(t)] \\ &= [\bar{\boldsymbol{\omega}}(t) \times \boldsymbol{r}(t)] \cdot \boldsymbol{s}(t) + [\boldsymbol{r}(t) \times \bar{\boldsymbol{\omega}}(t)] \cdot \boldsymbol{s}(t) = 0.\end{aligned} \tag{11.2.278}$$

Conclude that

$$\boldsymbol{r}(t) \cdot \boldsymbol{s}(t) = \boldsymbol{r}^0 \cdot \boldsymbol{s}^0. \tag{11.2.279}$$

The dot product of the two time-dependent vectors associated with two allowable paths (and also the dot product of such a vector with itself) remains constant as the paths are traversed.

Next consider the vector $\boldsymbol{r}(t) \times \boldsymbol{s}(t)$. Verify that

$$\begin{aligned}(d/dt)[\boldsymbol{r}(t) \times \boldsymbol{s}(t)] &= \dot{\boldsymbol{r}}(t) \times \boldsymbol{s}(t) + \boldsymbol{r}(t) \times \dot{\boldsymbol{s}}(t) \\ &= [\bar{\boldsymbol{\omega}}(t) \times \boldsymbol{r}(t)] \times \boldsymbol{s}(t) + \boldsymbol{r}(t) \times [\bar{\boldsymbol{\omega}}(t) \times \boldsymbol{s}(t)] \\ &= -\{\boldsymbol{s}(t) \times [\bar{\boldsymbol{\omega}}(t) \times \boldsymbol{r}(t)]\} + \{\boldsymbol{r}(t) \times [\bar{\boldsymbol{\omega}}(t) \times \boldsymbol{s}(t)]\} \\ &= -\{\bar{\boldsymbol{\omega}}(t)[\boldsymbol{r}(t) \cdot \boldsymbol{s}(t)] - \boldsymbol{r}(t)[\bar{\boldsymbol{\omega}}(t) \cdot \boldsymbol{s}(t)]\} \\ &\quad +\{\bar{\boldsymbol{\omega}}(t)[\boldsymbol{r}(t) \cdot \boldsymbol{s}(t)] - \boldsymbol{s}(t)[\bar{\boldsymbol{\omega}}(t) \cdot \boldsymbol{r}(t)]\} \\ &= \boldsymbol{r}(t)[\bar{\boldsymbol{\omega}}(t) \cdot \boldsymbol{s}(t)] - \boldsymbol{s}(t)[\bar{\boldsymbol{\omega}}(t) \cdot \boldsymbol{r}(t)] \\ &= \bar{\boldsymbol{\omega}}(t) \times [\boldsymbol{r}(t) \times \boldsymbol{s}(t)].\end{aligned} \tag{11.2.280}$$

You have shown, apart from normalization, that $\boldsymbol{r}(t) \times \boldsymbol{s}(t)$ is an allowable path if $\boldsymbol{r}(t)$ and $\boldsymbol{s}(t)$ are allowable paths. Argue, from the results of the previous paragraph, that there is also the result

$$|\boldsymbol{r}(t) \times \boldsymbol{s}(t)| = |\boldsymbol{r}^0 \times \boldsymbol{s}^0|. \tag{11.2.281}$$

Define three *fixed* vectors by the rules

$$\boldsymbol{e}_1 = \boldsymbol{r}^0, \tag{11.2.282}$$

$$\boldsymbol{e}_2 = (\boldsymbol{r}^0 \times \boldsymbol{s}^0)/|(\boldsymbol{r}^0 \times \boldsymbol{s}^0)|, \tag{11.2.283}$$

$$\boldsymbol{e}_3 = \boldsymbol{e}_1 \times \boldsymbol{e}_2. \tag{11.2.284}$$

Verify that together the $\boldsymbol{e}_j$ form a right-hand triad of orthonormal vectors. That is,

$$\boldsymbol{e}_j \cdot \boldsymbol{e}_k = \delta_{jk}, \tag{11.2.285}$$

$$\boldsymbol{e}_j \times \boldsymbol{e}_k = \boldsymbol{e}_\ell, \tag{11.2.286}$$

where j, k, ℓ is any cyclic permutation of $1, 2, 3$.

Define three *time-dependent* vectors by the rules

$$\boldsymbol{f}_1(t) = \boldsymbol{r}(t), \tag{11.2.287}$$

$$\boldsymbol{f}_2(t) = [\boldsymbol{r}(t) \times \boldsymbol{s}(t)]/|(\boldsymbol{r}^0 \times \boldsymbol{s}^0)|, \tag{11.2.288}$$

$$\boldsymbol{f}_3(t) = \boldsymbol{f}_1(t) \times \boldsymbol{f}_2(t). \tag{11.2.289}$$

Based on the work above, show that the $\boldsymbol{f}_j(t)$ have the initial conditions

$$\boldsymbol{f}_j(t^0) = \boldsymbol{e}_j \tag{11.2.290}$$

and are all allowable paths,

$$d\boldsymbol{f}_j/dt = \bar{\boldsymbol{\omega}}(t) \times \boldsymbol{f}_j. \tag{11.2.291}$$

Show that they also form a right-hand triad of orthonormal vectors at each instant t,

$$\boldsymbol{f}_j \cdot \boldsymbol{f}_k = \delta_{jk}, \tag{11.2.292}$$

$$\boldsymbol{f}_j \times \boldsymbol{f}_k = \boldsymbol{f}_\ell, \tag{11.2.293}$$

where j, k, ℓ is any cyclic permutation of $1, 2, 3$.

Next observe that since the $\boldsymbol{e}_j$ and $\boldsymbol{f}_j(t)$ both form triads of orthonormal vectors, they must be related by an orthogonal transformation $R(t)$,

$$\boldsymbol{f}_j(t) = R(t)\boldsymbol{e}_j, \tag{11.2.294}$$

with $R(t^0) = I$. Moreover, (2.294) specifies $R(t)$ uniquely. See Section 3.6.3. By continuity $R(t)$ must satisfy $\det R(t) = 1$, and therefore $R(t) \in SO(3, \mathbb{R})$. Thus, from a knowledge of two allowable paths, we are able to define a rigid body motion.

Also, from the work of Exercise 1.1, we know that there is the relation

$$\dot{R}(t)R^{-1}(t) = \boldsymbol{\omega}^{sf}(t) \cdot \boldsymbol{L}. \tag{11.2.295}$$

See (1.114). Are $\bar{\boldsymbol{\omega}}(t)$, $\boldsymbol{\omega}^{sf}(t)$, $\boldsymbol{\omega}^{bf}(t)$, and $\boldsymbol{\omega}(t)$ related? And, if so, how? From (2.294) and (2.295) show that

$$\dot{\boldsymbol{f}}_j = \dot{R}\boldsymbol{e}_j = \dot{R}R^{-1}R\boldsymbol{e}_j = [\boldsymbol{\omega}^{sf}\cdot\boldsymbol{L}]\boldsymbol{f}_j = \boldsymbol{\omega}^{sf}\times\boldsymbol{f}_j. \tag{11.2.296}$$

Upon comparing (2.291) and (2.296) we are led to make the conjecture

$$\bar{\boldsymbol{\omega}}(t) \overset{?}{=} \boldsymbol{\omega}^{sf}(t). \tag{11.2.297}$$

Note that in writing the far right side of (2.296) some explanation must be given as to what is meant by $\boldsymbol{\omega}^{sf}(t)$ when it is employed in a cross product. We will see that it is the vector with the components $\omega_j^{sf}(t)$ when the $\boldsymbol{e}_j$ are used as a basis. We also observe that, according to (1.103), $\boldsymbol{\omega}(t)$ is specified once the $\boldsymbol{f}_j(t)$ are known.

Because there are two basis sets involved, namely the $\boldsymbol{e}_j$ and the $\boldsymbol{f}_j(t)$, let us make this conjecture precise. From the definition (1.4), and the fact that the $\boldsymbol{f}_j(t)$ are all allowable paths, verify that

$$\begin{aligned}\omega_j^{bf}(t) &= -\boldsymbol{f}_k(t)\cdot\dot{\boldsymbol{f}}_\ell(t) = -\boldsymbol{f}_k(t)\cdot[\bar{\boldsymbol{\omega}}(t)\times\boldsymbol{f}_\ell(t)]\\ &= \boldsymbol{f}_k(t)\cdot[\boldsymbol{f}_\ell(t)\times\bar{\boldsymbol{\omega}}(t)] = [\boldsymbol{f}_k(t)\times\boldsymbol{f}_\ell(t)]\cdot\bar{\boldsymbol{\omega}}(t)]\\ &= \boldsymbol{f}_j(t)\cdot\bar{\boldsymbol{\omega}}(t)\end{aligned} \tag{11.2.298}$$

where j,k,ℓ is any cyclic permutation of $1,2,3$. Show, since the $\boldsymbol{f}_j(t)$ form an orthonormal basis, it follows that

$$\bar{\boldsymbol{\omega}}(t) = \sum_j[\boldsymbol{f}_j(t)\cdot\bar{\boldsymbol{\omega}}(t)]\boldsymbol{f}_j(t) = \sum_j\omega_j^{bf}(t)\boldsymbol{f}_j(t) = \sum_j\omega_j^{sf}(t)\boldsymbol{e}_j = \boldsymbol{\omega}(t). \tag{11.2.299}$$

Here we have again used the results and notation of Exercise 1.1. We conclude that a knowledge of two allowable paths completely specifies $\bar{\boldsymbol{\omega}}(t)$.

As an interesting side calculation, verify that there are also the relations

$$\begin{aligned}-\boldsymbol{f}_2(t)\cdot\{\boldsymbol{f}_1(t)\times\dot{\boldsymbol{f}}_2(t)\} &= -\boldsymbol{f}_2(t)\cdot\{\boldsymbol{f}_1(t)\times[\bar{\boldsymbol{\omega}}\times\boldsymbol{f}_2(t)]\}\\ &= -\boldsymbol{f}_2(t)\cdot\{-\boldsymbol{f}_2(t)[\boldsymbol{f}_1(t)\cdot\bar{\boldsymbol{\omega}}(t)]\}\\ &= \boldsymbol{f}_1(t)\cdot\bar{\boldsymbol{\omega}}(t),\end{aligned} \tag{11.2.300}$$

$$\begin{aligned}-\boldsymbol{f}_3(t)\cdot\{\boldsymbol{f}_2(t)\times\dot{\boldsymbol{f}}_3(t)\} &= -\boldsymbol{f}_3(t)\cdot\{\boldsymbol{f}_2(t)\times[\bar{\boldsymbol{\omega}}\times\boldsymbol{f}_3(t)]\}\\ &= -\boldsymbol{f}_3(t)\cdot\{-\boldsymbol{f}_3(t)[\boldsymbol{f}_2(t)\cdot\bar{\boldsymbol{\omega}}(t)]\}\\ &= \boldsymbol{f}_2(t)\cdot\bar{\boldsymbol{\omega}}(t),\end{aligned} \tag{11.2.301}$$

$$\begin{aligned}-\boldsymbol{f}_1(t)\cdot\{\boldsymbol{f}_3(t)\times\dot{\boldsymbol{f}}_1(t)\} &= -\boldsymbol{f}_1(t)\cdot\{\boldsymbol{f}_3(t)\times[\bar{\boldsymbol{\omega}}\times\boldsymbol{f}_1(t)]\}\\ &= -\boldsymbol{f}_1(t)\cdot\{-\boldsymbol{f}_1(t)[\boldsymbol{f}_3(t)\cdot\bar{\boldsymbol{\omega}}(t)]\}\\ &= \boldsymbol{f}_3(t)\cdot\bar{\boldsymbol{\omega}}(t).\end{aligned} \tag{11.2.302}$$

There is one last observation. Suppose $\boldsymbol{q}^0$ is any point in S^2. Since the $\boldsymbol{e}_j$ form a basis in E^3, any such $\boldsymbol{q}^0$ can be written as a linear combination of the $\boldsymbol{e}_j$,

$$\boldsymbol{q}^0 = \sum_{j=1}^{3} \alpha_j \boldsymbol{e}_j \tag{11.2.303}$$

with

$$\alpha_j = \boldsymbol{e}_j \cdot \boldsymbol{q}^0 \tag{11.2.304}$$

and

$$\sum_{j=1}^{3} \alpha_j^2 = 1. \tag{11.2.305}$$

It follows that $\boldsymbol{q}(t)$ given by

$$\boldsymbol{q}(t) = \sum_{j=1}^{3} \alpha_j \boldsymbol{f}_j(t) \tag{11.2.306}$$

will be an allowable path on S^2 with initial condition $\boldsymbol{q}^0$. Alternatively, we may write

$$\boldsymbol{q}(t) = R(t)\boldsymbol{q}^0. \tag{11.2.307}$$

We conclude that at $t = t^0$ every point on S^2 can be viewed as the starting point for a unique allowable path. Conversely, since $R(t)$ is in $SO(3,\mathbb{R})$ and therefore invertible at each time t, every point on S^2 at time t may be viewed as lying on a unique allowable path. Or, put another way, all allowable paths may be viewed as arising from a time-dependent rotation of S^2. Finally, we have learned that $R(t)$, and all allowable paths $\boldsymbol{q}(t)$, can be built from the two allowable paths $\boldsymbol{r}(t)$ and $\boldsymbol{s}(t)$.

11.2.2. Verify that the equations of motion (2.5) through (2.7) follow from assuming the validity of (2.8) through (2.10). That is, if (2.8) through (2.10) are satisfied, then (2.5) through (2.7) are satisfied.

11.2.3. Verify (2.11) through (2.20).

11.2.4. The aim of this exercise is to verify that (2.52) preserves $SU(2)$. Begin by assuming that $u(t)$ belongs to $SU(2)$ at some time $t = t^0$:

$$u^\dagger(t^0)u(t^0) = I \tag{11.2.308}$$

and

$$\det[u(t^0)] = 1. \tag{11.2.309}$$

First show that (2.52) preserves the determinant of u. Verify from (2.52) that

$$\begin{aligned} u(t+dt) &= u(t) + \dot{u}(t)dt + O[(dt)^2] \\ &= u(t) + (\bar{\boldsymbol{\omega}} \cdot \boldsymbol{K})u(t)dt + O[(dt)^2] \\ &= \exp[(dt)\bar{\boldsymbol{\omega}} \cdot \boldsymbol{K}]u(t) + O[(dt)^2]. \end{aligned} \tag{11.2.310}$$

Show that taking the determinant of both sides of (2.310) yields the result

$$\begin{aligned}\det[u(t+dt)] &= \det\{\exp[(dt)\bar{\boldsymbol{\omega}}\cdot\boldsymbol{K}]u(t)\}+O[(dt)^2]\\ &= \det\{\exp[(dt)\bar{\boldsymbol{\omega}}\cdot\boldsymbol{K}]\}\det[u(t)]+O[(dt)^2]\\ &= \exp[(dt)\,\mathrm{tr}(\bar{\boldsymbol{\omega}}\cdot\boldsymbol{K})]\det[u(t)]+O[(dt)^2]\\ &= \det[u(t)]+O[(dt)^2]. \end{aligned}\tag{11.2.311}$$

(Recall that the K^j are traceless.) Show from (2.311) that

$$\{\det[u(t+dt)]-\det[u(t)]\}/dt = O(dt)\tag{11.2.312}$$

and therefore

$$d\{\det[u(t)]\}/dt = 0.\tag{11.2.313}$$

Verify that the solution to the differential equation (2.313) with the initial condition (2.309) is the relation

$$\det[u(t)] = 1.\tag{11.2.314}$$

What remains to be verified is that (2.52) preserves unitarity. Show from (2.52) that

$$(d/dt)u^\dagger(t) = -u^\dagger(t)(\bar{\boldsymbol{\omega}}\cdot\boldsymbol{K}).\tag{11.2.315}$$

Next verify from (2.52) and (2.315) that

$$\begin{aligned}(d/dt)[u^\dagger(t)u(t)] &= [(d/dt)u^\dagger(t)]u(t)]+u^\dagger(t)(d/dt)u(t)\\ &= -u^\dagger(t)(\bar{\boldsymbol{\omega}}\cdot\boldsymbol{K})u(t)+u^\dagger(t)(\bar{\boldsymbol{\omega}}\cdot\boldsymbol{K})u(t)=0.\end{aligned}\tag{11.2.316}$$

Finally, verify that the solution to the differential equation (2.316) with the initial condition (2.308) is the relation

$$u^\dagger(t)u(t) = I.\tag{11.2.317}$$

11.2.5. The purpose of this exercise is to prove that S defined by (2.54) satisfies (2.43) and (2.44) providing u satisfies (2.52) and (2.53). You will need some of the ingredients of Exercise 2.4 above. Begin by verifying that (2.44) is satisfied because

$$(1/2)\mathrm{tr}[u^\dagger(t^0)\sigma^\alpha u(t^0)\sigma^\beta] = (1/2)\,\mathrm{tr}[\sigma^\alpha\sigma^\beta] = \delta_{\alpha\beta} = S_{\alpha\beta}(t^0).\tag{11.2.318}$$

Next work on proving that S satisfies (2.43). Show that

$$\dot{S}_{\alpha\beta}(t) = (1/2)\mathrm{tr}(\dot{u}^\dagger\sigma^\alpha u\sigma^\beta)+(1/2)\mathrm{tr}(u^\dagger\sigma^\alpha\dot{u}\sigma^\beta).\tag{11.2.319}$$

Verify that employing (2.52) and (2.315) in (2.319) yields the result

$$\begin{aligned}\dot{S}_{\alpha\delta}(t) &= -(1/2)\mathrm{tr}[u^\dagger(\bar{\boldsymbol{\omega}}\cdot\boldsymbol{K})\sigma^\alpha u\sigma^\delta]+(1/2)\mathrm{tr}[u^\dagger\sigma^\alpha(\bar{\boldsymbol{\omega}}\cdot\boldsymbol{K})u\sigma^\delta]\\ &= (1/2)\mathrm{tr}[u^\dagger\{\sigma^\alpha,(\bar{\boldsymbol{\omega}}\cdot\boldsymbol{K})\}u\sigma^\delta].\end{aligned}\tag{11.2.320}$$

Verify that

$$\{\sigma^\alpha,(\bar{\boldsymbol{\omega}}\cdot\boldsymbol{K})\} = \sum_\beta \bar{\omega}_\beta\{\sigma^\alpha,\sigma^\beta\} = \sum_{\beta\gamma}\bar{\omega}_\beta\epsilon_{\alpha\beta\gamma}\sigma^\gamma.\tag{11.2.321}$$

Recall (1.143) and (5.7.39). Insert (2.321) into (2.320) and use (3.7.181) to show that

$$\begin{aligned}\dot{S}_{\alpha\delta}(t) &= \sum_{\beta\gamma} \bar{\omega}_\beta \epsilon_{\alpha\beta\gamma}(1/2)\mathrm{tr}[u^\dagger \sigma^\gamma u \sigma^\delta] \\ &= \sum_{\beta\gamma} \bar{\omega}_\beta \epsilon_{\beta\alpha\gamma} S_{\gamma\delta} = \sum_{\beta\gamma} \bar{\omega}_\beta \epsilon_{\alpha\beta\gamma} S_{\gamma\delta} \\ &= \sum_{\beta\gamma} \bar{\omega}_\beta L^\beta_{\alpha\gamma} S_{\gamma\delta} = \sum_\gamma (\bar{\boldsymbol{\omega}} \cdot \boldsymbol{L})_{\alpha\gamma} S_{\gamma\delta}, \end{aligned} \tag{11.2.322}$$

or, in more compact matrix form,

$$\dot{S} = (\bar{\boldsymbol{\omega}} \cdot \boldsymbol{L}) S. \tag{11.2.323}$$

11.2.6. This exercise is in part a continuation of Exercises 1.7 and 1.8, which you should review. Our aim here is to explore the analogy between (2.40), which describes rotations in E^3, and (1.209), which will be seen to be related to rotations in E^4.

Verify that $A(\boldsymbol{\omega}^{bf})$ as given by (1.208) is 4×4 and antisymmetric, and therefore belongs to the Lie algebra $so(4, \mathbb{R})$. Recall that $(\bar{\boldsymbol{\omega}} \cdot \boldsymbol{L})$ is 3×3 and antisymmetric, and therefore belongs to the Lie algebra $so(3, \mathbb{R})$. Thus, (1.209) appears to be a four-dimensional analog of (2.40). Moreover because $A(\boldsymbol{\omega}^{bf}) \in so(4, \mathbb{R})$, (1.209) describes, as asserted, (at least some) rotations in E^4. Therefore it is no wonder that (1.209) preserves $w \cdot w$ (preserves S^3) just as (2.40) preserves S^2.

Verify, however, that $A(\boldsymbol{\omega}^{bf})$ is given by

$$A(\boldsymbol{\omega}^{bf}) = (1/2)(\omega_1^{bf} E^1 + \omega_2^{bf} E^2 + \omega_3^{bf} E^3) = (1/2)\boldsymbol{\omega}^{bf} \cdot \boldsymbol{E}, \tag{11.2.324}$$

and therefore A belongs to one of the $su(2)$ Lie algebras in $so(4, \mathbb{R})$. Thus the matrices A given by (1.208) do not span all of $so(4, \mathbb{R})$. [Recall that $so(4, \mathbb{R})$ is six dimensional.]

We have seen that the vector equation (2.40) has the corresponding matrix equation (2.43). Similarly, the vector equation (1.209) has the 4×4 matrix equation counterpart

$$\dot{T} = A(\boldsymbol{\omega}^{bf})T \tag{11.2.325}$$

with the initial condition

$$T(t^0) = I. \tag{11.2.326}$$

Therefore part of our task is also to explore the analogy between (2.43) and (2.325).

Verify that $T(t) \in SO(4, \mathbb{R})$, but that not all $SO(4, \mathbb{R})$ matrices can be produced in this fashion because the matrices A given by (1.208) do not span all of $so(4, \mathbb{R})$. Therefore the analogy between (2.40) and (1.209), and between (2.43) and (2.325), is not complete.[32]

The observation that A belongs to an $su(2)$ Lie algebra begs further exploration because we know that what counts, when solving differential equations, is not the matrix size of A but rather the Lie algebra to which it belongs. (The same reasoning led to the replacement

[32] Although the analogy between (2.40) and (1.209) is not complete, it can be shown that any path in S^3 can be produced by solution of (1.209) for some choice of $\boldsymbol{\omega}^{bf}(t)$ when employed in $A[\boldsymbol{\omega}^{bf}(t)]$. This is possible for two reasons: First, as manifolds, $SU(2)$ and S^3 are the same. Recall Exercise 5.10.3. Second, $SU(2)$ acts transitively on itself (as does any group) by both left and right multiplication.

of the L_j by the K_j in Exercise 1.5 and in Subsection 2.6.) Thus, we expect that there must be an $SU(2)$ formulation of the equations of motion (2.325). There is. Verify the commutation rules

$$\{(-E^j/2), (-E^k/2)\} = (-E^\ell/2) \tag{11.2.327}$$

where j, k, ℓ are any cyclic permutation of 1,2,3. Evidently these commutation rules are the same as those for the K^j, see (3.7.172, and therefore we may make the correspondence

$$(-E^j/2) \leftrightarrow K^j. \tag{11.2.328}$$

Consequently, in view of (2.324), there is also the correspondence

$$A(\boldsymbol{\omega}^{bf}) \leftrightarrow -\boldsymbol{\omega}^{bf} \cdot \boldsymbol{K} = \bar{\boldsymbol{\omega}} \cdot \boldsymbol{K} \tag{11.2.329}$$

with

$$\bar{\boldsymbol{\omega}} = -\boldsymbol{\omega}^{bf}. \tag{11.2.330}$$

It follows that the $SU(2)$ analog of (2.325) is the differential equation

$$\dot{u}(t) = (\bar{\boldsymbol{\omega}} \cdot \boldsymbol{K})u(t) \tag{11.2.331}$$

with the initial conditions

$$u(t^0) = I \tag{11.2.332}$$

where u is a complex 2×2 matrix.

We know from Exercise 2.4 that $u(t)$, the solution to (2.331) with the initial condition (2.332), will be in $SU(2)$ for all t. Moreover from Exercise 2.5 we know that, because of the homomorphism between $SU(2)$ and $SO(3, \mathbb{R})$, the S associated with u by the rule (2.54) satisfies the equation of motion (2.43) with the initial condition (2.44). Thus, we have come full circle back to (2.43).

There are other interesting paths we can take. For example, Exercise 2.15 below shows that there is a connection between quaternion parameters and solutions to the general Schroedinger equation in a two-dimensional complex Hilbert space.

11.2.7. The aim of this exercise is to explore the relations between matrix differential equations of the form

$$\dot{M}(t) = M(t)A(t) \tag{11.2.333}$$

with the initial condition

$$M(0) = I, \tag{11.2.334}$$

and matrix differential equations of the form

$$\dot{N}(t) = \bar{A}(t)N(t) \tag{11.2.335}$$

with the initial condition

$$N(0) = I. \tag{11.2.336}$$

Begin with $N(t)$, the solution to (2.335) with the initial condition (2.336). Define a matrix $M(t)$ by the rule

$$M = N^{-1}, \tag{11.2.337}$$

from which it follows that

$$M(0) = I \tag{11.2.338}$$

and

$$MN = I. \tag{11.2.339}$$

Verify that differentiating both sides of (2.339) gives the result

$$\dot{M}N + M\dot{N} = 0, \tag{11.2.340}$$

and therefore

$$\dot{M} = -M\dot{N}N^{-1} = -M\bar{A}NN^{-1} = -M\bar{A}. \tag{11.2.341}$$

Here we have also used (2.335). Evidently (2.341) agrees with (2.333) providing we make the identification

$$A = -\bar{A}. \tag{11.2.342}$$

Let us pause for a side check on consistency with previous results. For the case of $SO(3, \mathbb{R})$, verify that what we have just found is consistent with the relations (1.18) and (1.121). For the Cayley parameterization (1.293) in the case of quadratic groups, verify that the substitution

$$M \leftrightarrow M^{-1} \tag{11.2.343}$$

is equivalent to the substitution

$$V \leftrightarrow -V. \tag{11.2.344}$$

Verify that, under the substitution (2.344) and the substitution

$$A \leftrightarrow -\bar{A}, \tag{11.2.345}$$

the relation (1.313) is transformed into the relation (1.320).

Continue on. The only possible difficulty in making the transition between the two cases (2.333) and (2.335) is the inversion (2.337). It can be shown that matrices that satisfy differential equations of the form (2.333) and (2.335) must generally be invertible, and therefore the transition is generally possible. Your next task is to prove this result.

Review Exercise 1.4.6. Using the methods of that exercise show, starting with (2.335), that there is the relation

$$\det N(t) = [\det N(t^0)] \exp[\int_{t^0}^{t} dr' \operatorname{tr} \bar{A}(t')]. \tag{11.2.346}$$

Assume that $\bar{A}(t')$ is finite for all t'. Show it follows that the factor $\exp[\int_{t^0}^{t} dr' \operatorname{tr} \bar{A}(t')]$ is finite and nonzero. Verify it follows that if $N(t^0)$ is invertible at any time t^0, $\det N(t^0) \neq 0$, then $N(t)$ is invertible at all times t. Show, starting with (2.333), that an analogous result holds for $M(t)$.

Although matrix inversion is generally computationally intensive, it can be carried out easily if M (and consequently also N) belong to various groups. Let us recall some groups for which inverse elements are easily computed. The unitary group comes first to mind. If M is unitary, then

$$M^{-1} = M^{\dagger}. \tag{11.2.347}$$

And, if M is orthogonal,

$$M^{-1} = M^T. \tag{11.2.348}$$

Finally, if M belongs to a quadratic group, then

$$M^{-1} = L^{-1}M^T L. \tag{11.2.349}$$

See (3.12.32).

11.2.8. The purpose of this exercise is to prove that the exact solution to (2.99) preserves G. Verify that (2.103) solves, over the interval $[t^n, t^{n+1}]$, the equation (2.99) through terms of order h. It may make local errors of order h^2, but verify that it preserves G. In the terminology of Sections 2.1 and 2.2, let $h \to 0$ and correspondingly $N \to \infty$ to find $Y(t^0+T)$. Verify that this result is the exact solution of (2.99) evaluated at $t = t^0 + T$ and that this result is in G.

11.2.9. We have claimed that the two-stage Lie Runge Kutta routine (2.109) through (2.113) has order $m = 2$. Verify, for example, that this is true in the case of (2.43), when (2.100) and (2.101) hold, and the Butcher tableau (2.3.36) is employed.

11.2.10. Verify that the entries in the Butcher tableau (2.124) satisfy the consistency relation (2.3.16), the order conditions (2.3.42) through (2.3.45), and the additional order 3 condition (2.122).

11.2.11. Verify that (2.156) through (2.159) follow from (2.87) through (2.90) and (2.152) through (2.155).

11.2.12. Verify (2.168) using (2.163), (2.166), and (T.1.29).

11.2.13. Emboldened by the remarkable results in Subsection 2.8, achieved by combining everything into one Lie element using the exponential map, the purpose to the exercise is to explore what happens when everything is combined into one Lie element using the Cayley map. We will call the result *Cayley Lie Runge Kutta*. Again we will concentrate our efforts on $SU(2)$ for ease of computation.

Begin by assuming that $\hat{u}^v(\tau)$ has a *Cayley Taylor* approximation of the form

$$\hat{u}^v(\tau) \simeq \hat{u}^{\text{vcay}}(\tau) = (I + \boldsymbol{\mu}\cdot\boldsymbol{K})/(I - \boldsymbol{\mu}\cdot\boldsymbol{K}) \tag{11.2.350}$$

where

$$\boldsymbol{\mu}(\tau) = \boldsymbol{f}_0\tau + \boldsymbol{f}_1\tau^2 + \cdots + \boldsymbol{f}_M\tau^{M+1}, \tag{11.2.351}$$

$$\boldsymbol{\mu}(H) = \boldsymbol{f}_0 H + \boldsymbol{f}_1 H^2 + \cdots + \boldsymbol{f}_M H^{M+1}, \tag{11.2.352}$$

and the coefficients $\boldsymbol{f}_n$ are to be determined. See (3.12.71).

The coefficients $\boldsymbol{f}_n$ could be determined in terms of the $\boldsymbol{c}_m$ by integrating equations of the form (11.1.71). [.] Alternatively, we may assume that we already know $\hat{u}^v(H)$ in exponential form and therefore can use a relation of the form (3.12.73). Specifically, show that we may write

$$\begin{aligned}\boldsymbol{\mu} &= 2(\boldsymbol{\Omega}/|\boldsymbol{\Omega}|)\tan(|\boldsymbol{\Omega}|/4)\\ &= 2(\boldsymbol{\Omega}/|\boldsymbol{\Omega}|)[(|\boldsymbol{\Omega}|/4) + (1/3)(|\boldsymbol{\Omega}|/4)^3 + (2/15)(|\boldsymbol{\Omega}|/4)^5 + \cdots]\\ &= (1/2)(\boldsymbol{\Omega})[1 + *|\boldsymbol{\Omega}|^2 + *|\boldsymbol{\Omega}|^4 + \cdots]\end{aligned} \tag{11.2.353}$$

with $\boldsymbol{\Omega}$ given, through terms of order H^4, by (2.162). Show , through terms of order H^4, that

$$|\boldsymbol{\Omega}|^2 =, \tag{11.2.354}$$

$$|\boldsymbol{\Omega}|^4 =, \tag{11.2.355}$$

Use these relations and (2.312) to obtain the results

$$\boldsymbol{f}_0 =, \tag{11.2.356}$$

$$\boldsymbol{f}_1 =, \tag{11.2.357}$$

$$\boldsymbol{f}_2 =, \tag{11.2.358}$$

$$\boldsymbol{f}_3 = . \tag{11.2.359}$$

11.2.14. Subsection 2.9 treated the integration of (2.183) in its Lie algebra using both exponential and Cayley representations. Carry out the analogous tasks for the equation of motion

$$\dot{Y}(t) = Y(t)A(Y,t). \tag{11.2.360}$$

Hint: Use (*) in Appendix C and (1.312) in Exercise 1.12.

11.2.15. The Schroedinger equation reads

$$d\psi/dt = (-iH/\hbar)\psi. \tag{11.2.361}$$

Here the state vector ψ is supposed to belong to some Hilbert space (a vector space equipped with an inner product), and $(-iH/\hbar)$ is supposed to be an anti-Hermitian operator with respect to this inner product. Consider the case for which the Hilbert space is two dimensional. Then ψ is a two-component vector, and we may take the inner product $\langle *, * \rangle$ to be the usual complex inner product for finite-dimensional vector spaces over the complex field. Also, $(-iH/\hbar)$ is then a 2×2 anti-Hermitian matrix. Show that the most general such matrix is of the form

$$-iH/\hbar = i\delta(t)I + \breve{\boldsymbol{\omega}}(t) \cdot \boldsymbol{K} \tag{11.2.362}$$

where δ is any real possibly time-dependent number and $\breve{\boldsymbol{\omega}}$ is any real possibly time-dependent three-dimensional vector. Thus, for a two-dimensional Hilbert space, the most general Schroedinger equation reads

$$d\psi/dt = [i\delta(t)I + \breve{\boldsymbol{\omega}}(t) \cdot \boldsymbol{K}]\psi. \tag{11.2.363}$$

Out task is to find $\psi(t)$ given

$$\psi^0 = \psi(t^0). \tag{11.2.364}$$

Since (2.172) is linear, we may write $\psi(t)$ in the form

$$\psi(t) = v(t)\psi^0 \tag{11.2.365}$$

where v is a 2×2 matrix to be determined. Show that the general solution to (2.172) is of the form (2.174) providing v satisfies the equation

$$\dot{v} = [i\delta(t)I + \breve{\boldsymbol{\omega}}(t) \cdot \boldsymbol{K}]v \tag{11.2.366}$$

with the initial condition

$$v(t^0) = I. \tag{11.2.367}$$

Verify that v defined by (2.175) and (2.176) is an element in $U(2)$, and therefore $\langle \psi, \psi \rangle$ is preserved as is necessary for a probability interpretation of the state vector.[33]

Suppose that the δ term in (2.172) and (2.175) is omitted to yield differential equations of the form

$$\dot{\psi}' = \breve{\boldsymbol{\omega}} \cdot \boldsymbol{K} \psi', \tag{11.2.368}$$

with the same initial condition

$$\psi'(t^0) = \psi^0, \tag{11.2.369}$$

and

$$\dot{v}' = \breve{\boldsymbol{\omega}}(t) \cdot \boldsymbol{K} v', \tag{11.2.370}$$

with the same initial condition

$$v'(t^0) = I. \tag{11.2.371}$$

We already know from Exercise 2.3 that v' defined by (2.179) and (2.180) is an element in $SU(2)$. Verify that

$$\psi'(t) = v'(t)\psi^0. \tag{11.2.372}$$

How are ψ and ψ', and v and v', related? Define a quantity $\Delta(t)$ by the rule

$$\Delta(t) = \int_{t^0}^{t} dt' \delta(t'). \tag{11.2.373}$$

Verify the relations

$$\psi(t) = \exp[i\Delta(t)]\psi'(t) \tag{11.2.374}$$

and

$$v(t) = \exp[i\Delta(t)]v'(t). \tag{11.2.375}$$

We see that ψ and ψ', and also v and v', differ only by a phase factor. Since overall phase factors are supposed to be unobservable in quantum mechanics, we conclude there is no loss in generality in omitting the δ term. Thus, without loss of quantum-mechanical generality, we may drop primes and require that v satisfies the differential equation

$$\dot{v} = \breve{\boldsymbol{\omega}}(t) \cdot \boldsymbol{K} v \tag{11.2.376}$$

with the initial condition (2.176). Correspondingly, the Schroedinger equation now reads

$$\dot{\psi} = \breve{\boldsymbol{\omega}} \cdot \boldsymbol{K} \psi, \tag{11.2.377}$$

and has the general solution (2.174) with $v \in SU(2)$.

By construction, the relations (2.185) and (2.186) entail each other. But we now observe that the relations (2.185) and (2.176) have the same form as (2.52) and (2.53). Therefore, relations of the form (2.52) and (2.53) may also be viewed as arising from the most general Schroedinger equation in the case of a two-dimensional Hilbert space.

[33] That is why $(-iH/\hbar)$ is required to be anti-Hermitian.

What can be said about the nature of two-dimensional Hilbert space? Let $\psi^{\uparrow}$ and $\psi^{\downarrow}$ be the orthonormal basis vectors

$$\psi^{\uparrow} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{11.2.378}$$

$$\psi^{\downarrow} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{11.2.379}$$

Write the most general ψ in the form

$$\psi = \alpha\psi^{\uparrow} + \beta\psi^{\downarrow}. \tag{11.2.380}$$

Since both α and β are complex, they each have the topology of $\mathbb{C} = E^2$, and together they have the topology of $\mathbb{C} \times \mathbb{C} = E^4$. Show that requiring that ψ be a unit vector, which is necessary for a probability interpretation of quantum mechanics, yields the restriction

$$|\alpha|^2 + |\beta|^2 = (\Re\alpha)^2 + (\Im\alpha)^2 + (\Re\beta)^2 + (\Im\beta)^2 = 1. \tag{11.2.381}$$

Observe that (2.190) is the equation for S^3, the three-dimensional surface of a sphere in E^4. Thus the topology of unit vectors in two-dimensional Hilbert space is that of S^3.

Show that the general solution to the restriction (2.190) is given by

$$\alpha = \exp(i\gamma)\cos(\theta/2), \tag{11.2.382}$$

$$\beta = \exp(i\gamma)\exp(i\phi)\sin(\theta/2), \tag{11.2.383}$$

where γ, θ, and ϕ are real, but otherwise arbitrary. Actually, not all values of γ, θ, and ϕ are required for full generality. Show that all possibilities are covered by making the restrictions $\gamma \in [0, 2\pi]$, $\theta \in [0, \pi]$, $\phi \in [0, 2\pi]$. Moreover, since overall phase factors are unobservable, we may set $\gamma = 0$ if we wish, and without loss of quantum-mechanical generality. Thus, the set (collection of equivalence classes) of all *unit rays* can be written in the form

$$\psi = \cos(\theta/2)\psi^{\uparrow} + \exp(i\phi)\sin(\theta/2)\psi^{\downarrow} \tag{11.2.384}$$

with

$$\theta \in [0, \pi] \text{ and } \phi \in [0, 2\pi]. \tag{11.2.385}$$

(Recall that a unit ray is the equivalence class of a unit vector multiplied by an arbitrary phase factor.) The quantities θ and ϕ with the restrictions (2.194) may be regarded as the polar angles for points on S^2, the unit sphere in E^3. That is, in terms of Cartesian coordinates for E^3, we may write

$$x_1 = \sin(\theta)\cos(\phi), \tag{11.2.386}$$

$$x_2 = \sin(\theta)\sin(\phi), \tag{11.2.387}$$

$$x_3 = \cos(\theta). \tag{11.2.388}$$

Thus, the topology of unit rays in two-dimensional Hilbert space is that of S^2. In the context of quantum mechanics, this S^2 is frequently called the *Bloch sphere*. (It is the Poincaré sphere in the the context of describing polarized light.) Show that the north pole $(0, 0, 1)$ of the

Bloch sphere corresponds to the state $\psi^{\uparrow}$ and the south pole $(0, 0, -1)$ corresponds (up to an irrelevant phase factor) to the state $\psi^{\downarrow}$. Show that points on the equator correspond to the states $(1/\sqrt{2})(\psi^{\uparrow} + \exp(i\phi)\psi^{\downarrow})$. Show that the vectors ψ corresponding to diametrically opposite points on the Bloch sphere are orthogonal.

Let $\psi(\theta, \phi)$ denote the vector given by (2.384). Show that, in terms of the Euler-angle parameterization given by (3.7.194) and (3.7.195), there is the relation

$$\psi(\theta, \phi) = v(\phi, \theta, -\phi)\psi^{\uparrow}. \tag{11.2.389}$$

Let $\boldsymbol{n}$ denote the unit vector defined by the relation

$$\boldsymbol{n} = R(\phi, \theta, -\phi)\boldsymbol{e}_3 = \cos\phi\sin\theta\boldsymbol{e}_1 + \sin\phi\sin\theta\boldsymbol{e}_2 + \cos\theta\boldsymbol{e}_3. \tag{11.2.390}$$

See (3.7.208) and note the resemblance between (2.390) and (2.386) through (2.388). Show that there is the relation

$$(\boldsymbol{n} \cdot \boldsymbol{\sigma})\psi(\theta, \phi) = \psi(\theta, \phi). \tag{11.2.391}$$

That is, $\psi(\theta, \phi)$ is an eigenvector of $\boldsymbol{n} \cdot \boldsymbol{\sigma}$ with eigenvalue $+1$.

Verify the coset relation

$$SU(2)/U(1) = S^2. \tag{11.2.392}$$

See Subsection 5.12..4. Show that the correspondence between unit rays in a two-dimensional Hilbert space and points in S^2 is a consequence of this coset relation.

We are ready for a parenthetical remark about the field of quantum computing and quantum information. Let $|0\rangle$ and $|1\rangle$ be the *qubit* (quantum bit) states corresponding to the states 0 and 1 of a classical bit.[34] In this field it is conventional to define the vectors $|0\rangle$ and $|1\rangle$ by the relations

$$|0\rangle = \psi^{\uparrow}, \tag{11.2.393}$$

$$|1\rangle = \psi^{\downarrow}; \tag{11.2.394}$$

and a general qubit state $|\psi\rangle$ takes the superposition form

$$|\psi\rangle = \cos(\theta/2)|0\rangle + \exp(i\phi)\sin(\theta/2)|1\rangle. \tag{11.2.395}$$

Back to the main thread. There is still more to be said. Define a three-component vector $\boldsymbol{s}$ by the rule

$$\boldsymbol{s} = 2i\langle\psi, \boldsymbol{K}\psi\rangle = \langle\psi, \boldsymbol{\sigma}\psi\rangle. \tag{11.2.396}$$

Show that $\boldsymbol{s}$ is real, and does not depend on the phase of ψ. Show that using (2.193) for ψ and evaluating (2.201) gives the results (2.10) through (2.12). Thus, $\boldsymbol{s}$ is a unit vector with polar angles θ and ϕ. We see that $\boldsymbol{s}$ is a point on the Bloch sphere, and can be any point on the Bloch sphere. Conversely, given any point on the Bloch sphere, we can determine its polar angles θ and ϕ, and from these angles we can determine ψ up to an overall phase factor.

[34] A qubit is a quantum-mechanical system that is well described by a two-dimensional Hilbert space. Examples include spin 1/2 particles and polarized light, plus more complicated systems, including atoms and superconducting devices, that can be effectively regarded as two-state systems consisting of a ground state and first excited state, with higher excited states ignorable because they are well separated in energy from these states.

Therefore, for a two-dimensional Hilbert space, knowledge of a unit ray and knowledge of a unit vector $\boldsymbol{s}$ are equivalent.

Finally, suppose ψ evolves according to (2.186). How does the corresponding $\boldsymbol{s}$ evolve? Show, using (2.186) and (2.201), that in terms of components

$$\begin{aligned} \dot{s}_\alpha &= 2i[\langle\dot{\psi}, K^\alpha\psi\rangle + \langle\psi, K^\alpha\dot{\psi}\rangle] \\ &= 2i[\langle\breve{\boldsymbol{\omega}}\cdot\boldsymbol{K}\psi, K^\alpha\psi\rangle + \langle\psi, K^\alpha\breve{\boldsymbol{\omega}}\cdot\boldsymbol{K}\psi\rangle] \\ &= 2i[\langle\psi, (\breve{\boldsymbol{\omega}}\cdot\boldsymbol{K})^\dagger K^\alpha\psi\rangle + \langle\psi, K^\alpha\breve{\boldsymbol{\omega}}\cdot\boldsymbol{K}\psi\rangle] \\ &= 2i[\langle\psi, (-\breve{\boldsymbol{\omega}}\cdot\boldsymbol{K})K^\alpha\psi\rangle + \langle\psi, K^\alpha\breve{\boldsymbol{\omega}}\cdot\boldsymbol{K}\psi\rangle] \\ &= 2i\langle\psi, \{K^\alpha, \breve{\boldsymbol{\omega}}\cdot\boldsymbol{K}\}\psi\rangle. \end{aligned} \tag{11.2.397}$$

Next verify that

$$\{K^\alpha, \breve{\boldsymbol{\omega}}\cdot\boldsymbol{K}\} = \sum_\beta \{K^\alpha, K^\beta\}\breve{\omega}_\beta = \sum_{\beta\gamma}\epsilon_{\alpha\beta\gamma}\breve{\omega}_\beta K^\gamma. \tag{11.2.398}$$

Now combine (2.202) and (2.203) to show that

$$\dot{s}_\alpha = 2i\sum_{\beta\gamma}\epsilon_{\alpha\beta\gamma}\breve{\omega}_\beta\langle\psi, K^\gamma\psi\rangle = \sum_{\beta\gamma}\epsilon_{\alpha\beta\gamma}\breve{\omega}_\beta s_\gamma = (\breve{\boldsymbol{\omega}}\times\boldsymbol{s})_\alpha. \tag{11.2.399}$$

In vector notation, you have demonstrated that

$$\dot{\boldsymbol{s}} = \breve{\boldsymbol{\omega}}\times\boldsymbol{s}, \tag{11.2.400}$$

a relation of the form (2.1). Moreover, since a knowledge of $\boldsymbol{s}$ determines ψ up to a phase, given the equation (2.205) we may view it as arising from the Schroedinger equation (2.186). That is, given the $\breve{\boldsymbol{\omega}}$ appearing in (2.205), we may insert it into the Schroedinger equation (2.186) to find $\psi(t)$, and this $\psi(t)$ will yield $\boldsymbol{s}(t)$ by way of (2.201). Thus, we may also view (2.186) and (2.205) as entailing each other.

Your last task in this exercise is to verify a relation between components of ψ and quaternion parameters w. Let $\psi^u(t)$ and $\psi^d(t)$ be solutions to the Schroedinger equation (2.186) with the initial conditions

$$\psi^u(t^0) = \psi^\uparrow, \tag{11.2.401}$$

$$\psi^d(t^0) = \psi^\downarrow. \tag{11.2.402}$$

Here the superscript mnemonics u and d stand for *up* and *down*. Suppose v is parameterized in terms of quaternions in analogy to (1.132). Verify, using (1.174), (2.187), (2.188), and (5.10.29), that there are the relations

$$\langle\psi^\uparrow, \psi^u\rangle = v_{11} = w_0 + iw_3, \tag{11.2.403}$$

$$\langle\psi^\uparrow, \psi^d\rangle = v_{12} = iw_1 + w_2, \tag{11.2.404}$$

$$\langle\psi^\downarrow, \psi^u\rangle = v_{21} = iw_1 - w_2, \tag{11.2.405}$$

$$\langle\psi^\downarrow, \psi^d\rangle = v_{22} = w_0 - iw_3. \tag{11.2.406}$$

Solve (2.208) through (2.211) to yield the relations

$$w_0(t) = (1/2)(\langle\psi^\uparrow, \psi^u\rangle + \langle\psi^\downarrow, \psi^d\rangle), \tag{11.2.407}$$

$$w_1(t) = (-i/2)(\langle\psi^\uparrow, \psi^d\rangle + \langle\psi^\downarrow, \psi^u\rangle), \tag{11.2.408}$$

$$w_2(t) = (1/2)(\langle\psi^\uparrow, \psi^d\rangle - \langle\psi^\downarrow, \psi^u\rangle) \tag{11.2.409}$$

$$w_3(t) = (-i/2)(\langle\psi^\uparrow, \psi^u\rangle - \langle\psi^\downarrow, \psi^d\rangle). \tag{11.2.410}$$

11.2.16. Exact solutions to (2.52).

11.2.17. Consider a Stern-Gerlach apparatus in which the beam propagates in the y direction, the main magnetic field is in the z direction, and also has a gradient in the z direction. Such a field, when expanded about the beam axis (taken to be $x = z = 0$) and near the beam axis, consists essentially of a (skew) quadrupole field superimposed on a dipole field. That is, ignoring end effects, the magnetic field has, to the lowest nontrivial order, the expansion

$$\boldsymbol{B}(\boldsymbol{r}) = B^d\boldsymbol{e}_z + Q^q(z\boldsymbol{e}_z - x\boldsymbol{e}_x). \tag{11.2.411}$$

Here B^d is the strength of the main (dipole) field, and Q^q is the strength (field gradient) of the quadrupole field. Verify that this field is divergence and curl free as is desired. Note that there is a desired field gradient in the z direction, intended to produce a Stern-Gerlach force along the z direction, as well as an “undesirable” gradient along the x direction that will produce a Stern-Gerlach force along the x direction.

$$\Psi = f^\uparrow(\boldsymbol{r}, t)\psi^\uparrow + f^\downarrow(\boldsymbol{r}, t)\psi^\downarrow. \tag{11.2.412}$$

The Schroedinger equation reads

$$\partial\Psi/\partial t = (-i/\hbar)H\Psi \tag{11.2.413}$$

where

$$H = \boldsymbol{p}\cdot\boldsymbol{p}/(2m) + \mu\boldsymbol{B}(\boldsymbol{r})\cdot\boldsymbol{\sigma}. \tag{11.2.414}$$

Here m is the particle mass and μ is a measure of its magnetic moment. Indeed, suppose we ignore the kinetic energy term in (2.414). Then we find that

$$-iH/\hbar \approx -i(\mu/\hbar)\boldsymbol{B}(\boldsymbol{r})\cdot\boldsymbol{\sigma} = (2\mu/\hbar)\boldsymbol{B}(\boldsymbol{r})\cdot\boldsymbol{K}. \tag{11.2.415}$$

Upon comparing (2.362) with (2.415) we see that we should make the identification

$$\breve{\boldsymbol{\omega}} = (2\mu/\hbar)\boldsymbol{B}. \tag{11.2.416}$$

$$\partial f^\uparrow(\boldsymbol{r}, t)/\partial t = (-i/\hbar)\{[\hbar^2/(2m)]\nabla^2 f^\uparrow(\boldsymbol{r}, t) + \mu(B^d + Q^q z)f^\uparrow(\boldsymbol{r}, t) + \mu Q^q x f^\downarrow(\boldsymbol{r}, t)\}, \tag{11.2.417}$$

$$\partial f^\downarrow(\boldsymbol{r}, t)/\partial t = (-i/\hbar)\{[\hbar^2/(2m)]\nabla^2 f^\downarrow(\boldsymbol{r}, t) - \mu(B^d + Q^q z)f^\downarrow(\boldsymbol{r}, t) + \mu Q^q x f^\uparrow(\boldsymbol{r}, t)\}. \tag{11.2.418}$$

11.3 Numerical Integration on Manifolds: Charged Particle Motion in a Static Magnetic Field

Overview

Sections 1.1 and 1.2 treated the problem of integrating on manifolds largely either by projection or by integration in the Lie algebra or by parameterizing the manifold in question, finding the associated differential equation for the parameters, and integrating these equations using standard integration algorithms such as those described in Chapter 2. The only exception to this approach was the work of Subsections 2.6 through 2.8. The purpose of the section is to describe how manifold-preserving integration methods may be applied to the problem of charged-particle motion in a static magnetic field.

The reader may have been somewhat puzzled by the assertion (made at the beginning of Section 1.2) that (1.80) was like (1.79). The equation of motion (1.6.112) is equivalent to the pair

$$d\boldsymbol{r}/dt = \boldsymbol{v}, \tag{11.3.1}$$

$$d\boldsymbol{v}/dt = \bar{\boldsymbol{\omega}}(\boldsymbol{r}) \times \boldsymbol{v}, \tag{11.3.2}$$

with

$$\bar{\boldsymbol{\omega}}(\boldsymbol{r}) = -(q/m^*)\boldsymbol{B}(\boldsymbol{r}). \tag{11.3.3}$$

Since $\bar{\boldsymbol{\omega}}$ depends on $\boldsymbol{r}$ and, according to (1.148), $\boldsymbol{r}$ in turn depends on $\boldsymbol{v}$, the $\bar{\boldsymbol{\omega}}$ appearing in (1.149) is *not* a given function of t independent of everything else including $\boldsymbol{v}$. However, it is the case that the pair (1.148) and (1.149) does preserve the quantity $\boldsymbol{v} \cdot \boldsymbol{v}$. That is, the pair preserves the manifold

$$\Gamma = E^3 \times S^{2*} \tag{11.3.4}$$

with $\boldsymbol{r} \in E^3$ and $\boldsymbol{v} \in S^{2*}$. Here S^{2*} denotes a 2-sphere whose radius is given by $v^* = |\boldsymbol{v}^0|$, or equivalently, is determined by $\gamma = m^*/m$ with

$$\gamma = 1/\sqrt{1 - |\boldsymbol{v}^0|^2/c^2}. \tag{11.3.5}$$

See (1.6.113) and (1.6.114).

In Subsection 3.1 we will describe how the methods already developed in Sections 1 and 2 can be exploited to provide numerical integrators that preserve Γ. In Subsection 3.2 we will describe splitting methods that also preserve Γ but are more akin to some of the methods for symplectic integration to be described in Chapter 12.

11.3.1 Exploitation of Previous Results

The purpose of this subsection is to describe how the methods of Sections 1 and 2 can be applied to the computation of charged-particle motion in static magnetic fields. We will begin with the use of local tangent-space coordinates as illustrated in Subsection 2.3.

Use of Local Tangent-Space Coordinates

In analogy with the work of Subsection 2.3, let $\boldsymbol{v}^b$ be the velocity at the *beginning* of an integration step, and write

$$\boldsymbol{v}(t) = \boldsymbol{v}^b + \boldsymbol{v}^v(t) \tag{11.3.6}$$

with

$$\boldsymbol{v}^b = \boldsymbol{v}(t^b) \tag{11.3.7}$$

and

$$\boldsymbol{v}^v(t^b) = 0. \tag{11.3.8}$$

(Here we apologize for our notation: As before, when v appears as a superscript, it stands for *variable.* Elsewhere it denotes vector or scalar *velocity.*) It follows that we may also write

$$\boldsymbol{v}^v(t) = v_1^{vf}(t)\boldsymbol{f}_1 + v_2^{vf}(t)\boldsymbol{f}_2 + v_3^{vf}(t)\boldsymbol{f}_3 \tag{11.3.9}$$

and

$$\boldsymbol{v}(t) = [v^* + v_1^{vf}(t)]\boldsymbol{f}_1 + v_2^{vf}(t)\boldsymbol{f}_2 + v_3^{vf}(t)\boldsymbol{f}_3 \tag{11.3.10}$$

with

$$v_1^{vf}(t) = \{(v^*)^2 - [v_2^{vf}(t)]^2 - [v_3^{vf}(t)]^2\}^{1/2} - v^*. \tag{11.3.11}$$

Here, *mutatis mutandis*, the vectors $\boldsymbol{f}_j$ are constructed as in Subsection 2.3.

We have (locally) parameterized S^{2*} in terms of v_2^{vf} and v_3^{vf}. Put another way, we have changed variables from $\boldsymbol{v}$ to v_2^{vf} and v_3^{vf} in such a way that the S^{2*} manifold condition is built in. Corresponding to this change of variables, the $\boldsymbol{v}$ equations of motion (3.2) becomes the pair

$$\dot{v}_2^{vf}(t) = \bar{\omega}_3^f(\boldsymbol{r})\{(v^*)^2 - [v_2^{vf}(t)]^2 - [v_3^{vf}(t)]^2\}^{1/2} - \bar{\omega}_1^f(\boldsymbol{r})v_3^{vf}(t), \tag{11.3.12}$$

$$\dot{v}_3^{vf}(t) = \bar{\omega}_1^f(\boldsymbol{r})v_2^{vf}(t) - \bar{\omega}_2^f(\boldsymbol{r})\{(v^*)^2 - [v_2^{vf}(t)]^2 - [v_3^{vf}(t)]^2\}^{1/2}. \tag{11.3.13}$$

For the $\boldsymbol{r}$ variables make the decomposition

$$\boldsymbol{r}(t) = \boldsymbol{r}^b + \boldsymbol{r}^v(t) \tag{11.3.14}$$

with

$$\boldsymbol{r}^b = \boldsymbol{r}(t^b) \tag{11.3.15}$$

and

$$\boldsymbol{r}^v(t^b) = 0. \tag{11.3.16}$$

For $\boldsymbol{r}^v$ make the expansion

$$\boldsymbol{r}^v = r_1^{vf}\boldsymbol{f}_1 + r_2^{vf}\boldsymbol{f}_2 + r_3^{vf}\boldsymbol{f}_3. \tag{11.3.17}$$

Then, in view of (3.13) and (3.14), the equations of motion (3.1) for $\boldsymbol{r}$ become the set

$$\dot{r}_1^{vf} = v_1^{vf} = \{(v^*)^2 - [v_2^{vf}(t)]^2 - [v_3^{vf}(t)]^2\}^{1/2} - v^*, \tag{11.3.18}$$

$$\dot{r}_2^{vf} = v_2^{vf}, \tag{11.3.19}$$

$$\dot{r}_3^{vf} = v_3^{vf}. \tag{11.3.20}$$

It is these five equations that are to be numerically integrated from the time t^b to the time t^b+h (or perhaps t^b+kh) starting with the initial conditions $v_2^{vf}(t^b)=v_3^{vf}(t^b)=0$ and $r_1^{vf}(t^b)=r_2^{vf}(t^b)=r_3^{vf}(t^b)=0$. Then, once $\boldsymbol{v}^v(t^b+h)$ and $\boldsymbol{r}^v(t^b+h)$ [or perhaps $\boldsymbol{v}^v(t^b+kh)$ and $\boldsymbol{r}^v(t^b+kh)$] have been obtained, $\boldsymbol{v}(t^b+h)$ and $\boldsymbol{r}(t^b+h)$ [or perhaps $\boldsymbol{v}(t^b+kh)$ and $\boldsymbol{r}(t^b+kh)$] are given by (2.20) and * .[35] At this point, the whole process just described is repeated as often as desired. That is, the vectors $\boldsymbol{f}_j$ are reconstructed based on the most recently obtained $\boldsymbol{v}$, etc. Note that any numerical integration method may be used to carry out the desired integration. By construction, although the numerical integration may make local errors of order h^{m+1} in finding $\boldsymbol{v}(t^b+h)$ and $\boldsymbol{r}(t^b+h)$, these quantities are guaranteed to be in the manifold $\Gamma=E^3\times S^{2*}$ to machine precision.

Use of Connection with Rigid-Body Kinematics

An alternative to the use of local tangent-space coordinates to ensure that $\boldsymbol{v}$ and $\boldsymbol{r}$ lie in Γ is to make an Ansatz for $\boldsymbol{v}$ that involves rotations. We will seek to use the results of Subsections 2.4 and 2.5.

Again let $\boldsymbol{v}^b$ be the velocity at the *beginning* of an integration step,

$$\boldsymbol{v}^b=\boldsymbol{v}(t^b), \tag{11.3.21}$$

and write

$$\boldsymbol{v}(t)=R(t;\boldsymbol{v}^b,\boldsymbol{r}^b)\boldsymbol{v}^b \tag{11.3.22}$$

with and $R(t;\boldsymbol{v}^b,\boldsymbol{r}^b)$ being a rotation to be determined subject to the initial condition

$$R(t^b;\boldsymbol{v}^b,\boldsymbol{r}^b)=I. \tag{11.3.23}$$

At this point, explanations are in order both about the notation $R(t;\boldsymbol{v}^b,\boldsymbol{r}^b)$ and the generality of the Ansatz (3.27). The notation is meant to indicate that the relation (3.27) between $\boldsymbol{v}(t)$ and $\boldsymbol{v}^b$ need not (and generally will not) be linear because $R(t;\boldsymbol{v}^b,\boldsymbol{r}^b)$ can depend on $\boldsymbol{v}^b$ and $\boldsymbol{r}^b$.

11.3.2 Splitting: Exploitation of Future Results

Exercises

However, before doing so, let us consider the cost of implementing (1.156). The evaluation of $\exp[h\bar{\boldsymbol{\omega}}(\boldsymbol{r}^n)\cdot\boldsymbol{L}]$ can be performed using the Rodrigues formula (3.7.202). This evaluation is fairly expensive because it involves, among other the things, the evaluation of trigonometric functions. Alternatively, (1.153) could be rewritten in the form

$$\boldsymbol{v}^{n+1}=[I+h\bar{\boldsymbol{\omega}}(\boldsymbol{r}^n)\cdot\boldsymbol{L}]\boldsymbol{v}^n. \tag{11.3.24}$$

To the same accuracy, one could orthogonalize the matrix $[I+h\bar{\boldsymbol{\omega}}(\boldsymbol{r}^n)\cdot\boldsymbol{L}]$ using one of the methods of Subsection 3.6.4, and then apply this matrix to $\boldsymbol{v}^n$ to obtain $\boldsymbol{v}^{n+1}$. These

[35] If $k>1$ is attempted, one must monitor $[(v_2^{vf})^2+(v_3^{vf})^2]$ to ensure that the square root singularity in (2.30) is not approached too closely.

methods involve the computation of square roots. Finally, as already mentioned, at each step one could simply renormalize the $\boldsymbol{v}^{n+1}$ in (1.153) by simple scaling to project it back onto S^{2*}. So doing requires the evaluation of a square root.

11.3.1. Exercise on the growth of v dot v based on Euler result (1.154).

Bibliography

General References. See also the references for Chapter 12

[1] E. Celledoni and B. Owren, “Lie group methods for rigid body dynamics and time integration on manifolds”, *Computer Methods in Applied Mechanics and Engineering* **192**, 421–438 (2003).

[2] E. Celledoni, H. Marthinsen, and B. Owren, “An introduction to Lie group integrators—basics, new developments, and applications”, http://arxiv.org/pdf/1207.0069.pdf.

[3] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure Preserving Algorithms for Ordinary Differential Equations*, Second Edition, Springer (2006).

[4] H. Munthe-Kaas, A. Zanna, “Numerical integration of differential equations on homogeneous manifolds”, *Foundations of Computational Mathematics*, Springer Verlag (1997).

[5] H. Munthe-Kaas and B. Owren, “Computations in a free Lie algebra”, *Phil. Trans. R. Soc. Lond. A* **357**, 957–981 (1999).

[6] A. Iserles, H. Munthe-Kaas, S. Nørsett, and A. Zanna, “Lie-group methods”, *Acta Numerica* **14**, 1–148 (2005). Also available on the Web at http://www.damtp.cam.ac.uk/user/na/NA_papers/NA2000_03.pdf.

Factored Lie Runge Kutta

[7] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure Preserving Algorithms for Ordinary Differential Equations*, Second Edition, Springer (2006). See Section IV.8.

[8] P.E. Crouch and R. Grossman, “Numerical integration of ordinary differential equations on manifolds”, *J. Nonlinear Sci.* **3**, 1 (1993).

[9] B. Owren and A Marthinsen, “Runge-Kutta methods adapted to manifolds and based on rigid frames ”, *BIT* **39**, 116 (1999).

Magnus Lie Runge Kutta

[10] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure Preserving Algorithms for Ordinary Differential Equations*, Second Edition, Springer (2006). See Section IV.7.

[11] A. Zanna, "Collocation and relaxed collocation for the Fer and the Magnus expansions", *SIAM J. Numer. Anal* **36**, 1145 (1999).

[12] S. Blanes, F. Casas, and J. Ros, "Improved high order integrators based on the Magnus expansion", *BIT* **40**, 434 (2000).

Spin and Qubits

[13] S. Mane, Yu. Shatunov, and K. Yokoya, "Spin-polarized charged particle beams in high-energy accelerators", *Rep. Prog. Phys.* **68**, 1997 (2005).

[14] D.P. Barber, References to and lectures on spin: http://www.desy.de/~mpybar/, http://www.desy.de/~mpybar/psdump/Combined_CI_2012_16.pdf.

[15] M. Nielsen and I. Chuang, *Quantum Computation and Quantum Information*, Cambridge (2000).

Stern-Gerlach Effect

[16] N. Mott and H. Massey, *The Theory of Atomic Collisions*, Third Edition, Oxford (1965).

[17] J. Kessler, *Polarized Electrons*, Second Edition, Springer-Verlag (1985).

[18] H. Dehmelt, "Continuous Stern-Gerlach effect: Principle and idealized apparatus", *Proc. Natl. Acad. Sci. USA*, Vol. 83, pp. 2291-2294 (1986).

[19] H. Batelaan, T. Gay, and J. Schwendiman, "Stern-Gerlach Effect for Electron Beams", *Phys. Rev. Let.* **79**, pp. 4517-4521 (1997).

[20] G. Gallup, H. Batelaan, and T. Gay, "Quantum-Mechanical Analysis of a Longitudinal Stern-Gerlach Effect", *Phys. Rev. Let.* **86**, pp. 4508-4511 (2001).

[21] G. Werth, H. Häffner, and W. Quint, "Continuous Stern-Gerlach Effect on Atomic Ions", *Advances in Atomic, Molecular, and Optical Physics*, Vol. 48, pp. 191-217 (2002).

[22] W. Quint, J. Alonso, S. Djekić, H.-J. Kluge, S. Stahl, T. Valenzuela, J. Verdú, M. Vogel, and G. Werth, "Continuous Stern-Gerlach effect and the magnetic moment of the antiproton", *Nucl. Instru. Meth. Phys. Res.* B, Vol. 214, pp. 207-210 (2004).

[23] M. Kohda, S. Nakamura, Y. Nishihara, K. Lobayashi, T. Ono, J. Ohe, Y. Tokuram, T. Mineno, and J. Nitta, "Spin-orbit induced electronic spin separation in semiconductor nanostructures", *Nature Communications* **3** (2012).

Chapter 12

Geometric/Structure-Preserving Integration: Symplectic Integration

Overview

Imagine we wish to compute, by numerical integration, a trajectory governed by some Hamiltonian. Suppose z^i is an initial condition and z^f is an associated final condition. A numerical integrator is called a *symplectic integrator* for this Hamiltonian if the relation between z^i and z^f produced by use of this integrator is (to within roundoff errors) a symplectic map. Sometimes we will want an integrator to be a symplectic integrator for general Hamiltonians. Often, as we will see, it suffices to have a symplectic integrator for some class of Hamiltonians.

There are many things that might be said about symplectic integrators. Indeed, books have been and are being written on the subject. However, we must limit our discussion to a single chapter. Are symplectic integrators important, and if so, why? The answers to these questions are not fully known. As we have seen in Chapter 6, the production of symplectic maps is a key feature of Hamiltonian systems, and the preservation of this feature by any approximation scheme, including numerical integration, would appear to be highly desirable. To date, much experience with symplectic integrators, particularly when one is concerned with studying the long-term behavior of trajectories, seems to confirm this belief. A particular aspect of this subject is discussed in Chapter 34. For a broader discussion, see the references listed at the end of the present chapter. However, as might be feared, satisfying the symplectic condition comes at a cost. We will see that in general mth-order but exactly symplectic integrators require much more work (many more function evaluations) than the mth-order methods of Chapter 2, and are therefore considerably slower.[1]

What can be said about the numerical integration methods of Chapter 2 when applied to Hamiltonian systems? In general, they are not symplectic. Typically, at each step, they violate the symplectic condition by an amount of order h^{m+1} if the integration method is

[1] It is sometimes argued that this cost is compensated by the possibility of using larger step sizes in symplectic integration. Although the solution thereby obtained may not be particularly accurate, it is at least qualitatively correct whereas solutions obtained by other integration methods may exhibit spurious damping or spurious growth. Whether this trade-off can be realized or is acceptable depends on the problem being considered.

locally correct through terms of order h^m. Consequently, they are not exactly symplectic for any finite value of h.[2]

Are there modifications of the integration methods of Chapter 2 that make them symplectic? It is known that, for *general* Hamiltonians, there are no *explicit* Runge-Kutta methods that are symplectic. See Exercise 3.1. However, we will learn that there are *implicit* Runge-Kutta methods that are symplectic. Like corrector methods, at each step implicit Runge-Kutta methods require iteration, which may be slow, in order to solve a set of implicit equations.

What about the usual predictor-corrector finite-difference methods? It is known that they cannot be modified to be symplectic. However, as will be discussed, it is possible to use a predictor (that employs, as usual, previously stored trajectory data) along with an implicit symplectic Runge-Kutta method that serves the role of a corrector.

Finally, little seems to be known about possible symplectic modifications of extrapolation methods.

12.1 Splitting, $T + V$ Splitting, and Zassenhaus Formulas

This chapter is devoted to symplectic integration and the use of Zassenhaus formulas. In this section we begin with some background material, and then explore the case where the Hamiltonian is of the special but frequently encountered form $H = T + V$.

We have seen that if $H(z)$ is a time-independent (autonomous) Hamiltonian, the transfer map $\mathcal{M}$ associated with H can be written formally as

$$\mathcal{M} = \exp(t : -H :). \tag{12.1.1}$$

(Here, for convenience, we have taken the initial time to be $t = 0$, which can be done without loss of generality since H by assumption does not depend on t.) Suppose H is time dependent. Then we know that we may introduce a new independent variable τ, extend the phase space to include t and p_t as dynamical variables, and introduce the new Hamiltonian K defined by the relation

$$K(t, q; p_t, p) = p_t + H(q, p, t). \tag{12.1.2}$$

(See Exercise 1.6.5.) Since K does not depend on τ, we may write the transfer map associated with K formally as

$$\mathcal{M} = \exp(\tau : -K :). \tag{12.1.3}$$

At this point a remark is in order: Because (1.1) and (1.3) have identical structures, there seems to be no loss of generality in considering only the autonomous case. This assertion is correct if we have no particular concern about the form of the Hamiltonian. However, as seen in the last chapter, for the case (1.1) we are able to employ certain methods under the assumptions that H can be expanded about the origin $z = 0$ and the term H_1 is absent or

[2] Of course, in the limit $h \to 0$ they are symplectic because they are then exact. They are also symplectic to machine precision, ignoring round-off error, when h is small enough for the integrator to be accurate to machine precision.

small, and these methods make it possible to find expansions for $\mathcal{M}$ that can be evaluated explicitly. By contrast, these methods fail for the Hamiltonian (1.2) because the presence in K of the (linear) term p_t with a coefficient of *one* means that K_1 cannot be regarded as being small. Nevertheless, if a method is capable of handling sufficiently general Hamiltonians, then there is no loss in generality in considering only autonomous Hamiltonians. Such will be the case for the methods in this section and therefore, without loss in generality, we assume that H is time independent.

As in Section 2.1, let us subdivide the t or τ axis, whichever the case may be, into equal steps of duration h. Then, again employing the notation of Chapter 2, we have the exact marching rule

$$z^{n+1} = \exp(h : -H :)z^n. \tag{12.1.4}$$

Here H is to be viewed as a function of the variables z^n.

Equation (1.4) provides a stepping formula that can be used for numerical integration providing we have some method of computing or approximating $\exp(h : -H :)$. Suppose the Hamiltonian H can be *split* into (written as the sum of) two terms A and B,

$$H(q,p) = A(q,p) + B(q,p), \tag{12.1.5}$$

in such a way that both the maps $\exp(-h : A :)$ and $\exp(-h : B :)$ can be evaluated explicitly or have some other desirable property. For example, suppose that H can be written as a sum of kinetic and potential energies,

$$H(q,p) = T(p) + V(q). \tag{12.1.6}$$

(See Exercise 1.1 for a review of when this is possible.) Then we have the exact results

$$\exp(-h : T :)q_i = q_i + h\partial T/\partial p_i, \tag{12.1.7}$$

$$\exp(-h : T :)p_i = p_i, \tag{12.1.8}$$

$$\exp(-h : V :)q_i = q_i, \tag{12.1.9}$$

$$\exp(-h : V :)p_i = p_i - h\partial V/\partial q_i. \tag{12.1.10}$$

Consider making the simple Zassenhaus approximation

$$\exp(h : -H :) = \exp(-h : A : -h : B :) \simeq \exp(-h : A :)\exp(-h : B :). \tag{12.1.11}$$

See Section 8.8. From the BCH formula (3.7.34) we have the result

$$\exp(-h : A :)\exp(-h : B :) = \exp[-h : A : -h : B : +(h^2/2)\{: A :, : B :\} + O(h^3)]. \tag{12.1.12}$$

We see that, as a stepping formula, (1.4) with the approximation (1.11) makes local errors of order h^2. Thus, like the crude Euler method of Section 2.2, it can be used for numerical integration providing the step size is made sufficiently small. However, unlike crude Euler, the combination of (1.4) and (1.11) provides a method that is *exactly symplectic*. That is, the relation between z^f and z^i produced by this method is (apart from numerical roundoff errors) exactly a symplectic map for *any* value of h. See Exercise 1.2.

With this general background in mind, we now turn to the task of improving the approximation (1.11) to obtain integrators that are again exactly symplectic, but have higher order (in h) accuracy. We will seek improved formulas of the Zassenhaus type, but will discover a method that is, in fact, more general. From the BCH formula we find the result

$$\begin{aligned}&\exp[-(h/2) : A :] \exp(-h : B :) \exp[-(h/2) : A :]\\ &= \exp[h : -(A+B) : + h^3(\{: A :, \{: A :, : B :\}\}/24\\ &-\{: B :, \{: B :, : A :\}\}/12) + O(h^4)].\end{aligned} \tag{12.1.13}$$

Consequently, making the approximation

$$\exp(h : -H :) \simeq \exp[-(h/2) : A :] \exp(-h : B :) \exp[-(h/2) : A :], \tag{12.1.14}$$

sometimes called *Strang* splitting, produces a stepping formula that is again exactly symplectic (again assuming the individual factors can be evaluated exactly), but makes local errors of order h^3. This method, in somewhat different guise as the *leap-frog* algorithm, has been known for a long time, and long before the formal advent of symplectic integrators. See Exercise 1.3.

There is a reason why the coefficient of the h^2 term on the right side of (1.13) vanishes. Introduce the notation

$$\mathcal{S}_2(h) = \exp[-(h/2) : A :] \exp(-h : B :) \exp[-(h/2) : A :]. \tag{12.1.15}$$

It is easily verified that the map $\mathcal{S}_2$ satisfies the relation

$$\mathcal{S}_2^{-1}(h) = \mathcal{S}_2(-h). \tag{12.1.16}$$

A map that has the property

$$\mathcal{S}^{-1}(h) = \mathcal{S}(-h) \tag{12.1.17}$$

is called *symmetric*, and we have used the symbol $\mathcal{S}$ for this reason.[3] We have also appended a subscript of 2 in (1.15) because, as we have seen, the use of $\mathcal{S}_2$ furnishes us with an integrator

[3] Any integrator that is exact must satisfy (1.17) because exact integration backwards must send the final conditions back to the initial conditions. However, we are dealing with approximate integration that is only accurate through some power in h, and therefore (1.17) may or may not not be true depending on what integration method is employed. It is not true for the explicit Runge-Kutta methods or Adams predictor-corrector methods or extrapolation methods described in Chapter 2. For any such method (1.17) holds only through terms of order h^m assuming any such method is locally accurate through terms of order h^m. We also remark that the property (1.17) is also sometimes referred to as *reversibility* or *time reversibility*. We avoid this usage, which can lead to confusion, because, properly speaking, reversibility and time reversibility are properties of particular *dynamical systems*, and not others. See Chapter 36. Being symmetric is a property of an integrator, and being reversible or time reversible is a property of a dynamical system, and therefore also of its associated (and exact) transfer map. Of course, if we are integrating a reversible or time reversible system by some approximate integrator, the resulting approximate transfer map may or may not be reversible or time reversible. In this context, it can be shown that there is an interplay between the symmetry of the integrator and the reversibility or time reversibility of the resulting approximate transfer map. Finally, we note that some authors call an integration method symmetric if its global error has an expansion in h that contains only even powers, as is the case for the extrapolation method of Section 2.6. See (2.6.10). This terminology is also potentially confusing because the extrapolation method is not symmetric in the sense (1.17) used here.

that is locally correct through terms of order h^2. We claim that *any* map that satisfies (1.17), when written in exponential form, must have an exponential expansion that involves only *odd* powers of h. That is, if we write

$$\mathcal{S}(h) = \exp[\mathcal{C}(h)], \tag{12.1.18}$$

then $\mathcal{C}(h)$ must be odd in h

$$\mathcal{C}(-h) = -\mathcal{C}(h). \tag{12.1.19}$$

To see the truth of this assertion, make the expansion

$$\mathcal{C}(h) = \sum_{m=1}^{\infty} \mathcal{C}_m h^m. \tag{12.1.20}$$

The series (1.20) has no constant term because, as can be seen from (1.4) and (1.15), we want to impose the condition

$$\mathcal{S}(0) = \mathcal{I}. \tag{12.1.21}$$

Now insert the representation (1.18) into (1.17). Doing so gives the result

$$\exp[-\mathcal{C}(h)] = \exp[\mathcal{C}(-h)]. \tag{12.1.22}$$

We conclude from (1.22) that (1.19) must hold as a result of the uniqueness of the exponent. See Appendix C. [At this point it may be remarked, in passing, that the $O(h^4)$ term indicated in (1.13) must actually vanish so that the next possibly nonvanishing term must be of $O(h^5)$.]

We are now ready to describe a method, sometimes called the *triplet construction*, for parlaying a symmetric integrator of order $2k$ into a symmetric integrator of order $(2k+2)$. It is also sometimes called the *Yoshida trick* or *Yoshida construction* in recognition of one of its discoverers. Suppose an order $2k$ symmetric integrator $\mathcal{S}_{2k}$ is known. Then the map $\mathcal{S}_{2k+2}$ defined by writing

$$\mathcal{S}_{2k+2}(h) = \mathcal{S}_{2k}(\alpha h)\mathcal{S}_{2k}(\beta h)\mathcal{S}_{2k}(\alpha h) \tag{12.1.23}$$

is a symmetric integrator of order $(2k+2)$ providing α and β are the numbers

$$\alpha = 1/[2 - 2^{1/(2k+1)}], \tag{12.1.24}$$

$$\beta = -[2^{1/(2k+1)}]\alpha. \tag{12.1.25}$$

Why is this true? First we verify the relation

$$\begin{aligned} \mathcal{S}_{2k+2}^{-1}(h) &= \mathcal{S}_{2k}^{-1}(\alpha h)\mathcal{S}_{2k}^{-1}(\beta h)\mathcal{S}_{2k}^{-1}(\alpha h) \\ &= \mathcal{S}_{2k}(-\alpha h)\mathcal{S}_{2k}(-\beta h)\mathcal{S}_{2k}(-\alpha h) = \mathcal{S}_{2k+2}(-h), \end{aligned} \tag{12.1.26}$$

which follows from the assumed symmetry of $\mathcal{S}_{2k}$,

$$\mathcal{S}_{2k}^{-1}(h) = \mathcal{S}_{2k}(-h). \tag{12.1.27}$$

We see that $\mathcal{S}_{2k+2}$ is symmetric by construction. Next, because by hypothesis $\mathcal{S}_{2k}$ is an integrator of order $2k$, we have the result

$$\mathcal{S}_{2k}(h) = \exp(h : -H : + \, \mathcal{C}_{2k+1} h^{2k+1} + \cdots). \tag{12.1.28}$$

Consequently we conclude from (1.23), (1.28), the BCH series, and symmetry that $\mathcal{S}_{2k+2}$ must have the form

$$\mathcal{S}_{2k+2} = \exp[(2\alpha + \beta)h : -H : + (2\alpha^{2k+1} + \beta^{2k+1})\mathcal{C}_{2k+1} h^{2k+1} + O(h^{2k+3})]. \tag{12.1.29}$$

Evidently $\mathcal{S}_{2k+2}$ will be an integrator of order $(2k+2)$ if α and β satisfy the relations

$$2\alpha + \beta = 1, \tag{12.1.30}$$

$$2\alpha^{2k+1} + \beta^{2k+1} = 0. \tag{12.1.31}$$

Finally, the relations (1.30) and (1.31) have the solutions (1.24) and (1.25).

Note that nowhere does the demonstration just given depend on any property of $\mathcal{S}_{2k}$ other than symmetry and that it is indeed an integrator of order $2k$. Therefore, if we can produce a symmetric integrator of order $2k$ by any method whatsoever [not necessarily of Zassenhaus type and not necessarily making the splitting assumption (1.5)], then (1.23) through (1.25) produce from it a symmetric integrator of order $(2k+2)$. Finally, suppose $\mathcal{S}_{2k}$ is a symplectic integrator. Then, since any product of symplectic maps is again a symplectic map, (1.23) shows that $\mathcal{S}_{2k+2}$ will also be a symplectic integrator.

As an example of the use of (1.23), let us construct a 4th-order Zassenhaus integrator $\mathcal{S}_4$ from the known 2nd-order integrator $\mathcal{S}_2$ given by (1.15). From (1.23) through (1.25) we get the result

$$\mathcal{S}_4(h) = \mathcal{S}_2(\alpha h)\mathcal{S}_2(\beta h)\mathcal{S}_2(\alpha h) \tag{12.1.32}$$

with

$$\alpha = 1/(2 - 2^{1/3}), \tag{12.1.33}$$

$$\beta = -(2^{1/3})/(2 - 2^{1/3}). \tag{12.1.34}$$

Now carry out the multiplications indicated in (1.32) to obtain the final Zassenhaus relation

$$\begin{aligned} \mathcal{S}_4(h) = \quad & \exp(w_1 h : A :)\exp(w_2 h : B :)\exp(w_3 h : A :)\exp(w_4 h : B :) \times \\ & \exp(w_5 h : A :)\exp(w_6 h : B :)\exp(w_7 h : A :). \end{aligned} \tag{12.1.35}$$

Here the *weights* w_i have the values

$$w_1 = w_7 = -1/[2(2 - 2^{1/3})], \;\; w_3 = w_5 = (1 - 2^{1/3})w_1,$$

$$w_2 = w_6 = 2w_1, \;\; w_4 = -2^{1/3} w_2. \tag{12.1.36}$$

By construction they must have the remarkable property that

$$\mathcal{S}_4(h) = \exp[h : -H : + \, O(h^5)]. \tag{12.1.37}$$

The procedure (1.23) gives a general way of constructing even-order symmetric integrators, but we might also desire odd-order integrators; and we might be willing to give up the prescription (1.23) or even the symmetry condition in favor of having fewer factors in the Zassenhaus product. How does one find general Zassenhaus formulas of the form (1.35)? That is, having decided in advance how many factors we wish to allow in a Zassenhaus product, how do we select weights w_i to achieve a formula of maximum order? For example, there is the 3rd-order formula

$$\begin{aligned}\exp(-h:H:) = \quad & \exp(w_1 h:A:)\exp(w_2 h:B:)\exp(w_3 h:A:)\exp(w_4 h:B:)\times \\ & \exp(w_5 h:A:)\exp(w_6 h:B:)\times[1+O(h^4)] \qquad (12.1.38)\end{aligned}$$

with

$$\begin{aligned} w_1 = -7/24,\ w_3 = -3/4,\ w_5 = 1/24, \\ w_2 = -2/3,\ w_4 = 2/3,\ w_6 = -1. \qquad (12.1.39)\end{aligned}$$

The general answer to this question is difficult, but results are known through modest order, and are discussed in some of the references cited at the end of this chapter.[4]

We also note that if some commutators of the form $C = \{:A:, \{:A:, :B:\}\}$ or $D = \{:B:, \{:B:, :A:\}\}$ are readily computable, and if the maps $\exp(\tau C)$ or $\exp(\tau D)$ can be computed exactly, then one can also construct higher-order symplectic integrators using these quantities. Such methods are sometimes called *force gradient* algorithms. They are known for some examples to have superior accuracy compared to triplet constructed algorithms of the same order. Moreover, there are symmetric force gradient algorithms which can then be employed in a triplet construction to go to still higher order.

Exercises

12.1.1. Hamiltonians of the form $H = T(p) + V(q)$ commonly occur in the case of nonrelativistic motion in a force field derivable from a potential. However, they can also occur in some cases of relativistic motion in an electromagnetic field. Verify from (1.5.30) in Exercise 1.5.3 that H is of the $T+V$ form when there is only an electric field (no magnetic field so that $\boldsymbol{A} = 0$) and the time t is taken to be the independent variable. Verify from (1.6.16) in Exercise 1.6.1 that H is of the $T+V$ if $\psi = 0$, and the magnetic field $\boldsymbol{B}$ is of the form that it can be derived from a vector potential such that only $A_z \neq 0$, and the coordinate z is taken to be the independent variable.

12.1.2. Show that crude Euler is not symplectic. Show that integration using (1.4) and (1.11) is symplectic.

12.1.3. Leapfrog exercise.

12.1.4. Verify that $\mathcal{S}_2$, as given by (1.46), satisfies (1.16).

12.1.5. Verify (1.29).

[4] We remark that the methods we have been describing are sometimes referred to as *composition* methods because $\exp(-h:H:)$ is written as the composition of several factors or as *fractional-step* methods because a step of duration h is accomplished by taking several steps whose durations are sometimes fractions of h.

12.1.6. Verify that (1.30) and (1.31) have the solution (1.24) and (1.25).

12.1.7. Verify (1.35) and (1.36).

12.1.8. Exercise for case where V is time dependent.

12.1.9. For the $\mathcal{S}_2$ given by (1.15), show that

$$\mathcal{S}_2 H = H + h^3 G + O(h^4) \tag{12.1.40}$$

where

$$G = -[A,[A,[A,B]]]/24 - [B,[B,[A,B]]]/12 - [A,[B,[A,B]]]/8. \tag{12.1.41}$$

Evaluate G for the case

$$A = p^2/2 \ , \ \ B = q^2/2. \tag{12.1.42}$$

You have verified a particular case of the general theorem that symplectic integration does not conserve the value of the Hamiltonian.

12.1.10. Verify that the weights of the "A" and "B" terms in (1.36) and (1.39) sum separately to -1. Why should this be?

12.1.11. Review the discussion of backward error analysis that appears near the end of Section 2.7. For a two-dimensional phase space, let H be the Hamiltonian

$$H = (1/2)ap^2 + (1/2)bq^2 = T(p) + V(q), \tag{12.1.43}$$

and let $\mathcal{S}_2(H,h)$ be the symplectic integrator

$$\mathcal{S}_2(H,h) = \exp : -(h/2)T : \exp : -hV : \exp : -(h/2)T : . \tag{12.1.44}$$

Show that there are functions $\alpha(h)$ and $\beta(h)$ such that

$$\mathcal{S}_2(H,h) = \exp : -h\bar{H} : \tag{12.1.45}$$

where

$$\bar{H} = (1/2)\alpha p^2 + (1/2)\beta q^2. \tag{12.1.46}$$

Find $\alpha(h)$ and $\beta(h)$ explicitly. Thus, $\mathcal{S}_2(H,h)$ produces exact trajectories for the modified Hamiltonian $\bar{H}$. Find a Hamiltonian $\hat{H}$ of the form

$$\hat{H} = (1/2)A(h)p^2 + (1/2)B(h)q^2 \tag{12.1.47}$$

such that $\mathcal{S}_2(\hat{H},h)$ produces exact trajectories for the Hamiltonian H.

12.1.12. Exercise on using BCH to combine exponents and backward error analysis to find effective Hamiltonian.

12.2 Symplectic Runge-Kutta Methods for $T+V$ Split Hamiltonians: Partitioned Runge Kutta and Nyström Runge Kutta

The integration methods of the Zassenhaus type, for the *special* Hamiltonians of the form $T+V$, are explicit and symplectic. They can also be viewed as being Runge-Kutta-like in that they involve multiple evaluations of the effects of T and V in the course of making a single integration step. For these special Hamiltonians there are also other methods, called *partitioned* Runge Kutta and Nyström Runge Kutta, that are explicit and symplectic. $\cdots$.

12.3 Symplectic Runge-Kutta Methods for General Hamiltonians

So far, the construction of symplectic integrators has been based on the assumption that the Hamiltonian can be split in the $T+V$ form (1.6). This assumption is not true for a broad class of problems of interest for Accelerator Physics, namely motion in a general electromagnetic field. It is not true for the Hamiltonian (1.5.30), assuming $\boldsymbol{A} \neq 0$, and it is not true for the Hamiltonian (1.6.16) except in the special case $\psi = A_x = A_y = 0$. In this subsection we will describe briefly symplectic Runge-Kutta methods that are applicable to general Hamiltonians.[5]

12.3.1 Background

Let us begin by setting up a notation that is convenient for working with differential equations in Hamiltonian form. Suppose the canonical variables are ordered as in (1.7.9) and Hamilton's equations of motion are written in the form (2.1.1). That is, we make the identification

$$\boldsymbol{y} = (q_1, q_2, \cdots q_\ell; p_1, p_2, \cdots p_\ell), \tag{12.3.1}$$

and therefore, from $\dot{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{y}, t)$, it follows that

$$\boldsymbol{f} = J\partial H/\partial \boldsymbol{y}. \tag{12.3.2}$$

Write the Runge-Kutta procedure in terms of canonical variables by introducing the notation

$$\boldsymbol{y}^{n+1} = (Q_1, Q_2, \cdots Q_\ell; P_1, P_2, \cdots P_\ell) \tag{12.3.3}$$

and the slightly modified notation

$$\boldsymbol{y}^{n} = (q_1, q_2, \cdots q_\ell; p_1, p_2, \cdots p_\ell). \tag{12.3.4}$$

[5] We remark that the Hamiltonian for the restricted 3-body problem, when formulated in a rotating coordinate system in order to exploit the existence of the Jacobi integral that appears in this formulation, also cannot be split in the form $T+V$.

As in Sections 4.8 and 6.5.1, it is also convenient to employ the notation

$$z = (\boldsymbol{q}; \boldsymbol{p}), \tag{12.3.5}$$

$$Z = (\boldsymbol{Q}; \boldsymbol{P}). \tag{12.3.6}$$

In this terminology, performing a Runge-Kutta step from $t = t^n$ to $t = t^{n+1}$ sends the old pair $\boldsymbol{q}, \boldsymbol{p}$ to the new pair $\boldsymbol{Q}, \boldsymbol{P}$. Or, equivalently, it sends z to Z. In map notation, this transformation can be expressed in the form

$$Z = \mathcal{M}_{\text{RK}} z. \tag{12.3.7}$$

We would like the map $\mathcal{M}_{\text{RK}}$ to be exactly symplectic.

As further notation, decompose the vector $\boldsymbol{f}$ into q-like and p-like components,

$$\boldsymbol{f} = (\boldsymbol{f}^q; \boldsymbol{f}^p). \tag{12.3.8}$$

Also, introduce the ℓ-component vectors $\boldsymbol{H_q}$ and $\boldsymbol{H_p}$ by the rules

$$\boldsymbol{H_q}(\boldsymbol{q}, \boldsymbol{p}, t) = \partial H(\boldsymbol{q}, \boldsymbol{p}, t)/\partial \boldsymbol{q}, \tag{12.3.9}$$

$$\boldsymbol{H_p}(\boldsymbol{q}, \boldsymbol{p}, t) = \partial H(\boldsymbol{q}, \boldsymbol{p}, t)/\partial \boldsymbol{p}. \tag{12.3.10}$$

With these definitions, in view of (3.2), we have the relations

$$\boldsymbol{f}^q = \boldsymbol{H_p}, \tag{12.3.11}$$

$$\boldsymbol{f}^p = -\boldsymbol{H_q}. \tag{12.3.12}$$

Recall that a Runge-Kutta method is specified by a Butcher tableau. Review Section 2.3.4. From (2.3.6) we see that application of the Runge-Kutta stepping formula gives the relations

$$\boldsymbol{Q} = \boldsymbol{q} + h \sum_{i=1}^{s} b_i \boldsymbol{k}_i^q, \tag{12.3.13}$$

$$\boldsymbol{P} = \boldsymbol{p} + h \sum_{i=1}^{s} b_i \boldsymbol{k}_i^p. \tag{12.3.14}$$

Here we have introduced the notation

$$\boldsymbol{k}_i = (\boldsymbol{k}_i^q; \boldsymbol{k}_i^p) \tag{12.3.15}$$

to indicate that the $\boldsymbol{k}_i$ also have q-like and p-like components. In terms of this notation, (2.3.7), (3.11), and (3.12) yield the definitions

$$\boldsymbol{k}_i^q = \boldsymbol{f}^q(\boldsymbol{q}_i, \boldsymbol{p}_i, t_i) = \boldsymbol{H_p}(\boldsymbol{q}_i, \boldsymbol{p}_i, t_i), \tag{12.3.16}$$

$$\boldsymbol{k}_i^p = \boldsymbol{f}^p(\boldsymbol{q}_i, \boldsymbol{p}_i, t_i) = -\boldsymbol{H_q}(\boldsymbol{q}_i, \boldsymbol{p}_i, t_i). \tag{12.3.17}$$

For each value of i the right sides of (3.16) and (3.17) are to be evaluated at the points $(\boldsymbol{q}_i, \boldsymbol{p}_i, t_i)$ specified by the relations

$$\boldsymbol{q}_i = \boldsymbol{q} + h \sum_{j=1}^{s} a_{ij} \boldsymbol{k}_j^q, \tag{12.3.18}$$

$$\boldsymbol{p}_i = \boldsymbol{p} + h \sum_{j=1}^{s} a_{ij} \boldsymbol{k}_j^p, \tag{12.3.19}$$

$$t_i = t^n + c_i h. \tag{12.3.20}$$

We will see, in view of (3.16) and (3.17) and results to follow, that the relations (3.18) and (3.19) are implicit if the integrator is to be symplectic. They therefore must be solved by simple iteration or the more involved Newton's method.

12.3.2 Condition for Symplecticity

We will subsequently see that a necessary and sufficient condition for a Runge-Kutta method to be symplectic is that the entries in the matrix a and the vector b satisfy the relations

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0 \text{ for } i, j = 1, \cdots, s. \tag{12.3.21}$$

As usual, the entries c are given by (2.3.11). From the condition (3.21) it is easy to prove that there are no explicit symplectic Runge-Kutta methods. See Exercise 3.1.

Note that the condition (3.21) makes no mention of the number of equations being integrated. Of course, in the Hamiltonian case, one is always integrating an even number equations, say ℓ for the q's and ℓ for the p's. When the number of equations is odd, it makes no sense to talk about a symplectic condition. However, it can be shown that in general a Runge-Kutta integrator satisfying (3.21) has additional desirable stability properties compared to other Runge-Kutta integrators, and therefore the condition (3.21) is also of interest when any set of differential equations, including non-Hamiltonian or odd numbers of equations, is being integrated.

12.3.3 The Single-Stage Case

Before verifying the necessity and sufficiency of the condition (3.21), let us consider the simplest case, the one-stage case $s = 1$. In this case the Butcher tableau has the general form

$$\begin{array}{c|c} c_1 & a_{11} \\ \hline & b_1 \end{array}\,, \tag{12.3.22}$$

and use of (3.21) yields the relation

$$2a_{11}b_1 = (b_1)^2. \tag{12.3.23}$$

But, in order for the method to at least be of order 1, we must have b_1=1. See (2.3.31). It follows from (3.23) that $a_{11} = 1/2$. Thus, if (3.21) holds, the Butcher tableau for a one-stage symplectic Runge-Kutta method is

$$\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}\,. \tag{12.3.24}$$

Here we have also used (2.3.11).

We have seen this Butcher tableau before. Look at (2.3.12) and (2.3.13) to see that this method is the implicit midpoint rule. And, as we learned from Exercise 2.3.7, this method is of order two. We will soon observe that this method is related to Gaussian quadrature, and for this reason we will refer to it as *Gauss2*. And, given equations of motion, we will call $\mathcal{M}_{\mathrm{G2}}$ the transfer map that arises from integrating these equations of motion using Gauss2.

Let us now verify, in the Hamiltonian context, that $\mathcal{M}_{\mathrm{G2}}$ is a symplectic map. For the coefficients (3.24), and in the Hamiltonian case, the Runge-Kutta relations become

$$\boldsymbol{Q} = \boldsymbol{q} + h\boldsymbol{k}_1^q, \tag{12.3.25}$$

$$\boldsymbol{P} = \boldsymbol{p} + h\boldsymbol{k}_1^p, \tag{12.3.26}$$

where

$$\boldsymbol{k}_1^q = \boldsymbol{H}_{\boldsymbol{p}}(\boldsymbol{q}_1, \boldsymbol{p}_1, t_1), \tag{12.3.27}$$

$$\boldsymbol{k}_i^p = -\boldsymbol{H}_{\boldsymbol{q}}(\boldsymbol{q}_1, \boldsymbol{p}_1, t_1), \tag{12.3.28}$$

with

$$\boldsymbol{q}_1 = \boldsymbol{q} + (h/2)\boldsymbol{k}_1^q, \tag{12.3.29}$$

$$\boldsymbol{p}_1 = \boldsymbol{p} + (h/2)\boldsymbol{k}_1^p, \tag{12.3.30}$$

$$t_1 = t^n + h/2. \tag{12.3.31}$$

Suppose small changes $d\boldsymbol{q}$ and $d\boldsymbol{p}$ are made in $\boldsymbol{q}$ and $\boldsymbol{p}$. Then, there will be related small changes $d\boldsymbol{Q}$ and $d\boldsymbol{P}$ in $\boldsymbol{Q}$ and $\boldsymbol{P}$. What we wish to verify is that the matrix M_{G2} in the relation

$$dZ = M_{\mathrm{G2}}dz \tag{12.3.32}$$

is symplectic, and therefore the map $\mathcal{M}_{\mathrm{G2}}$ is symplectic.

According to (3.25) and (3.26), there will be the relations

$$d\boldsymbol{Q} = d\boldsymbol{q} + hd\boldsymbol{k}_1^q, \tag{12.3.33}$$

$$d\boldsymbol{P} = d\boldsymbol{p} + hd\boldsymbol{k}_1^p. \tag{12.3.34}$$

From (3.27) and (3.28) we find the relations

$$d\boldsymbol{k}_1^q = \mathcal{H}_{\boldsymbol{pq}}(\boldsymbol{q}_1, \boldsymbol{p}_1, t_1)d\boldsymbol{q}_1 + \mathcal{H}_{\boldsymbol{pp}}(\boldsymbol{q}_1, \boldsymbol{p}_1, t_1)d\boldsymbol{p}_1, \tag{12.3.35}$$

$$d\boldsymbol{k}_1^p = -\mathcal{H}_{\boldsymbol{qq}}(\boldsymbol{q}_1, \boldsymbol{p}_1, t_1)d\boldsymbol{q}_1 - \mathcal{H}_{\boldsymbol{qp}}(\boldsymbol{q}_1, \boldsymbol{p}_1, t_1)d\boldsymbol{p}_1. \tag{12.3.36}$$

Here $\mathcal{H}_{\boldsymbol{qq}}$, $\mathcal{H}_{\boldsymbol{pq}}$, etc. are the $\ell \times \ell$ Hessian block matrices,

$$\begin{aligned}
&\mathcal{H}_{\boldsymbol{qq}}(\boldsymbol{q}, \boldsymbol{p}, t) = \partial^2 H(\boldsymbol{q}, \boldsymbol{p}, t)/\partial\boldsymbol{q}\partial\boldsymbol{q}, \\
&\mathcal{H}_{\boldsymbol{pq}}(\boldsymbol{q}, \boldsymbol{p}, t) = \mathcal{H}_{\boldsymbol{qp}}(\boldsymbol{q}, \boldsymbol{p}, t) = \partial^2 H(\boldsymbol{q}, \boldsymbol{p}, t)/\partial\boldsymbol{p}\partial\boldsymbol{q}, \\
&\mathcal{H}_{\boldsymbol{pp}}(\boldsymbol{q}, \boldsymbol{p}, t) = \partial^2 H(\boldsymbol{q}, \boldsymbol{p}, t)/\partial\boldsymbol{p}\partial\boldsymbol{p}.
\end{aligned} \tag{12.3.37}$$

And, from (3.29) and (3.30), we find the relations

$$d\boldsymbol{q}_1 = d\boldsymbol{q} + (h/2)d\boldsymbol{k}_1^q, \tag{12.3.38}$$

$$d\boldsymbol{p}_1 = d\boldsymbol{p} + (h/2)d\boldsymbol{k}_1^p. \tag{12.3.39}$$

Now substitute (3.38) and (3.39) into (3.35) and (3.36) to yield the results

$$d\boldsymbol{k}_1^q = \mathcal{H}_{\boldsymbol{pq}}[d\boldsymbol{q} + (h/2)d\boldsymbol{k}_1^q] + \mathcal{H}_{\boldsymbol{pp}}[d\boldsymbol{p} + (h/2)d\boldsymbol{k}_1^p], \tag{12.3.40}$$

$$d\boldsymbol{k}_1^p = -\mathcal{H}_{\boldsymbol{qq}}[d\boldsymbol{q} + (h/2)d\boldsymbol{k}_1^q] - \mathcal{H}_{\boldsymbol{qp}}[d\boldsymbol{p} + (h/2)d\boldsymbol{k}_1^p], \tag{12.3.41}$$

which can be rewritten in the form

$$[1 - (h/2)\mathcal{H}_{\boldsymbol{pq}}]d\boldsymbol{k}_1^q - (h/2)\mathcal{H}_{\boldsymbol{pp}}d\boldsymbol{k}_1^p = \mathcal{H}_{\boldsymbol{pq}}d\boldsymbol{q} + \mathcal{H}_{\boldsymbol{pp}}d\boldsymbol{p}, \tag{12.3.42}$$

$$(h/2)\mathcal{H}_{\boldsymbol{qq}}d\boldsymbol{k}_1^q + [1 + (h/2)\mathcal{H}_{\boldsymbol{qp}}]d\boldsymbol{k}_1^p = -\mathcal{H}_{\boldsymbol{qq}}d\boldsymbol{q} - \mathcal{H}_{\boldsymbol{qp}}d\boldsymbol{p}. \tag{12.3.43}$$

Our goal is to solve the relations (3.42) and (3.43) for $d\boldsymbol{k}_1^q$ and $d\boldsymbol{k}_1^p$, and then substitute the results into (3.33) and (3.34). To do so it is convenient to rewrite (3.42) and (3.43) in matrix/vector form. Let A be the matrix

$$A = \begin{pmatrix} \mathcal{H}_{\boldsymbol{pq}} & \mathcal{H}_{\boldsymbol{pp}} \\ -\mathcal{H}_{\boldsymbol{qq}} & -\mathcal{H}_{\boldsymbol{qp}} \end{pmatrix}. \tag{12.3.44}$$

Then (3.42) and (3.43) can be written in the form

$$[I - (h/2)A]d\boldsymbol{k}_1 = Adz, \tag{12.3.45}$$

and therefore

$$d\boldsymbol{k}_1 = [I - (h/2)A]^{-1}Adz. \tag{12.3.46}$$

Correspondingly, (3.33) and (3.34) become

$$dZ = M_{\mathrm{G2}}dz \tag{12.3.47}$$

with

$$M_{\mathrm{G2}} = I + h[I - (h/2)A]^{-1}A = [I - (h/2)A]^{-1}[I + (h/2)A]. \tag{12.3.48}$$

We claim that M_{G2} is a symplectic matrix. Correspondingly, in this Hamiltonian case, $\mathcal{M}_{\mathrm{G2}}$ is a symplectic map. To verify this claim, observe that A can be written in the form

$$A = JS \tag{12.3.49}$$

where

$$S = \begin{pmatrix} \mathcal{H}_{\boldsymbol{qq}} & \mathcal{H}_{\boldsymbol{qp}} \\ \mathcal{H}_{\boldsymbol{pq}} & \mathcal{H}_{\boldsymbol{pp}} \end{pmatrix}. \tag{12.3.50}$$

As the notation is intended to indicate, and because of the equality of mixed partial derivatives, S is a symmetric matrix. Consequently M_{G2} can also written in the form

$$M_{\mathrm{G2}} = [I - (h/2)JS]^{-1}[I + (h/2)JS]. \tag{12.3.51}$$

Reference to Section 3.12 shows that (3.51) is a Cayley representation when we make the identification

$$W = (h/2)S, \tag{12.3.52}$$

and therefore M_{G2} is indeed a symplectic matrix.

12.3.4 Two-, Three-, and More-Stage Methods

In Subsection 3.3 we studied the single-stage symplectic Runge-Kutta method, found its Butcher tableau (3.24), and observed that this method is of order two. Remarkably, it is also known that, for s stages, there are Runge-Kutta methods of order $m = 2s$, and these methods are symplectic when applied to Hamiltonian systems. Butcher tableaux for these methods, for the cases of two and three stages, are given below.

$$\begin{array}{c|cc} 1/2-\sqrt{3}/6 & 1/4 & 1/4-\sqrt{3}/6 \\ 1/2+\sqrt{3}/6 & 1/4+\sqrt{3}/6 & 1/4 \\ \hline & 1/2 & 1/2 \end{array}\ , \tag{12.3.53}$$

$$\begin{array}{c|ccc} 1/2-\sqrt{15}/10 & 5/36 & 2/9-\sqrt{15}/15 & 5/36-\sqrt{15}/30 \\ 1/2 & 5/36+\sqrt{15}/24 & 2/9 & 5/36-\sqrt{15}/24 \\ 1/2+\sqrt{15}/10 & 5/36+\sqrt{15}/30 & 2/9+\sqrt{15}/15 & 5/36 \\ \hline & 5/18 & 8/18 & 5/18 \end{array}\ . \tag{12.3.54}$$

They have orders 4 and 6, respectively. Observe that, in tableaux (3.24), (3.53), and (3.54), the b_i are weights and the c_i are evaluation points for Gaussian quadrature. This circumstance arises from the fact that the Runge-Kutta methods based on these tableaux are related to Gaussian quadrature. See Appendix T. For this reason these methods are sometimes referred to as Gauss2, Gauss4, and Gauss6. Butcher tableaux for Gauss8 and Gauss10 are also known. See the book of Sanz-Serna and Calvo listed in the Bibliography for this chapter.

We also remark that these methods are symmetric,

$$\mathcal{M}_{\rm G2s}(-h) = [\mathcal{M}_{\rm G2s}(h)]^{-1}. \tag{12.3.55}$$

See the book of Hairer, Nørsett, and Wanner listed in the Bibliography for this chapter.

Exercises

12.3.1. The purpose of this exercise is to verify that a Runge-Kutta, in order to be symplectic, must be implicit. Set $j = i$ in (3.21) to obtain the condition

$$b_i a_{ii} + b_i a_{ii} - b_i b_i = 0 \text{ for } i = 1, \cdots, s, \tag{12.3.56}$$

from which it follows that

$$2 b_i a_{ii} = (b_i)^2 \text{ for } i = 1, \cdots, s. \tag{12.3.57}$$

Verify that, for a Runge-Kutta method to be explicit, the matrix a must be lower triangular. In particular, it must satisfy the condition

$$a_{ii} = 0 \text{ for } i = 1, \cdots, s. \tag{12.3.58}$$

Combine (3.56) and (3.57) to conclude that

$$b_i = 0 \text{ for } i = 1, \cdots, s, \tag{12.3.59}$$

in which case, according to (2.3.6), the associated Runge-Kutta method fails to advance the solution $\boldsymbol{y}$ at all.

12.3.2. Repeat the calculations of Subsection 3.3 for the case of general a_{11} and general b_1. Show that (3.21) must be satisfied for M_{G2} to be a symplectic matrix.

12.4 Study of Single-Stage Method

We have already confessed that the integration formulas associated with each of the Butcher tableaux (3.24), (3.53), and (3.54) are implicit, and therefore must be made explicit (repeatedly solved) at each integration step in order to actually produce a trajectory. Let us see what is involved by first examining the simplest case, that of Gauss2 specified by the single-stage Butcher tableau (3.24). Application of (2.3.6) and (2.3.7) shows that Gauss2 employs the stepping formula

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\boldsymbol{k}_1 \tag{12.4.1}$$

where, at each step,

$$\boldsymbol{k}_1 = \boldsymbol{f}[\boldsymbol{y}^n + h(1/2)\boldsymbol{k}_1,\ t^n + (1/2)h]. \tag{12.4.2}$$

Let us now examine how to deal with/solve the implicit relation (4.2). Suppose we have some initial guess, which we will call $\boldsymbol{k}_1^0$, for $\boldsymbol{k}_1$. It might be $\boldsymbol{f}(\boldsymbol{y}^n, t^n)$, but we could hope for something better. Let us convert (4.2) into a recursion relation by making the rule

$$\boldsymbol{k}_1^{j+1} = \boldsymbol{f}[\boldsymbol{y}^n + h(1/2)\boldsymbol{k}_1^j,\ t^n + (1/2)h]. \tag{12.4.3}$$

It can be verified that (4.3) is a contraction map for small enough h. Therefore, providing the initial guess $\boldsymbol{k}_1^0$ is in the basin of attraction, we have the result

$$\boldsymbol{k}_1 = \lim_{j\to\infty} \boldsymbol{k}_1^j. \tag{12.4.4}$$

Finally, having found $\boldsymbol{k}_1$, $\boldsymbol{y}^{n+1}$ is given by (4.1).

To learn a bit more about the nature of the iteration process (4.3) and (4.4) for the solution of (4.2), it is instructive to study a simple example, that of the harmonic oscillator with Hamiltonian

$$H = (p^2 + q^2)/2. \tag{12.4.5}$$

For this case, with $\boldsymbol{y} = (q, p)$, we find the results

$$f_1 = \dot{y}_1 = \dot{q} = \partial H/\partial p = p = y_2, \tag{12.4.6}$$

$$f_2 = \dot{y}_2 = \dot{p} = -\partial H/\partial q = -q = -y_1, \tag{12.4.7}$$

which can be written more compactly in the matrix form

$$\dot{\boldsymbol{y}} = \boldsymbol{f} = J\boldsymbol{y}. \tag{12.4.8}$$

Here J is the matrix J_2 given by (3.2.11). Correspondingly, application of (4.2) yields the relations

$$\{\boldsymbol{k}_1\}_1 = \{\boldsymbol{f}[\boldsymbol{y}^n + h(1/2)\boldsymbol{k}_1,\ t^n + (1/2)h]\}_1 = \{\boldsymbol{y}^n\}_2 + h(1/2)\{\boldsymbol{k}_1\}_2, \tag{12.4.9}$$

$$\{\boldsymbol{k}_1\}_2 = \{\boldsymbol{f}[\boldsymbol{y}^n + h(1/2)\boldsymbol{k}_1,\ t^n + (1/2)h]\}_2 = -\{\boldsymbol{y}^n\}_1 - h(1/2)\{\boldsymbol{k}_1\}_1, \tag{12.4.10}$$

which can be conveniently written together in the matrix form

$$\boldsymbol{k}_1 = J[\boldsymbol{y}^n + h(1/2)\boldsymbol{k}_1], \tag{12.4.11}$$

and solved for $\boldsymbol{k}_1$ to yield the result

$$\boldsymbol{k}_1 = (I - hJ/2)^{-1}J\boldsymbol{y}^n. \tag{12.4.12}$$

Note that in general we are only able to explicitly solve for $\boldsymbol{k}_1$ in the case that the right side of (4.2) is linear in $\boldsymbol{k}_1$, as is true for this special example.

At this point we cannot resist the urge to employ (4.12) in (4.1) to obtain the explicit stepping relation

$$\boldsymbol{y}^{n+1} = I\boldsymbol{y}^n + h(I - hJ/2)^{-1}J\boldsymbol{y}^n = (I + hJ/2)(I - hJ/2)^{-1}\boldsymbol{y}^n = M_{\mathrm{G2}}\boldsymbol{y}^n \tag{12.4.13}$$

with

$$M_{\mathrm{G2}} = (I + hJ/2)(I - hJ/2)^{-1}. \tag{12.4.14}$$

Observe that M_{G2} is in the Cayley form (3.12.5) with

$$W = (h/2)I. \tag{12.4.15}$$

Since W is symmetric, it follows that M_{G2} is symplectic, as expected. See (3.12.6).

We also know from Section 3.12 that M_{G2} can be written in the form

$$M_{\mathrm{G2}} = \exp(JS) \tag{12.4.16}$$

with S given by

$$S = -2J\tanh^{-1}(JW). \tag{12.4.17}$$

See (3.12.4). Use of (4.15) in (4.17) yields the result

$$\begin{aligned} S &= -2J\tanh^{-1}[(h/2)J] = (-2J)[(hJ/2) + (hJ/2)^3/3 + (hJ/2)^5/5 + \cdots] \\ &= 2[(h/2) - (h/2)^3/3 + (h/2)^5/5 + \cdots]I \\ &= 2\tan^{-1}(h/2)I = h(2/h)\tan^{-1}(h/2)I = h(1 - h^2/12 + h^4/80 - \cdots)I. \end{aligned} \tag{12.4.18}$$

It follows that, for H given by (4.5), M_{G2} can be written in the form

$$M_{\mathrm{G2}} = \exp(JS) = \exp[(2/h)\tan^{-1}(h/2)hJ]. \tag{12.4.19}$$

Consider the Hamiltonian H' defined by

$$H' = \omega(p^2 + q^2)/2. \tag{12.4.20}$$

It can be verified that the exact solution to the equation of motion generated by H' is given by the relation

$$\boldsymbol{y}_{\mathrm{true}}(t) = \exp[\omega(t - t^0)J]\boldsymbol{y}^0_{\mathrm{true}}, \tag{12.4.21}$$

and therefore

$$\boldsymbol{y}_{\rm true}^{n+1} = \exp(\omega hJ)\boldsymbol{y}_{\rm true}^{n}. \tag{12.4.22}$$

Upon comparing (4.13) and (4.19) with (4.22), we conclude that use of Gauss2 to integrate the equations of motion generated by the Hamiltonian H gives the exact solution to the equations of motion generated by H' with

$$\omega = (2/h)\tan^{-1}(h/2) = 1 - h^2/12 + h^4/80 - \cdots . \tag{12.4.23}$$

We observe that H' is conserved and therefore, since H and H' are proportional, H is also conserved by Gauss2. Finally, according to (4.23), the trajectory given by Gauss2, since it is the exact trajectory for H', differs from the exact trajectory for H only by a reparameterization of the time.

Also, here we have an instance of backward error analysis. *Approximately* but symplectically integrating the equations of motion generated by H yields the *exact* trajectory for the equations of motion generated by H' with H' being a small (when h is small) modification of H. See the discussion of backward error analysis in Section 2.7. Conversely, given H, it should be possible to find a related Hamiltonian H'' such that symplectically integrating the equations of motion generated by H'' yields the *exact* trajectory for the equations of motion generated by H. See Exercise 4.1.

With this diversion behind us, let us return to an analysis of the iteration process (4.3). By reasoning analogous to that which produced (4.11), the iteration process (4.3) for the case of H given by (4.5) yields the matrix relation

$$\boldsymbol{k}_1^{j+1} = J[\boldsymbol{y}^n + h(1/2)\boldsymbol{k}_1^j]. \tag{12.4.24}$$

We may view it as a mapping with fixed point $\boldsymbol{k}_1$. How do points near this fixed point behave under the influence of this map? Introduce deviation variables $\boldsymbol{\delta}^j$ by writing

$$\boldsymbol{k}_1^j = \boldsymbol{k}_1 + \boldsymbol{\delta}^j. \tag{12.4.25}$$

In terms of these variables, (4.24) takes the form

$$\boldsymbol{\delta}^{j+1} = (hJ/2)\boldsymbol{\delta}^j. \tag{12.4.26}$$

This recursion relation has the solution

$$\boldsymbol{\delta}^j = (hJ/2)^j\boldsymbol{\delta}^0. \tag{12.4.27}$$

The eigenvalues of J are $\pm i$, and therefore the eigenvalues of $hJ/2$ are $\pm ih/2$. These eigenvalues lie within the unit circle as long as the step size satisfies $|h/2| < 1$, and therefore the fixed point $\boldsymbol{k}_1$ is attracting under this condition. That is, if $|h/2| < 1$, then

$$\lim_{j\to\infty} \boldsymbol{\delta}^j = 0. \tag{12.4.28}$$

Moreover, examination of (4.27) shows that, for the H of this example, the basin of attraction is the entire $\boldsymbol{\delta}^0$ plane, and therefore also the entire $\boldsymbol{k}_1^0$ plane.

There remains the problem of constructing a good initial guess $\boldsymbol{k}_1^0$. Suppose we begin with the initial guess

$$\boldsymbol{k}_1^0 = J\boldsymbol{y}^n. \tag{12.4.29}$$

Doing so is equivalent to ignoring the order h terms in the argument of the right side of (4.2) thereby setting $\boldsymbol{k}_1^0 = \boldsymbol{f}(\boldsymbol{y}^n, t^n)$. We then find the results

$$\boldsymbol{k}_1^1 \;=\; J\boldsymbol{y}^n + (h/2)J\boldsymbol{k}_1^0 = J\boldsymbol{y}^n + Jh(1/2)J\boldsymbol{y}^n = (I + hJ/2)J\boldsymbol{y}^n, \tag{12.4.30}$$

$$\begin{aligned} \boldsymbol{k}_1^2 &= J\boldsymbol{y}^n + (h/2)J\boldsymbol{k}_1^1 = J\boldsymbol{y}^n + (hJ/2)(I + hJ/2)J\boldsymbol{y}^n \\ &= [I + (hJ/2) + (hJ/2)^2]J\boldsymbol{y}^n, \text{ etc.} \end{aligned} \tag{12.4.31}$$

Observe that (4.12) has the expansion

$$\boldsymbol{k}_1 = (I - hJ/2)^{-1}J\boldsymbol{y}^n = [I + (hJ/2) + (hJ/2)^2 + \cdots]J\boldsymbol{y}^n. \tag{12.4.32}$$

Evidently, the iterative process, with the initial guess (4.29), reproduces this expansion in such a way that each iteration produces one more term in the expansion.

Suppose instead we begin with the guess

$$\boldsymbol{k}_1^0 = (I + hJ/2)J\boldsymbol{y}^n. \tag{12.4.33}$$

Then we find

$$\boldsymbol{k}_1^1 = J\boldsymbol{y}^n + (h/2)J\boldsymbol{k}_1^0 = J\boldsymbol{y}^n + (hJ/2)(I + hJ/2)J\boldsymbol{y}^n = [I + (hJ/2) + (hJ/2)^2]J\boldsymbol{y}^n. \tag{12.4.34}$$

Evidently, this is a better guess because it moves us one further step down the chain of iterations.

How could we have anticipated that this would be a better guess? We have remarked that the Butcher tableaux (3.24), (3.53), and (3.54) are related to Gaussian quadrature. Integrate both sides of (2.1.1) over the interval $[t^n, t^{n+1}]$. So doing yields the result

$$\boldsymbol{y}^{n+1} - \boldsymbol{y}^n = \int_{t^n}^{t^{n+1}} d\tau \dot{\boldsymbol{y}}(\tau) = \int_{t^n}^{t^{n+1}} d\tau \boldsymbol{f}[\boldsymbol{y}(\tau),\ \tau]. \tag{12.4.35}$$

Estimate the integral on the right side of (4.35) using lowest-order Gaussian quadrature, which amounts to the midpoint rule, to find the approximation

$$\boldsymbol{y}^{n+1} - \boldsymbol{y}^n = \int_{t^n}^{t^{n+1}} d\tau \boldsymbol{f}[\boldsymbol{y}(\tau),\ \tau] \simeq h\boldsymbol{f}[\boldsymbol{y}(t^n + h/2),\ t^n + h/2] = h\boldsymbol{f}[\boldsymbol{y}(t^n + c_1 h),\ t^n + c_1 h]. \tag{12.4.36}$$

Comparison of (4.36) with (4.1) suggests that a good first guess in the one-stage case would be

$$\boldsymbol{k}_1^0 = \boldsymbol{f}[\boldsymbol{y}(t^n + h/2),\ t^n + h/2]. \tag{12.4.37}$$

Here $\boldsymbol{y}$ is the exact solution to (2.1.1). But, of course, we do not know the exact solution. However, we can imagine having computed and stored $\boldsymbol{f}^n$ as given by (2.1.4). Then, in predictor-corrector terminology with $N = 0$ (see Section 2.4), we can construct a predictor formula using jet formulation (see Section 2.5.3) that will produce $\boldsymbol{y}_{\rm pred}(t^n + h/2)$ with a local error of order h^2 [see (2.4.38)].[6] With a knowledge of $\boldsymbol{y}_{\rm pred}(t^n + h/2)$ we can define $\boldsymbol{k}_1^0$ by the rule

$$\boldsymbol{k}_1^0 = \boldsymbol{f}[\boldsymbol{y}_{\rm pred}(t^n + h/2),\ t^n + h/2]. \qquad (12.4.38)$$

Use of the $N = 0$ predictor in jet formulation gives the result

$$\boldsymbol{y}_{\rm pred}(t^n + h/2) = \boldsymbol{y}^n + (h/2)\boldsymbol{f}^n. \qquad (12.4.39)$$

For the current example,

$$\boldsymbol{y}^n + (h/2)\boldsymbol{f}^n = (I + hJ/2)\boldsymbol{y}^n. \qquad (12.4.40)$$

Combining (4.38) through (4.40) yields the result

$$\begin{aligned} \boldsymbol{k}_1^0 &= \boldsymbol{f}[\boldsymbol{y}_{\rm pred}(t^n + h/2),\ t^n + h/2] \\ &= \boldsymbol{f}[\boldsymbol{y}^n + (h/2)\boldsymbol{f}^n,\ t^n + h/2] \\ &= \boldsymbol{f}[(I + hJ/2)\boldsymbol{y}^n] \\ &= (I + hJ/2)J\boldsymbol{y}^n, \end{aligned} \qquad (12.4.41)$$

in agreement with (4.33). With this $\boldsymbol{k}_1^0$ in hand, we can proceed to carry out the iterations (4.3) to yield, depending on the number of iterations made, some approximation to $\boldsymbol{k}_1$, and then finally determine $\boldsymbol{y}^{n+1}$ using (4.1). Note that, if we wish, we may view (4.3) as a kind of corrector formula.

Let us make two last comments about the solution of (4.2). First, as we have seen from the harmonic oscillator example, its solution by iteration, as in (4.3) and (4.4), requires an infinite number of iterations. Therefore if we, as we must, make only make a finite number iterations, the result of the integration method will not be exactly symplectic. It will only be symplectic to some high power of h depending on how many iterations are made at each step. Each iteration makes the method exactly symplectic through terms of yet one order higher in h. Of course, no matter how many iterations are made, the result is only locally accurate through terms of order h^2. That is, although the result may be highly symplectic, depending on the number of iterations, an error of order h^3 is still made at each step. Second, in order to speed convergence, we might attempt to solve (5.2) by Newton's method. This is possible at the cost of extra programming.[7] For an introduction to Newton's method, see Section 29.4.3.

[6] One might wonder about storing more previous $\boldsymbol{f}$ values so that $\boldsymbol{y}_{\rm pred}(t^n + h/2)$ would be given with yet higher order accuracy. The use of an additional $\boldsymbol{f}$ value would yield an h^2 contribution to $\boldsymbol{k}_1^0$ and an h^3 contribution to $\boldsymbol{k}_1^1$. However, there would seem to be no point in doing so. We know that the predictor is attempting to integrate the trajectories associated with H and the symplectic Runge-Kutta procedure integrates the trajectories associated with H', and these Hamiltonians differ by terms of order h^2. Therefore, the higher-order terms produced by a higher-order predictor would not be expected to improve the guess of $\boldsymbol{k}_1^0$.

[7] Of course, whatever method is used to solve (4.2), it can, if desired, be solved to machine precision in a finite number of steps.

Exercises

12.4.1. This is a study in backward error analysis for the harmonic oscillator when its equations of motion are integrated using Gauss2. We have learned, when integrating the equations of motion associated with the Hamiltonian H given by (4.5), that use of Gauss2 yields exact trajectories for the nearby Hamiltonian H' given by (4.20) and (4.23). Find a Hamiltonian H'' whose equations of motion, when integrated using Gauss2, produces the exact trajectories for the Hamiltonian H.

12.4.2. This is a study in backward error analysis for the general static quadratic Hamiltonian when its associated equations of motion are integrated using Gauss2. Review Subsection 3.3. There we were able to find $M_{\rm G2}$, the matrix for the linear part of $\mathcal{M}_{\rm G2}$, for a general trajectory generated by a general Hamiltonian. Verify that if the equations of motion are linear, as will be the case if the Hamiltonian is quadratic, then the various implicit equations that have to be solved using Gauss2 are all linear. Moreover, again because the Hamiltonian is assumed to be quadratic, the matrix S given by (3.50) will have constant entries. Correspondingly, the map $\mathcal{M}_{\rm G2}$ will be linear, and will be completely represented by its linear part $M_{\rm G2}$ with $M_{\rm G2}$ given by (3.51).

Verify, at least when H is quadratic and static, that $\mathcal{M}_{\rm G2}$ is symmetric,

$$\mathcal{M}_{\rm G2}(-h) = [\mathcal{M}_{\rm G2}(h)]^{-1}. \tag{12.4.42}$$

As stated earlier, an analogous result is known to hold in general when any Gauss2s is used to integrate any set of differential equations.

How does $\mathcal{M}_{\rm G2}$ compare with the exact map $\mathcal{M}$? In the case that H is static, we know that

$$\mathcal{M} = \exp(-h : H :). \tag{12.4.43}$$

Recall (1.4). Show that if H is quadratic, then $\mathcal{M}$ will be linear and will be described by the matrix M given by

$$M = \exp(hJS). \tag{12.4.44}$$

We can now compare $\mathcal{M}_{\rm G2}$ and $\mathcal{M}$. From (3.12.1) through (3.12.5) show that

$$M_{\rm G2} = \exp(hJS') \tag{12.4.45}$$

with

$$JS' = (2/h)\tanh^{-1}[(h/2)JS], \tag{12.4.46}$$

and therefore

$$S' = -J(2/h)\tanh^{-1}[(h/2)JS]. \tag{12.4.47}$$

Consequently, show that use of Gauss2 to integrate the equations of motion associated with the quadratic Hamiltonian H given by

$$H(z) = (1/2)(z, Sz) \tag{12.4.48}$$

produces the exact trajectory for the equations of motion associated with the Hamiltonian H' given by

$$H'(z) = (1/2)(z, S'z). \tag{12.4.49}$$

(Here, as before, we assume that S is time independent.)

Use the results of Exercise 3.12.1, the machinery of (5.5.1) through (5.5.13), and (4.45) to show that

$$[H, H'] = 0, \tag{12.4.50}$$

and therefore $\mathcal{M}_{\mathrm{G2}}$ conserves H.

Given a quadratic time independent H, find a Hamiltonian H'' such that integrating its associated equations of motion using Gauss2 produces the exact trajectories for the Hamiltonian H.

12.5 Study of Two-Stage Method

Now that we have explored the behavior of the single-stage method Gauss2, let us make a similar exploration of the two-stage method Gauss4. Doing so will give us a general understanding of what to expect in the multi-stage case. For the two-stage case use of (2.3.6) and (2.3.7) and the Butcher tableau (3.53) provides the stepping formula

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h(1/2)\boldsymbol{k}_1 + h(1/2)\boldsymbol{k}_2 \tag{12.5.1}$$

where, at each step,

$$\boldsymbol{k}_1 = \boldsymbol{f}[\boldsymbol{y}^n + h(1/4)\boldsymbol{k}_1 + h(1/4 - \sqrt{3}/6)\boldsymbol{k}_2,\ t^n + h(1/2 - \sqrt{3}/6)], \tag{12.5.2}$$

$$\boldsymbol{k}_2 = \boldsymbol{f}[\boldsymbol{y}^n + h(1/4 + \sqrt{3}/6)\boldsymbol{k}_1 + h(1/4)\boldsymbol{k}_2,\ t^n + h(1/2 + \sqrt{3}/6)]. \tag{12.5.3}$$

Like (4.2) in the single-stage case, the relations (5.2) and (5.3) are implicit. To solve them numerically, we may make initial guesses $\boldsymbol{k}_1^0$ and $\boldsymbol{k}_2^0$ for $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$, respectively, and set up the recursion relations

$$\boldsymbol{k}_1^{j+1} = \boldsymbol{f}[\boldsymbol{y}^n + h(1/4)\boldsymbol{k}_1^j + h(1/4 - \sqrt{3}/6)\boldsymbol{k}_2^j,\ t^n + h(1/2 - \sqrt{3}/6)], \tag{12.5.4}$$

$$\boldsymbol{k}_2^{j+1} = \boldsymbol{f}[\boldsymbol{y}^n + h(1/4 + \sqrt{3}/6)\boldsymbol{k}_1^j + h(1/4)\boldsymbol{k}_2^j,\ t^n + h(1/2 + \sqrt{3}/6)]. \tag{12.5.5}$$

It can be shown that the relations (5.4) and (5.5) constitute a contraction map for sufficiently small h. Consequently, assuming that the $\boldsymbol{k}_i^0$ are in the basin of attraction, there will be the result

$$\boldsymbol{k}_i = \lim_{j\to\infty} \boldsymbol{k}_i^j. \tag{12.5.6}$$

Alternatively, to achieve more rapid convergence at the expense of a more involved procedure, we may solve (5.2) and (5.3) by Newton's method, which will also be convergent for sufficiently small h and a sufficiently good guess for the $\boldsymbol{k}_i^0$. Finally, having found the $\boldsymbol{k}_i$ by whatever method, $\boldsymbol{y}^{n+1}$ is given by (5.1).

It is again instructive to apply this method to the harmonic oscillator example with the Hamiltonian (4.5) and the equations of motion (4.8). In this case the relations (5.2) and (5.3) take the specific form

$$\boldsymbol{k}_1 = J[\boldsymbol{y}^n + h(1/4)\boldsymbol{k}_1 + h(1/4 - \sqrt{3}/6)\boldsymbol{k}_2], \tag{12.5.7}$$

$$\boldsymbol{k}_2 = J[\boldsymbol{y}^n + h(1/4 + \sqrt{3}/6)\boldsymbol{k}_1 + h(1/4)\boldsymbol{k}_2]. \tag{12.5.8}$$

As they stand these relations, because of their vector/matrix form, comprise four linear equations in four unknowns, namely the components of the two two-dimensional vectors $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$. With sufficient effort, linear equations can always be solved. In this case there is a trick that simplifies the problem. First rewrite (5.8) to bring all terms involving $\boldsymbol{k}_2$ to the left and all other terms to the right. So doing yields the relation

$$(I - hJ/4)\boldsymbol{k}_2 = J[\boldsymbol{y}^n + h(1/4 + \sqrt{3}/6)\boldsymbol{k}_1]. \tag{12.5.9}$$

Next solve (5.9) for $\boldsymbol{k}_2$ in terms of everything else to find the result

$$\boldsymbol{k}_2 = (I - hJ/4)^{-1}J[\boldsymbol{y}^n + h(1/4 + \sqrt{3}/6)\boldsymbol{k}_1]. \tag{12.5.10}$$

Now substitute (5.10) into (5.7) to obtain a relation involving only $\boldsymbol{k}_1$,

$$\boldsymbol{k}_1 = J\{\boldsymbol{y}^n + h(1/4)\boldsymbol{k}_1 + h(1/4 - \sqrt{3}/6)(I - hJ/4)^{-1}J[\boldsymbol{y}^n + h(1/4 + \sqrt{3}/6)\boldsymbol{k}_1]\}. \tag{12.5.11}$$

Manipulate this relation to find, as intermediate steps, the results

$$\begin{aligned}&\boldsymbol{k}_1 - Jh(1/4)\boldsymbol{k}_1 + h(1/4 - \sqrt{3}/6)(I - hJ/4)^{-1}h(1/4 + \sqrt{3}/6)\boldsymbol{k}_1\\ &= J[\boldsymbol{y}^n + h(1/4 - \sqrt{3}/6)(I - hJ/4)^{-1}J\boldsymbol{y}^n],\end{aligned} \tag{12.5.12}$$

or,

$$\begin{aligned}&[I - Jh(1/4) - h^2(1/48)(I - hJ/4)^{-1}]\boldsymbol{k}_1\\ &= [J - h(1/4 - \sqrt{3}/6)(I - hJ/4)^{-1}]\boldsymbol{y}^n,\end{aligned} \tag{12.5.13}$$

and, as a last step, the explicit solution

$$\boldsymbol{k}_1 = [I - hJ/4 - (h^2/48)(I - hJ/4)^{-1}]^{-1}[J - h(1/4 - \sqrt{3}/6)(I - hJ/4)^{-1}]\boldsymbol{y}^n. \tag{12.5.14}$$

(Here we have used the relation $J^2 = -I$.) Finally, insert (5.14) into (5.10) to yield an explicit result for $\boldsymbol{k}_2$,

$$\boldsymbol{k}_2 = . \tag{12.5.15}$$

Exercises

12.5.1.

12.5.2.

12.6 Numerical Examples for One- and Two-Stage Methods

12.7 Proof of Condition for Symplecticity

There are at least three ways to verify that (3.21) is a necessary and sufficient condition for $\mathcal{M}_{\rm RK}$ to be symplectic. The first requires making a brute force calculation of the Jacobian matrix M_{RK} associated with $\mathcal{M}_{\rm RK}$ followed by a verification that M_{RK} is a symplectic matrix. That is what we have just done for the simplest case $s = 1$. See also Exercise 3.2. The second, far more elegant and compact, makes use of differential forms. The third, which we will employ here, uses the contents of the Butcher tableau to define a generating function F_2. It then demonstrates that the generating produced by F_2 reproduces the Runge-Kutta step for the Butcher tableau provided the contents of the Butcher tableau satisfy (3.21).

Consider the generating function $F_2(\boldsymbol{q},\boldsymbol{P},t^n;h)$ defined by writing

$$F_2(\boldsymbol{q},\boldsymbol{P},t^n;h) = \boldsymbol{q}\cdot\boldsymbol{P} + G_2(\boldsymbol{q},\boldsymbol{P},t^n;h), \tag{12.7.1}$$

where

$$G_2 = G_2^1 + G_2^2 \tag{12.7.2}$$

with

$$G_2^1 = h\sum_{i=1}^{s} b_i H(\boldsymbol{q}_i,\boldsymbol{p}_i,t_i), \tag{12.7.3}$$

$$G_2^2 = -h^2\sum_{i,j=1}^{s} b_i a_{ij}[\boldsymbol{H_q}(\boldsymbol{q}_i,\boldsymbol{p}_i,t_i)\cdot\boldsymbol{H_p}(\boldsymbol{q}_j,\boldsymbol{p}_j,t_j)]. \tag{12.7.4}$$

The thoughtful reader may find the definition given by (7.1) through (7.4) puzzling because the terms on the right sides of (7.3) and (7.4) are functions of the old phase-space variables variables $\boldsymbol{q},\boldsymbol{p}$ while the phase-space arguments of G_2 are specified as being the mixed pair $\boldsymbol{q},\boldsymbol{P}$. Here is what is meant: The relation (3.14) specifies $\boldsymbol{P}$ as a function of $\boldsymbol{q},\boldsymbol{p},t^n$; and h,

$$\boldsymbol{P} = \boldsymbol{P}(\boldsymbol{q},\boldsymbol{p},t^n;h). \tag{12.7.5}$$

This relation is to be partially inverted to yield $\boldsymbol{p}$ as a function of $\boldsymbol{q},\boldsymbol{P},t^n$ and h,

$$\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{q},\boldsymbol{P},t^n;h). \tag{12.7.6}$$

From the form of (3.14) we see, by the inverse function theorem, that such an inversion is possible for small enough h because then $\boldsymbol{p}\simeq\boldsymbol{P}$. Finally, the right side of (7.6) is to be substituted for $\boldsymbol{p}$ in the right sides of (7.3) and (7.4) to yield $G_2(\boldsymbol{q},\boldsymbol{P},t^n;h)$.

Suppose we use $F_2(\boldsymbol{q},\boldsymbol{P},t^n;h)$ to produce a transformation that sends $\boldsymbol{q},\boldsymbol{p}$ to $\boldsymbol{Q},\boldsymbol{P}$ by the standard rules (6.5.5),

$$p_k = \partial F_2/\partial q_k\ ,\ Q_k = \partial F_2/\partial P_k. \tag{12.7.7}$$

From the work of Section 6.5.1 we know that so doing will produce a symplectic map, which we will call $\mathcal{M}$. With a view to implementing (7.7), let us compute (save for holding t^n fixed) the total differential of $G_2(\boldsymbol{q},\boldsymbol{P},t^n;h)$.

12.8 Symplectic Integration of General Hamiltonians Using Generating Functions

Section 6.7.3 on the Hamilton-Jacobi equation studied the relation between Hamiltonians and generating functions. There it was shown, once a Darboux matrix α has been selected, that there is a unique relation between the Hamiltonian $H(Z,t)$ and the source function $g(u,t)$. Given H, some time t^n, some phase-space point z^n, and a time step h, we wish to find z^{n+1} at time $t^{n+1} = t^n + h$ in such a way that the relation between z^{n+1} and z^n is symplectic and z^{n+1} is very nearly (with error of order h^{m+1}) equal to the result found by exactly integrating the equations generated by H starting with initial conditions z^n at $t = t^n$ and integrating to $t = t^n + h$. We assume that the trajectory generated by H is analytic in t, which will be the case if $H(z,t)$ is analytic in the phase-space variables z and the time t. (See Poincaré's theorem in Section 1.3.) Then $z(t^n + h)$ will have a Taylor expansion in h.

12.9 Special Symplectic Integrator for Motion in General Electromagnetic Fields

We have seen that there are *implicit* symplectic Runge-Kutta integrators for general Hamiltonians, and hence also for motion in general electromagnetic fields. Remarkably, there are *explicit* symplectic integrators for the Hamiltonian (1.6.192) and, by extension, for the Hamiltonian (1.6.77).

We begin with the simpler case, the Hamiltonian (1.6.192), which can be written in the form

$$H = H_x + H_y + H_z \tag{12.9.1}$$

where

$$H_x = (p_x - qA_x)^2/(2m^*), \tag{12.9.2}$$

$$H_y = (p_y - qA_y)^2/(2m^*), \tag{12.9.3}$$

$$H_z = (p_z - qA_z)^2/(2m^*). \tag{12.9.4}$$

Since H is time independent, the relation (1.1) still holds. Moreover, we may again subdivide the time axis into equal steps of duration h to obtain the exact marching rule (1.4). However, since in the present context the symbol z is being used to denote a coordinate, we rewrite (1.4) in the form

$$w^{n+1} = \exp(h : -H :)w^n \tag{12.9.5}$$

where w denotes the collection of phase-space variables

$$w = (x, y, z, ; p_x, p_y, p_z). \tag{12.9.6}$$

We next make the approximation

$$\exp(h : -H :) \cong \mathcal{S}_2(h) \tag{12.9.7}$$

where $\mathcal{S}_2(k)$ is now defined by the rule

$$\begin{aligned} \mathcal{S}_2(h) \;=\; & \exp[-(h/2) : H_x :] \exp[-(h/2) : H_y :] \times \\ & \exp[-h : H_z :] \exp[-(h/2) : H_y :] \exp[-(h/2) : H_x :]. \end{aligned} \tag{12.9.8}$$

Upon combining the exponents on the right side of (9.8) to first order in h, we see that the exponent on the left side of (9.7) is regained. Also, by construction, $\mathcal{S}_2$ satisfies (1.16). Therefore, as the notation is intended to indicate, $\mathcal{S}_2$ is a symmetric integrator that is locally correct through terms of order h^2. Note other permutations of H_x, H_y, and H_z could have been used in the definition of $\mathcal{S}_2$. There are thus 3! possible formulas of the kind (9.8).

We are still faced with the problem of evaluating the action of the individual factors on the right side of (9.8). Define functions $U_x(x, y, z)$, $U_y(x, y, z)$, and $U_z(x, y, z)$ by the requirements

$$A_x = \partial U_x/\partial x \;,\; A_y = \partial U_y/\partial y \;,\; A_z = \partial U_z/\partial z. \tag{12.9.9}$$

There are many such functions, and we may choose among them at will at each integration step. For example, we may write

$$U_x = \int^x A_x(x', y, z)dx' \tag{12.9.10}$$

and add to it any function of y and z. Use the U's to make symplectic maps $\mathcal{A}_x$, $\mathcal{A}_y$, and $\mathcal{A}_z$ defined by the relations

$$\mathcal{A}_x = \exp(-q : U_x :) \;,\; \mathcal{A}_y = \exp(-q : U_y :) \;,\; \mathcal{A}_z = \exp(-q : U_z :). \tag{12.9.11}$$

These maps produce gauge transformations. See Exercise 6.2.8. It is easily verified, for example, that $\mathcal{A}_x$ and $\mathcal{A}_x^{-1}$ have the phase-space actions

$$\mathcal{A}_x x = x \;,\; \mathcal{A}_x y = y \;,\; \mathcal{A}_x z = z, \tag{12.9.12}$$

$$\mathcal{A}_x^{-1} x = x \;,\; \mathcal{A}_x^{-1} y = y \;,\; \mathcal{A}_x^{-1} z = z, \tag{12.9.13}$$

$$\mathcal{A}_x p_x = p_x - q : U_x : p_x = p_x - q[U_x, p_x] = p_x - q(\partial U_x/\partial x) = p_x - qA_x, \tag{12.9.14}$$

$$\mathcal{A}_x^{-1} p_x = p_x + q : U_x : p_x = p_x + qA_x, \tag{12.9.15}$$

$$\mathcal{A}_x p_y = p_y - q : U_x : p_y = p_y - q[U_x, p_y] = p_y - q(\partial U_x/\partial y), \tag{12.9.16}$$

$$\mathcal{A}_x^{-1} p_y = p_y + q : U_x : p_y = p_y + q(\partial U_x/\partial y), \tag{12.9.17}$$

$$\mathcal{A}_x p_z = p_z - q : U_x : p_z = p_z - q[U_x, p_z] = p_z - q(\partial U_x/\partial z), \tag{12.9.18}$$

$$\mathcal{A}_x^{-1} p_z = p_z + q : U_x : p_z = p_z + q(\partial U_x/\partial z). \tag{12.9.19}$$

In particular, it follows from (9.14) and (9.15) that $\mathcal{A}_x$ has the property

$$\begin{aligned} & \mathcal{A}_x \exp[-(h/2) : p_x^2/(2m^*) :]\mathcal{A}_x^{-1} = \exp[-(h/2) : (p_x - qA_x)^2/(2m^*)] \\ & = \exp[-(h/2) : H_x :]. \end{aligned} \tag{12.9.20}$$

We note at this point that it does not matter what special choice is made for U_x, see (9.10), because from (9.20) it is evident that all allowed choices yield the same net result.

As a consequence of (9.20) and similar relations, $\mathcal{S}_2$ can be rewritten in the factored product form

$$\begin{aligned}\mathcal{S}_2 = \quad & \mathcal{A}_x \exp[-(h/2) : \bar{H}_x :]\mathcal{A}_x^{-1} \times \\ & \mathcal{A}_y \exp[-(h/2) : \bar{H}_y :]\mathcal{A}_y^{-1} \times \\ & \mathcal{A}_z \exp[-h : \bar{H}_z :]\mathcal{A}_z^{-1} \times \\ & \mathcal{A}_y \exp[-(h/2) : \bar{H}_y :]\mathcal{A}_y^{-1} \times \\ & \mathcal{A}_x \exp[-(h/2) : \bar{H}_x :]\mathcal{A}_x^{-1}, \end{aligned} \tag{12.9.21}$$

where we have used the notation

$$\bar{H}_x = p_x^2/(2m^*), \tag{12.9.22}$$

$$\bar{H}_y = p_y^2/(2m^*), \tag{12.9.23}$$

$$\bar{H}_z = p_z^2/(2m^*). \tag{12.9.24}$$

We have already seen that the phase-space actions of the $\mathcal{A}$'s and $\mathcal{A}^{-1}$'s can be evaluated exactly using relations of the form (9.12) through (9.19). Evidently the actions of the maps $\exp[-(h/2) : \bar{H}_x :]$, $\exp[-(h/2) : \bar{H}_y :]$, and $\exp[-h : \bar{H}_z :]$ can also be evaluated exactly. See Exercises 5.4.1 and 5.4.2. Therefore, use of the approximation (9.7) with $\mathcal{S}_2$ given by (9.21) produces a symmetric integrator that is locally correct through terms of order h^2 and is exactly symplectic.

At this point, two remarks are in order. The first is that, with $\mathcal{S}_2(k)$ in hand, the triplet construction can be used to produce higher-order symmetric and symplectic integrators. For example, $\mathcal{S}_4$ is given by (1,32) through (1.34).

The second remark is less triumphant. For a symmetric integrator there is the general relation

$$\mathcal{S}_{2k}(h) = \exp[h : -H : +O(h^{2k+1})]. \tag{12.9.25}$$

Moreover, the error term does not commute with $: H :$ so that for each integration step there is the result

$$\mathcal{S}_{2k}(h)H = H + O(h^{2k+1}) \tag{12.9.26}$$

where the error term is *nonzero*. In fact *Ge* and *Marsden* have essentially shown that it is impossible to construct an integrator that is exactly symplectic and also exactly conserves H.[8] In some applications this may not much matter. Indeed, it is sometimes argued that a symplectic integrator can be used with a larger time step h than a nonsymplectic integrator of the same order because the symplectic integrator at least respects the underlying structure of any Hamiltonian system. And the fact (hope) that a larger time step can be used compensates for the relatively large amount of work associated with each time step.

[8] Ge and Marsden have shown that if a symplectic integrator conserves H, it must be exact or, at worst, produce exact trajectories up to a reparameterization of the time. (See Section 4 for an example where this happens.) Now in general a symplectic integrator cannot produce exact trajectories because in general it makes errors of some order in h. Moreover, it is unlikely that these errors only amount to a reparameterization of the time. Therefore, generally H is not conserved. Of course, one can easily check during the course of an integration to see if the value H is changing, and in general, as expected, one finds that it is.

[Moreover, the variation in H during symplectic integration is often observed to be essentially periodic when the trajectory being integrated is essentially periodic, and this good behavior can be understood using the BCH series. By contrast, the value of H typically grows (or damps) linearly or quadratically, or eventually even exponentially, in time when nonsymplectic integration is employed.] However, for the Hamiltonian (1.6.192) we know that the only physically meaningful trajectories are those for which H has the value (1.6.193. Therefore in this case it is necessary to use a time step h that is sufficiently small to ensure that over the course of integration H obeys (1.6.193) to high accuracy. In particular, one should monitor the value of H during the course of integration to verify that (1.6.193) is met with sufficient accuracy.

While in the mode of exploring difficulties associated with this approach, we should also consider how much effort is required to carry out the integrations of the form (9.10) required to compute the functions $U_x(x, y, z)$ through $U_z(x, y, z)$. If the vector potential $\boldsymbol{A}(\boldsymbol{r})$ is known in analytic form and has a sufficiently simple structure, then these integrals can be evaluated analytically in terms of elementary functions prior to any use of the symplectic integrator. However for most if not all realistic applications, these integrations yield higher transcendental functions that are expensive to evaluate, or these integrations must be carried out numerically. And these evaluations/numerical integrations must be performed with high accuracy if the symplecticity of the overall integration process is to be assured. These evaluations/integrations further add to the already high computational overhead associated with symplectic integration. (See Exercise 9.3.) The considerations of this and the previous paragraph make one wonder if the computational burden for symplectic integration of the kind just described for trajectories in realistic electromagnetic fields is so high as to make nonsymplectic, but high-order, integration superior to symplectic integration. The answer to this question is presumably problem dependent. Its answer for any realistic problem would require the comparison of symplectic integration and high accuracy (probably not Runge-Kutta) nonsymplectic integration. When such explorations are made, one should also consider symplectic (but implicit) Runge-Kutta methods and symplectic generating function methods of the kind described in the previous subsections.

We close this section with an analogous discussion of the Hamiltonian (1.6.77), which can be written in the form

$$H_R = H_x + H_y + H_z + H_t \tag{12.9.27}$$

where

$$H_x = (p_x - qA_x)^2/(2mc), \tag{12.9.28}$$

$$H_y = (p_y - qA_y)^2/(2mc), \tag{12.9.29}$$

$$H_z = (p_z - qA_z)^2/(2mc), \tag{12.9.30}$$

$$H_t = -(p_4 + qA^4)^2/(2mc), \tag{12.9.31}$$

and w becomes the collection of phase-space variables

$$w = (x, y, z, x^4; p_x, p_y, p_z, p_4). \tag{12.9.32}$$

Since H_R is τ independent, the transfer map associated with H_R is given by the relation

$$\mathcal{M} = \exp(\tau : -H_R :). \tag{12.9.33}$$

As expected, we subdivide the τ axis into steps of equal amount h to obtain the exact marching rule (9.5).

In analogy with (9.7) and (9.8), we next make the approximation

$$\exp(h : -H_R :) \cong \mathcal{S}_2(h) \tag{12.9.34}$$

where $\mathcal{S}_2(h)$ is now defined by the rule

$$\begin{aligned}\mathcal{S}_2(h) \quad = \quad & \exp[-(h/2) : H_x :] \exp[-(h/2) : H_y :] \times \\ & \exp[-(h/2) : H_z :] \exp[-h : H_t :] \exp[-(h/2) : H_z :] \times \\ & \exp[-(h/2) : H_y :] \exp[-(h/2) : H_x :].\end{aligned} \tag{12.9.35}$$

By construction this $\mathcal{S}_2$ is a symmetric integrator that is locally correct through terms of order h^2. Other permutations of H_x through H_t could have been used in the definition of $\mathcal{S}_2$, and there are therefore 4! possible integrators of this kind.

We again define functions U_x through U_z by the requirements (9.10), and we add to their collection the function U^4 defined by the requirement

$$A^4 = -\partial U^4/\partial x^4. \tag{12.9.36}$$

The symplectic maps $\mathcal{A}_x$ through $\mathcal{A}_z$ are also again defined by (9.11), and to their collection we add the symplectic map $\mathcal{A}_4$ defined by

$$\mathcal{A}_4 = \exp : -qU^4 : . \tag{12.9.37}$$

This map has the property

$$\mathcal{A}_4 p_4 = p_4 - q : U^4 : p_4 = p_4 - q[U^4, p_4] = p_4 - q\partial U^4/\partial x^4 = p_4 + qA^4, \tag{12.9.38}$$

from which it follows that

$$\exp(-h : H_t :) = \mathcal{A}_4 \exp[h : p_4^2/(2mc) :]\mathcal{A}_4^{-1}. \tag{12.9.39}$$

We are now able to proceed as before to express $\mathcal{S}_2$ as a product of maps, all of which can be evaluated explicitly,

$$\begin{aligned}\mathcal{S}_2 \quad = \quad & \mathcal{A}_x \exp[-(h/2) : \bar{H}_x :]\mathcal{A}_x^{-1} \times \\ & \mathcal{A}_y \exp[-(h/2) : \bar{H}_y :]\mathcal{A}_y^{-1} \times \\ & \mathcal{A}_z \exp[-(h/2) : \bar{H}_z :]\mathcal{A}_z^{-1} \times \\ & \mathcal{A}_4 \exp[-h : \bar{H}_t :]\mathcal{A}_4^{-1} \times \\ & \mathcal{A}_z \exp[-(h/2) : \bar{H}_z :]\mathcal{A}_z^{-1} \times \\ & \mathcal{A}_y \exp[-(h/2) : \bar{H}_y :]\mathcal{A}_y^{-1} \times \\ & \mathcal{A}_x \exp[-(h/2) : \bar{H}_x :]\mathcal{A}_x^{-1}.\end{aligned} \tag{12.9.40}$$

Here we have used the notation

$$\bar{H}_x = p_x^2/(2mc),$$

$$\bar{H}_y = p_y^2/(2mc),$$

$$\bar{H}_z = p_z^2/(2mc),$$

$$\bar{H}_t = -p_4^2/(2mc). \tag{12.9.41}$$

We must again be aware that trajectories generated by H_R are only physically meaningful when H_R has a special value, namely that given by the (mass shell) condition (1.6.92), and that this value cannot be maintained exactly by any symplectic integrator. Therefore it is again necessary to choose h sufficiently small to ensure that over the course of integration H_R obeys (1.6.92) to high accuracy. Moreover there is again added overhead. Now we must compute the functions U_x through U^4. See Exercise 9.3.

Exercises

12.9.1. Exercise on what happens when, in H_R, $A^4 = 0$ and $\boldsymbol{A}$ is static.

12.9.2. Show that in the nonrelativistic approximation the Lagrangian (1.5.1) may be replaced by the Langrangian

$$L_{NR} = (1/2)mv^2 - q\psi(\boldsymbol{r}, t) + q\boldsymbol{v} \cdot \boldsymbol{A}(\boldsymbol{r}, t). \tag{12.9.42}$$

Find the associated Hamiltonian H_{NR}. Show that H_{NR} is conserved if the electromagnetic fields are static. Construct a symplectic integrator for H_{NR}. Show that this integrator does not conserve H_{NR}. Perhaps this nonconservation is not so important in the nonrelativistic case because it might be argued that H_{NR} has no *fundamental* significance. Moreover, H_{NR} is not conserved anyway for the exact motion if the electromagnetic fields are time dependent.

12.9.3. Exercise on what exactly is involved in computing the integrals U_x through U_z or U_x through U^4.

12.10 Zassenhaus Formulas and Map Computation

The discussion so far has dealt mostly with the use of Zassenhaus formulas of the kinds (1.14), (1.35), (1.38), (9.21), and (9.40) as *symplectic integrators*. However, Zassenhaus formulas can also be used to *compute maps* both in Taylor and factored product form.

12.10.1 Case of $T + V$ or General Electromagnetic Field Hamiltonians

As a simple example, suppose that H has the $T + V$ decomposition (1.6). Then, the Taylor maps for $\exp(\sigma : T :)$ and $\exp(\sigma : V :)$, where σ is some parameter, can be computed exactly by formulas analogous to (1.7) through (1.10); and therefore we can find the net Taylor map for any factored map of the kinds (1.14), (1.35), or (1.38). Also, we can expand T and V in homogeneous polynomials. For example, we can write

$$T = \sum_{m=0}^{\infty} T_m \tag{12.10.1}$$

with a similar expansion for V. Then we have the factored product representation

$$\exp(\sigma : T :) = \exp(\sigma : T_1 :) \exp(\sigma : T_2 :) \exp(\sigma : T_3 :) \cdots , \tag{12.10.2}$$

with a similar representation for $\exp(\sigma : V :)$. It follows that each of the factors appearing in any approximation of the kinds (1.14), (1.35), or (1.38) for the map $\exp(h : -H :)$ can be written in factored product form. The resulting factors can then be concatenated together to yield for the map $\exp(h : -H :)$ a final approximation that is also in factored product form.

After some reflection, we see that the same procedure can be applied to the symplectic integrators (9.21) and (9.40). Each of the factors can be be expanded in Taylor form or written in factored product form, and these maps can then be concatenated to yield a net map either in Taylor or factored product form.

There is one last remark to be made: As explained in Section 1.6, it is often convenient to have maps for which some coordinate is the independent variable, and in this subsection we have been using the time t or some world-line parameter τ as the independent variable. If we wish to compute maps rather than trajectories (apart from the reference trajectory) this problem can be overcome with the use of *matching* maps. When the map for some element has been computed using t or τ as an independent variable, the necessary conversion can be made by preceding the map with a matching map that transforms from phase-space variables for which some coordinate is the independent variable to phase-space variables for which t or τ is the independent variable, and following the map by a second matching map that transforms back to the phase-space variables for which some coordinate is the independent variable. See Section *.*.

12.10.2 Case of Hamiltonians Expanded in Homogeneous Polynomials

Zassenhaus formulas can also be used to provide factored product approximations for $\mathcal{M} = \exp(h : -H :)$ when H is decomposed into homogeneous polynomials as in (10.9.1). Here we will consider the autonomous case. The nonautonomous case is best treated using the methods of Sections 10.5.2 and 10.6.2.

Derivation

Define A and B by writing the equations

$$A = H_1 + H_2, \tag{12.10.3}$$

$$B = H_r = H_3 + H_4 + \cdots . \tag{12.10.4}$$

Evidently, any map of the kind $\exp(-h : A :)$ with A given by (10.3) can be written in factored product form using the methods of Section 9.2. See (9.2.4), (9.2.7), and (9.2.9). What about maps of the kind $\exp(-h : B :)$ with B given by (10.4)? How do we find generators f_m such that

$$\exp(-h : B :) = \exp(-h : H_3 + H_4 + \cdots :) = \exp(: f_3 :) \exp(: f_4 :) \cdots ? \tag{12.10.5}$$

We note that, since there is no H_2 term in (10.4), we may use the methods of Section 10.6.2 with the understanding that

$$\mathcal{M}_2 = \mathcal{I} \tag{12.10.6}$$

and therefore

$$H_m^{\text{int}} = H_m. \tag{12.10.7}$$

It follows from (10.6.14) through (10.6.20) that

$$f_3 = -hH_3, \tag{12.10.8}$$

$$f_4 = -hH_4, \tag{12.10.9}$$

$$f_5 = -hH_5 - (h^2/2)[H_3, H_4], \tag{12.10.10}$$

$$f_6 = -hH_6 - (h^2/2)[H_3, H_5] - (h^3/6)[H_3, [H_3, H_4]], \tag{12.10.11}$$

$$\begin{aligned} f_7 &= -hH_7 - (h^2/2)([H_3, H_6] + [H_4, H_5]) - h^3(H_3, [H_3, H_5]]/6 + [H_4, [H_3, H_4]/3) \\ &- (h^4/24)[H_3, [H_3, [H_3, H_4], \end{aligned} \tag{12.10.12}$$

$$\begin{aligned} f_8 &= -hH_8 - (h^2/2)([H_3, H_7] + [H_4, H_6]) \\ &- h^3([H_3, [H_3, H_6]]/6 + [H_4, [H_3, H_5]]/3 + [H_5, [H_3, H_4]]/12) \\ &- h^4([H_3, [H_3, [H_3, H_5]]]/24 + [H_4, [H_3, [H_3, H_4]]]/8) \\ &- (h^5/120)[H_3, [H_3, [H_3, [H_3, H_4]]]], \end{aligned} \tag{12.10.13}$$

$$f_m = \text{ expression involving } H_m \text{ and the } H_\ell \text{ with } \ell < m. \tag{12.10.14}$$

See Exercise 10.1.

We conclude that all the factors in a Zassenhaus representation can themselves be written in factored product form. These maps can now be concatenated together to yield a final approximation for $\exp(h : -H :)$ that is also in factored product form and is accurate through some order in h. (Note that if H_1 terms are present in H, then we must at this point assume they are small in order to use the concatenation formulas of Section 9.3.) The net result of our discussion is that the use of Zassenhaus symplectic integrator formulas makes it possible to find "linear" maps $\mathcal{R}(h)$ and generators $f_m(h)$ such that $\exp(-h : H :)$ has the factorization

$$\exp(-h : H :) = \exp[: f_1(h) :]\mathcal{R}(h) \exp[: f_3(h) :] \exp[: f_4(h) :] \cdots \times [1 + O(h^{N+1})] \tag{12.10.15}$$

where $(N + 1)$ is the order of the error in the Zassenhaus formula.

Application to Scaling, Splitting, and Squaring

For the autonomous case that we have just been considering, we are now able to use scaling, splitting, and squaring (as in Section 10.7) with (10.17) now playing the role of a splitting formula. As before, we define τ by writing

$$\tau = t/2^n, \tag{12.10.16}$$

and find the approximation

$$\begin{aligned}\mathcal{M} &= \exp(t : -H :) = [\exp(\tau : -H :)]^{2^n}\\ &= \{\cdots\{\{\exp[: f_1(\tau) :]\mathcal{R}(\tau)\exp[: f_3(\tau) :]\exp[: f_4(\tau) :]\cdots\}^2\}^2\cdots\}^2\\ &(n \text{ squarings}).\end{aligned} \tag{12.10.17}$$

[Note: The quantity τ as given by (10.16) should not be confused with that used in (1.3) or (9.33.]

What will be the relative error for this approximation? We expect that it will scale as $(1/2^n)^N$. For a more precise result we need an estimate for the error term in the underlying Zassenhaus formula. Suppose, for example, we use the Zassenhaus formula $\mathcal{S}_4$ as given by (1.35). Although we have discussed Zassenhaus formulas in the context of symplectic integrators, they are really *operator identities* that hold for any set of linear operators. (Consequently, they may be used in other contexts including the construction of integrators designed to preserve group properties for any Lie group. For example, they may be used in the context of rigid-body motion to preserve the orthogonality condition and in the context of quantum dynamics to preserve the unitarity condition.) To emphasize this fact, let us introduce the notation

$$\begin{aligned}\mathcal{S}_4(h\mathcal{A}, h\mathcal{B}) \;\;=\;\; & \exp(w_1 h\mathcal{A})\exp(w_2 h\mathcal{B})\exp(w_3 h\mathcal{A})\exp(w_4 h\mathcal{B})\,\times\\ & \exp(w_5 h\mathcal{A})\exp(w_6 h\mathcal{B})\exp(w_7 h\mathcal{A}),\end{aligned} \tag{12.10.18}$$

where $\mathcal{A}$ and $\mathcal{B}$ are any pair of linear operators. Then we have a result of the form

$$\mathcal{S}_4(h\mathcal{A}, h\mathcal{B}) = \exp[-h(\mathcal{A}+\mathcal{B}) + \mathcal{C}_5 h^5 + O(h^7)]. \tag{12.10.19}$$

What we need for error analysis is an estimate for the term $\mathcal{C}_5$. In analogy to the error term in (9.13) and based on the general properties of the BCH series, we expect that $\mathcal{C}_5$ will be made of multiple commutators of $\mathcal{A}$ and $\mathcal{B}$. First there will be a term with four $\mathcal{A}$'s and one $\mathcal{B}$. Next there will be terms with three $\mathcal{A}$'s and two $\mathcal{B}$'s, etc. Finally, there will be a term with one $\mathcal{A}$ and four $\mathcal{B}$'s. Consequently, $\mathcal{C}_5$ can be written in the form

$$\mathcal{C}_5(\mathcal{A}, \mathcal{B}) = d_1 \#\mathcal{A}\#^4\mathcal{B} + \cdots + d_{\text{last}}\#\mathcal{B}\#^4\mathcal{A}. \tag{12.10.20}$$

Here, in accord with the notation of Chapter 8, $\#\mathcal{A}\#$ denotes the adjoint of $\mathcal{A}$ as defined in terms of the commutator,

$$\#\mathcal{A}\#\mathcal{B} = \{\mathcal{A}, \mathcal{B}\}. \tag{12.10.21}$$

Thus, to specify $\mathcal{C}_5$, we need to determine the coefficients $d_1 \cdots d_{\text{last}}$.

The determination of all the coefficients $d_1 \cdots d_{\text{last}}$ is a sizable algebraic task. However, we can find d_1 and d_{last} fairly easily. Since (10.16) is an operator identity, as we have just discussed, it must hold for any linear operators $\mathcal{A}$ and $\mathcal{B}$. In particular, it must hold for 2×2 matrices. Left F and G be the matrices

$$F = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{12.10.22}$$

$$G = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}. \tag{12.10.23}$$

They satisfy the commutation rules

$$\#F\#G = \{F, G\} = 2G, \tag{12.10.24}$$

$$\#G\#F = \{G, F\} = -\{F, G\} = -2G. \tag{12.10.25}$$

It follows from these rules that

$$\mathcal{C}_5(F, G) = d_1\#F\#^4G + \cdots + d_{\text{last}}\#G\#^4F = 2^4 d_1 G, \tag{12.10.26}$$

$$\mathcal{C}_5(G, F) = d_1\#G\#^4F + \cdots + d_{\text{last}}\#F\#^4G = 2^4 d_{\text{last}} G. \tag{12.10.27}$$

That is, only the first term in the expansion (10.17) contributes to $\mathcal{C}_5(F, G)$, and only the last term contributes to $\mathcal{C}_5(G, F)$. Consequently, we have the matrix identity

$$\begin{aligned} \mathcal{S}_4(hF, hG) &= \exp[-h(F + G) + 16d_1Gh^5 + O(h^7)] \\ &= \exp[-h(F + G)] \exp[16d_1Gh^5 + O(h^6)] \\ &= \exp[-h(F + G)][1 + 16d_1Gh^5 + O(h^6)]. \end{aligned} \tag{12.10.28}$$

From (10.25) it follows that

$$16d_1Gh^5 = \mathcal{S}_4(hF, hG) - \exp[-h(F + G)] + O(h^6). \tag{12.10.29}$$

Similarly, we have the result

$$16d_{\text{last}}Gh^5 = \mathcal{S}_4(hG, hF) - \exp[-h(F + G)] + O(h^6). \tag{12.10.30}$$

At this point we observe that the right sides of (10.26) and (10.27) can be evaluated exactly. First, it is easily verified that the matrix $(F + G)$ has the property

$$(F + G)^2 = I. \tag{12.10.31}$$

Consequently, we have the relation

$$\begin{aligned} \exp[-h(F + G)] &= \cosh[h(F + G)] - \sinh[h(F + G)] = I\cosh(h) - (F + G)\sinh(h) \\ &= \begin{pmatrix} \exp(-h) & -\sinh(h) \\ 0 & \exp(h) \end{pmatrix}. \end{aligned} \tag{12.10.32}$$

Now take 1,2 matrix elements of both sides of (10.26) and (10.27) to get the results

$$16d_1h^5 = [\mathcal{S}_4(hF, hG)]_{12} + \sinh(h), \tag{12.10.33}$$

$$16d_{\text{last}}h^5 = [\mathcal{S}_4(hG, hF)]_{12} + \sinh(h). \tag{12.10.34}$$

Next, we have the relations

$$\exp(\sigma F) = \begin{pmatrix} \exp(\sigma) & 0 \\ 0 & \exp(-\sigma) \end{pmatrix}, \tag{12.10.35}$$

$$\exp(\sigma G) = \begin{pmatrix} 1 & \sigma \\ 0 & 1 \end{pmatrix}. \tag{12.10.36}$$

Consequently, all the factors in $\mathcal{S}_4(hF, hG)$ and $\mathcal{S}_4(hG, hF)$ are known, and the multiplications indicated in (10.15) can be carried out exactly. Finally, we can expand the right sides of (10.30) and (10.31) as power series in h and extract the terms of degree 5. Doing so gives the results

$$d_1 = (1/16)[18(2)^{2/3} - 40(2^{1/3}) + 22]/[9360(2)^{2/3} - 7200(2)^{1/3} - 5760] \simeq 4.14 \times 10^{-4}, \tag{12.10.37}$$

$$d_{\text{last}} = (1/16)(1/12)\{(1/10) + [14(2^{1/3}) - 11(2^{2/3})]/[2 - (2^{1/3})]^5 \simeq 4.68 \times 10^{-3}. \tag{12.10.38}$$

We are now prepared to make an error analysis similar to that of Section 10.7. For A and B given by (10.3) and (10.4), and assuming $H_1 = 0$ for simplicity, we find for $\mathcal{C}_5$ the result

$$\mathcal{C}_5 = d_1\#H_2\#^4 : H_r : + \cdots + d_{\text{last}}\#H_r\#^4 : H_2 : . \tag{12.10.39}$$

We know, in this context, that $\mathcal{C}_5$ must be a Lie operator. Let C be the function associated with $\mathcal{C}_5$,

$$\mathcal{C}_5 =: C : . \tag{12.10.40}$$

With this notation, (10.39) is equivalent to the relation

$$C = d_1 : H_2 :^4 H_r + \cdots + d_{\text{last}} : H_r :^4 H_2. \tag{12.10.41}$$

We will also focus our attention on the first term in (10.41), which is equivalent to the assumption that H_2 has a larger effect than H_r. Then, in view of (9.36) and in line with our assumption, we want to make the comparison

$$hH_r \overset{?}{\leftrightarrow} h^5 d_1 : H_2 :^4 H_r \tag{12.10.42}$$

where, according to (10.13) and (10.14),

$$h = \tau = t/2^n. \tag{12.10.43}$$

By making use of (7.52) and (7.55), and looking at (10.39), we see that the relative error can be written in the form

$$\text{relative error } \sim d_1(m\lambda/2^n)^4. \tag{12.10.44}$$

Suppose, as in Section 10.7, we limit our attention to the case $m \leq 8$ and select n so that (7.57) is satisfied. Then we find the result

$$\text{relative error } \sim d_1(1/20)^4 \simeq 3 \times 10^{-9}. \tag{12.10.45}$$

This error is somewhat larger than that given by (7.59). However, the relations (10.6) through (10.11) are simpler than the relations (7.28) through (7.32). Consequently, the use of Zassenhaus formulas for splitting is easier to program. Of course both errors can be decreased substantially, if desired, by a modest increase in the number of squarings n.

Exercises

12.10.1. The aim of this exercise is to show that the results (10.8) through (10.14) can be obtained from (10.6.14) through (10.6.20). What you are to do is to integrate the equations (10.6.14) through (10.6.20) with respect to t over the interval $t = 0$ to $t = h$. Let us begin. Verify that integrating (10.6.14), remembering (10.7) and that H_3 is assumed to be time independent, gives the result

$$f_3 = \int_0^h dt\ \dot{f}_3 = -\int_0^h dt\ H_3 = -hH_3. \tag{12.10.46}$$

Therefore (10.8) follows from (10.6.14). Next consider f_4. From (10.46) we know that

$$f_3(t) = -tH_3. \tag{12.10.47}$$

With regard to the ingredients of (10.5.15), show, in view of (10.7) and (10.47), that in our case

$$(: f_3 : /2)(-H_3^{\text{int}}) = 0. \tag{12.10.48}$$

Verify, therefore, that integration of (10.6.15) in our case yields the result

$$f_4 = \int_0^h dt\ \dot{f}_4 = -\int_0^h dt\ H_4 = -hH_4, \tag{12.10.49}$$

in agreement with (10.9). Move on to the case of f_5. Verify that in our case some of the ingredients in (10.6.16) also vanish so that there is the result

$$f_5 = \int_0^h dt\ \dot{f}_5 = -\int_0^h dt\ H_5 + \int_0^h dt\ t[H_4, H_3] = -hH_5 - (h^2/2)[H_3, H_4], \tag{12.10.50}$$

in agreement with (10.10). What about f_6? Verify that in our case several of the ingredients in (10.6.17) vanish so that there is the result

$$\begin{aligned} f_6 &= \int_0^h dt\ \dot{f}_6 = -\int_0^h dt\ H_6 + \int_0^h dt\ \ : f_5 : (-H_3) \\ &= -hH_6 - \int_0^h dt\ \ : tH_5 + (t^2/2)[H_3, H_4] : (-H_3) \\ &= -hH_6 - (h^2/2)[H_3, H_5] - (h^3/6)[H_3, [H_3, H_4]], \end{aligned} \tag{12.10.51}$$

in agreement with (10.11). Show that (10.12) and (10.13) can be obtained analogously.

12.10.2.

12.11 Other Zassenhaus Formulas and Their Use

There is a somewhat different class of Zassenhaus formulas that also merits discussion. Suppose we rewrite (9.13) in the more general form

$$\exp[(s/2)\mathcal{A}]\exp(t\mathcal{B})\exp[(s/2)\mathcal{A}] = \exp[(s\mathcal{A}+t\mathcal{B})+\mathcal{C}(s\mathcal{A},t\mathcal{B})] \tag{12.11.1}$$

where

$$\mathcal{C} = (s^2t/24)\{\mathcal{A},\{\mathcal{A},\mathcal{B}\}\} - (st^2/12)\{\mathcal{B},\{\mathcal{B},\mathcal{A}\}\} + \cdots . \tag{12.11.2}$$

Here, as before, $\mathcal{A}$ and $\mathcal{B}$ are any pair of linear operators, and s and t are expansion parameters. We see that the left side of (11.1) produces the desired result $\exp(s\mathcal{A}+t\mathcal{B})$ save for an error $\mathcal{C}$ that contains, among other things, terms linear in s (but higher order in t) and terms linear in t (but higher order is s). Examination of (10.15) through (10.17) shows that the higher order Zassenhaus integrator formulas have similar properties. For example, reference to (10.16) and (10.17) shows that the $\mathcal{S}_4$ given by (10.15) has errors linear in t that are proportional to s^4, errors quadratic in t that are proportional to s^3, etc.

Suppose we set for ourselves what will turn out to be an easier goal: find Zassenhaus approximations that are only correct through terms linear in t, but the term that is independent of t and the term that is linear in t should be correct to *high* order in s. Our starting point is the relation (8.8.13). See Section 8.8. This relation has the generalization

$$\exp(s\mathcal{A}+t\mathcal{B}) = \exp[O(t^2)]\exp[\text{iex}(s\#\mathcal{A}\#)(t\mathcal{B})]\exp(s\mathcal{A}) \tag{12.11.3}$$

where

$$\text{iex}(s\#\mathcal{A}\#)(t\mathcal{B}) = \int_0^1 d\tau\exp(\tau s\#\mathcal{A}\#)(t\mathcal{B}). \tag{12.11.4}$$

By construction we know that the term in (11.3) that is independent of t and the term that is linear in t are both *exact* in s.

The next step is to convert the integral (11.4) into a finite sum with the aid of a *quadrature formula*. Suppose we wish to integrate some function (operator) $\mathcal{G}(\tau)$. A quadrature formula is a set of k successive *sampling points* τ_i in the interval $[0,1]$ and *weights* w_i such that

$$\int_0^1 d\tau\mathcal{G}(\tau) \simeq \sum_{i=1}^k w_i\mathcal{G}(\tau_i). \tag{12.11.5}$$

In our case

$$\mathcal{G}(\tau) = \exp(\tau s\#\mathcal{A}\#)(t\mathcal{B}). \tag{12.11.6}$$

Shortly we will consider how the τ_i and w_i might be chosen. First, let us see how a quadrature formula can be used. With the aid of (11.5) and (11.6) we find the result

$$\begin{aligned}\exp[\text{iex}(s\#\mathcal{A}\#)(t\mathcal{B}) &= \exp[\int_0^1 d\tau\mathcal{G}(\tau)] \simeq \exp\left[\sum_{i=1}^k w_i\mathcal{G}(\tau_i)\right]\\ = \quad &\exp[O(t^2)]\exp[w_1\mathcal{G}(\tau_1)]\exp[w_2\mathcal{G}(\tau_2)]\cdots\exp[w_k\mathcal{G}(\tau_k)]. \end{aligned}\tag{12.11.7}$$

Note that the operators $\mathcal{G}(\tau_i)$ generally do not commute. Therefore the conversion of the exponential of a sum into a product of exponentials, which occurs as the last step in (11.7), produces correction terms that involve commutators. However, since the $\mathcal{G}(\tau_i)$ are linear in t, these commutators are $O(t^2)$ as indicated in (11.7).

For each factor in (11.7) we have the result

$$\begin{aligned}\exp[w_i\mathcal{G}(\tau_i)] &= \exp[\exp(\tau_i s\#\mathcal{A}\#)(w_i t\mathcal{B})]\\ &= \exp[\exp(\tau_i s\mathcal{A})(w_i t\mathcal{B})\exp(-\tau_i s\mathcal{A})]\\ &= \exp(\tau_i s\mathcal{A})\exp(w_i t\mathcal{B})\exp(-\tau_i s\mathcal{A}). \end{aligned} \tag{12.11.8}$$

Putting all these results together gives the relation

$$\begin{aligned}&\exp(s\mathcal{A}+t\mathcal{B}) \simeq \exp[O(t^2)]\exp(\tau_1 s\mathcal{A})\exp(w_1 t\mathcal{B})\exp(-\tau_1 s\mathcal{A}) \times\\ &\exp(\tau_2 s\mathcal{A})\exp(w_2 t\mathcal{B})\exp(-\tau_2 s\mathcal{A})\cdots \times\\ &\exp(\tau_k s\mathcal{A})\exp(w_k t\mathcal{B})\exp(-\tau_k s\mathcal{A})\exp(s\mathcal{A}).\end{aligned} \tag{12.11.9}$$

Finally, carrying out the indicated multiplications gives the result

$$\begin{aligned}&\exp(s\mathcal{A}+t\mathcal{B}) \simeq \exp[O(t^2)]\exp(\tau_1 s\mathcal{A})\exp(w_1 t\mathcal{B})\\ &\exp[(\tau_2-\tau_1)s\mathcal{A}]\exp(w_2 t\mathcal{B})\exp[(\tau_3-\tau_2)s\mathcal{A}]\cdots\\ &\exp(w_k t\mathcal{B})\exp[(1-\tau_k)s\mathcal{A}].\end{aligned} \tag{12.11.10}$$

This is the desired Zassenhaus approximation.

We must still consider how to select the τ_i and w_i. One possibility is to space the τ_i evenly with $\tau_1 = 0$ and $\tau_k = 1$,

$$\tau_i = (i-1)/(k-1). \tag{12.11.11}$$

In this case we should use *Newton-Cotes* weights. See Appendix T. For example, for the case $k = 3$ we have the celebrated *Simpson's rule* 1-4-1 formula

$$\int_0^1 d\tau\mathcal{G}(\tau) \simeq (1/6)\mathcal{G}(0) + (4/6)\mathcal{G}(1/2) + (1/6)\mathcal{G}(1). \tag{12.11.12}$$

Another appealing possibility is not to space the τ_i evenly, but rather to select them (as well as the weights w_i) in such a way that (for a fixed k) the order is maximized. This choice produces the family of *Legendre-Gauss* quadrature formulas. Again see Appendix T. For example, for $k = 3$ there is the formula

$$\int_0^1 d\tau\mathcal{G}(\tau) \simeq (5/18)\mathcal{G}(1/2-\sqrt{15}/10) + (8/18)\mathcal{G}(1/2) + (5/18)\mathcal{G}(1/2+\sqrt{15}/10). \tag{12.11.13}$$

As another example, consider the case $k = 2$. Then there is the formula

$$\int_0^1 d\tau\mathcal{G}(\tau) \simeq (1/2)\mathcal{G}(1/2-\sqrt{3}/6) + (1/2)\mathcal{G}(1/2+\sqrt{3}/6). \tag{12.11.14}$$

Because (11.5) replaces an integral by a sum, the term in the Zassenhaus approximation (11.10) that is linear in t is no longer exact in s. (However, the term independent of t still is exact.) We can estimate the error made in s from formulas of the kind (11.14), (11.16), and (11.18). Taylor series expansion of $G(\tau)$ as given by (11.6) provides the result

$$\sum_{i=1}^{k} w_i G(\tau_i) = \sum_{n=0}^{\infty} (s^n/n!)\#\mathcal{A}\#^n t\mathcal{B} \sum_{i=1}^{k} w_i \tau_i^n. \tag{12.11.15}$$

Let c_k be the error term in the relation

$$\sum_{i=1}^{k} w_i (\tau_i)^{\ell_{\max}+1} = 1/(\ell_{\max}+2) + c_k. \tag{12.11.16}$$

Again see Appendix T, and the example relations (T.1.6), (T.1.12), and (T.1.15). From (12.15) and (12.16) we find the error estimate

$$\sum_{i=1}^{k} w_i G(\tau_i) = \mathrm{iex}(s\#\mathcal{A}\#)(t\mathcal{B}) + [c_k/(\ell_{\max}+1)!] s^{(\ell_{\max}+1)} (\#\mathcal{A}\#)^{(\ell_{\max}+1)}(t\mathcal{B}) + O[t(s)^{(\ell_{\max}+2)}]. \tag{12.11.17}$$

Consequently, to examine relative error, we must make the comparison

$$t\mathcal{B} \overset{?}{\leftrightarrow} [c_k/(\ell_{\max+1}+1)!] s^{(\ell_{\max}+1)} (\#\mathcal{A}\#)^{\ell_{\max}+1}(t\mathcal{B}). \tag{12.11.18}$$

For example, suppose that $\mathcal{A}$ is the Lie operator $: H_2 :$ for a quadratic Hamiltonian, and $(t\mathcal{B})$ is the Lie operator $: H_r :$ for the remaining piece as in (10.4). Let JS be the matrix associated with H_2. In analogy to (7.56), define λ by the relation

$$\lambda = \|sJS\|. \tag{12.11.19}$$

Also, suppose that H_r does not contain terms beyond degree m. Then, in analogy to (6.5.2), we have the estimate

$$s^{\ell_{\max}+1} (\#\mathcal{A}\#)^{\ell_{\max}+1}(t\mathcal{B}) \sim (m\lambda)^{\ell_{\max}+1}(t\mathcal{B}). \tag{12.11.20}$$

Consequently, we conclude that the relative error in the Zassenhaus approximation (11.10) has the estimate

$$\text{relative error } \sim [c_k/(\ell_{\max}+1)!](m\lambda)^{(\ell_{\max}+1)}. \tag{12.11.21}$$

What uses can be made of Zasshaus approximations of the form (11.10)? In the context of Accelerator Physics, we will see in Chapters 12 and 12 that H_r becomes small in the limit of high energies. Therefore, at least three possible uses come to mind.

First, in the autonomous case, these Zassenhaus approximations can be used as splitting formulas for map computation by scaling, splitting, and squaring. Second, they can be used as symplectic integrators. The autonomous case can be treated using the form (11.10), and the nonautonomous case can be treated using related formulas. See Exercise 11.*. In the context of symplectic integrators, employing the Gaussian sampling points and weights seems particularly attractive because doing so minimizes (for a given k) the number of operators $\exp(w_i t\mathcal{B})$, whose evaluation is relatively expensive.

Finally, these Zassenhaus approximations can be used as the basis of an accelerator lattice *correction* scheme. Suppose we find H_r, or some terms in H_r, to be offensive. Then we can counter the effect of these terms on the performance of an accelerator lattice by placing, at the sampling points, local correctors having these same offensive properties. These correctors should be powered with strengths proportional to $(-w_i)$ in such a way that the *net* effect of the correctors and the offensive terms in H_r *cancel* to first order in t and high order in s. See Exercise *.*. In this case perhaps, although not necessarily, use of the Newton-Cotes sampling points might be more convenient from an engineering perspective.

Exercises

12.11.1.

Bibliography

General References

[1] E. Forest, *Beam Dynamics: A New Attitude and Framework*, Harwood Academic Press (1998).

[2] E. Hairer, S. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I. Nonstiff Problems*, Springer (1993).

[3] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II. Stiff and Differential Algebraic Problems*, Springer (2002).

Geometric/Symplectic Integration and Splitting Methods

[4] R.D. Ruth, *IEEE Trans. Nucl. Sci.* **30**, 2669 (1983).

[5] F. Neri, "Lie algebras and canonical integration", Department of Physics Technical Report, University of Maryland, 1988 (unpublished). This paper initiated the use of Lie methods to find symplectic integrators.

[6] J.M. Sanz-Serna and M.P. Calvo, *Numerical Hamiltonian Problems*, Chapman and Hall (1994).

[7] J.E. Marsden, G.W. Patrick, W.F. Shadwick, eds., *Integration Algorithms and Classical Mechanics*, American Mathematical Society (1996).

[8] A.M. Stuart and A.R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge Univeristy Press (1996).

[9] H. Yoshida, "Recent Progress in the Theory and Application of Symplectic Integrators", *Celestial Mechanics and Dynamical Astronomy* **56**, 27 (1993).

[10] E. Forest and R.D. Ruth, "Fourth-Order Symplectic Integration", *Physica D* **43**, 105 (1990).

[11] E. Forest, "Sixth-Order Lie Group Integrators", *J. Comp. Phys.* **99**, 209 (1992).

[12] E. Forest, J. Bengtsson, and M.F. Reusch,"Application of the Yoshida-Ruth techniques to implicit integration and multi-map explicit integration", *Phys. Lett. A* **158**, 99 (1991).

[13] Y. K. Wu, E. Forest, and D. Robin, “Explicit symplectic integrator for s-dependent static magnetic field”, *Physical Review E*, p. 046502 (2003).

[14] E. Forest, “Geometric Integration for Particle Accelerators”, *Journal of Physics A* **39**, 5321-5377 (2006).

[15] P.-V. Koseleff, “Recherche d’approximants d’exponentielles”, preprint, École Polytechnique (1993).

[16] P.J. Channell and C. Scovel, “Symplectic integration of Hamiltonian systems”, *Nonlinearity* **3**, 231 (1990).

[17] M. Campostrini and P. Rossi, “A comparison of numerical algorithms for dynamical fermions”, *Nuclear Physics B* **329**, 753 (1990).

[18] M. Suzuki, “On the convergence of exponential operators – the Zassenhaus formula, BCH formula and systematic approximants”, *Commun. Math. Phys.* **57**, 193-200 (1977).

[19] M. Suzuki, “Symmetry and Coherent Approximations in Non-Equilibrium Systems”, preprint, University of Tokyo.

[20] M. Suzuki, “General Theory of higher-order decomposition of exponential operators and symplectic integrators”, *Phys. Lett. A* **165**, 387 (1992).

[21] M. Suzuki, “General Nonsymmetric Higher-Order Decomposition of Exponential Operators and Symplectic Integrators”, *J. Phys. Soc. of Japan* **61**, 3015 (1992).

[22] M. Suzuki, “Fractal Decompostion of Exponential Operators with Applications to Many-body Theories and Monte Carlo Simulation”, *Phys. Lett. A* **146**, 319 (1990).

[23] M. Suzuki, “General theory of fractal path integrals with applications to many-body theories and statistical physics”, *J. Math. Phys.* **32**, 400 (1991). http://chaosbook.org/library/SuzukiJMP91.pdf

[24] M. Suzuki, “Symmetry and Systematics”, preprint, University of Japan.

[25] M. Suzuki, “Improved Trotter-like Formula”, *Phys. Lett. A* **180**, 232 (1993).

[26] M. Suzuki and K. Umeno, “Higher-Order Decomposition Theory of Exponential Operators and Its Applications to QMC and Nonlinear Dynamics”, *Computer Simulation Studies in Condensed Matter Physics VI*, D.P. Landau, K.K. Mon, and H.-B. Schüttler, eds., pp. 74-86, Springer-Verlag (1993).

[27] M. Glasner, D. Yevick, and B. Hermansson, “Sixth Order generalized propagation techniques”, *Electron. Lett.*, vol. 27, pp. 475-478, 1991.

[28] M. Glasner, D. Yevick, and B. Hermansson, “High Order generalized propagation techniques”, *J. Opt. Soc. B*, vol. 8, pp. 413-415, 1991.

[29] M. Glasner, D. Yevick, and B. Hermansson, “Generalized propagation formulas of arbitrarily high order”, *J. Chem. Phys.*, vol. 95, pp. 8266-8272, 1991.

[30] M. Glasner, D. Yevick, and B. Hermansson, “Computer generated generalized propagation techniques”, *Appl. Math. Lett.*, vol. 4, pp. 85-90, 1991.

[31] M. Glasner, D. Yevick, and B. Hermansson, “Generalized propagation techniques for longitudinally varying refractive index distributions”, *Math. and Comp. Modelling*, vol. 16, pp. 179-184, 1992.

[32] P. Saha and S. Tremaine, “Symplectic integrators for solar system dynamics”, *Astron. J.* **104**, 1633 (1992).

[33] M. Austin, P.S. Krishnaprasad, and L.-S. Wang, “On Symplectic and Almost Poisson Integration of Rigid Body Systems”, University of Maryland Systems Research Center, technical report TR91-45 (1991).

[34] Qin Meng-Zhao and Zhu Wen-Jie, “Construction of Higher Order Schemes for Ordinary Differential Equations by Composing Self-adjoint Lower Order Ones”, preprint, Computing Center, Academia Sinica, Beijing.

[35] Qin Meng-Zhao, “A Difference Scheme for the Hamiltonian Equation”, *J. Comp. Math.* **5**, 203 (1987).

[36] Feng Kang and Qin Meng-Zhao, “The Symplectic Methods for the Computation of Hamiltonian Equations”, in *Numerical Methods for partial differential equations*, Lect. Notes Math **1297**, 1, Springer (1987).

[37] Feng Kang and Qin Meng-Zhao, “Hamiltonian algorithms for Hamiltonian systems and a comparative numerical study”, *Computer Physics Communications* **65**, 173-187 (1991).

[38] H. De Raedt, *Product Formula Algorithms for Solving the Time Dependent Schrödinger Equation*, Computer Physics Reports 7, (North Holland, Amsterdam 1987).

[39] H. De Raedt and B. De Raedt, *Phys. Rev. A* **28**, 3575 (1983).

[40] H. De Raedt, “Quantum Dynamics in Nanoscale Devices”, in *Computational Physics*, K.H. Hoffmann and M. Schreiber, Eds., Springer (1996).

[41] J. Candy and W. Rozmus, “A Symplectic Integration Algorithm for Separable Hamiltonian Functions”. *J. Comput. Phys.* **92**, 230 (1991).

[42] S. Mikkola and P. Wiegert, “Regularizing time transformations in symplectic and composite integration”, *Celest. Mech. Dyn. Astron.* **82**, 375 (2002).

[43] J. Wisdom and M. Holman, “Symplectic maps for the n-body problem: Stability analysis”, *Astron. J.* **104**, 2022 (1992).

[44] R.I. McLachlan, "On the numerical integration of ordinary differential equations by symmetric composition methods", *SIAM J. Sci. Comp.* **16**, 151-168, (1995). Also available on the Web at http://www.massey.ac.nz/~rmclachl/sisc95.pdf.

[45] R.I. McLachlan, "Composition methods in the presence of small parameters", *BIT* **35**, 258-268, (1995).

[46] R. I. McLachlan and P. Atela, "The accuracy of symplectic integrators", *Nonlinearity* **5**, 541-562 (1992).

[47] R.I. McLachlan and G.R.W. Quispel, "Splitting methods", *Acta Numerica* **11**, 341-434 (2002).

[48] R.I. McLachlan and G.R.W. Quispel, "Explicit geometric integration of polynomial vector fields", *BIT* **44**, 515-538, (2004).

[49] R.I. McLachlan and G.R.W. Quispel, "Geometric integrators for ODEs", *Journal of Physics A* **39**, 5251-5285, (2006).

[50] G.R.W. Quispel and R.I. McLachlan, Eds., Special Issue on Geometric Numerical Integration of Differential Equations, *Journal of Physics A* **39** (2006).

[51] S. Blanes, F. Casas, and J. Ros, "Symplectic Integration with Processing: A General Study", *SIAM J. Sci. Comput.* (1998).

[52] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure Preserving Algorithms for Ordinary Differential Equations*, Springer (2002), corrected second printing 2004.

[53] B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge University Press (2004).

[54] S. Reich, "Backward error analysis for numerical integrators", *SIAM J. Numer. Anal.* **36**, 1549-1570 (1999).

[55] A. Iserles, H. Munthe-Kaas, S. Nørsett, and A. Zanna, "Lie-group methods", *Acta Numerica* **14**, 1–148, (2005). Also available on the Web at http://www.damtp.cam.ac.uk/user/na/NA_papers/NA2000_03.pdf.

[56] I.P. Omelyan, I.M. Mryglod, and R. Folk, "Symplectic Analytically Integrable Decomposition Algorithms: Classification, Derivation, and Application to Molecular Dynamics, Quantum and Celestial Mechanics Simulations", *Computer Physics Communications*, vol. 151, 273-314 (2003).

[57] S.-H. Tsai, H.K. Lee, and D.P. Landau, "Molecular and Spin Dynamics Simulations using Modern Integration Methods", *Am. J. Physics*, vol. 73, 615-624 (2005).

[58] L. Gauckler, E. Hairer, and C. Lubich, "Dynamics, Numerical Analysis, and some Geometry", preprint (10 October 2017) https://arxiv.org/abs/1710.03946

[59] S. Blanes and F. Casas, *A Concise Introduction to Geometric Numerical Integration*, CRC Press (2016).

Triplet Construction

[60] M. Creutz and A. Gocksch, "Higher-order Hybrid Monte Carlo Algorithms", *Phys. Rev. Lett.* **63**, 9 (1989).

[61] H. Yoshida, "Construction of higher order symplectic integrators", *Phys. Lett. A* **150**, 262 (1990).

[62] E. Forest, J. Bengtsson, and M. Reusch, "Application of the Yoshida-Ruth techniques to implicit integration and multi-map explicit integration", *Phys. Lett. A* **158**, 99 (1991).

Force Gradient Algorithms

[63] S.A. Chin and D.W. Kidwell, "Higher-order Force Gradient Symplectic Algorithms", *Phys. Rev. E* **62**, 8746 (2000).

[64] S.A. Chin and C.R. Chen, "Forward Symplectic Integrators for Solving Gravitational Few-Body Problems", *Celestial Mechanics and Dynamical Astronomy* **91**, 301-322 (2005).

Partitioned Runge Kutta and Symplectic Runge Kutta

[65] B. Blanes and P.C. Moan, "Practical symplectic partitioned Runge-Kutta and Runge-Kutta-Nyström methods", *Journal of Computational and Applied Mathematics* **142**, 313-330 (2002).

[66] A. Iserles, "Efficient Runge-Kutta Methods for Hamiltonian Equations", in *Advances on Computer Mathematics and its Applications*, E. Lipitakis, Ed., World Scientific (1993).

[67] M. Sofroniou and W. Oevel, "Symplectic Runge-Kutta Schemes I: Order Conditions", *SIAM Journal of Numerical Analysis* **34(5)** (1997).

[68] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure Preserving Algorithms for Ordinary Differential Equations*, Springer (2002), corrected second printing 2004.

[69] X. Tan, "Almost symplectic Runge-Kutta schemes for Hamiltonian systems", *Journal of Computational Physics* 203, 250-273 (2005).

[70] R. McLachlan, "A New Implementation of Symplectic Runge-Kutta Methods", *SIAM Journal on Scientific Computing*, Volume 29, Issue 4, 1637-1649 (2007).

[71] J.C. Butcher, *Numerical Methods for Ordinary Differential Equations*, Second Edition, John Wiley (2008). http://www.math.auckland.ac.nz/~butcher/ODE-book-2008/.

[72] D. Okunbor and R.D. Skeel, "Explicit Canonical Methods for Hamiltonian Systems", *Mathematics of Computation* **59**, 439-455 (1992).

[73] Lin-Yi Chou and P. W. Sharp, "On order 5 symplectic explicit Runge-Kutta Nyström methods", *Journal of Applied Mathematics & Decision Sciences* **4**, 143-150 (2000).

Symplectic Integrator for Motion in a Magnetic Field

[74] Y.K. Wu, E. Forest, and D.S. Robin, "Explicit symplectic integrator for s-dependent static magnetic field", *Phys. Rev. E* **68**, 046502, (2003).

[75] M. Aichinger, S.A. Chin, and E. Krotscheck, "Fourth-order Algorithms for Solving Local Schroedinger Equations in a Strong Magnetic Field", *Comp. Phys. Comm.* **171**, 197 (2005).

[76] E. Chacon-Golcher and F. Neri, "A symplectic integrator with arbitrary vector and scalar potentials", *Physics Letters A* **372**, 4661-4666, (2008).

Conservation of H

[77] Z. Ge and J. Marsden, "Lie-Poisson Hamilton-Jacobi theory and Lie-Poisson integrators", *Phys. Lett. A*, **133**, 134-139, (1988).

[78] G. Benettin and A. Giorgilli, "On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms", *J. Statist. Phys.* **74**, 1117 (1994).

[79] S. Reich, "Backward error analysis for numerical integrators", *SIAM J. Numer. Anal.* **36**, 1549-1570 (1999).

Symplectic Integration Using Generating Functions

[80] Feng Kang and Qin Meng-Zhao, "The Symplectic Methods for the Computation of Hamiltonian Equations", in *Numerical Methods for partial differential equations*, Lect. Notes Math **1297**, 1, Springer (1987).

[81] Feng Kang, Wu Hua-mo, Qin Meng-shao, and Wang Dao-liu, "Construction of Canonical Difference Schemes for Hamiltonian Formalism via Generating Functions", *Journal of Computational Mathematics* **11**, p. 71 (1989).

[82] Feng Kang, "The Calculus of Generating Functions and the Formal Energy for Hamiltonian Algorithms", *Journal of Computational Mathematics* **16**, p. 481 (1998).

[83] Feng Kang and Mengzhao Qin, *Symplectic Geometric Algorithms for Hamiltonian Systems*, Zhejiang Publishing and Springer-Verlag (2010).

[84] P. J. Channell and C. Scovel, "Symplectic integration of Hamiltonian systems", *Nonlinearity* **3**, 231-259 (1990).

Local Correction

[85] D. Neuffer and E. Forest, Phys. Lett. A **135**, 197, (1989).

[86] D. Neuffer, Nucl. Instr. and Meth. A **274**, 400, (1989).

[87] D. Neuffer, Proceedings of the Workshop on Effects of Errors in Accelerators, Corpus Christi, TX, 1991, A. Chao, Ed., AIP Conf. Proc. **255**, 215, (1992).

[88] E. Forest, *Beam Dynamics*, Section 11.4, Harwood Academic (1998).

Chapter 13

Transfer Maps for Idealized Straight Beam-Line Elements

13.1 Background

In this chapter we will describe the computation of transfer maps for idealized straight beam-line elements. Here we make two assumptions: First, the element geometry is such that Cartesian coordinates can conveniently be employed. Second, the design orbit is a straight line, which we take to be the z axis. For simplicity we will treat only magnetic elements, but the case of straight electric elements or straight electromagnetic elements can be handled similarly. Since the design orbit is assumed to be a straight line, the case of magnetic dipole fields is excluded.

We further assume that the design orbit is traversed in time with *constant* velocity v_z^0. In particular, to ensure that the velocity v_z^0 is indeed constant we also assume that the electric scalar potential vanishes, $\psi = 0$. For the same reason, the vector potential $\boldsymbol{A}$ is taken to be *time independent*. See (1.5.2). More complicated situations where v_z^0 is not constant, such as occurs in RF accelerating cavities, can be treated in an analogous way.

Since the design orbit is taken to lie along the z axis, it is convenient to take z as the independent variable. In this case, according to (1.6.16), the Hamiltonian K is given by the relation

$$K = -[(p_t^{\text{can}})^2/c^2 - m^2c^2 - (p_x^{\text{can}} - qA_x)^2 - (p_y^{\text{can}} - qA_y)^2]^{1/2} - qA_z. \tag{13.1.1}$$

Here we have taken care to indicate that the momenta employed are *canonical*. The idealizations we will make in this chapter are that the vector potential $\boldsymbol{A}$ is z *independent* and fringe-field effects can be neglected. These restrictions will be removed in the subsequent Chapters 15 through 21.

13.1.1 Specification of Design Orbit

Under the assumptions made about the *design* orbit, we may write the relations

$$x^d = y^d = 0. \tag{13.1.2}$$

Also, the transverse velocities and therefore the transverse mechanical momenta will vanish on the design orbit. We further assume that the fields and selected gauge for the vector potential are such that the transverse components of the vector potential vanish on the design orbit. It follows that in this case the transverse canonical momenta will also vanish on the design orbit,

$$p_x^{\text{can d}} = p_x^{\text{can d}} = 0. \tag{13.1.3}$$

With regard to the the momentum p_t^{can}, since we have assumed that $\psi = 0$ and v_z^0 is constant, we may write

$$p_t^{\text{can d}} = p_t^0 \tag{13.1.4}$$

where p_t^0 is the value of p_t^{can} on the design orbit.

To complete our description of the design orbit, we need to find the time t as a function of z on this orbit. We may write

$$(dt/dz)|_{\text{design orbit}} = 1/[(dz/dt)|_{\text{design orbit}}] = 1/v_z^0. \tag{13.1.5}$$

Here, as described earlier, v_z^0 is the design velocity on the design orbit, from which it follows that there is the relation

$$v_z^0 = (c^2\gamma m v_z^0)/(\gamma m c^2) = -c^2 p^0/p_t^0 \tag{13.1.6}$$

where γ denotes the usual relativistic factor and p^0 is the value of p_z^{mech} on the design orbit. [This same result can be obtained from (1.1) by observing that p_x^{can}, p_y^{can}, A_x, and A_y all vanish on the design orbit. See Exercise *.] Note also that there are the relations

$$p_t^0 = -[m^2c^4 + (p^0)^2c^2]^{1/2}, \tag{13.1.7}$$

$$p^0 = [(p_t^0/c)^2 - m^2c^2]^{1/2}. \tag{13.1.8}$$

Since v_z^0 is constant, integration of (1.5) yields the relation

$$t^d(z) = z/v_z^0 \tag{13.1.9}$$

where we have taken the origin in time to be such that the design orbit passes through $z = 0$ at the time $t = 0$.

13.1.2 Deviation Variables

Under the assumptions just made about the design orbit, we may introduce transverse coordinate deviation variables ξ, η by the definitions

$$\xi = x - x^d = x, \tag{13.1.10}$$

$$\eta = y - y^d = y. \tag{13.1.11}$$

We also introduce transverse momentum deviation variables p_ξ, p_η by the definitions

$$p_\xi = p_x^{\text{can}} - p_x^{\text{can d}} = p_x^{\text{can}}, \tag{13.1.12}$$

$$p_\eta = p_y^{\mathrm{can}} - p_y^{\mathrm{can\ d}} = p_y^{\mathrm{can}}. \tag{13.1.13}$$

With the results and definitions (1.2) through (1.4) and (1.9) through (1.13) in hand, we are now able to introduce a full set of of deviation variables $(\xi, \eta, T; p_\xi, p_\eta, p_T)$ by adding to the definitions (1.10) through (1.13) the definitions

$$T = t - t^d = t - z/v_z^0, \tag{13.1.14}$$

$$p_T = p_t^{\mathrm{can}} - p_t^{\mathrm{can\ d}} = p_t^{\mathrm{can}} - p_t^0. \tag{13.1.15}$$

Note that by construction the deviation variables all vanish on the design orbit as desired. Since the relations (1.9) through (1.14) simply amount to a phase-space translation, it follows that the relation between the original variables and the deviation variables is a canonical transformation.

13.1.3 Deviation Variable Hamiltonian

Since the transformation given by (1.9) through (1.14) is canonical, the equations of motion for the deviation variables must also arise from a Hamiltonian, which we will call $K^{\mathrm{new}}(\xi, \eta, T, p_\xi, p_\eta, p_T; z)$. Our task is to find K^{new} in terms of K. To do so, we will use the machinery of Subsection 10.4.1.

Following (10.4.2), define a function $\bar{K}(\xi, \eta, T, p_\xi, p_\eta, p_T; z)$ by the rule

$$\bar{K}(\xi, \eta, T, p_\xi, p_\eta, p_T; z) = K(\xi, \eta, z/v_z^0 + T, p_\xi, p_\eta, p_t^0 + p_T; z). \tag{13.1.16}$$

Then, according to (10.4.20), the new Hamiltonian is given by the rule

$$\begin{aligned} &K^{\mathrm{new}}(\xi, \eta, T, p_\xi, p_\eta, p_T; z) = \\ &\quad \bar{K}(\xi, \eta, T, p_\xi, p_\eta, p_T; z) - \bar{K}_1(\xi, \eta, T, p_\xi, p_\eta, p_T; z). \end{aligned} \tag{13.1.17}$$

Let us work out the implications of this rule for the case where K is given by (1.1). For $\bar{K}$ we find the result

$$\begin{aligned} &\bar{K}(\xi, \eta, T, p_\xi, p_\eta, p_T; z) = \\ &\quad -[(p_t^0 + p_T)^2/c^2 - m^2c^2 - (p_\xi - qA_x)^2 - (p_\eta - qA_y)^2]^{1/2} - qA_z. \end{aligned} \tag{13.1.18}$$

We next assume that $\boldsymbol{A}$ is time independent, the expansions of A_x and A_y begin with linear terms, and the expansion of A_z begins with quadratic terms. In that case $\boldsymbol{A}$ does not contribute to $\bar{K}_1(\xi, \eta, T, p_\xi, p_\eta, p_T; z)$, and we find from (1.17) the result

$$\begin{aligned} \bar{K}_1(\xi, \eta, T, p_\xi, p_\eta, p_T; z) &= -(p_t^0/c^2)[(p_t^0/c)^2 - m^2c^2]^{-1/2} p_T \\ &= -(p_t^0/c^2)(1/p^0) p_T \\ &= p_T/v_z^0. \end{aligned} \tag{13.1.19}$$

Here we have used (1.6) and (1.8). Finally, combining (1.17) through (1.19) gives the result

$$\begin{aligned} &K^{\mathrm{new}}(\xi, \eta, T, p_\xi, p_\eta, p_T; z) = \\ &\quad -[(p_t^0 + p_T)^2/c^2 - m^2c^2 - (p_\xi - qA_x)^2 - (p_\eta - qA_y)^2]^{1/2} - qA_z - p_T/v_z^0. \end{aligned} \tag{13.1.20}$$

13.1.4 Dimensionless Scaled Deviation Variables

For many purpose it is useful to describe trajectories and maps in terms of dimensionless variables. Let ℓ be some convenient scale length.[1] Introduce dimensionless variables $(X, Y, \tau; P_x, P_y, P_\tau)$, defined in terms of the deviation variables and the scale length, by the rules

$$X = \xi/\ell, \tag{13.1.21}$$

$$Y = \eta/\ell, \tag{13.1.22}$$

$$\tau = cT/\ell; \tag{13.1.23}$$

$$P_x = p_\xi/p^0, \tag{13.1.24}$$

$$P_y = p_\eta/p^0, \tag{13.1.25}$$

$$P_\tau = p_T/(p^0 c). \tag{13.1.26}$$

At this point it is useful to relate P_τ, which may be viewed as a scaled energy deviation (with a minus sign), to the momentum deviation parameter δ of Exercise 1.7.6. They are connected by the relations

$$\begin{aligned} P_\tau &= -(1/\beta_0)\{[1 + (2\delta + \delta^2)\beta_0^2]^{1/2} - 1\} \\ &= -\beta_0\delta + (\delta^2/2)(\beta_0^3 - \beta_0) - (\delta^3/2)(\beta_0^5 - \beta_0^3) + \cdots , \end{aligned} \tag{13.1.27}$$

$$\begin{aligned} \delta &= (1 - 2P_\tau/\beta_0 + P_\tau^2)^{1/2} - 1 \\ &= -P_\tau/\beta_0 + (P_\tau^2/2)(1 - \beta_0^{-2}) + \cdots . \end{aligned} \tag{13.1.28}$$

Here β_0 is the usual relativistic factor evaluated on the design orbit,

$$\beta_0 = v_z^0/c = -cp^0/p_t^0. \tag{13.1.29}$$

Note that in the ultra relativistic limit $\beta_0 \to 1$ there are the relations

$$P_\tau = -\delta, \tag{13.1.30}$$

$$\delta = -P_\tau. \tag{13.1.31}$$

See Exercise *.

13.1.5 Scaled Deviation-Variable Hamiltonian

We will now seek equations of motion for the scaled deviation variables. We will learn that they also can be derived from what we will call *scaled* deviation-variable Hamiltonian and will denote by the symbol H^s. To do so requires some care.

Although the Poisson brackets of the new coordinates with each other and the Poisson brackets of the new momenta with each other all vanish, the transformation given by (2.54)

[1] Here the coordinate scale factor ℓ is not to be confused with the "path length" ℓ of Exercise 1.7.6. Note also that in this subsection we use the notation p^0 to denote the quantity $p_0^{\rm mech}$ of Exercise 1.7.6.

through (2.59) is not canonical because the Poisson brackets of the new coordinates with their corresponding new momenta do not have the value 1. Instead they have the common value

$$[X, P_x] = [Y, P_y] = [\tau, P_\tau] = (p^0\ell)^{-1}. \tag{13.1.32}$$

Nevertheless, it is still possible to obtain the equations of motion for the deviation variables from a Hamiltonian providing the Hamiltonian, which we now denote by H, is taken to be the function

$$H^s = K^{\rm new}/(p^0\ell), \tag{13.1.33}$$

and we treat the deviation variables as being canonically conjugate. That is, the Poisson brackets of the new coordinates with each other and the Poisson brackets of the new momenta with each other all vanish, and the Poisson brackets of the new coordinates with their corresponding new momenta are taken to have the value 1,

$$[X, P_x] = [Y, P_y] = [\tau, P_\tau] = 1. \tag{13.1.34}$$

See Exercise *.

Let us apply the Ansatz (2.66) to the Hamiltonian $\bar{H}$ given by (2.53) and (1.1). So doing gives the preliminary result

$$\begin{aligned} H(X, Y, \tau, P_x, P_y, P_\tau; z) &= [1/(p^0\ell)][K - (p_T + p_t^0)/v_z^0] \\ &= [1/(p^0\ell)]K - [1/(p^0\ell)](p_T + p_t^0)/v_z^0. \end{aligned} \tag{13.1.35}$$

We will work separately on each of the two terms appearing on the far right side of (2.68).

Save for the $1/\ell$ factor the first term takes the form

$$(1/p^0)K = -\{[(p_t^{\rm can}/p^0c)^2 - (mc/p^0)^2 - (p_x^{\rm can}/p^0 - A_x^s)^2 - (p_y^{\rm can}/p^0 - A_y^s)^2]^{1/2} + A_z^s\} \tag{13.1.36}$$

where $\boldsymbol{A}^s$ is a *scaled* vector potential given by

$$\boldsymbol{A}^s(X, Y, z) = (q/p^0)\boldsymbol{A}(\ell X, \ell Y, z). \tag{13.1.37}$$

Further manipulation employing (2.45) and (2.59) produces the relation

$$\begin{aligned} [p_t^{\rm can}/(p^0c)]^2 - (mc/p^0)^2 &= [(p_T + p_t^0)/(p^0c)]^2 - (mc/p^0)^2 \\ &= [(p^0cP_\tau + p_t^0)/(p^0c)]^2 - (mc/p^0)^2 \\ &= [P_\tau + p_t^0/(p^0c)]^2 - (mc/p^0)^2. \end{aligned} \tag{13.1.38}$$

From (2.62) there is the relation

$$p_t^0/(p^0c) = -1/\beta_0. \tag{13.1.39}$$

There is also the relation

$$mc/p^0 = mc/(m\gamma_0\beta_0c) = 1/(\gamma_0\beta_0) \tag{13.1.40}$$

where γ_0 is the usual relativistic factor evaluated on the design orbit. It follows that there is the relation

$$\begin{aligned}[p_t^{\rm can}/(p^0c)]^2-(mc/p^0)^2 &= (P_\tau-1/\beta_0)^2-1/(\gamma_0\beta_0)^2\\ &= P_\tau^2-(2P_\tau/\beta_0)+1/\beta_0^2-1/(\gamma_0\beta_0)^2. \end{aligned} \tag{13.1.41}$$

But, there is the identity

$$1/\beta_0^2-1/(\gamma_0\beta_0)^2=(1/\beta_0)^2[1-1/\gamma_0^2]=(1/\beta_0)^2[1-(1-\beta_0^2)]=1. \tag{13.1.42}$$

Consequently there is the net result

$$[p_t^{\rm can}/(p^0c)]^2-(mc/p^0)^2=P_\tau^2-(2P_\tau/\beta_0)+1. \tag{13.1.43}$$

We conclude that

$$[1/(p_0\ell)]K=-(1/\ell)\{[1-(2P_\tau/\beta_0)+P_\tau^2-(P_x-A_x^s)^2-(P_y-A_y^s)^2]^{1/2}+A_z^s\} \tag{13.1.44}$$

where we have also used (2.38), (2.39), (2.57), and (2.58).

What remains is to work on he second term on the right side of (2.68). Save for a $(-1/\ell)$ factor it takes the form

$$\begin{aligned}(1/p^0)(p_T+p_t^0)/v_z^0 &= (1/p^0)(p_0cP_\tau+p_t^0)[-p_t^0/(p_0c^2)]\\ &= -[p_t^0/(p^0c)]P_\tau-(p_t^0)^2/(p^0c)^2\\ &= (P_\tau/\beta_0)-(1/\beta_0^2). \end{aligned} \tag{13.1.45}$$

Here we have used (2.59) and (2.72). We conclude that

$$-[1/(p^0\ell)](p_T+p_t^0)/v_z^0=(1/\ell)[(P_\tau/\beta_0)-(1/\beta_0^2)]. \tag{13.1.46}$$

We are, at last, ready to compute H. Upon combining (2.68), (2.77), and (2.79) we find the final result

$$H=-(1/\ell)\{[1-(2P_\tau/\beta_0)+P_\tau^2-(P_x-A_x^s)^2-(P_y-A_y^s)^2]^{1/2}+A_z^s+(P_\tau/\beta_0)\}. \tag{13.1.47}$$

In the case of no magnetic field (1.45) takes the form

$$H=-(1/\ell)\{[1-(2P_\tau/\beta_0)+P_\tau^2-P_x^2-P_y^2]^{1/2}+(P_\tau/\beta_0)\}. \tag{13.1.48}$$

Let us use this Hamiltonian to compute x'. From Hamilton's equations of motion we find the result

$$\begin{aligned} x' &= dx/dz=\ell dX/dz=\ell\partial H/\partial P_x\\ &= P_x[1-(2P_\tau/\beta_0)+P_\tau^2-P_x^2-P_y^2]^{-1/2}\\ &= P_x+P_xP_\tau/\beta_0+(1/2)P_x[P_\tau^2(3\beta_0^{-2}-1)+P_x^2+P_y^2]+\cdots. \end{aligned} \tag{13.1.49}$$

We see that x' agrees with P_x in lowest order; but there are second-order chromatic differences, and third- and higher-order geometric and chromatic differences. Also, X and x' are *not* canonically conjugate, $[X,x']\neq 1$.

Exercises

13.2 Axial Rotation

13.3 Drift

In this subsection we will compute the transfer map for a drift. To do so we begin with the Hamiltonian (1.1) and employ the vector potential given by (2.7) through (2.10). We then introduce deviation variables followed by scaled deviation variables. Next we find the scaled deviation-variable Hamiltonian. Finally, we expand the scaled deviation-variable Hamiltonian in a Taylor series, and employ this Taylor series to compute the transfer map.

13.4 Solenoid

$$R = *. \tag{13.4.1}$$

13.5 Wiggler/Undulator

13.6 Quadrupole

13.7 Sextupole

13.8 Octupole

13.9 Higher-Order Multipoles

13.10 Thin Lens Multipoles

13.11 Combined Function Quadrupole

13.12 Radio Frequency Cavity

Bibliography

Chapter 14

Transfer Maps for Idealized Curved Beam-Line Elements

Chapter 15

Taylor and Spherical and Cylindrical Harmonic Expansions

15.1 Introduction

Chapters 13 and 14 treated idealized beam-line elements for which variations in the field with position along the beam-line element, and fringe-field effects, were neglected. There are situations for which these neglected effects can be important when accurate modeling is desired. By developing various mathematical tools, this chapter prepares the way for Chapters 16 through 21 that describe the calculation of realistic transfer maps for straight beam-line elements, and Chapter 22 that describes the calculation of realistic transfer maps for general curved beam-line elements.

Restrictions Discovered by Hamilton (Symplecticity)

In previous chapters we learned that the motion of charged particles through any beam-line element can be described by the transfer map $\mathcal{M}$ for that element. We also learned that the equations of motion for charged particle motion can be derived from a Hamiltonian, and therefore $\mathcal{M}$ cannot be an arbitrary map, but must be a symplectic map. Consequently, through aberrations of order $(n-1)$, such a map has the Lie representation

$$\mathcal{M} = \mathcal{R}_2 \exp(:f_3:) \exp(:f_4:) \cdots \exp(:f_n:) \tag{15.1.1}$$

where $\mathcal{R}_2$ describes the linear part of the map. The linear map $\mathcal{R}_2$ and the Lie generators f_ℓ are determined by solving the equation of motion

$$\dot{\mathcal{M}} = \mathcal{M} : -H : \tag{15.1.2}$$

where

$$H = H_2 + H_3 + H_4 + \cdots \tag{15.1.3}$$

is the Hamiltonian expressed in terms of deviation variables and expanded in a homogeneous polynomial series. See Sections 10.1 and 10.4. The deviation variable Hamiltonian

H is determined in turn by the Hamiltonian K for which some coordinate is the independent variable. For example, in Cartesian coordinates and with z taken as the independent variable, K is given by the relation

$$K = -[(p_t + q\psi)^2/c^2 - m^2c^2 - (p_x - qA_x)^2 - (p_y - qA_y)^2]^{1/2} - qA_z. \tag{15.1.4}$$

Here ψ and $\boldsymbol{A}$ are the electric scalar and magnetic vector potentials, respectively. See (1.6.16). We conclude that (in the case of no electric fields, $\psi = 0$) what we need are Taylor expansions for the vector potential components A_x, A_y, A_z in the deviation variables x and y.

Restrictions Associated with Maxwell's Equations

For common beam-line elements the charged particles move in an evacuated beam pipe, and therefore the electric and magnetic fields controlling particle motion are source free in the vicinity of the beam. Correspondingly, the source-free Maxwell equations impose restrictions on what fields can be employed. This chapter begins with a discussion of harmonic functions and the use of spherical coordinates to obtain *spherical harmonic* expansions thereby leading to suitable Taylor expansions for source-free magnetic fields and their scalar and vector potentials. For our purposes this material is of particular use in terminating end fields. See Sections 16.7 and 22.8. It has other general uses as well including the treatment of ambient fields.

Next we will find, for the case of straight beam-line elements, expressions for the required Taylor expansions in terms of *on-axis gradients* with the aid of *cylindrical harmonic* expansions.[1] The on-axis gradients themselves are generally unspecified functions of z. In some simple cases they can be found analytically, as illustrated in Chapter 16. However, in general they must be determined numerically. Chapters 17 through 21 describe how this can be done in terms of magnetic field or magnetic potential values determined numerically at points on some regular 3-dimensional grid with the aid of some electromagnetic code.

In this part of the present chapter we will first learn how to characterize the magnetic scalar potential in terms of cylindrical harmonics described by on-axis gradients. Next we will find field expansions in terms of cylindrical harmonics. Then we will relate vector potentials to on-axis gradients. The work of this part of the chapter concludes with the treatment of an analytically soluble model problem, that of the magnetic monopole doublet, which will be used in Chapter 19 to benchmark the methods to be developed in Chapters 17 through 21.

The chapter ends with some closing remarks meant to provide added perspective.

[1] In this chapter we will use the word *cylindrical* to mean *circular* cylindrical. In subsequent chapters we will distinguish between circular, elliptic, and rectangular cylinders.

15.2 Spherical Expansion

15.2.1 Harmonic Functions and Absolute and Expansion Coordinates

In a current-free region the magnetic field $\boldsymbol{B}$ is curl free, and can therefore be described most simply in terms of a *magnetic scalar potential* Ψ with

$$\boldsymbol{B} = +\nabla\Psi. \tag{15.2.1}$$

Because $\boldsymbol{B}$ is also divergence free, Ψ must obey the Laplace equation,

$$\nabla^2\Psi = \nabla \cdot \boldsymbol{B} = 0. \tag{15.2.2}$$

Functions Ψ that obey the Laplace equation are said to be *harmonic.*

At this point it is convenient to introduce an "absolute" coordinate $\boldsymbol{R}$ and an "expansion" coordinate $\boldsymbol{r}$ about some "reference/expansion" point $\boldsymbol{R}_0$ by writing

$$\boldsymbol{R} = \boldsymbol{R}_0 + \boldsymbol{r}. \tag{15.2.3}$$

In terms of these variables we may define a related scalar potential ψ by writing

$$\psi(x, y, z; \boldsymbol{R}_0) = \Psi(\boldsymbol{R}_0 + \boldsymbol{r}) \tag{15.2.4}$$

where

$$\boldsymbol{r} = x\boldsymbol{e}_x + y\boldsymbol{e}_y + z\boldsymbol{e}_z. \tag{15.2.5}$$

Like Ψ, the related scalar potential ψ satisfies the relations

$$\boldsymbol{B} = +\nabla\psi, \tag{15.2.6}$$

and

$$\nabla^2\psi = 0. \tag{15.2.7}$$

Here, ψ is not to be confused with the ψ that was used in other sections to describe an electric field. Note also the + sign in (2.1 and (2.6) compared to the − sign in (1.4.2). These sign choices are a matter of convention. Also, strictly speaking, the derivatives in (2.1) and (2.2) are to be taken with respect to the components of $\boldsymbol{R}$, and the derivatives in (2.6) and (2.7) are to be taken with respect to the components of $\boldsymbol{r}$. Moreover, in subsequent work, we will sometimes suppress the dependence of ψ on $\boldsymbol{R}_0$ and simply write $\psi(x, y, z)$.

Finally we note that, since we will assume that $\boldsymbol{R}_0$ is in a current free region, $\boldsymbol{B}$ will be analytic is this region. Consequently, Ψ will be analytic in this region. Correspondingly, $\boldsymbol{B}(\boldsymbol{r})$ and $\psi(\boldsymbol{r})$ will be analytic about $\boldsymbol{r} = 0$. See Chapter 35 and Appendix F.

15.2.2 Spherical and Cylindrical Coordinates

Introduce spherical coordinates r,θ,ϕ by the usual rules

$$r^2 = x^2 + y^2 + z^2, \tag{15.2.8}$$

$$x = r\sin(\theta)\cos(\phi), \tag{15.2.9}$$

$$y = r\sin(\theta)\sin(\phi), \tag{15.2.10}$$

$$z = r\cos\theta. \tag{15.2.11}$$

Also, for future use, introduce cylindrical coordinates ρ, ϕ, and z by the usual rules

$$\rho^2 = x^2 + y^2, \tag{15.2.12}$$

$$x = \rho\cos\phi, \tag{15.2.13}$$

$$y = \rho\sin\phi. \tag{15.2.14}$$

Note these two coordinate systems have the coordinate ϕ in common. In both cases there is the ϕ-defining pair of relations

$$\sin\phi = y/\sqrt{x^2 + y^2}, \tag{15.2.15}$$

$$\cos\phi = x/\sqrt{x^2 + y^2}. \tag{15.2.16}$$

The other coordinates are related by (2.11) and the equations

$$r^2 = \rho^2 + z^2, \tag{15.2.17}$$

$$\rho = r\sin\theta. \tag{15.2.18}$$

We also record results for the orthonormal triads $\boldsymbol{e}_r$, $\boldsymbol{e}_\theta$, $\boldsymbol{e}_\phi$ and $\boldsymbol{e}_\rho$, $\boldsymbol{e}_\phi$, $\boldsymbol{e}_z$, and their relation to $\boldsymbol{r}$. For the spherical orthonormal triad there are the results

$$\begin{aligned} \boldsymbol{e}_r &= \sin(\theta)\cos(\phi)\boldsymbol{e}_x + \sin(\theta)\sin(\phi)\boldsymbol{e}_y + \cos(\theta)\boldsymbol{e}_z \\ &= (1/r)(x\boldsymbol{e}_x + y\boldsymbol{e}_y + z\boldsymbol{e}_z) = \boldsymbol{r}/r, \\ \boldsymbol{e}_\theta &= \cos(\theta)\cos(\phi)\boldsymbol{e}_x + \cos(\theta)\sin(\phi)\boldsymbol{e}_y - \sin(\theta)\boldsymbol{e}_z, \\ \boldsymbol{e}_\phi &= -\sin(\phi)\boldsymbol{e}_x + \cos(\phi)\boldsymbol{e}_y, \\ \boldsymbol{r} &= r\boldsymbol{e}_r. \end{aligned} \tag{15.2.19}$$

For the cylindrical orthonormal triad there are the results

$$\begin{aligned} &\boldsymbol{e}_\rho = \cos\phi\ \boldsymbol{e}_x + \sin\phi\ \boldsymbol{e}_y = (1/\rho)(x\boldsymbol{e}_x + y\boldsymbol{e}_y), \\ &\boldsymbol{e}_\phi = -\sin\phi\ \boldsymbol{e}_x + \cos\phi\ \boldsymbol{e}_y = (1/\rho)(-y\boldsymbol{e}_x + x\boldsymbol{e}_y), \\ &\boldsymbol{r} = \rho\boldsymbol{e}_\rho + z\boldsymbol{e}_z. \end{aligned} \tag{15.2.20}$$

See Exercises 2.1 and 2.2.

Finally, with regard to the relation between Cartesian and cylindrical coordinates, if one defines Cartesian and cylindrical components for any vector $\boldsymbol{A}$ by writing

$$\boldsymbol{A} = A_x\boldsymbol{e}_x + A_y\boldsymbol{e}_y + A_z\boldsymbol{e}_z = A_\rho\boldsymbol{e}_\rho + A_\phi\boldsymbol{e}_\phi + A_z\boldsymbol{e}_z, \tag{15.2.21}$$

then there are the component relations

$$A_\rho = \boldsymbol{e}_\rho \cdot \boldsymbol{A} = \cos\phi\ A_x + \sin\phi\ A_y, \tag{15.2.22}$$

$$A_\phi = \boldsymbol{e}_\phi \cdot \boldsymbol{A} = -\sin\phi\ A_x + \cos\phi\ A_y, \tag{15.2.23}$$

and their inverses

$$A_x = \boldsymbol{e}_x \cdot \boldsymbol{A} = \cos\phi\ A_\rho - \sin\phi\ A_\phi, \tag{15.2.24}$$

$$A_y = \boldsymbol{e}_y \cdot \boldsymbol{A} = \sin\phi\ A_\rho + \cos\phi\ A_\phi. \tag{15.2.25}$$

15.2.3 Harmonic Polynomials, Harmonic Polynomial Expansions, and General Spherical Polynomials

Polynomials in x, y, and z that are harmonic are called *harmonic polynomials* or *solid harmonics*. In *complex* form these polynomials, call them H_ℓ^m, can be defined in terms of the spherical harmonics $Y_\ell^m(\theta,\phi)$ and the associated Legendre functions P_ℓ^m by the rule

$$\begin{aligned} H_\ell^m(\boldsymbol{r}) &= r^\ell Y_\ell^m(\theta,\phi) \\ &= r^\ell\{[(2\ell+1)(\ell-m)!]/[4\pi(\ell+m)!]\}^{1/2}P_\ell^m(\cos\theta)\exp(im\phi) \\ &\quad \text{with } -\ell \le m \le \ell. \end{aligned} \tag{15.2.26}$$

The H_ℓ^m are homogeneous polynomials of degree ℓ in the variables x, y, and z.[2] For example, there are the definitions

$$H_0^0(\boldsymbol{r}) = 1/\sqrt{4\pi}; \tag{15.2.27}$$

$$\begin{aligned} &H_1^1(\boldsymbol{r}) = \sqrt{3/(4\pi)}(-1/\sqrt{2})(x+iy) = -\sqrt{3/(8\pi)}(x+iy), \\ &H_1^0(\boldsymbol{r}) = \sqrt{3/(4\pi)}z, \\ &H_1^{-1}(\boldsymbol{r}) = \sqrt{3/(4\pi)}(1/\sqrt{2})(x-iy) = \sqrt{3/(8\pi)}(x-iy); \end{aligned} \tag{15.2.28}$$

$$\begin{aligned} &H_2^2(\boldsymbol{r}) = \sqrt{15/(32\pi)}(x+iy)^2, \\ &H_2^1(\boldsymbol{r}) = -\sqrt{15/(8\pi)}(x+iy)z, \\ &H_2^0(\boldsymbol{r}) = \sqrt{5/(16\pi)}(2z^2-x^2-y^2), \\ &H_2^{-1}(\boldsymbol{r}) = \sqrt{15/(8\pi)}(x-iy)z, \\ &H_2^{-2}(\boldsymbol{r}) = \sqrt{15/(32\pi)}(x-iy)^2. \end{aligned} \tag{15.2.29}$$

See Appendix U. Because thy are defined in terms of the Y_ℓ^m and powers of r, the harmonic polynomials H_ℓ^m have well-defined properties under the action of the rotation group $SO(3)$.[3]

From potential theory we know that any harmonic function analytic at the origin $\boldsymbol{r}=0$ can be expanded in harmonic polynomials. Thus, under the assumption that a harmonic function ψ is analytic at the origin, it has the expansion

$$\psi(x,y,z) = \sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} g_{\ell m}H_\ell^m(\boldsymbol{r}) \tag{15.2.30}$$

where the coefficients $g_{\ell m}$ are arbitrary.

We can also define *real* versions of harmonic polynomials, call them $H_\ell^{m,\alpha}$ where $m \ge 0$ and $\alpha = c$ or s, by writing

$$\begin{aligned} &H_\ell^{m,c}(x,y,z) = \{[(2\ell+1)(\ell-m)!]/[4\pi(l+m)!]\}^{1/2}r^\ell P_\ell^m(\cos\theta)\cos(m\phi) \\ &\text{with } \ell = 0,1,\cdots,\infty \text{ and } m = 0,1,\cdots,\ell \end{aligned} \tag{15.2.31}$$

[2] Note that although we initially work with spherical coordinates, the final result is in the form of power series in *Cartesian* coordinates.

[3] We note that the spherical harmonics Y_ℓ^m could better be called *surface* harmonics. Then we could use the name *spherical harmonics* to refer to the functions H_ℓ^m.

and

$$H_\ell^{m,s}(x,y,z) = \{[(2\ell+1)(\ell-m)!]/[4\pi(l+m)!]\}^{1/2} r^\ell P_\ell^m(\cos\theta)\sin(m\phi)$$
$$\text{with } \ell = 1,2,\cdots,\infty \text{ and } m = 1,2,\cdots,\ell. \tag{15.2.32}$$

These definitions yield, for example, the results

$$H_0^{0,c} = 1/\sqrt{4\pi}; \tag{15.2.33}$$

$$\begin{aligned} H_1^{1,c} &= -[3/(8\pi)]^{1/2}x, \\ H_1^{0,c} &= [3/(4\pi)]^{1/2}z, \\ H_1^{1,s} &= -[3/(8\pi)]^{1/2}y; \end{aligned} \tag{15.2.34}$$

$$\begin{aligned} H_2^{2,c} &= (1/4)[15/(2\pi)]^{1/2}(x^2-y^2), \\ H_2^{1,c} &= -[15/(8\pi)]^{1/2}xz, \\ H_2^{0,c} &= (1/2)[5/(4\pi)]^{1/2}(2z^2-x^2-y^2), \\ H_2^{2,s} &= (1/2)[15/(2\pi)]^{1/2}xy, \\ H_2^{1,s} &= -[15/(8\pi)]^{1/2}yz; \end{aligned} \tag{15.2.35}$$

$$\begin{aligned} H_3^{3,c} &= -(1/4)[35/(4\pi)]^{1/2}(x^3-3xy^2), \\ H_3^{2,c} &= (1/4)[105/(2\pi)]^{1/2}[z(x^2-y^2)], \\ H_3^{1,c} &= -(1/4)[21/(2\pi)]^{1/2}[x(4z^2-x^2-y^2)], \\ H_3^{0,c} &= (1/2)[7/(4\pi)]^{1/2}[2z^3-3z(x^2+y^2)], \\ H_3^{3,s} &= (1/4)[35/(4\pi)]^{1/2}(y^3-3x^2y), \\ H_3^{2,s} &= (1/2)[105/(2\pi)]^{1/2}(xyz), \\ H_3^{1,s} &= -(1/4)[21/(4\pi)]^{1/2}[y(4z^2-x^2-y^2)]. \end{aligned} \tag{15.2.36}$$

In terms of these polynomials any harmonic function ψ analytic at the origin has an expansion of the form

$$\psi(x,y,z) = \sum_{\ell=0}^{\infty}\sum_{m=0}^{\ell} g_{\ell,m,c}H_\ell^{m,c}(\boldsymbol{r}) + \sum_{\ell=1}^{\infty}\sum_{m=1}^{\ell} g_{\ell,m,s}H_\ell^{m,s}(\boldsymbol{r}), \tag{15.2.37}$$

where the coefficients $g_{\ell,m,\alpha}$ are arbitrary.

It is also convenient to define general *spherical polynomials* $S_{n\ell}^m(\boldsymbol{r})$ in terms of the spherical harmonics and powers of r by making the definition

$$S_{n\ell}^m(\boldsymbol{r}) = r^n Y_\ell^m(\theta,\phi). \tag{15.2.38}$$

See Subsection U.2.4. The $S_{n\ell}^m(\boldsymbol{r})$ are evidently of degree n. They form a basis for the space of all functions that are anaytic at the origin, and harmonic polynomials comprise special cases for which $\ell = n$,

$$H_n^m(\boldsymbol{r}) = S_{nn}^m(\boldsymbol{r}). \tag{15.2.39}$$

Therefore, we may also write (2.30) in the form

$$\psi(x,y,z) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} g_{nm} S_{nn}^{m}(\boldsymbol{r}). \tag{15.2.40}$$

Finally, because they are constructed from the Y_ℓ^m and powers of r, the $S_{n\ell}^m(\boldsymbol{r})$ also have well-defined properties under the action of $SO(3)$.

15.2.4 Spherical Polynomial Vector Fields

We have been working with *scalar* fields such as $\psi(\boldsymbol{r})$. Just as it is possible to construct scalar polynomial fields $S_{n\ell}^m(\boldsymbol{r})$ having well-defined properties under the action of $SO(3)$, it is also possible to construct polynomial *vector* fields that have well-defined properties under the action of $SO(3)$. We call such polynomial vector fields *spherical polynomial* vector fields and denote them by the symbols $\boldsymbol{S}_{n\ell J}^{M}(\boldsymbol{r})$. These are vector fields whose components are homogenous polynomials of degree n in the components of $\boldsymbol{r}$. See Subsections U.3.2 and U.3.3 for their definition and some examples.

Any vector field analytic at the origin can be expanded in terms of spherical polynomial vector fields. In particular, both the magnetic field $\boldsymbol{B}(\boldsymbol{r})$ and any associated vector potential $\boldsymbol{A}(\boldsymbol{r})$ can be expanded in terms of spherical polynomial vector fields.

Because of their properties under the action of $SO(3)$, there are well-organized relations between spherical polynomials and spherical polynomial vector fields. For example, there is the relation

$$\nabla H_n^m(\boldsymbol{r}) = \nabla S_{nn}^m(\boldsymbol{r}) = \sqrt{n(2n+1)}\boldsymbol{S}_{n-1,n-1,n}^{m}(\boldsymbol{r}). \tag{15.2.41}$$

See (U.5.6). Upon combining (2.40) and (2.41) we see that a general source-free field $\boldsymbol{B}$ has a spherical polynomial vector field expansion of the form

$$\boldsymbol{B}(\boldsymbol{r}) = \nabla\psi = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} g_{nm}\sqrt{n(2n+1)}\boldsymbol{S}_{n-1,n-1,n}^{m}(\boldsymbol{r}). \tag{15.2.42}$$

Note that the $n = 0$ term does not actually contribute, as expected.

15.2.5 Determination of Minimum Vector Potential: the Poincaré-Coulomb Gauge

Suppose we are given a magnetic field, specified either by a scalar potential presented in the forms (2.30) or (2.40), or equivalently by a spherical polynomial vector field expansion of the form (2.42). And suppose, in order to treat charged-particle motion in this field using a Hamiltonian formulation, we wish to find an associated vector potential. We know that in principle there are many such vector potentials, all of which are related by gauge transformations. Given $\boldsymbol{B}(\boldsymbol{r})$ in some region, the goal of this subsection is to find an associated vector potential $\boldsymbol{A}^{\min}(\boldsymbol{r})$ that is, at least *locally*, as *minimal*/small as possible in the sense that $\boldsymbol{A}^{\min}(\boldsymbol{r})$ is small if $\boldsymbol{B}(\boldsymbol{r})$ is small. The reason for this goal is that, according to (1.5.30), mechanical and canonical momenta differ by the vector potential; and there are situations where we would like this difference to be as small as possible. See Sections 16.7 and 22.8.

Conceptually, our plan is as follows: Make Taylor expansions, with initially unknown coefficients, for the Cartesian components of $\boldsymbol{A}^{\min}(\boldsymbol{r})$, organize these expansions into homogeneous polynomials, and then further organize them as spherical polynomial vector fields. Then use this representation to compute and organize $\nabla \times \boldsymbol{A}^{\min}$ in terms of spherical polynomial vector fields. At the same time parameterize $\boldsymbol{B}(\boldsymbol{r})$ in terms of a scalar potential ψ expanded in harmonic polynomials. Finally, compare the two expansions for $\boldsymbol{B}(\boldsymbol{r})$ given by $\boldsymbol{B} = \nabla\psi$ and $\boldsymbol{B} = \nabla \times \boldsymbol{A}^{\min}$, equate coefficients of like terms, and thereby determine the coefficients in the Taylor expansion for the components of $\boldsymbol{A}^{\min}$ in terms of the coefficients in the expansion for ψ. Do all this while keeping the minimal/small goal in mind. For the notation and machinery required for the execution of this plan, again see Appendix U. What lies ahead may seem complicated, but the final result will prove to be remarkably simple.

Construction of Minimum Vector Potential

We begin with the harmonic polynomial expansion (2.40) for the scalar potential ψ, which we rewrite in the form

$$\psi(\boldsymbol{r}) = \sum_{n=1}^{n_{\max}} \sum_{m} g_{nm} S_{nn}^{m}(\boldsymbol{r}). \tag{15.2.43}$$

Here we assume an expansion through terms of degree $n_{\max}$, and omit $n = 0$ terms since constant terms make no contribution to $\boldsymbol{B}$ as given by (2.42).

For the associated vector potential $\boldsymbol{A}^{\min}$ we make the spherical polynomial vector field expansion

$$\boldsymbol{A}^{\min}(\boldsymbol{r}) = \sum_{n=1}^{n_{\max}} \sum_{\ell} \sum_{J} \sum_{M} f_{n\ell JM} \boldsymbol{S}_{n\ell J}^{M}(\boldsymbol{r}). \tag{15.2.44}$$

Again see Appendix U. Given the coefficients g_{nm}, our task is to use the equality

$$\nabla \times \boldsymbol{A}^{\min}(\boldsymbol{r}) = \nabla \times \sum_{n=1}^{n_{\max}} \sum_{\ell} \sum_{J} \sum_{M} f_{n\ell JM} \boldsymbol{S}_{n\ell J}^{M}(\boldsymbol{r}) = \nabla \sum_{n=1}^{n_{\max}} \sum_{m} g_{nm} S_{nn}^{m}(\boldsymbol{r}) = \nabla\psi(\boldsymbol{r}) \tag{15.2.45}$$

to find the coefficients $f_{n\ell JM}$ in terms of the g_{nm}.

We already know the result of evaluating the right side of (2.45). Use of (2.42) gives the result

$$\boldsymbol{B}(\boldsymbol{r}) = \nabla\psi(\boldsymbol{r}) = \nabla \sum_{n=1}^{n_{\max}} \sum_{m} g_{nm} S_{nn}^{m}(\boldsymbol{r}) = \sum_{n=1}^{n_{\max}} \sum_{m} g_{nm} \sqrt{n(2n+1)} \boldsymbol{S}_{n-1,n-1,n}^{m}(\boldsymbol{r}). \tag{15.2.46}$$

Next work on evaluating the left side of (2.45). This is a more complicated task. In accord with the range rules (U.3.7) and (U.3.8), we decompose the expansion into the sum of four pieces with each containing a particular kind of term:

a) All terms for which $\ell = 0$ and hence $J = 1$. Also, therefore, $n = 2k$ with $k > 0$. The associated spherical polynomial vectors are of the form $\boldsymbol{S}_{2k,0,1}^{M}(\boldsymbol{r})$.

b) All terms for which $\ell > 0$ and $J = \ell + 1$. The associated spherical polynomial vectors are of the form $\boldsymbol{S}^M_{n,\ell,\ell+1}(\boldsymbol{r})$.

c) All terms for which $\ell > 0$ and $J = \ell$. The associated spherical polynomial vectors are of the form $\boldsymbol{S}^M_{n,\ell,\ell}(\boldsymbol{r})$.

d) All terms for which $\ell > 0$ and $J = \ell - 1$. The associated spherical polynomial vectors are of the form $\boldsymbol{S}^M_{n,\ell,\ell-1}(\boldsymbol{r})$.

Thus, we write

$$\boldsymbol{A}^{\min} = \boldsymbol{A}^{\min a} + \boldsymbol{A}^{\min b} + \boldsymbol{A}^{\min c} + \boldsymbol{A}^{\min d} \tag{15.2.47}$$

where

$$\boldsymbol{A}^{\min a}(\boldsymbol{r}) = \sum_{k=1}^{k_{\max}} \sum_{M} f_{2k,0,1,M} \boldsymbol{S}^M_{2k,0,1}(\boldsymbol{r}), \tag{15.2.48}$$

$$\boldsymbol{A}^{\min b}(\boldsymbol{r}) = \sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} f_{n,\ell,\ell+1,M} \boldsymbol{S}^M_{n,\ell,\ell+1}(\boldsymbol{r}), \tag{15.2.49}$$

$$\boldsymbol{A}^{\min c}(\boldsymbol{r}) = \sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} f_{n,\ell,\ell,M} \boldsymbol{S}^M_{n,\ell,\ell}(\boldsymbol{r}), \tag{15.2.50}$$

$$\boldsymbol{A}^{\min d}(\boldsymbol{r}) = \sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} f_{n,\ell,\ell-1,M} \boldsymbol{S}^M_{n,\ell,\ell-1}(\boldsymbol{r}). \tag{15.2.51}$$

We are ready to proceed. For the $\boldsymbol{A}^{\min a}$ term we find, using (U.5.20), the result

$$\nabla \times \boldsymbol{A}^{\min a}(\boldsymbol{r}) = \nabla \times \sum_{k=1}^{k_{\max}} \sum_{M} f_{2k,0,1,M} \boldsymbol{S}^M_{2k,0,1}(\boldsymbol{r}) = \sum_{k=1}^{k_{\max}} \sum_{M} f_{2k,0,1,M} [i(\sqrt{2/3})(2k)] \boldsymbol{S}^M_{2k-1,1,1}(\boldsymbol{r}). \tag{15.2.52}$$

For the $\boldsymbol{A}^{\min b}$ term we find, using (U.5.17), the result

$$\nabla \times \boldsymbol{A}^{\min b}(\boldsymbol{r}) = \nabla \times \sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} f_{n,\ell,\ell+1,M} \boldsymbol{S}^M_{n,\ell,\ell+1}(\boldsymbol{r}) =$$
$$\sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} f_{n,\ell,\ell+1,M} [i\sqrt{(\ell+2)/(2\ell+3)}(n-\ell)] \boldsymbol{S}^M_{n-1,\ell+1,\ell+1}(\boldsymbol{r}). \tag{15.2.53}$$

For the $\boldsymbol{A}^{\min c}$ term we find, using (U.5.18), the result

$$\nabla \times \boldsymbol{A}^{\min c}(\boldsymbol{r}) = \nabla \times \sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} f_{n,\ell,\ell,M} \boldsymbol{S}^M_{n,\ell,\ell}(\boldsymbol{r}) =$$
$$\sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} f_{n,\ell,\ell,M} [i\sqrt{(\ell+1)/(2\ell+1)}(n+\ell+1)] \boldsymbol{S}^M_{n-1,\ell-1,\ell}(\boldsymbol{r})$$
$$+ \sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} f_{n,\ell,\ell,M} [i\sqrt{\ell/(2\ell+1)}(n-\ell)] \boldsymbol{S}^M_{n-1,\ell+1,\ell}(\boldsymbol{r}). \tag{15.2.54}$$

Finally, for the $\boldsymbol{A}^{\min d}$ term we find, using (U.5.19), the result

$$\nabla \times \boldsymbol{A}^{\min d}(\boldsymbol{r}) = \nabla \times \sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} f_{n,\ell,\ell-1,M} \boldsymbol{S}^{M}_{n,\ell,\ell-1}(\boldsymbol{r}) =$$
$$\sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} f_{n,\ell,\ell-1,M} [i\sqrt{(\ell-1/(2\ell-1)}(n+\ell+1)] \boldsymbol{S}^{M}_{n-1,\ell-1,\ell-1}(\boldsymbol{r}). \quad (15.2.55)$$

We can now equate coefficients of like terms. Let us begin with the first few corresponding to small values of n. The first of these, corresponding to $n=0$, is $\boldsymbol{S}^{M}_{0,0,1}$. For the right side of (2.45) we see from (2.46) that

$$\text{coefficient of } \boldsymbol{S}^{M}_{0,0,1} \text{ in } \nabla\psi = \sqrt{3}\, g_{1M}. \quad (15.2.56)$$

Next, for the left side, examine the terms in $\nabla \times \boldsymbol{A}^{\min}$: From (2.52) we see that there are no terms of the desired kind, namely terms involving $\boldsymbol{S}^{M}_{0,0,1}$, in $\nabla \times \boldsymbol{A}^{\min a}$. From (2.53) we see that there are no terms of the desired kind in $\nabla \times \boldsymbol{A}^{\min b}$. From (2.54) we see that there are terms of the desired kind in $\nabla \times \boldsymbol{A}^{\min c}$, and find the result

$$\text{coefficient of } \boldsymbol{S}^{M}_{0,0,1} \text{ in } \nabla \times \boldsymbol{A}^{\min c} = i\sqrt{6}\, f_{1,1,1,M}. \quad (15.2.57)$$

Finally, from (2.55) we see that there are no terms of the desired kind in $\nabla \times \boldsymbol{A}^{\min d}$.

Upon comparing (2.56) and (2.57) we conclude that there must be the relation

$$i\sqrt{6}\, f_{1,1,1,M} = \sqrt{3}\, g_{1M}, \quad (15.2.58)$$

and therefore

$$f_{1,1,1,M} = -i\sqrt{1/2}\, g_{1M}. \quad (15.2.59)$$

Note that this relation is consistent with (U.6.39). Moreover, we conclude that the six remaining $n=1$ coefficients in $\boldsymbol{A}^{\min}$, namely $f_{1,1,0,0}$ and the $f_{1,1,2,M}$, can be anything since there are the relations (U.6.38) and (U.6.40). In pursuit of our minimal/small goal, we set these coefficients to zero. Then, so far, we have the result

$$\boldsymbol{A}^{\min}(\boldsymbol{r}) = \sum_{M} (-i)\sqrt{1/2}\, g_{1M}\ \boldsymbol{S}^{M}_{111}(\boldsymbol{r}) + \text{ terms of degree } > 1. \quad (15.2.60)$$

In terms of Cartesian components, (2.60) yields the relation

$$\boldsymbol{A}^{\min}(\boldsymbol{r}) = -(1/2)\boldsymbol{r} \times \boldsymbol{B}(0) + \text{ terms of degree } > 1. \quad (15.2.61)$$

Here we have used (2.42), (2.43), and (U.6.25) evaluated for $n=1$.

Let us push on to the case $n=1$; in which case there are the spherical polynomial vector fields $\boldsymbol{S}^{0}_{110}$, $\boldsymbol{S}^{M}_{111}$ with $-1 \le M \le 1$, and $\boldsymbol{S}^{M}_{112}$ with $-2 \le M \le 2$. First see where/how they occur in $\nabla\psi$. Examination of (2.46) shows that the only such term in $\nabla\psi$ is $\boldsymbol{S}^{M}_{112}$, and we have the relation

$$\text{coefficient of } \boldsymbol{S}^{M}_{1,1,2} \text{ in } \nabla\psi = \sqrt{10}\, g_{2M}. \quad (15.2.62)$$

We next examine the terms in $\nabla \times \boldsymbol{A}^{\min}$: From (2.52) we see that there are no terms of the desired kind, namely terms involving $\boldsymbol{S}^M_{1,1,2}$, in $\nabla \times \boldsymbol{A}^{\min a}$. From (2.53) we see that there are no terms of the desired kind in $\nabla \times \boldsymbol{A}^{\min b}$. From (2.54) we see that there are terms of the desired kind in $\nabla \times \boldsymbol{A}^{\min c}$, and find the relation

$$\text{coefficient of } \boldsymbol{S}^M_{1,1,2} \text{ in } \nabla \times \boldsymbol{A}^{\min c} = i\sqrt{15}\, f_{2,2,2,M}. \tag{15.2.63}$$

Finally, from (2.55) we see that there are no terms of the desired kind in $\nabla \times \boldsymbol{A}^{\min d}$.

Upon comparing (2.62) and (2.63) we conclude that there must be the relation

$$i\sqrt{15}\, f_{2,2,2,M} = \sqrt{10}\, g_{2M}, \tag{15.2.64}$$

and therefore

$$f_{2,2,2,M} = -i\sqrt{2/3}\, g_{2M}. \tag{15.2.65}$$

What can be said about the thirteen remaining $n = 2$ coefficients in $\boldsymbol{A}^{\min}$, namely the coefficients f_{201M}, $f_{2,2,3,M}$, and $f_{2,2,1,M}$? It can be shown that $\nabla \times \boldsymbol{S}^M_{223}(\boldsymbol{r}) = 0$, and therefore the terms with coefficients $f_{2,2,3,M}$ make no contribution to $\boldsymbol{B}(\boldsymbol{r})$. See Exercise (U.6.21). In further pursuit of our minimal/small goal, we set these coefficients to zero. It can be shown that terms with the coefficients f_{201M} and $f_{2,2,1,M}$ produce terms in $\boldsymbol{B}(\boldsymbol{r})$ having nonzero curl. Again see Exercise (U.6.21). We also set these coefficients to zero to ensure that $\boldsymbol{B}(\boldsymbol{r})$ is curl free. Putting everything together we have learned so far yields the result

$$\boldsymbol{A}^{\min}(\boldsymbol{r}) = \sum_M (-i)\sqrt{1/2}\, g_{1M}\, \boldsymbol{S}^M_{111}(\boldsymbol{r}) + \sum_M (-i)\sqrt{2/3}\, g_{2M}\, \boldsymbol{S}^M_{222}(\boldsymbol{r}) + \text{ terms of degree } > 2. \tag{15.2.66}$$

The pattern should now be clear. There are the general relations

$$\nabla S^M_{nn}(\boldsymbol{r}) = \sqrt{n(2n+1)}\boldsymbol{S}^M_{n-1,n-1,n}(\boldsymbol{r}) \tag{15.2.67}$$

and

$$\nabla \times \boldsymbol{S}^M_{n,n,n}(\boldsymbol{r}) = i\sqrt{(n+1)(2n+1)}\boldsymbol{S}^M_{n-1,n-1,n}(\boldsymbol{r}). \tag{15.2.68}$$

Therefore there is the general relation

$$\boldsymbol{A}^{\min}(\boldsymbol{r}) = \sum_{n=1}^{n_{\max}} \sum_{M=-n}^{n} (-i)\sqrt{n/(n+1)}\, g_{nM}\boldsymbol{S}^M_{nnn}(\boldsymbol{r}). \tag{15.2.69}$$

We have found a formula for the vector potential $\boldsymbol{A}^{\min}(\boldsymbol{r})$ in terms of the harmonic expansion coefficients for the scalar potential $\psi(\boldsymbol{r})$.

Properties of Minimum Vector Potential

It can be verified that this particular choice of $\boldsymbol{A}^{\min}(\boldsymbol{r})$ has the property

$$\nabla \cdot \boldsymbol{A}^{\min}(\boldsymbol{r}) = 0. \tag{15.2.70}$$

See (U.5.11). Therefore $\boldsymbol{A}^{\min}(\boldsymbol{r})$ is in a *Coulomb/solenoidal* gauge.[4] It also has the property

$$\boldsymbol{r} \cdot \boldsymbol{A}^{\min}(\boldsymbol{r}) = 0. \tag{15.2.71}$$

See (U.6.9). This is the condition that $\boldsymbol{A}^{\min}(\boldsymbol{r})$ be in what is called a *Poincaré* gauge.[5] Taken together, we will say that a vector potential that satisfies both (2.70) and (2.71) is in the *Poincaré-Coulomb* gauge.[6]

At this point we observe that any vector potential $\boldsymbol{A}(\boldsymbol{r})$ that obeys the Poincaré gauge condition

$$\boldsymbol{r} \cdot \boldsymbol{A}(\boldsymbol{r}) = 0 \tag{15.2.72}$$

must vanish at the origin,

$$\boldsymbol{A}(0) = 0. \tag{15.2.73}$$

That is, it has no constant part. Note that, according to (2.3), $\boldsymbol{r} = 0$ corresponds to the expansion point $\boldsymbol{R} = \boldsymbol{R}_0$, and we again assume analyticity so that $\boldsymbol{A}$ is analytic at the origin.

To see that $\boldsymbol{A}(\boldsymbol{r})$ has no constant part, expand it in homogeneous polynomials of degree n by writing

$$\boldsymbol{A}(\boldsymbol{r}) = \sum_{n=0}^{\infty} \boldsymbol{A}^n(\boldsymbol{r}). \tag{15.2.74}$$

Combining (2.72) and (2.74) gives the relation

$$\boldsymbol{r} \cdot \boldsymbol{A}(\boldsymbol{r}) = \sum_{n=0}^{\infty} \boldsymbol{r} \cdot \boldsymbol{A}^n(\boldsymbol{r}) = 0, \tag{15.2.75}$$

from which it follows, upon equating terms of like degree, that

$$\boldsymbol{r} \cdot \boldsymbol{A}^n(\boldsymbol{r}) = 0 \tag{15.2.76}$$

for all n. In particular, there is the result

$$\boldsymbol{r} \cdot \boldsymbol{A}^0(\boldsymbol{r}) = \boldsymbol{r} \cdot \boldsymbol{A}(0) = 0 \tag{15.2.77}$$

for all $\boldsymbol{r}$, from which (2.73) follows.

Evaluation of Work

Have we achieved our goal of finding a *minimal* vector potential? We have, in the following sense: Inspection of (2.46) shows that it provides an expansion of $\boldsymbol{B}(\boldsymbol{r})$ in terms of spherical polynomial vector fields $\boldsymbol{S}^m_{n-1,n-1,n}(\boldsymbol{r})$ with expansion coefficients proportional to the g_{nm}. Inspection of (2.69) shows that it provides an expansion of $\boldsymbol{A}^{\min}(\boldsymbol{r})$ in terms of spherical

[4] Fields that are divergence free are also called *solenoidal*.

[5] A Poincaré gauge is also sometimes called a *multipolar* gauge.

[6] It is interesting to observe that these two conditions have related "Fourier analogs". In Fourier space the condition (2.70) becomes $\boldsymbol{k} \cdot \tilde{\boldsymbol{A}}^{\min}(\boldsymbol{k}) = 0$ and the condition (2.71) becomes $\nabla_{\boldsymbol{k}} \cdot \tilde{\boldsymbol{A}}^{\min}(\boldsymbol{k}) = 0$. Here a tilde denotes a Fourier transform. See Exercise 22.2.23.

polynomial vector fields $\boldsymbol{S}_{nnn}^{M}(\boldsymbol{r})$ with expansion coefficients again proportional to the g_{nM}. As shown above, the vector potential $\boldsymbol{A}^{\min}(\boldsymbol{r})$ has no constant part. We also see that its non-constant parts are directly proportional to the coefficients g_{nm} that describe the constant and non-constant parts of $\boldsymbol{B}(\boldsymbol{r})$. Moreover, there is an order-by-order relation. Terms of order n in $\boldsymbol{A}^{\min}(\boldsymbol{r})$ are proportional to terms of order $n-1$ in $\boldsymbol{B}(\boldsymbol{r})$. Thus, $\boldsymbol{A}^{\min}(\boldsymbol{r})$ is small if $\boldsymbol{B}(\boldsymbol{r})$ is small. In particular, if high-order terms in $\boldsymbol{B}(\boldsymbol{r})$ are negligible, they will also be negligible in $\boldsymbol{A}^{\min}(\boldsymbol{r})$.

There is yet another sense in which the vector potential we have found is minimal. Suppose, for example, that we confine our attention to the case of a vector potential that is homogeneous of degree 1, which is the case we need to produce a constant magnetic field. When $n = 1$ we see from Table U.3.1 that $\ell = 1$ and $J = 0, 1, 2$. Therefore, such a vector potential, call it $\boldsymbol{A}^1$, can be written in the form

$$\boldsymbol{A}^1(\boldsymbol{r}) = \sum_J \sum_M f_{11JM} \boldsymbol{S}_{11J}^{M}(\boldsymbol{r}). \tag{15.2.78}$$

Recall (2.44). Let us compute the *norm* of $\boldsymbol{A}^1$ as defined by the rule

$$||\boldsymbol{A}^1(\boldsymbol{r})||^2 = \int d\Omega\ [\boldsymbol{A}^1(\boldsymbol{r})]^* \cdot \boldsymbol{A}^1(\boldsymbol{r}). \tag{15.2.79}$$

Since the $\boldsymbol{S}_{11J}^{M}(\boldsymbol{r})$ are mutually orthogonal under angular integration, we find from (2.78), (U.3.18), and (U.4.3) the result

$$||\boldsymbol{A}^1(\boldsymbol{r})||^2 = r^2 \sum_J \sum_M |f_{11JM}|^2. \tag{15.2.80}$$

We know the value of f_{111M} is fixed by (2.59), and we have chosen to set the remaining f_{11JM} to zero. We now see, since (2.80) is a sum of squares, that doing so *minimizes* $||\boldsymbol{A}^1(\boldsymbol{r})||$. Similar computations may be made for other values of n. The result is that the choice we have made for $\boldsymbol{A}^{\min}$ minimizes $||\boldsymbol{A}^n(\boldsymbol{r})||$ for each value of n.

A Further Simplification

We close this subsection by observing that the relation (2.69) can be further manipulated using (U.6.25). Doing so gives the pleasing result

$$\begin{aligned} \boldsymbol{A}^{\min}(\boldsymbol{r}) &= -\sum_{n=1}^{n_{\max}} \sum_{M=-n}^{n} [1/(n+1)] g_{nM} [\boldsymbol{r} \times \nabla S_{nn}^{M}(\boldsymbol{r})] \\ &= -\sum_{n=1}^{n_{\max}} \sum_{M=-n}^{n} [1/(n+1)] g_{nM} [\boldsymbol{r} \times \nabla H_{n}^{M}(\boldsymbol{r})]. \end{aligned} \tag{15.2.81}$$

Note that this result has the virtue that none of the extensive machinery of Appendix U is required for its evaluation.

15.2.6 Uniqueness of Poincaré-Coulomb Gauge

Is a vector potential in the Poincaré-Coulomb gauge unique? It is. Suppose $\boldsymbol{A}$ and $\boldsymbol{A}'$ are two vector potentials associated with the same field $\boldsymbol{B}$. Then we know they are related by a gauge transformation of the form

$$\boldsymbol{A}'(\boldsymbol{r}) = \boldsymbol{A}(\boldsymbol{r}) + \nabla\chi(\boldsymbol{r}). \tag{15.2.82}$$

If we require that both $\boldsymbol{A}$ and $\boldsymbol{A}'$ be in the Coulomb gauge, then χ must be harmonic: taking the divergence of both sides of (2.82) yields the result

$$\nabla^2\chi = 0. \tag{15.2.83}$$

See Section 15.6. If we further require that both $\boldsymbol{A}$ and $\boldsymbol{A}'$ be in the Poincaré-Coulomb gauge, then there must be the additional relations

$$\boldsymbol{r}\cdot\boldsymbol{A}(\boldsymbol{r}) = 0 \text{ and } \boldsymbol{r}\cdot\boldsymbol{A}'(\boldsymbol{r}) = 0. \tag{15.2.84}$$

Requiring (2.84) of (2.82 yields the result that χ must also obey the condition

$$\boldsymbol{r}\cdot\nabla\chi = 0. \tag{15.2.85}$$

Suppose χ is decomposed into homogeneous polynomials of degree n by writing

$$\chi = \sum_{n=0}^{\infty}\chi^n. \tag{15.2.86}$$

Then, by Euler's relation for homogeneous functions, it follows that

$$\boldsymbol{r}\cdot\nabla\chi = \sum_{n=0}^{\infty}n\chi^n. \tag{15.2.87}$$

Comparison of (2.85) and (2.87) and equating terms of like degree yields the result

$$n\chi^n = 0, \tag{15.2.88}$$

from which it follows that $\chi^n = 0$ for $n \neq 0$. We see that all that is left in the sum (2.86) and in the relation (2.82) is the *constant* term χ^0, and this term does not contribute to (2.82). We therefore conclude that $\boldsymbol{A}'(\boldsymbol{r}) = \boldsymbol{A}(\boldsymbol{r})$.

15.2.7 Direct Construction of Poincaré-Coulomb Gauge Vector Potential

Subsection 2.5 obtained the final result (2.81) using the machinery of Appendix U. The purpose of this subsection is to proceed in reverse. After some stage setting, we will make an Ansatz that is essentially equivalent to (2.81), and then verify that the vector potential produced by this Ansatz yields $\boldsymbol{B}(\boldsymbol{r})$ as desired, and has other desired/interesting properties.

With reference to (2.37), define scalar fields $\psi_{\ell,m,\alpha}$ by the rule

$$\psi_{\ell,m,\alpha} = H_\ell^{m,\alpha} \tag{15.2.89}$$

so that we may write

$$\psi(x,y,z) = \sum_{\ell=0}^{\infty}\sum_{m=0}^{\ell} g_{\ell,m,c}\psi_{\ell,m,c} + \sum_{\ell=1}^{\infty}\sum_{m=1}^{\ell} g_{\ell,m,s}\psi_{\ell,m,s}. \tag{15.2.90}$$

Next define related vector fields $\boldsymbol{B}^{\ell,m,\alpha}$ by the rule

$$\boldsymbol{B}^{\ell,m,\alpha} = \nabla\psi_{\ell,m,\alpha}. \tag{15.2.91}$$

Then, with the aid of (2.90) and (2.91), we may write

$$\boldsymbol{B} = \nabla\psi = \sum_{\ell=1}^{\infty}\sum_{m=0}^{\ell} g_{\ell,m,c}\boldsymbol{B}^{\ell,m,c} + \sum_{\ell=1}^{\infty}\sum_{m=1}^{\ell} g_{\ell,m,s}\boldsymbol{B}^{\ell,m,s}. \tag{15.2.92}$$

We now seek individual vector potentials $\boldsymbol{A}^{\ell,m,\alpha}$ such that

$$\nabla\times\boldsymbol{A}^{\ell,m,\alpha} = \boldsymbol{B}^{\ell,m,\alpha}. \tag{15.2.93}$$

Simple calculation shows that If we can find them, then we may write

$$\boldsymbol{B} = \nabla\times\boldsymbol{A} \tag{15.2.94}$$

with

$$\boldsymbol{A} = \sum_{\ell=1}^{\infty}\sum_{m=0}^{\ell} g_{\ell,m,c}\boldsymbol{A}^{\ell,m,c} + \sum_{\ell=1}^{\infty}\sum_{m=1}^{\ell} g_{\ell,m,s}\boldsymbol{A}^{\ell,m,s}. \tag{15.2.95}$$

We claim that a solution to (2.93) is given by the Ansatz

$$\boldsymbol{A}^{\ell,m,\alpha} = [-1/(\ell+1)][\boldsymbol{r}\times\boldsymbol{B}^{\ell,m,\alpha}] = [-1/(\ell+1)][\boldsymbol{r}\times\nabla\psi_{\ell,m,\alpha}]. \tag{15.2.96}$$

Let us check this claim. Recall the vector identity

$$\nabla\times(\boldsymbol{a}\times\boldsymbol{b}) = \boldsymbol{a}(\nabla\cdot\boldsymbol{b}) - \boldsymbol{b}(\nabla\cdot\boldsymbol{a}) + (\boldsymbol{b}\cdot\nabla)\boldsymbol{a} - (\boldsymbol{a}\cdot\nabla)\boldsymbol{b}. \tag{15.2.97}$$

Then, with

$$\boldsymbol{a} = \boldsymbol{r} \tag{15.2.98}$$

and

$$\boldsymbol{b} = \nabla\psi_{\ell,m,\alpha} = \boldsymbol{B}^{\ell,m,\alpha}, \tag{15.2.99}$$

the identity (2.97) yields the result

$$\nabla\times(\boldsymbol{r}\times\nabla\psi_{\ell,m,\alpha}) = \boldsymbol{r}(\nabla^2\psi_{\ell,m,\alpha}) - \boldsymbol{B}^{\ell,m,\alpha}(\nabla\cdot\boldsymbol{r}) + (\boldsymbol{B}^{\ell,m,\alpha}\cdot\nabla)\boldsymbol{r} - (\boldsymbol{r}\cdot\nabla)\boldsymbol{B}^{\ell,m,\alpha}. \tag{15.2.100}$$

Moreover, there are the relations

$$\nabla^2 \psi_{\ell,m,\alpha} = 0, \tag{15.2.101}$$

$$\nabla \cdot \boldsymbol{r} = 3, \tag{15.2.102}$$

$$(\boldsymbol{B}^{\ell,m,\alpha} \cdot \nabla)\boldsymbol{r} = \boldsymbol{B}^{\ell,m,\alpha}, \tag{15.2.103}$$

and

$$(\boldsymbol{r} \cdot \nabla)\boldsymbol{B}^{\ell,m,\alpha} = (\ell - 1)\boldsymbol{B}^{\ell,m,\alpha}. \tag{15.2.104}$$

This last relation follows from the fact that the Cartesian components of $\boldsymbol{B}^{\ell,m,\alpha}$ are homogenous polynomials of degree $(\ell - 1)$. We conclude that

$$\nabla \times (\boldsymbol{r} \times \nabla \psi_{\ell,m,\alpha}) = [-3 + 1 - (\ell - 1)]\boldsymbol{B}^{\ell,m,\alpha} = [-(\ell + 1)]\boldsymbol{B}^{\ell,m,\alpha}. \tag{15.2.105}$$

Therefore, the $\boldsymbol{A}^{\ell,m,\alpha}$ defined by (2.96) satisfy (2.93). We also note that the Cartesian components of the $\boldsymbol{A}^{\ell,m,\alpha}$ are homogeneous polynomials of degree ℓ.

In addition there is the vector identity

$$\nabla \cdot (\boldsymbol{a} \times \boldsymbol{b}) = \boldsymbol{b} \cdot (\nabla \times \boldsymbol{a}) - \boldsymbol{a} \cdot (\nabla \times \boldsymbol{b}). \tag{15.2.106}$$

From this identity and from (2.96), (2.98), and (2.99) it follows that

$$\nabla \cdot \boldsymbol{A}^{\ell,m,\alpha} = 0. \tag{15.2.107}$$

Thus, the $\boldsymbol{A}^{\ell,m,\alpha}$ are in the Coulomb gauge. Moreover, from (2.96), there is the relation

$$\boldsymbol{r} \cdot \boldsymbol{A}^{\ell,m,\alpha}(\boldsymbol{r}) = 0. \tag{15.2.108}$$

Therefore the $\boldsymbol{A}^{\ell,m,\alpha}$ are also in the Poincaré gauge, and thus in the Poincaré-Coulomb gauge.

Three final points: First, suppose the magnetic field $\boldsymbol{B}(\boldsymbol{r})$ is decomposed into homogeneous polynomials by writing

$$\boldsymbol{B}(\boldsymbol{r}) = \sum_{n=0}^{\infty} \boldsymbol{B}^n(\boldsymbol{r}). \tag{15.2.109}$$

The vector potential can also be decomposed into homogeneous polynomials by writing

$$\boldsymbol{A}(\boldsymbol{r}) = \sum_{n=1}^{\infty} \boldsymbol{A}^n(\boldsymbol{r}). \tag{15.2.110}$$

From (2.96) we see that there is the relation

$$\boldsymbol{A}^n(\boldsymbol{r}) = -[1/(n+1)][\boldsymbol{r} \times \boldsymbol{B}^{n-1}(\boldsymbol{r})] \text{ for } n = 1, 2, \cdots . \tag{15.2.111}$$

Second, suppose that (2.109) and (2.110) are *truncated* by writing

$$\boldsymbol{B}^{\text{trunc}}(\boldsymbol{r}) = \sum_{n=0}^{N} \boldsymbol{B}^n(\boldsymbol{r}), \tag{15.2.112}$$

$$\boldsymbol{A}^{\rm trunc}(\boldsymbol{r})=\sum_{n=1}^{N+1}\boldsymbol{A}^n(\boldsymbol{r}), \tag{15.2.113}$$

with (2.111) continuing to hold for $n=1,2,\cdots N+1$. It is easy to verify that $\boldsymbol{B}^{\rm trunc}(\boldsymbol{r})$ is curl and divergence free if $\boldsymbol{B}(\boldsymbol{r})$ is. It is also true that

$$\boldsymbol{B}^{\rm trunc}(\boldsymbol{r})=\nabla\times\boldsymbol{A}^{\rm trunc}(\boldsymbol{r}). \tag{15.2.114}$$

Thus, truncation by degree does not violate the Maxwell equations.

Finally, we note that the relation (2.111) specifies the vector potential order-by-order in terms of the order-by-order magnetic field. There is an equivalent integral relation that specifies the full vector potential in terms of the full magnetic field. It is given by the relation

$$\boldsymbol{A}(\boldsymbol{r})=-\boldsymbol{r}\times\int_0^1 d\lambda\;\lambda\boldsymbol{B}(\lambda\boldsymbol{r}). \tag{15.2.115}$$

See Exercise 2.4.

Exercises

15.2.1. The purpose of this exercise is to verify (2.19). In terms of the spherical coordinates r,θ,ϕ defined in Subsection 15.2.2 there is the result

$$\boldsymbol{r}=x\boldsymbol{e}_x+y\boldsymbol{e}_y+z\boldsymbol{e}_z=r\sin(\theta)\cos(\phi)\boldsymbol{e}_x+r\sin(\theta)\sin(\phi)\boldsymbol{e}_y+r\cos(\theta)\boldsymbol{e}_z. \tag{15.2.116}$$

Verify that there is an associated orthonormal triad $\boldsymbol{e}_r$, $\boldsymbol{e}_\theta$, $\boldsymbol{e}_\phi$ obeying the relations

$$\begin{aligned}\boldsymbol{e}_r &= [\partial\boldsymbol{r}/\partial r]/||[\partial\boldsymbol{r}/\partial r]||=\sin(\theta)\cos(\phi)\boldsymbol{e}_x+\sin(\theta)\sin(\phi)\boldsymbol{e}_y+\cos(\theta)\boldsymbol{e}_z\\ &= (1/r)(x\boldsymbol{e}_x+y\boldsymbol{e}_y+z\boldsymbol{e}_z)=\boldsymbol{r}/r,\end{aligned} \tag{15.2.117}$$

$$\boldsymbol{e}_\theta = [\partial\boldsymbol{r}/\partial\theta]/||[\partial\boldsymbol{r}/\partial\theta]||=\cos(\theta)\cos(\phi)\boldsymbol{e}_x+\cos(\theta)\sin(\phi)\boldsymbol{e}_y-\sin(\theta)\boldsymbol{e}_z, \tag{15.2.118}$$

$$\boldsymbol{e}_\phi=[\partial\boldsymbol{r}/\partial\phi]/||[\partial\boldsymbol{r}/\partial\phi]||=-\sin(\phi)\boldsymbol{e}_x+\cos(\phi)\boldsymbol{e}_y, \tag{15.2.119}$$

$$\boldsymbol{r}=r\boldsymbol{e}_r. \tag{15.2.120}$$

15.2.2. The purpose of this exercise is to verify (2.20) and (2.22) through (2.25). In terms of the cylindrical coordinates ρ,ϕ,z defined in Section 15.2 there is the result

$$\boldsymbol{r}=x\boldsymbol{e}_x+y\boldsymbol{e}_y+z\boldsymbol{e}_z=\rho\cos\phi\;\boldsymbol{e}_x+\rho\sin\phi\;\boldsymbol{e}_y+z\boldsymbol{e}_z. \tag{15.2.121}$$

Verify that there is an associated orthonormal triad $\boldsymbol{e}_\rho$, $\boldsymbol{e}_\phi$, $\boldsymbol{e}_z$ obeying the relations

$$\boldsymbol{e}_\rho=[\partial\boldsymbol{r}/\partial\rho]/||[\partial\boldsymbol{r}/\partial\rho]||=\cos\phi\;\boldsymbol{e}_x+\sin\phi\;\boldsymbol{e}_y=(1/\rho)(x\boldsymbol{e}_x+y\boldsymbol{e}_y), \tag{15.2.122}$$

$$\boldsymbol{e}_\phi=[\partial\boldsymbol{r}/\partial\phi]/||[\partial\boldsymbol{r}/\partial\phi]||=-\sin\phi\;\boldsymbol{e}_x+\cos\phi\;\boldsymbol{e}_y=(1/\rho)(-y\boldsymbol{e}_x+x\boldsymbol{e}_y), \tag{15.2.123}$$

$$\boldsymbol{e}_z=[\partial\boldsymbol{r}/\partial z]/||[\partial\boldsymbol{r}/\partial z]||, \tag{15.2.124}$$

$$x\boldsymbol{e}_x+y\boldsymbol{e}_y=\rho\boldsymbol{e}_\rho, \tag{15.2.125}$$

$$\boldsymbol{r}=\rho\boldsymbol{e}_\rho+z\boldsymbol{e}_z. \tag{15.2.126}$$

Verify (2.22) through (2.25).

15.2.3. Define an operator $\boldsymbol{L}$ by the rule

$$\boldsymbol{L} = \boldsymbol{r} \times \nabla. \tag{15.2.127}$$

Show that it has the components

$$L_x = y\partial_z - z\partial_y, \tag{15.2.128}$$

$$L_y = z\partial_x - x\partial_z, \tag{15.2.129}$$

$$L_z = x\partial_y - y\partial_x. \tag{15.2.130}$$

Verify that the components of $\boldsymbol{L}$ satisfy the commutation rules

$$\{L_x, L_y\} = -L_z, \text{ etc.}, \tag{15.2.131}$$

which are a variant of the commutation rules for $so(3, \mathbb{R})$.

Show that (2.96) can be rewritten in the form

$$\boldsymbol{A}^{\ell,m,\alpha} = [-1/(\ell+1)][\boldsymbol{L}\psi_{\ell,m,\alpha}]. \tag{15.2.132}$$

Verify that $\boldsymbol{L}$ and ∇^2 commute, and use this fact to show (as expected, see Section 5) that all the Cartesian components of $\boldsymbol{A}^{\ell,m,\alpha}$ are harmonic functions,

$$\nabla^2 \boldsymbol{A}^{\ell,m,\alpha} = 0. \tag{15.2.133}$$

Verify that $\boldsymbol{L}$ and r commute,

$$\{\boldsymbol{L}, r\} = 0. \tag{15.2.134}$$

Of course, this result is to be expected since r is invariant under rotations. Next, in view of (2.31) and (2.32), we may define functions $h_\ell^{m,\alpha}(\theta, \phi)$ by writing

$$\psi_{\ell,m,\alpha} = H_\ell^{m.\alpha} = r^\ell h_\ell^{m,\alpha}(\theta, \phi). \tag{15.2.135}$$

Show it follows from (2.134) and (2.135) that

$$\boldsymbol{L}\psi_{\ell,m,\alpha} = r^\ell \boldsymbol{L} h_\ell^{m,\alpha}(\theta, \phi). \tag{15.2.136}$$

Therefore, if we wish, we can evaluate (2.132) using a raising and lowering operator formalism.

15.2.4. The purpose of this exercise is to verify (2.115) thus showing that the relations (2.109) through (2.111) can be written in a more compact form. By the definition of homogeneity, there is the relation

$$\boldsymbol{B}^n(\lambda \boldsymbol{r}) = \lambda^n \boldsymbol{B}^n(\boldsymbol{r}) \tag{15.2.137}$$

where λ is a scalar. Show from (2.109) that

$$\boldsymbol{B}(\lambda \boldsymbol{r}) = \sum_{n=0}^{\infty} \lambda^n \boldsymbol{B}^n(\boldsymbol{r}). \tag{15.2.138}$$

Next integrate both sides of (2.138) to demonstrate that

$$\int_0^1 d\lambda\, \lambda \boldsymbol{B}(\lambda \boldsymbol{r}) = \sum_{n=0}^{\infty}[1/(n+2)]\boldsymbol{B}^n(\boldsymbol{r}) = \sum_{n=1}^{\infty}[1/(n+1)]\boldsymbol{B}^{n-1}(\boldsymbol{r}). \tag{15.2.139}$$

Finally, using (2.110), (2.111), and (2.139), verify that there is the integral relation

$$\boldsymbol{A}(\boldsymbol{r}) = -\boldsymbol{r} \times \int_0^1 d\lambda\, \lambda \boldsymbol{B}(\lambda \boldsymbol{r}). \tag{15.2.140}$$

15.2.5. Subsection 2.6 showed that the Poincaré-Coulomb gauge is unique. Accordingly, starting from the requirement $\boldsymbol{B} = \nabla \times \boldsymbol{A}^{\rm min}$ and the requirements (2.70) and (2.71) and the assumption that $\boldsymbol{B}(\boldsymbol{r})$ is analytic in a neighborhood of $\boldsymbol{r} = 0$, it should be possible to derive the relations (2.111) and (2.115). Do so! Acknowledgement: This exercise was motivated by a suggestion of Sateesh Mane.

15.2.6. The relations (2.111) and(2.115) specify the minimum vector potential $\boldsymbol{A}^{\rm min}$in terms of the magnetic field $\boldsymbol{B}$. The purpose of this exercise is to derive relations that specify the scalar potential ψ in terms of $\boldsymbol{B}$.

Since B is assumed to be analytic and curl free, show that ψ may be defined by the rule

$$\psi(\boldsymbol{r}) = \int_0^r \boldsymbol{B}(\boldsymbol{r}') \cdot d\boldsymbol{r}' \tag{15.2.141}$$

where the integral (because $\boldsymbol{B}$ is curl free) may be carried out over any path joining 0 and $\boldsymbol{r}$. Note that with this definition

$$\psi(0) = 0. \tag{15.2.142}$$

Choose the path to be the straight line joining 0 and $\boldsymbol{r}$ by making the Ansatz

$$\boldsymbol{r}' = \lambda \boldsymbol{r} \text{ with } \lambda \in [0, 1]. \tag{15.2.143}$$

Show that so doing yields the result

$$\psi(\boldsymbol{r}) = \boldsymbol{r} \cdot \int_0^1 d\lambda\, \boldsymbol{B}(\lambda \boldsymbol{r}). \tag{15.2.144}$$

Assume that $\boldsymbol{B}(\boldsymbol{r})$ is decomposed into homogeneous polynomials as in (2.109). Show, using (2.138) and (2.144), that

$$\psi(\boldsymbol{r}) = \boldsymbol{r} \cdot \sum_{n=0}^{\infty}[1/(n+1)]\boldsymbol{B}^n(\boldsymbol{r}). \tag{15.2.145}$$

Suppose that $\psi(\boldsymbol{r})$ is also decomposed into homogeneous polynomials by writing

$$\psi(\boldsymbol{r}) = \sum_{n=1}^{\infty} \psi^n(\boldsymbol{r}). \tag{15.2.146}$$

Show that

$$\psi^n(\boldsymbol{r}) = (1/n)\boldsymbol{r} \cdot \boldsymbol{B}^{n-1}(\boldsymbol{r}) \text{ for } n = 1, 2, \cdots . \tag{15.2.147}$$

Verify directly that

$$\nabla\psi^n(\boldsymbol{r}) = \boldsymbol{B}^{n-1}(\boldsymbol{r}). \tag{15.2.148}$$

Hint: Use the vector identity

$$\nabla(\boldsymbol{a} \cdot \boldsymbol{b}) = (\boldsymbol{a} \cdot \nabla)\boldsymbol{b} + (\boldsymbol{b} \cdot \nabla)\boldsymbol{a} + \boldsymbol{a} \times (\nabla \times \boldsymbol{b}) + \boldsymbol{b} \times (\nabla \times \boldsymbol{a}) \tag{15.2.149}$$

with

$$\boldsymbol{a} = \boldsymbol{r} \tag{15.2.150}$$

and

$$\boldsymbol{b} = \boldsymbol{B}^{n-1}(\boldsymbol{r}). \tag{15.2.151}$$

15.2.7. Review Exercise 1.5.7. There it is found that a uniform vertical magnetic field $\boldsymbol{B} = B\boldsymbol{e}_y$ can be derived from the vector potential

$$\boldsymbol{A} = -Bx\boldsymbol{e}_z. \tag{15.2.152}$$

Recall the definition (2.79), the relations (2.9) through (2.11), and the definition

$$\int d\Omega = \int_0^{\pi} \int_0^{2\pi} \sin\theta \; d\theta d\phi. \tag{15.2.153}$$

Show that

$$\int x^2 d\Omega = \int y^2 d\Omega = \int z^2 d\Omega = (4/3)\pi r^2, \tag{15.2.154}$$

and therefore

$$||\boldsymbol{A}(\boldsymbol{r})||^2 = (4/3)\pi B^2 r^2. \tag{15.2.155}$$

Let $\boldsymbol{A}^{PC}$ be the associated vector potential in the Poincaré-Coulomb gauge. Verify that

$$\boldsymbol{A}^{PC} = -(B/2)(x\boldsymbol{e}_z - z\boldsymbol{e}_x). \tag{15.2.156}$$

Show that

$$||\boldsymbol{A}^{PC}(\boldsymbol{r})||^2 = (2/3)\pi B^2 r^2. \tag{15.2.157}$$

Comparison of (2.155) and (2.157) shows that the Poincaré-Coulomb gauge vector potential has a smaller norm, as expected. Show that

$$\boldsymbol{A}^{PC} = \boldsymbol{A} + \nabla\chi \tag{15.2.158}$$

with

$$\chi = (B/2)xz. \tag{15.2.159}$$

Verify that χ is harmonic, as expected.

15.2.8. Review Exercise 1.5.9. There it is found that a quadrupole magnetic field with midplane symmetry,

$$\boldsymbol{B} = Qy\boldsymbol{e}_x + Qx\boldsymbol{e}_y, \tag{15.2.160}$$

can be derived from the vector potential

$$\boldsymbol{A} = -(Q/2)(x^2 - y^2)\boldsymbol{e}_z. \tag{15.2.161}$$

Recall the definition (2.79), the relations (2.9) through (2.11), and the definition (2.153). Show that

$$\int (x^2 - y^2)^2 d\Omega = (16/15)\pi r^4, \tag{15.2.162}$$

and therefore

$$||\boldsymbol{A}(\boldsymbol{r})||^2 = (4/15)\pi Q^2 r^4. \tag{15.2.163}$$

Let $\boldsymbol{A}^{PC}$ be the associated vector potential in the Poincaré-Coulomb gauge. Verify that

$$\boldsymbol{r} \times \boldsymbol{B} = Q[(x^2 - y^2)\boldsymbol{e}_z + zy\boldsymbol{e}_y - zx\boldsymbol{e}_x], \tag{15.2.164}$$

and therefore

$$\boldsymbol{A}^{PC} = -(Q/3)[-zx\boldsymbol{e}_x + zy\boldsymbol{e}_y + (x^2 - y^2)\boldsymbol{e}_z]. \tag{15.2.165}$$

Show that

$$\int z^2x^2 d\Omega = \int z^2y^2 d\Omega = \int x^2y^2 d\Omega = (4/15)\pi r^4, \tag{15.2.166}$$

$$\int x^4 d\Omega = \int y^4 d\Omega = \int z^4 d\Omega = (4/5)\pi r^4, \tag{15.2.167}$$

and therefore

$$||\boldsymbol{A}^{PC}(\boldsymbol{r})||^2 = (2/3)(4/15)\pi Q^2 r^4. \tag{15.2.168}$$

Comparison of (2.163) and (2.168) shows that the Poincaré-Coulomb gauge vector potential has a smaller norm, as expected. Show that

$$\boldsymbol{A}^{PC} = \boldsymbol{A} + \nabla\chi \tag{15.2.169}$$

with

$$\chi = (Q/6)z(x^2 - y^2). \tag{15.2.170}$$

Verify that χ is harmonic, as expected.

15.2.9. Demonstration that harmonic functions take their extrema on boundaries.

15.3 Cylindrical Harmonic Expansion

In the previous section we employed spherical coordinates to find *local* expansions (expansions about a point $\boldsymbol{R}_0$) for the scalar potential ψ and the associated magnetic field $\boldsymbol{B}$. We also found a suitable vector potential $\boldsymbol{A}^{\min}$. The goal of this section is to show that it is possible to obtain *semi-global* expansions for ψ and $\boldsymbol{B}$ in the case of a *straight* geometry, the case of straight beam-line elements. By semi-global we mean that an expansion holds all along the vicinity of the beam-line axis (which we take to be the z axis). That is, while the variables x and y are treated as being small, the variable z need not be small.

For this purpose is convenient to work in cylindrical coordinates ρ, ϕ, and z as given by (2.12) through (2.14). We also note, for future use, that (2.13) and (2.14) can be written in the form

$$x + iy = \rho \exp(i\phi). \tag{15.3.1}$$

From this form it follows that

$$\rho^{2\ell} = (x^2 + y^2)^\ell \tag{15.3.2}$$

and, for $m \geq 0$,

$$\rho^m \cos m\phi = \Re[(x + iy)^m], \tag{15.3.3}$$

$$\rho^m \sin m\phi = \Im[(x + iy)^m]. \tag{15.3.4}$$

We see that *even* powers of ρ and the combinations $\rho^m \cos m\phi$ and $\rho^m \sin m\phi$ are *analytic* (in fact, *polynomial*) functions of x and y.

15.3.1 Complex Cylindrical Harmonic Expansion

To find the general ψ in cylindrical coordinates that satisfies Laplace's equation, recall that the functions $\exp(im\phi)$ form a complete set for the Hilbert space of functions over the interval $\phi \in [0, 2\pi]$, and the functions $\exp(ikz)$ form a complete set for the Hilbert space of functions over the interval $z \in [-\infty, \infty]$. Therefore any function ψ in the product Hilbert space can be written as a superposition of functions of the form $\Omega_m(k, \rho) \exp(ikz) \exp(im\phi)$ where the functions $\Omega_m(k, \rho)$ are yet to be determined. In cylindrical coordinates the Laplacian has the form

$$\nabla^2 = (1/\rho)(\partial/\partial\rho)(\rho\partial/\partial\rho) + (1/\rho^2)(\partial^2/\partial\phi^2) + \partial^2/\partial z^2. \tag{15.3.5}$$

Thus if the product $\Omega_m(k, \rho) \exp(ikz) \exp(im\phi)$ is to satisfy Laplace's equation, the functions $\Omega_m(k, \rho)$ must satisfy the modified Bessel equation,

$$(1/\rho)(\partial/\partial\rho)(\rho\partial\Omega_m/\partial\rho) - (m^2/\rho^2)\Omega_m - k^2\Omega_m = 0. \tag{15.3.6}$$

The solutions to this equation (that are regular for small ρ) are the modified Bessel functions $I_m(k\rho)$. Consequently, in cylindrical coordinates, a general ψ satisfying Laplace's equation and analytic in x,y near the z axis has the expansion

$$\psi(x, y, z) = \sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} dk\; G_m(k) \exp(ikz) \exp(im\phi) I_m(k\rho) \tag{15.3.7}$$

where the functions $G_m(k)$ are arbitrary. We remark, for future use, that the modified Bessel functions $I_m(w)$ have the property

$$I_{-m}(w) = I_m(w). \tag{15.3.8}$$

The representation (3.7) is a *cylindrical harmonic* or *cylindrical multipole* expansion, where m is related to the order of the multipole. For example, $m = 0$ for a 'monopole' source (including a solenoid), $m = 1$ for a dipole, $m = 2$ for a quadrupole, etc. We also remark that these are what we will call *pure* multipoles. For example, a real/physical quadrupole (even one with perfect four-fold symmetry) will have primarily $m = 2$ components plus smaller higher-order pure multipole components that are not forbidden by symmetry. Both pole/coil shape and rotational symmetry matter. See Subsection 3.5. Finally we should remark that (3.7) could more accurately be called a *circular* cylinder harmonic expansion. In Section 17.4 we will extend this work to include the case of cylinders with *elliptic* cross sections, and in Section 17.5 we will treat the case of cylinders with *rectangular* cross sections.

As stated at the beginning of this chapter, our ultimate goal is a Taylor expansion of the vector potential $\boldsymbol{A}$ in the variables x, y. To do this, we first seek an expansion of ψ as a Taylor series in the variables x, y with coefficients that depend on z. This can be achieved, by using the Taylor expansions for $I_m(w)$, as follows: Using (3.7) we may write

$$\begin{aligned}
\psi(x,y,z) &= \sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} dk\ G_m(k)\exp(ikz)\exp(im\phi)I_m(k\rho) \\
&= \sum_{m=-\infty}^{\infty} \exp(im\phi) \int_{-\infty}^{\infty} dk\ G_m(k)\exp(ikz)I_m(k\rho) \\
&= \sum_{m=-\infty}^{\infty} \exp(im\phi)\Psi_m(\rho,z) \qquad (15.3.9)
\end{aligned}$$

where

$$\Psi_m(\rho,z) = \int_{-\infty}^{\infty} dk\ G_m(k)\exp(ikz)I_m(k\rho). \qquad (15.3.10)$$

The modified Bessel functions have the expansions

$$I_m(w) = (1/2)^{|m|}w^{|m|}\sum_{\ell=0}^{\infty} w^{2\ell}/[2^{2\ell}\ell!(\ell+|m|)!]. \qquad (15.3.11)$$

Therefore we may also write

$$\begin{aligned}
&\Psi_m(\rho,z) = \int_{-\infty}^{\infty} dk\ G_m(k)\exp(ikz)I_m(k\rho) = \\
&\int_{-\infty}^{\infty} dk\ G_m(k)\exp(ikz)(1/2)^{|m|}(k\rho)^{|m|}\sum_{\ell=0}^{\infty}(k\rho)^{2\ell}/[2^{2\ell}\ell!(\ell+|m|)!] = \\
&\sum_{\ell=0}^{\infty}\{1/[2^{2\ell}\ell!(\ell+|m|)!]\}\rho^{2\ell+|m|}(1/2)^{|m|}\int_{-\infty}^{\infty} dk\ k^{2\ell+|m|}G_m(k)\exp(ikz).
\end{aligned} \qquad (15.3.12)$$

Define functions $C_m^{[0]}(z)$ by writing

$$C_m^{[0]}(z) \overset{\text{def}}{=} (1/2)^{|m|}(1/|m|!)\int_{-\infty}^{\infty} dk\ k^{|m|}G_m(k)\exp(ikz). \qquad (15.3.13)$$

Also, define functions $C_m^{[n]}(z)$ by writing

$$C_m^{[n]}(z) = (\partial_z)^n C_m^{[0]}(z). \qquad (15.3.14)$$

Then, by differentiating under the integral sign, we have the result

$$C_m^{[n]}(z) = (\partial_z)^n C_m^{[0]}(z) = i^n (1/2)^{|m|}(1/|m|!)\int_{-\infty}^{\infty} dk\ k^{n+|m|} G_m(k)\exp(ikz) \tag{15.3.15}$$

and, in particular,

$$C_m^{[2\ell]}(z) = (-1)^\ell (1/2)^{|m|}(1/|m|!)\int_{-\infty}^{\infty} dk\ k^{2\ell+|m|} G_m(k)\exp(ikz). \tag{15.3.16}$$

Thus, we may also write the relation

$$(1/2)^{|m|}\int_{-\infty}^{\infty} dk\ k^{2\ell+|m|} G_m(k)\exp(ikz) = (-1)^\ell |m|! C_m^{[2\ell]}(z). \tag{15.3.17}$$

Putting everything together gives the result

$$\Psi_m(\rho,z) = \sum_{\ell=0}^{\infty} (-1)^\ell \frac{|m|!}{2^{2\ell}\ell!(\ell+|m|)!} C_m^{[2\ell]}(z)\rho^{2\ell+|m|}. \tag{15.3.18}$$

Consequently, $\psi(x,y,z)$ has the representation

$$\psi(x,y,z) = \sum_{m=-\infty}^{\infty} \exp(im\phi) \sum_{\ell=0}^{\infty} (-1)^\ell \frac{|m|!}{2^{2\ell}\ell!(\ell+|m|)!} C_m^{[2\ell]}(z)\rho^{2\ell+|m|}. \tag{15.3.19}$$

Note that, in view of (3.2) through (3.4), the terms appearing on the right side of (3.19) are polynomial in the variables x and y.

From (3.18) we see that

$$C_m^{[0]}(z) = \lim_{\rho\to 0}(1/\rho^{|m|})\Psi_m(\rho,z). \tag{15.3.20}$$

For this reason, the functions $C_m^{[0]}(z)$ are called the *generalized on-axis gradients*.[7] Note that the generalized gradients depend on the longitudinal variable z. However we will soon see that, for fields produced by long well-made magnets, the z dependence will be significant only at the ends.

15.3.2 Real Cylindrical Harmonic Expansion

So far, for mathematical convenience, we have worked with a possibly complex representation for ψ. We will now convert our results into equivalent real forms suitable for physical applications. We begin with the relation (3.19). Suppose we require that $\psi(x,y,z)$ be real. Forming the complex conjugate of (3.19) gives the result

$$\bar{\psi}(x,y,z) = \sum_{m=-\infty}^{\infty} \exp(-im\phi) \sum_{\ell=0}^{\infty} (-1)^\ell \frac{|m|!}{2^{2\ell}\ell!(\ell+|m|)!} \bar{C}_m^{[2\ell]}(z)\rho^{2\ell+|m|}. \tag{15.3.21}$$

[7] Although (3.20) is mathematically correct, it is not a good way to actually compute the on-axis gradients due to the delicate nature of the limiting process. Indeed, one of the aims of Chapters 17 through 21 is to provide reliable ways of computing the on-axis gradients.

The right side of (3.21) can be rewritten to give the relation

$$\begin{aligned}\sum_{m=-\infty}^{\infty} \exp(-im\phi) \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{|m|!}{2^{2\ell}\ell!(\ell+|m|)!}\bar{C}_m^{[2\ell]}(z)\rho^{2\ell+|m|} &= \\ \sum_{m=-\infty}^{\infty} \exp(im\phi) \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{|m|!}{2^{2\ell}\ell!(\ell+|m|)!}\bar{C}_{-m}^{[2\ell]}(z)\rho^{2\ell+|m|}. \qquad (15.3.22)\end{aligned}$$

Therefore requiring

$$\bar{\psi}(x,y,z) = \psi(x,y,z) \qquad (15.3.23)$$

is equivalent to the requirement

$$\bar{C}_{-m}^{[2\ell]}(z) = C_m^{[2\ell]}(z), \qquad (15.3.24)$$

or

$$C_{-m}^{[2\ell]}(z) = \bar{C}_m^{[2\ell]}(z). \qquad (15.3.25)$$

Let us now use this information to rewrite ψ. From (3.19) we have, in the general case,

$$\begin{aligned}\psi(x,y,z) &= \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_0^{[2\ell]}(z)\rho^{2\ell} \\ &+ \sum_{m\neq 0}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{|m|!}{2^{2\ell}\ell!(\ell+|m|)!}C_m^{[2\ell]}(z)\rho^{2\ell+|m|} \\ &+ i\sum_{m\neq 0}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{|m|!}{2^{2\ell}\ell!(\ell+|m|)!}C_m^{[2\ell]}(z)\rho^{2\ell+|m|}.\end{aligned} \qquad (15.3.26)$$

The second sum over m in (3.26) can be rewritten as

$$\begin{aligned}\sum_{m\neq 0}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{|m|!}{2^{2\ell}\ell!(\ell+|m|)!}C_m^{[2\ell]}(z)\rho^{2\ell+|m|} &= \\ \sum_{m=1}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}[C_m^{[2\ell]}(z)+C_{-m}^{[2\ell]}(z)]\rho^{2\ell+m}. \qquad (15.3.27)\end{aligned}$$

Now define functions $C_{m,c}^{[2\ell]}(z)$ by the rule

$$C_{m,c}^{[2\ell]}(z) = C_m^{[2\ell]}(z) + C_{-m}^{[2\ell]}(z) \text{ for } m \geq 1, \qquad (15.3.28)$$

so we may also write

$$\begin{aligned}\sum_{m\neq 0}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{|m|!}{2^{2\ell}\ell!(\ell+|m|)!}C_m^{[2\ell]}(z)\rho^{2\ell+|m|} &= \\ \sum_{m=1}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell+m}. \qquad (15.3.29)\end{aligned}$$

According to (3.25), the functions $C_{m,c}^{[2\ell]}(z)$ will be real if ψ is real. The third sum over m in (3.26) can be rewritten as

$$i\sum_{m\neq 0}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{|m|!}{2^{2\ell}\ell!(\ell+|m|)!}C_m^{[2\ell]}(z)\rho^{2\ell+|m|} =$$
$$i\sum_{m=1}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}[C_m^{[2\ell]}(z)-C_{-m}^{[2\ell]}(z)]\rho^{2\ell+m}. \tag{15.3.30}$$

Now define functions $C_{m,s}^{[2\ell]}(z)$ by the rule

$$C_{m,s}^{[2\ell]}(z) = i[C_m^{[2\ell]}(z)-C_{-m}^{[2\ell]}(z)] \text{ for } m\geq 1, \tag{15.3.31}$$

so we may also write

$$i\sum_{m\neq 0}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{|m|!}{2^{2\ell}\ell!(\ell+|m|)!}C_m^{[2\ell]}(z)\rho^{2\ell+|m|} =$$
$$\sum_{m=1}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell+m}. \tag{15.3.32}$$

According to (3.25), the functions $C_{m,s}^{[2\ell]}(z)$ with $m\geq 1$ will be real if ψ is real. Combining the various results so far gives the representation

$$\begin{aligned}\psi(x,y,z) &= \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_0^{[2\ell]}(z)\rho^{2\ell}\\ &+ \sum_{m=1}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell+m}\\ &+ \sum_{m=1}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell+m}.\end{aligned} \tag{15.3.33}$$

Observe that, according to (3.25), the $C_0^{[2\ell]}(z)$ will be real if ψ is real,

$$C_0^{[0]}(z) = \bar{C}_0^{[0]}(z). \tag{15.3.34}$$

Thus, all the quantities appearing in (3.33) are real if ψ is real. For future use it will be convenient to extend the definitions (3.28) and (3.31) to the $m=0$ case by writing

$$C_{0,c}^{[0]}(z) = C_0^{[0]}(z), \tag{15.3.35}$$

$$C_{0,s}^{[0]}(z) = 0. \tag{15.3.36}$$

We note that all the functions $C_0^{[0]}(z)$, $C_{m,c}^{[0]}(z)$, and $C_{m,s}^{[0]}(z)$ may be chosen independently, and any such choice produces a harmonic function when employed in (3.33). Finally, we observe that all the terms in (3.33) are sums of quantities of the form $\rho^m\cos(m\phi)$ or $\rho^m\sin(m\phi)$

multiplied by powers of ρ^2 with z-dependent coefficients $C_0^{[2\ell]}(z)$, $C_{m,c}^{[2\ell]}(z)$, and $C_{m,s}^{[2\ell]}(z)$. Thus, in view of (3.2) through (3.4), we have achieved our goal of finding a Taylor expansion for $\psi(x,y,z)$ in powers of x,y with coefficients that depend on z.

We close this subsection by introducing some further notation that will be of future use. Define quantities $\Psi_0(\rho,z)$, $\Psi_{m,c}(\rho,z)$, and $\Psi_{m,s}(\rho,z)$ by the equations

$$\Psi_0(\rho,z)=\Psi_{0,c}(\rho,z)=\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_0^{[2\ell]}(z)\rho^{2\ell}, \tag{15.3.37}$$

$$\Psi_{m,c}(\rho,z)=\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell+m}, \tag{15.3.38}$$

$$\Psi_{m,s}(\rho,z)=\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell+m}, \tag{15.3.39}$$

so that, in view of (3.33), we may write

$$\begin{aligned}\psi(x,y,z) &= \Psi_0(\rho,z)+\sum_{m=1}^{\infty}\Psi_{m,c}(\rho,z)\cos m\phi+\sum_{m=1}^{\infty}\Psi_{m,s}(\rho,z)\sin m\phi\\ &= \psi_0(x,y,z)+\sum_{m=1}^{\infty}\psi_{m,c}(x,y,z)+\sum_{m=1}^{\infty}\psi_{m,s}(x,y,z) \qquad (15.3.40)\end{aligned}$$

where

$$\psi_0(x,y,z)=\psi_{0,c}(x,y,z)=\Psi_0(\rho,z), \tag{15.3.41}$$

$$\psi_{m,c}(x,y,z)=\psi_{m,c}(\rho,\phi,z)=\cos(m\phi)\Psi_{m,c}(\rho,z), \tag{15.3.42}$$

$$\psi_{m,s}(x,y,z)=\psi_{m,s}(\rho,\phi,z)=\sin(m\phi)\Psi_{m,s}(\rho,z). \tag{15.3.43}$$

Finally, there is a variant coefficient relation that will be of subsequent use. Begin by rewriting (3.07) in the form

$$\begin{aligned}\psi(x,y,z) &= \sum_{m=-\infty}^{\infty}\int_{-\infty}^{\infty}dk\ G_m(k)\exp(ikz)\cos(m\phi)I_m(k\rho)\\ &+ i\sum_{m=-\infty}^{\infty}\int_{-\infty}^{\infty}dk\ G_m(k)\exp(ikz)\sin(m\phi)I_m(k\rho), \qquad (15.3.44)\end{aligned}$$

from which it follows that

$$\begin{aligned}\psi(x,y,z) &= \int_{-\infty}^{\infty}dk\ G_0(k)\exp(ikz)I_0(k\rho)\\ &+ \sum_{m=1}^{\infty}\int_{-\infty}^{\infty}dk\ [G_m(k)+G_{-m}(k)]\exp(ikz)\cos(m\phi)I_m(k\rho)\\ &+ \sum_{m=1}^{\infty}\int_{-\infty}^{\infty}dk\ [iG_m(k)-iG_{-m}](k)\exp(ikz)\sin(m\phi)I_m(k\rho). \qquad (15.3.45)\end{aligned}$$

Next, with $m \geq 1$, introduce the notation

$$G_{m,c}(k) = G_m(k) + G_{-m}(k), \tag{15.3.46}$$

$$G_{m,s}(k) = iG_m(k) - iG_{-m}(k), \tag{15.3.47}$$

with the conventions

$$G_{0,c}(k) = G_0(k), \tag{15.3.48}$$

$$G_{0,s}(k) = 0. \tag{15.3.49}$$

In terms of this notation, ψ has the representation

$$\begin{aligned} \psi(x,y,z) &= \sum_{m=0}^{\infty} \int_{-\infty}^{\infty} dk\ G_{m,c}(k)\exp(ikz)\cos(m\phi)I_m(k\rho) \\ &+ \sum_{m=1}^{\infty} \int_{-\infty}^{\infty} dk\ G_{m,s}(k)\exp(ikz)\sin(m\phi)I_m(k\rho). \end{aligned} \tag{15.3.50}$$

We observe that (3.28) and (3.31) can now also be written in the forms

$$C_{m,c}^{[n]}(z) = (\partial_z)^n C_{m,c}^{[0]}(z) = i^n(1/2)^m(1/m!)\int_{-\infty}^{\infty} dk\ k^{n+m}G_{m,c}(k)\exp(ikz), \tag{15.3.51}$$

$$C_{m,s}^{[n]}(z) = (\partial_z)^n C_{m,s}^{[0]}(z) = i^n(1/2)^m(1/m!)\int_{-\infty}^{\infty} dk\ k^{n+m}G_{m,s}(k)\exp(ikz). \tag{15.3.52}$$

15.3.3 Some Simple Examples: $m = 0, 1, 2$

Let us seek a physical interpretation for the functions $C_0^{[0]}(z)$, $C_{m,c}^{[0]}(z)$, and $C_{m,s}^{[0]}(z)$ by computing the associated magnetic fields using (2.6).

The Case $m = 0$

If only the $C_0^{[0]}(z)$ related terms (the $m = 0$ terms) are present in (3.33), ψ has the expansion

$$\psi_0(x,y,z) = C_0^{[0]}(z) - (1/4)(x^2+y^2)C_0^{[2]}(z) + \cdots, \tag{15.3.53}$$

and therefore

$$B_x = \partial_x\psi_0 = -(1/2)xC_0^{[2]}(z) + \cdots, \tag{15.3.54}$$

$$B_y = \partial_y\psi_0 = -(1/2)yC_0^{[2]}(z) + \cdots, \tag{15.3.55}$$

$$B_z = \partial_z\psi_0 = C_0^{[1]}(z) - (1/4)(x^2+y^2)C_0^{[3]}(z) + \cdots. \tag{15.3.56}$$

We see that $\boldsymbol{B}$ is primarily in the z direction and that $B_z(0,0,z)$, the on-axis z component of $\boldsymbol{B}$, has a profile given by $C_0^{[1]}(z)$. As long as $C_0^{[1]}(z)$ is nearly constant, $C_0^{[2]}(z)$ will be small, and therefore the transverse field components B_x and B_y will be small. Such would be the case for the field of a solenoid where the field is primarily longitudinal and only has transverse components in the fringe-field regions at each end where $C_0^{[1]}(z)$ is changing. (See

Sections 16.1 and 21.1.) We know that for any solenoid-like element (an element having a nonvanishing $m=0$ component in the cylindrical harmonic expansion of its scalar potential and a nonvanishing longitudinal field somewhere on axis) $C_0^{[1]}(z)$ must depend on z because this on-axis gradient must be nonzero somewhere for such an element and must vanish far outside any such element because $\boldsymbol{B}$ vanishes there. Therefore the functions $C_0^{[2]}(z)$, $C_0^{[3]}(z)$, etc., must be nonzero, at least near the end and fringe-field regions of any such element. We conclude that, as a consequence of Maxwell's equations, the scalar potential ψ_0 (and as we will see, the associated vector potential) for any such element must contain terms beyond degree two in the variables x, y. Correspondingly, the transfer map for any real solenoid must contain nonlinear terms.

The same set up could also describe some portion of the field due to an off-center dipole (or any other off-center higher-order multipole) since such magnets would also have an on-axis B_z component somewhere in the fringe-field regions. In all cases we know that $C_0^{[1]}(z)$ must depend on z because this on-axis gradient must be nonzero somewhere in or near the element and must vanish far outside the element, again because $\boldsymbol{B}$ vanishes there.

The Case $m = 1$

Next, suppose that only the $C_{1,s}^{[0]}(z)$ related terms are present in (3.33). In this $m=1$ case, ψ has an expansion of the form

$$\begin{aligned}\psi_{1,s}(x,y,z) &= \rho\sin(\phi)[C_{1,s}^{[0]}(z)-(1/8)(x^2+y^2)C_{1,s}^{[2]}(z)+\cdots]\\ &= y[C_{1,s}^{[0]}(z)-(1/8)(x^2+y^2)C_{1,s}^{[2]}(z)+\cdots]\end{aligned}\tag{15.3.57}$$

and therefore

$$B_x=\partial_x\psi_{1,s}=-(1/4)xyC_{1,s}^{[2]}(z)+\cdots\ ,\tag{15.3.58}$$

$$B_y=\partial_y\psi_{1,s}=C_{1,s}^{[0]}(z)-(1/8)(x^2+3y^2)C_{1,s}^{[2]}(z)+\cdots\ ,\tag{15.3.59}$$

$$B_z=\partial_z\psi_{1,s}=y[C_{1,s}^{[1]}(z)-(1/8)(x^2+y^2)C_{1,s}^{[3]}(z)+\cdots].\tag{15.3.60}$$

We see that $\boldsymbol{B}$ is primarily in the y direction with a profile given by $C_{1,s}^{[0]}(z)$. As long as $C_{1,s}^{[0]}(z)$ is nearly constant, $C_{1,s}^{[1]}(z)$ and higher derivatives of $C_{1,s}^{[0]}(z)$ will be small, and therefore the other field components B_x and B_z will be small. Such would be the case for the field of a (normal) dipole where the field is primarily vertical and only has x and z components in the fringe-field regions at each end where $C_{1,s}^{[0]}(z)$ is changing. (See Exercise 1.5.7.) However, there will always be nonlinear terms at the ends where $C_{1,s}^{[0]}(z)$ and higher derivatives cannot be constant. Correspondingly, the transfer map for any real dipole must contain nonlinear terms. We end the discussion of the $m=1$ case by remarking that the $C_{1,c}^{[0]}(z)$ related terms describe the field of a *skew* dipole. See Exercise 4.1.

The Case $m = 2$

As a last example, suppose that only the $C^{[0]}_{2,s}(z)$ related terms are present in (3.33). In this $m = 2$ case, ψ has an expansion of the form

$$\begin{aligned}\psi_{2,s}(x,y,z) &= \rho^2 \sin(2\phi)[C^{[0]}_{2,s}(z) - (1/24)(x^2+y^2)C^{[2]}_{2,s}(z) + \cdots] \\ &= 2xy[C^{[0]}_{2,s}(z) - (1/24)(x^2+y^2)C^{[2]}_{2,s}(z) + \cdots] \qquad (15.3.61)\end{aligned}$$

and therefore

$$B_x = \partial_x \psi_{2,s} = 2yC^{[0]}_{2,s}(z) - (1/12)(3x^2y + y^3)C^{[2]}_{2,s}(z) + \cdots , \qquad (15.3.62)$$

$$B_y = \partial_y \psi_{2,s} = 2xC^{[0]}_{2,s}(z) - (1/12)(x^3 + 3xy^2)C^{[2]}_{2,s}(z) + \cdots , \qquad (15.3.63)$$

$$B_z = \partial_z \psi_{2,s} = 2xy[C^{[1]}_{2,s}(z) - (1/24)(x^2+y^2)C^{[3]}_{2,s}(z) + \cdots]. \qquad (15.3.64)$$

We see that $\boldsymbol{B}$ is primarily the field of a (normal) quadrupole with a profile given by $Q = 2C^{[0]}_{2,s}(z)$. See Exercise 1.5.9. As long as $C^{[0]}_{2,s}(z)$ is nearly constant, $C^{[1]}_{2,s}(z)$ and higher derivatives of $C^{[0]}_{2,s}(z)$ will be small, and therefore the other field components will be small. Such would be the case for the field of a (normal) quadrupole where the field is primarily of the form given by (1.5.62) through (1.5.64) and only has z components in the fringe-field regions at each end where $C^{[0]}_{2,s}(z)$ is changing. Again, the transfer map for a real quadrupole must contain nonlinear terms because $C^{[0]}_{2,s}(z)$ must have nonzero derivatives in the fringe-field regions. We close the discussion of the $m = 2$ case by remarking that the $C^{[0]}_{2,c}(z)$ related terms describe the field of a skew quadrupole. See Exercise 4.2.

15.3.4 Magnetic Field Expansions for the General Case

General Results

Since ψ is a harmonic function and the operators $\partial_x, \partial_y, \partial_z$ commute with ∇^2, it follows from (2.6) that the (Cartesian) components of $\boldsymbol{B}$ must also be harmonic functions. Consequently each component of $\boldsymbol{B}$ must have a cylindrical multipole expansion of the form (3.33). Indeed, if ψ has the expansion (3.33), then it can be shown that the components of $\boldsymbol{B}$ have the expansions

$$
\begin{aligned}
B_x = \partial_x \psi(x,y,z) = &\sum_{\ell=0}^{\infty}(-1)^\ell \frac{1}{2^{2\ell}\ell!\ell!} C_{1,c}^{[2\ell]}(z)\rho^{2\ell} \\
&+ \cos(\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{2}{2^{2\ell}\ell!(\ell+1)!}[C_{2,c}^{[2\ell]}(z) - (1/4)C_0^{[2\ell+2]}(z)]\rho^{2\ell+1} \\
&+ \sum_{m=2}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,c}^{[2\ell]}(z) - [1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\}\rho^{2\ell+m} \\
&+ \sin(\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{2}{2^{2\ell}\ell!(\ell+1)!} C_{2,s}^{[2\ell]}(z)\rho^{2\ell+1} \\
&+ \sum_{m=2}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,s}^{[2\ell]}(z) - [1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\}\rho^{2\ell+m},
\end{aligned}
\tag{15.3.65}
$$

$$
\begin{aligned}
B_y = \partial_y \psi(x,y,z) = &\sum_{\ell=0}^{\infty}(-1)^\ell \frac{1}{2^{2\ell}\ell!\ell!} C_{1,s}^{[2\ell]}(z)\rho^{2\ell} \\
&+ \cos(\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{2}{2^{2\ell}\ell!(\ell+1)!} C_{2,s}^{[2\ell]}(z)\rho^{2\ell+1} \\
&+ \sum_{m=2}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,s}^{[2\ell]}(z) + [1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\}\rho^{2\ell+m} \\
&- \sin(\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{2}{2^{2\ell}\ell!(\ell+1)!}[C_{2,c}^{[2\ell]}(z) + (1/4)C_0^{[2\ell+2]}(z)]\rho^{2\ell+1} \\
&- \sum_{m=2}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,c}^{[2\ell]}(z) + [1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\}\rho^{2\ell+m},
\end{aligned}
\tag{15.3.66}
$$

$$
\begin{aligned}
B_z = \partial_z \psi(x,y,z) = &\sum_{\ell=0}^{\infty}(-1)^\ell \frac{1}{2^{2\ell}\ell!\ell!} C_0^{[2\ell+1]}(z)\rho^{2\ell} \\
+ &\sum_{m=1}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!} C_{m,c}^{[2\ell+1]}(z)\rho^{2\ell+m} \\
+ &\sum_{m=1}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!} C_{m,s}^{[2\ell+1]}(z)\rho^{2\ell+m}.
\end{aligned}
\tag{15.3.67}
$$

See Appendix H.[8]

[8] That the components of $\boldsymbol{B}$ should depend on the coordinates ρ, ϕ, z and the coefficients $C_{m,c}^{[n]}, C_{m,s}^{[n]}$ in some fashion is a consequence of (2.6). That they should do so in the particular combinations (3.65) through

Leading Behavior in Body

Let us compute the leading behavior of the various components of $\boldsymbol{B}$. For the monopole (solenoid) case, for which $m = 0$, we assume $C_0^{[1]}(z)$ is *constant* and all other coefficients are zero. Then we find from (3.65) through (3.67) the results

$$B_x^0 = 0, \tag{15.3.68}$$

$$B_y^0 = 0, \tag{15.3.69}$$

$$B_z^0 = C_0^{[1]}. \tag{15.3.70}$$

For the dipole ($m = 1$) cases, assuming $C_{1,c}^{[0]}(z)$ or $C_{1,s}^{[0]}(z)$ is constant and all other coefficients are zero, we find the results

$$B_x^{1,c} = C_{1,c}^{[0]}, \tag{15.3.71}$$

$$B_y^{1,c} = 0, \tag{15.3.72}$$

$$B_z^{1,c} = 0; \tag{15.3.73}$$

$$B_x^{1,s} = 0, \tag{15.3.74}$$

$$B_y^{1,s} = C_{1,s}^{[0]}, \tag{15.3.75}$$

$$B_z^{1,s} = 0. \tag{15.3.76}$$

For the $m \geq 2$ cases, assuming $C_{m,c}^{[0]}(z)$ or $C_{m,s}^{[0]}(z)$ is constant and therefore all other coefficients $C_{m,c}^{[n]}(z)$ and $C_{m,s}^{[n]}(z)$ vanish for $n > 0$, we find the results

$$B_x^{m,c} = m\cos[(m-1)\phi]\rho^{m-1}C_{m,c}^{[0]} = mC_{m,c}^{[0]}\,\Re[(x+iy)^{m-1}], \tag{15.3.77}$$

$$B_y^{m,c} = -m\sin[(m-1)\phi]\rho^{m-1}C_{m,c}^{[0]} = -mC_{m,c}^{[0]}\,\Im[(x+iy)^{m-1}], \tag{15.3.78}$$

$$B_z^{m,c} = 0; \tag{15.3.79}$$

$$B_x^{m,s} = m\sin[(m-1)\phi]\rho^{m-1}C_{m,s}^{[0]} = mC_{m,s}^{[0]}\,\Im[(x+iy)^{m-1}], \tag{15.3.80}$$

$$B_y^{m,s} = m\cos[(m-1)\phi]\rho^{m-1}C_{m,s}^{[0]} = mC_{m,s}^{[0]}\,\Re[(x+iy)^{m-1}], \tag{15.3.81}$$

$$B_z^{m,s} = 0. \tag{15.3.82}$$

Here we have used (3.3) and (3.4). Note that, if the relations (3.77) through (3.82) are evaluated for $m = 1$, they reproduce the results (3.71) through (3.76). They therefore actually hold for all $m > 0$.

When both $C_{m,c}^{[0]}(z)$ and $C_{m,s}^{[0]}(z)$ are present and constant, and all other coeffciients are zero, we may write

$$\boldsymbol{B}^m = \boldsymbol{B}^{m,c} + \boldsymbol{B}^{m,s}. \tag{15.3.83}$$

(3.67) is in part a consequence of the components of $\boldsymbol{B}$ being harmonic. See Exercise 3.4.

With this notation, the relations (3.71), (3.72), (3.74), (3.75), (3.77), (3.78), (3.80), and (3.81) can be rewritten in the compact (but complex) form

$$B_y^m + iB_x^m = m(C_{m,s}^{[0]} + iC_{m,c}^{[0]})(x+iy)^{m-1}. \tag{15.3.84}$$

It must be emphasized, however, that (3.84) holds only in the *body* of a pure multipole magnet, and not in the fringe-field regions at the ends. Moreover, fringe fields and their nonlinear contributions to transfer maps are inescapable consequences of Maxwell's equations.

15.3.5 Symmetry and Allowed and Forbidden Multipoles

There are restrictions on multipole content dictated by symmetry conditions. As a first example, consider a rectangular bending magnet such as that shown in Figures 1.6.1 and 1.6.2. Suppose the magnet is rotated by 180° about the z axis, and simultaneously the strength of the magnet is reversed in sign (so that ψ is replaced by $-\psi$). Assuming perfect symmetry, doing so should produce the same magnetic field as before. Correspondingly, the scalar potential ψ should remain unchanged. Suppose ψ as given by (3.33) is regarded as a function of ρ, ϕ, and z. Then we demand that

$$-\psi(\rho, \phi - \pi, z) = \psi(\rho, \phi, z). \tag{15.3.85}$$

Inspection of (3.33) shows that the requirement (3.85) forces all the coefficients $C_{m,c}^{[2\ell]}$ and $C_{m,s}^{[2\ell]}$ to be zero save those for which $m = 1, 3, 5, \cdots$.

Next consider a quadrupole magnet. Again assuming perfect symmetry, its magnet field should be unchanged if it is rotated by 90° about the z axis, and simultaneously the strength of the magnet is reversed in sign. In this case we demand that

$$-\psi(\rho, \phi - \pi/2, z) = \psi(\rho, \phi, z). \tag{15.3.86}$$

Now inspection of (3.33) shows that the requirement (3.86) forces all the coefficients $C_{m,c}^{[2\ell]}$ and $C_{m,s}^{[2\ell]}$ to be zero save those for which $m = 2, 6, 10, \cdots$.

Finally, consider a perfectly symmetric $2n$-pole magnet for $n = 1, 2, 3, \cdots$. In this case a rotation by $(360/2n)°$ and reversing the strength should leave the field unchanged. Now we conclude all multipole coefficients must vanish save possibly those for which

$$m = n(2j+1) \text{ with } j = 0, 1, 2, 3, \cdots . \tag{15.3.87}$$

In addition to rotational symmetry, there is the consideration of *midplane* symmetry in which one observes what happens when $y \to -y$, or equivalently, in view of (2.13) and (2.14), $\phi \to -\phi$. From (3.2) and (3.37) through (3.39) we see that the functions Ψ_0, $\Psi_{m,c}$, and $\Psi_{m,s}$ are invariant under this operation. It follows from (3.41) through (3.43) that there are the relations

$$\psi_0(x, -y, z) = \psi_0(x, y, z), \tag{15.3.88}$$

$$\psi_{m,c}(x, -y, z) = \psi_{m,c}(x, y, z), \tag{15.3.89}$$

$$\psi_{m,s}(x, -y, z) = -\psi_{m,s}(x, y, z). \tag{15.3.90}$$

We will say that a magnetic field $\boldsymbol{B}$ has *midplane* symmetry if it arises from a scalar potential that only has terms of the form $\psi_{m,s}$,

$$\boldsymbol{B} = \nabla \sum_{m=1}^{\infty} \psi_{m,s}. \tag{15.3.91}$$

Such a field is also said to be *normal* or produced by normal multipoles. Correspondingly, fields of the form

$$\boldsymbol{B} = \nabla \sum_{m=1}^{\infty} \psi_{m,c} \tag{15.3.92}$$

are said to be *skew* or produced by skew multipoles. From (2.6) we see that normal fields have the symmetry property

$$\boldsymbol{B}_x^{\rm normal}(x,-y,z) = -\boldsymbol{B}_x^{\rm normal}(x,y,z), \tag{15.3.93}$$

$$\boldsymbol{B}_y^{\rm normal}(x,-y,z) = \boldsymbol{B}_y^{\rm normal}(x,y,z), \tag{15.3.94}$$

$$\boldsymbol{B}_z^{\rm normal}(x,-y,z) = -\boldsymbol{B}_z^{\rm normal}(x,y,z); \tag{15.3.95}$$

and skew fields have the property

$$\boldsymbol{B}_x^{\rm skew}(x,-y,z) = \boldsymbol{B}_x^{\rm skew}(x,y,z), \tag{15.3.96}$$

$$\boldsymbol{B}_y^{\rm skew}(x,-y,z) = -\boldsymbol{B}_y^{\rm skew}(x,y,z), \tag{15.3.97}$$

$$\boldsymbol{B}_z^{\rm skew}(x,-y,z) = \boldsymbol{B}_z^{\rm skew}(x,y,z). \tag{15.3.98}$$

The same conclusions can be drawn from (3.65) through (3.67). Observe that, by these definitions, the field arising from ψ_0, for example the field of a solenoid, is also skew.

We note that, multipole by multipole, skew elements are related to normal elements and vice versa by rotations about the z axis. From (3.42) and (3.43) we find the relations

$$\psi_{m,c}[\rho,\phi-\pi/(2m),z] = \sin(m\phi)\Psi_{m,c}(\rho,z), \tag{15.3.99}$$

$$\psi_{m,s}[\rho,\phi-\pi/(2m),z] = -\cos(m\phi)\Psi_{m,s}(\rho,z). \tag{15.3.100}$$

We see that a skew element is converted into a normal element, but with on-axis gradients $C_{m,c}^{[2\ell]}(z)$; and a normal element is converted into a skew element, but with on-axis gradients $-C_{m,s}^{[2\ell]}(z)$.

Finally, we observe that a similar discussion could be given to the possible symmetry operation $x \to -x$, for which the properties of $\psi_{m,c}$ and $\psi_{m,s}$ are interchanged.

15.3.6 Relation between Harmonic Polynomials in Spherical and Cylindrical Coordinates

Equation (2.26) defined Harmonic polynomials in terns of a radius r and the spherical harmonics $Y_\ell^m(\theta,\phi)$, thereby providing a description in terms of spherical coordinates; and the relations (2.27) through (2.29) illustrate that the results of this definition are indeed polynomials in the Cartesian coordinates x, y, z. Suppose these polynomials are re-expressed in terms of cylindrical coordinates. This could be done using the z axis as the axis of the cylinder as in the relations (2.13) and (2.14), or with some other axis taken to be the axis of the cylinder as in Exercise 1.5.4.. What would be the appearance of such expansions? And how are such expansions related to the cylindrical harmonic expansions found in (3.19) and (3.33)? The answers to these questions are the subject of this section.

Suppose the substitutions (2.13) and (2.14) are made in the relations (2.27) through (2.29). So doing and employing (3.1) yields the results

$$H_0^0(\boldsymbol{r}) = 1/\sqrt{4\pi}; \tag{15.3.101}$$

$$\begin{aligned} H_1^1(\boldsymbol{r}) &= -\sqrt{3/(8\pi)}\rho\exp(i\phi), \\ H_1^0(\boldsymbol{r}) &= \sqrt{3/(4\pi)}z, \\ H_1^{-1}(\boldsymbol{r}) &= \sqrt{3/(8\pi)}\rho\exp(-i\phi); \end{aligned} \tag{15.3.102}$$

$$\begin{aligned} H_2^2(\boldsymbol{r}) &= \sqrt{15/(32\pi)}\rho^2\exp(2i\phi), \\ H_2^1(\boldsymbol{r}) &= -\sqrt{15/(8\pi)}z\rho\exp(i\phi), \\ H_2^0(\boldsymbol{r}) &= \sqrt{5/(16\pi)}(2z^2-\rho^2), \\ H_2^{-1}(\boldsymbol{r}) &= \sqrt{15/(8\pi)}z\rho\exp(-i\phi), \\ H_2^{-2}(\boldsymbol{r}) &= \sqrt{15/(32\pi)}\rho^2\exp(-2i\phi). \end{aligned} \tag{15.3.103}$$

How are these results related to cylindrical harmonic expansions? For the expansion (3.19) consider the special case in which $C_m^0(z) \neq 0$ for only one value of m, and suppose for this value of m that $C_m^{[0]}(z)$ has the special form

$$C_m^{[0]}(z) = a_n^m f_n(z) \tag{15.3.104}$$

with

$$f_n(z) = z^n. \tag{15.3.105}$$

Call the result $\psi_n^m(x, y, z)$. That is, make the Ansatz

$$\psi_n^m(x,y,z) = a_n^m\rho^{|m|}\exp(im\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{|m|!}{2^{2\ell}\ell!(\ell+|m|)!}f_n^{[2\ell]}(z)\rho^{2\ell}. \tag{15.3.106}$$

What are the properties of this Ansatz?

We begin by observing that the combination $f_n^{[2\ell]}(z)\rho^{2\ell+|m|}$ is a monomial in the variables z and ρ for each value of ℓ, with $2\ell \leq n$, and all these monomials are of degree $n + |m|$. For example, there are the relations

$$f_0^{[0]} = 1, \tag{15.3.107}$$

$$f_0^{[2]} = 0; \tag{15.3.108}$$

$$f_1^{[0]} = z, \tag{15.3.109}$$

$$f_1^{[2]} = 0; \tag{15.3.110}$$
$$f_2^{[0]} = z^2, \tag{15.3.111}$$
$$f_2^{[2]} = 2, \tag{15.3.112}$$
$$f_2^{[4]} = 0; \tag{15.3.113}$$
$$f_3^{[0]} = z^3, \tag{15.3.114}$$
$$f_3^{[2]} = 6z, \tag{15.3.115}$$
$$f_3^{[4]} = 0, \text{ etc.} \tag{15.3.116}$$

It follows that there are the results

$$f_0^{[2\ell]}(z)\rho^{2\ell} = 1 \text{ when } \ell = 0; \tag{15.3.117}$$

$$f_1^{[2\ell]}(z)\rho^{2\ell} = z \text{ when } \ell = 0; \tag{15.3.118}$$

$$f_2^{[2\ell]}(z)\rho^{2\ell} = z^2 \text{ when } \ell = 0, \tag{15.3.119}$$
$$f_2^{[2\ell]}(z)\rho^{2\ell} = 2\rho^2 \text{ when } \ell = 1; \tag{15.3.120}$$

$$f_3^{[2\ell]}(z)\rho^{2\ell} = z^3 \text{ when } \ell = 0, \tag{15.3.121}$$
$$f_3^{[2\ell]}(z)\rho^{2\ell} = 6z\rho^2 \text{ when } \ell = 1, \text{ etc.} \tag{15.3.122}$$

Now let us use these results to work out the first few $\psi_n^m(x, y, z)$. So doing gives the relations

$$\psi_0^0(x, y, z) = a_0^0 \propto H_0^0(\boldsymbol{r}); \tag{15.3.123}$$

$$\psi_0^1(x, y, z) = a_0^1 \rho \exp(i\phi) \propto H_1^1(\boldsymbol{r}), \tag{15.3.124}$$
$$\psi_1^0(x, y, z) = a_1^0 z \propto H_1^0(\boldsymbol{r}), \tag{15.3.125}$$
$$\psi_0^{-1}(x, y, z) = a_0^{-1} \rho \exp(-i\phi) \propto H_1^{-1}(\boldsymbol{r}); \tag{15.3.126}$$

$$\psi_0^2(x, y, z) = a_0^2 \rho^2 \exp(2i\phi) \propto H_2^2(\boldsymbol{r}), \tag{15.3.127}$$
$$\psi_1^1(x, y, z) = a_1^1 z\rho \exp(i\phi) \propto H_2^1(\boldsymbol{r}), \tag{15.3.128}$$
$$\psi_2^0(x, y, z) = a_2^0 (z^2 - \rho^2/2) \propto H_2^0(\boldsymbol{r}), \tag{15.3.129}$$
$$\psi_1^{-1}(x, y, z) = a_1^{-1} z\rho \exp(-i\phi) \propto H_2^{-1}(\boldsymbol{r}), \tag{15.3.130}$$
$$\psi_0^{-2}(x, y, z) = a_0^{-2} \rho^2 \exp(-2i\phi) \propto H_2^{-2}(\boldsymbol{r}). \tag{15.3.131}$$

These examples illustrate that the Ansatz specified by (3.105) and (3.106) produces the harmonic polynomials expressed in cylindrical coordinates. There is the general relation

$$\psi_n^m(x, y, z) \propto H_{n+|m|}^m(\boldsymbol{r}) \tag{15.3.132}$$

where both sides of (3.132) are to be expressed in terms of cylindrical coordinates. Moreover, we note that the functions (3.105) provide a basis for the set of functions $C_m^{[0]}(z)$. Therefore, as we already know from other arguments, harmonic polynomials form a basis for the set of harmonic functions.

Exercises

15.3.1. Given (3.65) through (3.67), verify (3.68) through (3.84).

15.3.2. Suppose that some beam line element is described by the magnetic scalar potential $\psi(x, y, z) = \psi(\rho, \phi, z)$ and suppose this element is rotated by angle θ about the z axis. With regard to sign convention, look down the z axis in the direction of increasing z and suppose the rotation is made in the clockwise direction by an angle θ when θ is positive. Let $\hat{\psi}$ be the magnetic scalar potential for this rotated element. Show that there is the relation

$$\hat{\psi}(\rho, \phi, z) = \psi(\rho, \phi - \theta, z). \tag{15.3.133}$$

Suppose that ψ has the expansion given by the first line of (3.40) and that $\hat{\psi}$ has an expansion of the form

$$\hat{\psi}(x, y, z) = \hat{\Psi}_0(\rho, z) + \sum_{m=1}^{\infty} \hat{\Psi}_{m,c}(\rho, z) \cos m\phi + \sum_{m=1}^{\infty} \hat{\Psi}_{m,s}(\rho, z) \sin m\phi. \tag{15.3.134}$$

Show that

$$\hat{\Psi}_0(\rho, z) = \Psi_0(\rho, z), \tag{15.3.135}$$

$$\hat{\Psi}_{m,c}(\rho, z) = \cos(m\theta)\ \Psi_{m,c}(\rho, z) - \sin(m\theta)\ \Psi_{m,s}(\rho, z), \tag{15.3.136}$$

$$\hat{\Psi}_{m,s}(\rho, z) = \sin(m\theta)\ \Psi_{m,c}(\rho, z) + \cos(m\theta)\ \Psi_{m,s}(\rho, z). \tag{15.3.137}$$

With regard to on-axis gradients, suppose the original on-axis gradients are the functions $C^{[n]}_{m,\alpha}(z)$. See (3.33). Suppose that the on-axis gradients for the rotated element are the functions $\hat{C}^{[n]}_{m,\alpha}(z)$. Show that there the relations

$$\hat{C}^{[n]}_0(z) = C^{[n]}_0(z), \tag{15.3.138}$$

$$\hat{C}^{[n]}_{m,c}(z) = \cos(m\theta)\ C^{[n]}_{m,c}(z) - \sin(m\theta)\ C^{[n]}_{m,s}(z), \tag{15.3.139}$$

$$\hat{C}^{[n]}_{m,s}(z) = \sin(m\theta)\ C^{[n]}_{m,c}(z) + \cos(m\theta)\ C^{[n]}_{m,s}(z). \tag{15.3.140}$$

15.3.3. Show that the definitions (3.28) and (3.31) can be inverted to give the relations

$$C^{[0]}_m(z) = (1/2)[C^{[0]}_{m,c}(z) - iC^{[0]}_{m,s}(z)] \text{ for } m \geq 1, \tag{15.3.141}$$

$$C^{[0]}_{-m}(z) = (1/2)[C^{[0]}_{m,c}(z) + iC^{[0]}_{m,s}(z)] \text{ for } m \geq 1. \tag{15.3.142}$$

15.3.4. The relation (3.5) displays the Laplacian in cylindrical coordinates. Verify that one may write

$$\nabla^2 = \nabla^2_\perp + \partial^2/\partial z^2 \tag{15.3.143}$$

where

$$\begin{aligned} \nabla^2_\perp &= \partial^2/\partial x^2 + \partial^2/\partial y^2 = (1/\rho)(\partial/\partial\rho)(\rho\partial/\partial\rho) + (1/\rho^2)(\partial^2/\partial\phi^2) \\ &= \partial^2/\partial\rho^2 + (1/\rho)\partial/\partial\rho + (1/\rho^2)(\partial^2/\partial\phi^2). \end{aligned} \tag{15.3.144}$$

Define functions χ_c and χ_s by the rules

$$\chi_c = \rho^{2\ell+m} \cos m\phi, \tag{15.3.145}$$

$$\chi_s = \rho^{2\ell+m} \sin m\phi. \tag{15.3.146}$$

Show that they have the property

$$\nabla_\perp^2 \chi_\alpha = 4\ell(\ell+m)\chi_\alpha/\rho^2 \tag{15.3.147}$$

where $\alpha = c, s$. Use this property to show that ψ as given by (3.33) satisfies the Laplace equation (2.7).

15.3.5. Review Exercise 3.4. Next, note the resemblance between the functional forms of ψ as given by (3.33) and the Cartesian components of the gradient of ψ as given by (3.65) through (3.67). Looking forward, observe that the Cartesian components of the associated vector potential in a Coulomb gauge, see (5.89) through (5.94), also have analogous functional forms. Why should this be?

15.3.6. Suppose the Fourier coefficient $G_m(k)$ appearing in (3.7) has the form

$$G_m(k) = \lambda \delta_{m,m'} \delta(k)/k^{|m|}. \tag{15.3.148}$$

Show, using (3.11), that in this case

$$\psi(x, y, z) = [\lambda/(2^{|m'|}|m'|!)]\rho^{|m'|} \exp(im'\phi). \tag{15.3.149}$$

Verify, by direct calculation, that ψ as given by (3.149) is harmonic.

15.4 Determination of the Vector Potential: Azimuthal-Free Gauge

Although the description of magnetic fields $\boldsymbol{B}$ in terms of the scalar potential ψ is convenient, it is not what we ultimately need. What we need, if we wish to exploit the symplectic structure of Hamiltonian dynamics, is a description of $\boldsymbol{B}$ in terms of a vector potential $\boldsymbol{A}$ such that

$$\boldsymbol{B} = \nabla \times \boldsymbol{A}. \tag{15.4.1}$$

We recall that in cylindrical coordinates the radial and azimuthal components of a vector, in this case the vector potential $\boldsymbol{A}$, are related to the transverse Cartesian components by equations (2.22) through (2.25). Since there is gauge freedom in the choice of a vector potential, it is sometimes convenient, if possible, to work in a gauge for which the azimuthal component vanishes,

$$A_\phi = 0. \tag{15.4.2}$$

According to (2.24) and (2.25), in this gauge we have the relations

$$A_x = A_\rho \cos\phi, \tag{15.4.3}$$

$$A_y = A_\rho \sin\phi. \tag{15.4.4}$$

We call this gauge the *azimuthal-free* gauge. We will see that it is possible to find a vector potential in the azimuthal-free gauge for the magnetic field of any multipole save for $m = 0$. (In subsequent sections, we will find a vector potential for the $m = 0$ case in a Coulomb gauge.)

15.4.1 Derivation

We will employ the notation (3.38) and (3.39). With this notation in mind, define vector potentials $\boldsymbol{A}^{m,c}$ and $\boldsymbol{A}^{m,s}$ by the rules

$$A_\rho^{m,c} = -\frac{\sin(m\phi)}{m}\rho\frac{\partial}{\partial z}\Psi_{m,c}, \tag{15.4.5}$$

$$A_\phi^{m,c} = 0, \tag{15.4.6}$$

$$A_z^{m,c} = \frac{\sin(m\phi)}{m}\rho\frac{\partial}{\partial \rho}\Psi_{m,c}; \tag{15.4.7}$$

$$A_\rho^{m,s} = \frac{\cos(m\phi)}{m}\rho\frac{\partial}{\partial z}\Psi_{m,s}, \tag{15.4.8}$$

$$A_\phi^{m,s} = 0, \tag{15.4.9}$$

$$A_z^{m,s} = -\frac{\cos(m\phi)}{m}\rho\frac{\partial}{\partial \rho}\Psi_{m,s}. \tag{15.4.10}$$

(Note that these definitions fail for the $m = 0$ case. See Exercise 4.5.) Then, it is easily verified that

$$\nabla \times \boldsymbol{A}^{m,c} = \nabla\psi_{m,c}, \tag{15.4.11}$$

$$\nabla \times \boldsymbol{A}^{m,s} = \nabla\psi_{m,s}. \tag{15.4.12}$$

See Exercise 4.6. Correspondingly, if we define $\boldsymbol{A}$ by the sum

$$\boldsymbol{A} = \sum_{m=1}^{\infty} \boldsymbol{A}^{m,c} + \sum_{m=1}^{\infty} \boldsymbol{A}^{m,s}, \tag{15.4.13}$$

we have, by linearity and again omitting the $m = 0$ term, the result,

$$\nabla \times \boldsymbol{A} = \nabla\psi = \boldsymbol{B}. \tag{15.4.14}$$

At this point we make an important observation. We know that $\boldsymbol{B}$ falls to zero for large $|z|$ because for large $|z|$ the observation point must be well outside the beam-line element in question. From (3.65) through (3.67) and the definitions of $\Psi_{m,c}$ and $\Psi_{m,c}$, we see that these Ψ must also fall to zero for large $|z|$. Correspondingly, from (4.5) through (4.7), we see that $\boldsymbol{A}^{m,c}$ and $\boldsymbol{A}^{m,s}$ must fall to zero for large $|z|$. This behavior is important because it guarantees that, for the azimuthal-free gauge, the canonical and mechanical momenta will be *equal* far outside any beam-line element. See (1.5.30).

We close this subsection by presenting explicit formulas for the cylindrical and Cartesian components of $\boldsymbol{A}^{m,c}$ and $\boldsymbol{A}^{m,s}$ for general $m \geq 1$. From (4.5) through (4.10) and the expansions (3.38) and (3.39) we find the results

$$A_\rho^{m,c} = -\frac{\sin(m\phi)}{m}\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell+1]}(z)\rho^{2\ell+m+1}, \tag{15.4.15}$$

$$A_\phi^{m,c} = 0, \tag{15.4.16}$$

$$A_z^{m,c} = \frac{\sin(m\phi)}{m} \sum_{\ell=0}^{\infty} (-1)^\ell \frac{(2\ell+m)(m!)}{2^{2\ell}\ell!(\ell+m)!} C_{m,c}^{[2\ell]}(z)\rho^{2\ell+m}; \tag{15.4.17}$$

$$A_\rho^{m,s} = \frac{\cos(m\phi)}{m} \sum_{\ell=0}^{\infty} (-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!} C_{m,s}^{[2\ell+1]}(z)\rho^{2\ell+m+1}, \tag{15.4.18}$$

$$A_\phi^{m,s} = 0, \tag{15.4.19}$$

$$A_z^{m,s} = -\frac{\cos(m\phi)}{m} \sum_{\ell=0}^{\infty} (-1)^\ell \frac{(2\ell+m)(m!)}{2^{2\ell}\ell!(\ell+m)!} C_{m,s}^{[2\ell]}(z)\rho^{2\ell+m}. \tag{15.4.20}$$

From (4.3), (4.4), and (4.15) through (4.17) we find the results

$$A_x^{m,c} = -(1/m)x\Im[(x+iy)^m] \sum_{\ell=0}^{\infty} (-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!} C_{m,c}^{[2\ell+1]}(z)(x^2+y^2)^\ell, \tag{15.4.21}$$

$$A_y^{m,c} = -(1/m)y\Im[(x+iy)^m] \sum_{\ell=0}^{\infty} (-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!} C_{m,c}^{[2\ell+1]}(z)(x^2+y^2)^\ell, \tag{15.4.22}$$

$$A_z^{m,c} = -(1/m)\Im[(x+iy)^m] \sum_{\ell=0}^{\infty} (-1)^\ell \frac{(2\ell+m)(m!)}{2^{2\ell}\ell!(\ell+m)!} C_{m,c}^{[2\ell]}(z)(x^2+y^2)^\ell. \tag{15.4.23}$$

From (4.3), (4.4), and (4.18) through (4.20) we find the results

$$A_x^{m,s} = -(1/m)x\Re[(x+iy)^m] \sum_{\ell=0}^{\infty} (-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!} C_{m,s}^{[2\ell+1]}(z)(x^2+y^2)^\ell, \tag{15.4.24}$$

$$A_y^{m,s} = -(1/m)y\Re[(x+iy)^m] \sum_{\ell=0}^{\infty} (-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!} C_{m,s}^{[2\ell+1]}(z)(x^2+y^2)^\ell, \tag{15.4.25}$$

$$A_z^{m,s} = -(1/m)\Re[(x+iy)^m] \sum_{\ell=0}^{\infty} (-1)^\ell \frac{(2\ell+m)(m!)}{2^{2\ell}\ell!(\ell+m)!} C_{m,s}^{[2\ell]}(z)(x^2+y^2)^\ell. \tag{15.4.26}$$

Note that (4.21) through (4.26) provide expansions of the vector potential in terms of homogeneous polynomials in the variables x, y with z-dependent coefficients $C_{m,\alpha}^{[n]}(z)$, and that the *minimum* degree of these polynomials is m. Therefore, if the design orbit is on the z axis, as it will be for all beam-line elements not having $m = 1$ (dipole) content, only a finite number of m and ℓ values are required to to compute the transfer map through some finite order in the Lie generators.

15.4.2 Some Simple Examples: $m = 1, 2$

As a first example of a vector potential in the azimuthal-free gauge, suppose all terms in (3.33) vanish save for the 'pure' dipole terms $C_{1,s}^{[n]}(z)$. Then, using (4.24) through (4.26), we find through terms of degree five that $\boldsymbol{A}^{1,s}$ has the expansion

$$A_x^{1,s} = x^2 C_{1,s}^{[1]}(z) - (1/8)(x^4 + x^2y^2)C_{1,s}^{[3]}(z) + \cdots , \tag{15.4.27}$$

$$A_y^{1,s} = xyC_{1,s}^{[1]}(z) - (1/8)(x^3y + xy^3)C_{1,s}^{[3]}(z) + \cdots , \tag{15.4.28}$$

$$\begin{aligned} A_z^{1,s} &= -xC_{1,s}^{[0]}(z) + (3/8)(x^3 + xy^2)C_{1,s}^{[2]}(z) \\ &- (5/192)(x^5 + 2x^3y^2 + xy^4)C_{1,s}^{[4]}(z) + \cdots . \end{aligned} \tag{15.4.29}$$

Note that the results (4.27) through (4.29) agree with (1.5.77) if we make the identification $B = C_{1,s}^{[0]}$. However, we know that $C_{1,s}^{[0]}(z)$ must depend on z because the on-axis gradients must vanish far outside any magnet. Therefore the functions $C_{1,s}^{[1]}(z)$, $C_{1,s}^{[2]}(z)$, $C_{1,s}^{[3]}(z)$, etc. must be nonzero, at least near the end and fringe-field regions of any (rectangular) dipole magnet. We conclude that, as a consequence of Maxwell's equations, the vector potential must contain terms beyond degree two in the variables x, y. Correspondingly, as already stated earlier, the transfer map for any real dipole must contain nonlinear terms.

As a second example of a vector potential in the azimuthal-free gauge, suppose all terms in (3.33) vanish save for the 'pure' (normal) quadrupole terms $C_{2,s}^{[n]}(z)$. Then, again using (4.24) through (4.26), we find through terms of degree four that $\boldsymbol{A}^{2,s}$ has the expansion

$$A_x^{2,s} = (1/2)(x^3 - xy^2)C_{2,s}^{[1]}(z) + \cdots , \tag{15.4.30}$$

$$A_y^{2,s} = -(1/2)(y^3 - yx^2)C_{2,s}^{[1]}(z) + \cdots , \tag{15.4.31}$$

$$A_z^{2,s} = -(x^2 - y^2)C_{2,s}^{[0]}(z) + (1/6)(x^4 - y^4)C_{2,s}^{[2]}(z) + \cdots . \tag{15.4.32}$$

Note that the results (4.30) through (4.32) agree with (1.5.83) if we make the identification $Q/2 = C_{2,s}^{[0]}$. However, we know that $C_{2,s}^{[0]}(z)$ must depend on z because the on-axis gradients must vanish far outside any magnet. Therefore the functions $C_{2,s}^{[1]}(z)$, $C_{2,s}^{[2]}(z)$, etc. must be nonzero, at least near the end and fringe-field regions of any quadrupole magnet. We conclude again that, as a consequence of Maxwell's equations, the vector potential must contain terms beyond degree two in the variables x, y. Correspondingly, the transfer map for any real quadrupole must contain nonlinear terms. Analogous results hold for the skew case corresponding to nonvanishing $C_{2,c}^{[n]}(z)$.

Exercises

15.4.1. Show that the scalar potential $\psi_{1,c}$ produces a skew dipole magnetic field that is primarily in the x direction. Assuming the magnet has iron pole faces, sketch the pole faces and windings required to produce such a field, and label the pole faces N and S. Also sketch the magnetic field lines and the directions the current must flow in the windings. Compare your results with those of Exercise 1.5.7.

15.4.2. Show that the scalar potential $\psi_{2,c}$ produces a skew quadrupole magnetic field. Assuming the magnet has iron pole faces, sketch the pole faces and windings required to produce such a field, and label the pole faces N and S. Also sketch the magnetic field lines and the directions the current must flow in the windings. Compare your results with those of Exercise 1.5.9.

15.4.3. Verify (3.84).

15.4.4. Consider, as a model, the field of an iron-dominated dipole with very wide (in the x direction) pole faces. See Figures 1.6.1 and 1.6.2. Based on symmetry one might imagine that the field of such a magnet would have no B_x component and, correspondingly, ψ for such a magnet would have no x dependence. Let us make the Ansatz

$$\begin{aligned}\psi(y,z) &= \sum_{n=0}^{\infty}(-1)^n[1/(2n+1)!]y^{2n+1}O^{[2n]}(z)\\ &= yO^{[0]}(z)-(1/6)y^3O^{[2]}(z)+(1/120)y^5O^{[4]}(z)+\cdots \qquad (15.4.33)\end{aligned}$$

where $O^{[0]}(z)$ is, in principle, an arbitrary function, but required in our case to go to zero as $|z|\to\infty$. Show that this ψ is harmonic. Hint: See Appendix H.

Next, show that this ψ will produce a magnetic field $\boldsymbol{B}^{\text{iwd}}$, the field of an *infinite-width dipole*, with components

$$B_x^{\text{iwd}}=0, \qquad (15.4.34)$$

$$\begin{aligned}B_y^{\text{iwd}} &= \sum_{n=0}^{\infty}(-1)^n[1/(2n)!]y^{2n}O^{[2n]}(z)\\ &= O^{[0]}(z)-(1/2)y^2O^{[2]}(z)+(1/24)y^4O^{[4]}(z)+\cdots, \qquad (15.4.35)\end{aligned}$$

$$\begin{aligned}B_z^{\text{iwd}} &= \sum_{n=0}^{\infty}(-1)^n[1/(2n+1)!]y^{2n+1}O^{[2n+1]}(z)\\ &= yO^{[1]}(z)-(1/6)y^3O^{[3]}(z)+(1/120)y^5O^{[5]}(z)+\cdots. \qquad (15.4.36)\end{aligned}$$

Thus, $\boldsymbol{B}^{\text{iwd}}$ is primarily in the y direction, has no x component, but does have a z component in the fringe-field regions where $O^{[n]}(z)\neq 0$ for $n>0$.

How is the expansion (4.33) related to a cylindrical multipole expansion? Note the identities

$$y=\rho\sin\phi, \qquad (15.4.37)$$

$$y^3=\rho^3\sin^3\phi=\rho^3[(3/4)\sin\phi-(1/4)\sin 3\phi], \qquad (15.4.38)$$

$$y^5=\rho^5\sin^5\phi=\rho^5[(10/16)\sin\phi-(5/16)\sin 3\phi+(1/16)\sin 5\phi],\ \text{etc.} \qquad (15.4.39)$$

Show from (3.43) that there the relations

$$\Psi_{1,s}(\rho,z)=O^{[0]}(z)\rho-(1/8)O^{[2]}(z)\rho^3+(1/192)O^{[4]}(z)\rho^5+\cdots, \qquad (15.4.40)$$

$$\Psi_{3,s}(\rho,z)=(1/24)O^{[2]}(z)\rho^3-(1/384)O^{[4]}(z)\rho^5+\cdots, \qquad (15.4.41)$$

$$\Psi_{5,s}(\rho,z) = (1/1920)O^{[4]}(z)\rho^5 + \cdots\ , \text{ etc}, \tag{15.4.42}$$

Using (3.39), show that there are the expansions

$$\Psi_{1,s}(\rho,z) = C^{[0]}_{1,s}(z)\rho - (1/8)C^{[2]}_{1,s}(z)\rho^3 + (1/192)C^{[4]}_{1,s}\rho^5 + \cdots\ , \tag{15.4.43}$$

$$\Psi_{3,s}(\rho,z) = C^{[0]}_{3,s}(z)\rho^3 - (1/16)C^{[2]}_{3,s}(z)\rho^5 + \cdots\ , \tag{15.4.44}$$

$$\Psi_{5,s}(\rho,z) = C^{[0]}_{5,s}(z)\rho^5 + \cdots\ . \tag{15.4.45}$$

Now derive the relations

$$C^{[n]}_{1,s}(z) = O^{[n]}(z), \tag{15.4.46}$$

$$C^{[n]}_{3,s}(z) = (1/24)O^{[n+2]}(z), \tag{15.4.47}$$

$$C^{[n]}_{5,s}(z) = (1/1920)O^{[n+4]}(z), \text{ etc.} \tag{15.4.48}$$

Thus the infinite-width dipole has a major contribution from $\psi_{1,s}(x,y,z)$, and further contributions, in the fringe-field regions, from all the $\psi_{m,s}(x,y,z)$ with $m = 3, 5, \cdots$; and all the relevant $C^{[n]}_{m,s}(z)$ are determined by the $O^{[n+m-1]}(z)$.

What is the azimuthal-free vector potential for this model field? The vector potential for $\psi_{1,s}$ has already been found. It is given by (4.27) through (4.29). Show that an analogous calculation for $\psi_{3,s}$ gives the result

$$A^{3,s}_x = (1/3)x(x^3 - 3xy^2)C^{[1]}_{3,s}(z) + \cdots\ , \tag{15.4.49}$$

$$A^{3,s}_y = (1/3)y(x^3 - 3xy^2)C^{[1]}_{3,s}(z) + \cdots\ , \tag{15.4.50}$$

$$A^{3,s}_z = -(x^3 - 3xy^2)C^{[0]}_{3,s}(z) + (5/48)(x^5 - 2x^3y^2 - 3xy^4)C^{[2]}_{3,s}(z) + \cdots\ ; \tag{15.4.51}$$

and an analogous calculation for $\psi_{5,s}$ gives the result

$$A^{5,s}_x = (1/5)x(x^5 - 10x^3y^2 + 5xy^4)C^{[1]}_{5,s}(z) + \cdots\ , \tag{15.4.52}$$

$$A^{5,s}_y = (1/5)y(x^5 - 10x^3y^2 + 5xy^4)C^{[1]}_{5,s}(z) + \cdots\ , \tag{15.4.53}$$

$$A^{5,s}_z = -(x^5 - 10x^3y^2 + 5xy^4)C^{[0]}_{5,s}(z) + \cdots\ . \tag{15.4.54}$$

Add all these vector potentials together and use (4.46) through (4.48) to show that the azimuthal-free vector potential for the model field is given by the relations

$$A_x = x^2O^{[1]}(z) - (1/18)(2x^4 + 3x^2y^2)O^{[3]}(z) + \cdots\ , \tag{15.4.55}$$

$$A_y = xyO^{[1]}(z) - (1/18)(2x^3y + 3xy^3)O^{[3]}(z) + \cdots\ , \tag{15.4.56}$$

$$\begin{aligned} A_z = &-xO^{[0]}(z) + (1/6)(2x^3 + 3xy^2)O^{[2]}(z) \\ &- (1/360)(8x^5 + 20x^3y^2 + 15xy^4)O^{[4]}(z) + \cdots\ . \end{aligned} \tag{15.4.57}$$

We have learned, as inspection of (4.55) through (4.57) illustrates, that in the azimuthal-free gauge the vector potential associated with even a fairly simple magnetic field, such as that of our infinite-width dipole model, is quite complicated. Perhaps other gauges would give simpler results? Indeed, consider the infinite-width-dipole vector potential $\boldsymbol{A}^{\rm iwd}$ defined by the equations

$$A_x^{\rm iwd} = \sum_{n=1}^{\infty}(-1)^n[1/(2n)!]y^{2n}O^{[2n-1]}(z) = -(1/2)y^2O^{[1]}(z)+(1/24)y^4O^{[3]}(z)+\cdots, \quad (15.4.58)$$

$$A_y^{\rm iwd} = 0, \quad (15.4.59)$$

$$A_z^{\rm iwd} = -xO^{[0]}(z). \quad (15.4.60)$$

Show that use of this much simpler vector potential also yields the field $\boldsymbol{B}^{\rm iwd}$ given by (4.34) through (4.36). The question of other gauges will be explored in Sections 15.5 and 15.6. Also, see Exercise 6.2.

Finally note that, for the vector potentials given either by (4.55) through (4.57) or by (4.58) through (4.60), the primary component is in the z direction with additional components in some transverse direction significant only in the fringe-field regions. This feature is advantageous when using z as the independent variable, see Exercise 1.6.1, because then the z component of the vector potential appears only outside the square root that is ubiquitous in all canonical and relativistic formulations.

15.4.5. Let $\mathcal{C}$ be a circle in some plane of constant z and centered on $x = y = 0$. Show that for any vector potential $\boldsymbol{A}$ in the azimuthal-free gauge there is the result

$$\int_{\mathcal{C}} \boldsymbol{A}\cdot d\boldsymbol{r} = 0. \quad (15.4.61)$$

But, by Stokes' theorem, there is also the result

$$\int_{\mathcal{C}} \boldsymbol{A}\cdot d\boldsymbol{r} = \int_{\mathcal{D}} (\nabla\times\boldsymbol{A})\cdot d\boldsymbol{S} = \int_{\mathcal{D}} \boldsymbol{B}\cdot d\boldsymbol{S} = \int_{\mathcal{D}} B_z dS \quad (15.4.62)$$

where $\mathcal{D}$ is the disc in the constant z plane surrounded by $\mathcal{C}$. For the monopole case ($m = 0$) we know that the magnetic field is predominantly in the z direction when x, y are near zero. Therefore the integral (4.62) cannot vanish. Correspondingly, an $m = 0$ magnetic field cannot be derived from an azimuthal-free vector potential.

15.4.6. The goal of this exercise is to verify (4.11) and (4.12). Begin with the fact that, by construction, $\psi_{m,c}$ and $\psi_{m,s}$ are harmonic. Show, using (3.5), that

$$[(1/\rho)(\partial/\partial\rho)(\rho\partial/\partial\rho) + \partial^2/\partial z^2]\Psi_{m,c} = (m/\rho)^2\Psi_{m,c}, \quad (15.4.63)$$

with an analogous result for $\Psi_{m,s}$. Using these results, verify (4.11) and (4.12) by employing the formulas for $\nabla\times$ and ∇ in cylindrical coordinates.

15.4.7. Assume that (3.77) through (3.83) hold in the *body* of a pure multipole. Let

$$\boldsymbol{B}^m = \nabla\times\boldsymbol{A}^m \quad (15.4.64)$$

with

$$\boldsymbol{A}^m = \boldsymbol{A}^{m,c} + \boldsymbol{A}^{m,s}. \tag{15.4.65}$$

Show from (4.21) through (4.26) that in this case (the azimuthal-free gauge case)

$$A_x^m = A_y^m = 0, \tag{15.4.66}$$

and

$$\begin{aligned} A_z^m &= C_{m,c}^{[0]}\Im[(x+iy)^m] - C_{m,s}^{[0]}\Re[(x+iy)^m] \\ &= -\Re[(C_{m,s}^{[0]} + iC_{m,c}^{[0]})(x+iy)^m] = \Im[(C_{m,c}^{[0]} - iC_{m,s}^{[0]})(x+iy)^m]. \end{aligned} \tag{15.4.67}$$

Here the quantities $C_{m,\alpha}^{[0]}$ are assumed to be *constant* (z independent).

15.5 Determination of the Vector Potential: Symmetric Coulomb Gauge

Sometimes it is convenient to work in a Coulomb gauge rather than the azimuthal-free gauge. In this section we will find such a vector potential which, for reasons that will become apparent, we will call the *symmetric* Coulomb gauge vector potential. Before doing so, there is an important fact to be noted about Coulomb-gauge vector potentials for source-free magnetic fields. Suppose that $\hat{\boldsymbol{A}}$ is a vector potential for $\boldsymbol{B}$ that satisfies the Coulomb gauge condition

$$\nabla \cdot \hat{\boldsymbol{A}} = 0. \tag{15.5.1}$$

We know by construction that

$$\nabla \times (\nabla \times \hat{\boldsymbol{A}}) = \nabla \times \boldsymbol{B} = 0. \tag{15.5.2}$$

(Here we have assumed that $\boldsymbol{B}$ is source free.) But there is also the vector identity

$$\nabla \times (\nabla \times \hat{\boldsymbol{A}}) = \nabla(\nabla \cdot \hat{\boldsymbol{A}}) - \nabla^2 \hat{\boldsymbol{A}}. \tag{15.5.3}$$

It follows from (5.1) through (5.3) that there is the relation

$$\nabla^2 \hat{\boldsymbol{A}} = 0. \tag{15.5.4}$$

That is, each *Cartesian* component of a Coulomb-gauge vector potential is a harmonic function. (It need not be true of other components such as spherical and cylindrical components.) This fact will be useful in the next section.

15.5.1 The $m = 0$ Case

We now turn to the task of computing vector potentials in a Coulomb gauge. We begin with the $m = 0$ case, for which there is no azimuthal-free gauge vector potential. In the $m = 0$ case ψ takes the form

$$\psi_0(x, y, z) = \sum_{\ell=0}^{\infty} (-1)^\ell \frac{1}{2^{2\ell}\ell!\ell!} C_0^{[2\ell]}(z)\rho^{2\ell}. \tag{15.5.5}$$

See (3.33). Define a function $U(\rho, z)$ by the rule

$$U(\rho, z) = (1/2)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!(\ell+1)!}C_0^{[2\ell+1]}(z)\rho^{2\ell}. \tag{15.5.6}$$

Now define a vector potential $\hat{\boldsymbol{A}}^0$ by the Ansatz

$$\hat{A}_x^0 = -yU, \tag{15.5.7}$$

$$\hat{A}_y^0 = xU, \tag{15.5.8}$$

$$\hat{A}_z^0 = 0. \tag{15.5.9}$$

We will soon see that this vector potential produces $\boldsymbol{B}$ and is in a Coulomb gauge. Because the two transverse components of this vector potential involve the same master function U in analogous ways, we will refer to this vector potential as the *symmetric* $m = 0$ Coulomb gauge vector potential.

Let us verify that this vector potential produces $\boldsymbol{B}$. First we have to answer the question

$$(\nabla \times \hat{\boldsymbol{A}}^0)_z = \partial_x \hat{A}_y^0 - \partial_y \hat{A}_x^0 = \partial_z \psi_0? \tag{15.5.10}$$

From (5.8) we find the result

$$\partial_x \hat{A}_y^0 = U + x\partial_x U. \tag{15.5.11}$$

But, by the chain rule we have the result

$$\partial_x U = (\partial U/\partial\rho)(\partial\rho/\partial x) = (x/\rho)(\partial U/\partial\rho). \tag{15.5.12}$$

Here we have used the relation

$$\partial\rho/\partial x = x/\rho. \tag{15.5.13}$$

See Appendix H. From (5.11) and (5.12) we conclude that

$$\partial_x \hat{A}_y^0 = U + (x^2/\rho)(\partial U/\partial\rho). \tag{15.5.14}$$

Similarly, we find that

$$-\partial_y \hat{A}_x^0 = U + (y^2/\rho)(\partial U/\partial\rho), \tag{15.5.15}$$

and therefore

$$\partial_x \hat{A}_y^0 - \partial_y \hat{A}_x^0 = 2U + \rho(\partial U/\partial\rho). \tag{15.5.16}$$

But, from the definition (5.6), we see that

$$\rho(\partial U/\partial\rho) = (1/2)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2\ell}{2^{2\ell}\ell!(\ell+1)!}C_0^{[2\ell+1]}(z)\rho^{2\ell} \tag{15.5.17}$$

and consequently,

$$\begin{aligned}\partial_x \hat{A}_y^0 - \partial_y \hat{A}_x^0 = 2U + \rho(\partial U/\partial\rho) &= (1/2)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2\ell+2}{2^{2\ell}\ell!(\ell+1)!}C_0^{[2\ell+1]}(z)\rho^{2\ell} \\ = \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_0^{[2\ell+1]}(z)\rho^{2\ell} &= \partial_z\psi_0 = B_z.\end{aligned} \tag{15.5.18}$$

Next examine the question

$$(\nabla \times \hat{\boldsymbol{A}}^0)_x = \partial_y \hat{A}_z^0 - \partial_z \hat{A}_y^0 = -\partial_z \hat{A}_y^0 = \partial_x \psi_0? \tag{15.5.19}$$

From (5.8) we see that

$$-\partial_z \hat{A}_y^0 = -x(\partial U/\partial z) = -(x/2)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{1}{2^{2\ell}\ell!(\ell+1)!}C_0^{[2\ell+2]}(z)\rho^{2\ell}. \tag{15.5.20}$$

But we also see from the form (5.5) for ψ_0 that

$$\partial_x \psi_0 = (\partial\psi_0/\partial\rho)(\partial\rho/\partial x) = (x/\rho)(\partial\psi_0/\partial\rho). \tag{15.5.21}$$

Similarly, and for future use, there is the relation

$$\partial_y \psi_0 = (\partial\psi_0/\partial\rho)(\partial\rho/\partial y) = (y/\rho)(\partial\psi_0/\partial\rho). \tag{15.5.22}$$

But from the representation (5.5) we see that

$$\begin{aligned}(1/\rho)(\partial\psi_0/\partial\rho) &= \sum_{\ell=0}^{\infty}(-1)^\ell \frac{2\ell}{2^{2\ell}\ell!\ell!}C_0^{[2\ell]}(z)\rho^{2\ell-2} \\ &= \sum_{\ell=1}^{\infty}(-1)^\ell \frac{2\ell}{2^{2\ell}\ell!\ell!}C_0^{[2\ell]}(z)\rho^{2\ell-2} = \sum_{n=0}^{\infty}(-1)^{n+1}\frac{(2n+2)}{2^{(2n+2)}(n+1)!(n+1)!}C_0^{[2n+2]}(z)\rho^{2n} \\ &= (-1/2)\sum_{n=0}^{\infty}(-1)^n \frac{1}{2^{2n}n!(n+1)!}C_0^{[2n+2]}(z)\rho^{2n},\end{aligned} \tag{15.5.23}$$

and consequently

$$\partial_x \psi_0 = (x/\rho)(\partial\psi_0/\partial\rho) = (-x/2)\sum_{n=0}^{\infty}(-1)^n \frac{1}{2^{2n}n!(n+1)!}C_0^{[2n+2]}(z)\rho^{2n}. \tag{15.5.24}$$

Comparison of (5.20) and (5.24) shows that (5.19) is satisfied.

The last question to examine is

$$(\nabla \times \hat{\boldsymbol{A}}^0)_y = \partial_z \hat{A}_x^0 - \partial_x \hat{A}_z^0 = \partial_z \hat{A}_x^0 = \partial_y \psi_0? \tag{15.5.25}$$

From (5.7) we see that

$$\partial_z \hat{A}_x^0 = -y(\partial U/\partial z) = -(y/2)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{1}{2^{2\ell}\ell!(\ell+1)!}C_0^{[2\ell+2]}(z)\rho^{2\ell}. \tag{15.5.26}$$

But from (5.22) and (5.23) we have

$$\partial_y \psi_0 = (y/\rho)(\partial\psi_0/\partial\rho) = (-y/2)\sum_{n=0}^{\infty}(-1)^n \frac{1}{2^{2n}n!(n+1)!}C_0^{[2n+2]}(z)\rho^{2n}. \tag{15.5.27}$$

Comparison of (5.26) and (5.27) shows that (5.25) is satisfied.

We can also check that $\hat{\boldsymbol{A}}^0$ is divergence free. From (5.7) through (5.9) we see that

$$\nabla \cdot \hat{\boldsymbol{A}}^0 = \partial_x \hat{A}^0_x + \partial_y \hat{A}^0_y + \partial_z \hat{A}^0_z = -y\partial_x U + x\partial_y U. \tag{15.5.28}$$

Now use (5.12) and its y analog to find the result

$$-y\partial_x U + x\partial_y U = [(-yx/\rho) + (xy/\rho)](\partial U/\partial\rho) = 0. \tag{15.5.29}$$

From (5.6) through (5.8), we see that $\hat{A}^0_x$ and $\hat{A}^0_y$ can be written as

$$\hat{A}^0_x = -\sin(\phi)(1/2)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!(\ell+1)!}C_0^{[2\ell+1]}(z)\rho^{2\ell+1} \tag{15.5.30}$$

and

$$\hat{A}^0_y = \cos(\phi)(1/2)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!(\ell+1)!}C_0^{[2\ell+1]}(z)\rho^{2\ell+1}. \tag{15.5.31}$$

Comparison of these expressions with (3.33) shows that both $\hat{A}^0_x$ and $\hat{A}^0_y$ are harmonic functions, as expected.

Inserting (2.13) and (2.14) into (5.30) and (5.31) gives the even more explicit results

$$\begin{aligned}\hat{A}^0_x &= -(y/2)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!(\ell+1)!}C_0^{[2\ell+1]}(z)(x^2+y^2)^{\ell}\\ &= -(y/2)[C_0^{[1]} - (1/8)C_0^{[3]}(x^2+y^2) + \cdots],\end{aligned} \tag{15.5.32}$$

$$\begin{aligned}\hat{A}^0_y &= (x/2)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!(\ell+1)!}C_0^{[2\ell+1]}(z)(x^2+y^2)^{\ell}\\ &= (x/2)[C_0^{[1]} - (1/8)C_0^{[3]}(x^2+y^2) + \cdots],\end{aligned} \tag{15.5.33}$$

$$\hat{A}^0_z = 0. \tag{15.5.34}$$

From the relation

$$B_z = \partial_z\psi_0 = \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_0^{[2\ell+1]}(z)\rho^{2\ell}, \tag{15.5.35}$$

we see that

$$C_0^{[1]}(z) = B_z(0,0,z). \tag{15.5.36}$$

Since we assume that $B_z(0,0,z)$ falls of for large $|z|$, the same will be true for $C_0^{[1]}(z)$ and for the $C_0^{[2\ell+1]}(z)$, and hence, according to (5.32) through (5.34), also for $\hat{\boldsymbol{A}}^0$.

For future work it is also useful to have expressions for $\hat{\boldsymbol{A}}^0$ in cylindrical components. Using (2.22), (2.23), and (5.7) through (5.9) gives the results

$$\hat{A}^0_\rho = 0, \tag{15.5.37}$$

$$\hat{A}^0_\phi = \rho U(\rho, z) = (1/2)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{1}{2^{2\ell}\ell!(\ell+1)!}C_0^{[2\ell+1]}(z)\rho^{2\ell+1}, \tag{15.5.38}$$

$$\hat{A}^0_z = 0. \tag{15.5.39}$$

Note that, in contrast to the azimuthal-free gauge of Section 15.4, this vector potential has *only* an azimuthal component.

15.5.2 The $m \geq 1$ Cases

Derivation

Let us begin with some notation: Since we will often be dealing with both the skew and normal cases simultaneously, we will use the symbol α to denote either c or s. For example, we will use $\psi_{m,\alpha}$ to denote either $\psi_{m,c}$ or $\psi_{m,s}$. With this convention in mind, the purpose of the present subsection is to find vector potentials $\hat{\boldsymbol{A}}^{m,\alpha}$ that are in a Coulomb gauge,

$$\nabla \cdot \hat{\boldsymbol{A}}^{m,\alpha} = 0, \tag{15.5.40}$$

and also produce the $\boldsymbol{B}$ fields associated with the $\psi_{m,\alpha}$,

$$\nabla \times \hat{\boldsymbol{A}}^{m,\alpha} = \nabla\psi_{m,\alpha}. \tag{15.5.41}$$

The requirements (5.41), when combined with the relations (4.11) and (4.12), yield the conditions

$$\nabla \times (\hat{\boldsymbol{A}}^{m,\alpha} - \boldsymbol{A}^{m,\alpha}) = 0, \tag{15.5.42}$$

from which it follows that there are functions $\chi_{m,\alpha}$ such that

$$\hat{\boldsymbol{A}}^{m,\alpha} = \boldsymbol{A}^{m,\alpha} + \nabla\chi_{m,\alpha}. \tag{15.5.43}$$

Of course, the relations (5.43) are simply gauge transformations. Our strategy will be to find the functions $\chi_{m,\alpha}$ and then use (5.43) to yield the desired vector potentials.

Upon taking the divergence of both sides of (5.43), and using the Coulomb conditions (5.40), we find that the $\chi_{m,\alpha}$ must satisfy the equations

$$\nabla^2\chi_{m,\alpha} = -\nabla \cdot \boldsymbol{A}^{m,\alpha}. \tag{15.5.44}$$

For an azimuthal-free $\boldsymbol{A}$, see (4.2), we have the relation

$$\nabla \cdot \boldsymbol{A} = (1/\rho)(\partial/\partial\rho)(\rho A_\rho) + (\partial/\partial z)A_z. \tag{15.5.45}$$

Using the representations (4.5) through (4.10) for the right sides of (5.44) gives the results

$$\nabla \cdot \boldsymbol{A}^{m,\alpha} = (2/\rho)A^{m,\alpha}_\rho. \tag{15.5.46}$$

Consequently, the $\chi_{m,\alpha}$ must satisfy the relations

$$\nabla^2\chi_{m,\alpha} = -(2/\rho)A^{m,\alpha}_\rho. \tag{15.5.47}$$

To find the $\chi_{m,\alpha}$, let us make the Ansätze

$$\chi_{m,c} = -(\sin m\phi) d_{m,c}(\rho, z), \tag{15.5.48}$$

$$\chi_{m,s} = (\cos m\phi) d_{m,s}(\rho, z), \tag{15.5.49}$$

where the functions $d_{m,\alpha}(\rho, z)$ are yet to be determined. From the representations (5.48) and (5.49) we find the relations

$$\nabla^2 \chi_{m,c} = -(\sin m\phi)[(1/\rho)(\partial/\partial\rho)(\rho\partial/\partial\rho) - m^2/\rho^2 + (\partial/\partial z)^2] d_{m,s}(\rho, z), \tag{15.5.50}$$

$$\nabla^2 \chi_{m,s} = (\cos m\phi)[(1/\rho)(\partial/\partial\rho)(\rho\partial/\partial\rho) - m^2/\rho^2 + (\partial/\partial z)^2] d_{m,s}(\rho, z). \tag{15.5.51}$$

See (3.5). Upon using (5.47) and comparing (5.50) and (5.51) with (4.5) and (4.8), we find the relations

$$[(1/\rho)(\partial/\partial\rho)(\rho\partial/\partial\rho) - m^2/\rho^2 + (\partial/\partial z)^2] d_{m,\alpha}(\rho, z) = (-2/m)(\partial/\partial z)\Psi_{m,\alpha}(\rho, z). \tag{15.5.52}$$

Next assume that each $d_{m,\alpha}(\rho, z)$ has an expansion of the form

$$d_{m,\alpha}(\rho, z) = \sum_{\ell=0}^{\infty} D_{m,\alpha}^{2\ell}(z)\rho^{2\ell+m+2} \tag{15.5.53}$$

where the functions $D_{m,\alpha}^{2\ell}(z)$ are yet to be determined. It easily verified that there is the relation

$$[(1/\rho)(\partial/\partial\rho)(\rho\partial/\partial\rho) - m^2/\rho^2]\rho^n = (n^2 - m^2)\rho^{n-2}. \tag{15.5.54}$$

It follows, by using the expansion (5.53), that there is the relation

$$\begin{aligned} &[(1/\rho)(\partial/\partial\rho)(\rho\partial/\partial\rho) - m^2/\rho^2 + (\partial/\partial z)^2] d_{m,\alpha}(\rho, z) \\ = \ &\sum_{\ell=0}^{\infty}[(2\ell + m + 2)^2 - m^2] D_{m,\alpha}^{2\ell}(z)\rho^{2\ell+m} + \sum_{\ell=0}^{\infty}[(\partial/\partial z)^2 D_{m,\alpha}^{2\ell}(z)]\rho^{2\ell+m+2}. \end{aligned} \tag{15.5.55}$$

The sum consisting of the second set of terms on the right side of (5.55) can be rewritten in the form

$$\sum_{\ell=0}^{\infty}[(\partial/\partial z)^2 D_{m,\alpha}^{2\ell}(z)]\rho^{2\ell+m+2} = \sum_{n=1}^{\infty}[(\partial/\partial z)^2 D_{m,\alpha}^{2n-2}(z)]\rho^{2n+m} \tag{15.5.56}$$

or, equivalently, in the form

$$\sum_{\ell=0}^{\infty}[(\partial/\partial z)^2 D_{m,\alpha}^{2\ell}(z)]\rho^{2\ell+m+2} = \sum_{\ell=0}^{\infty}[(\partial/\partial z)^2 D_{m,\alpha}^{2\ell-2}(z)]\rho^{2\ell+m} \tag{15.5.57}$$

with the understanding that

$$D_{m,\alpha}^{-2} = 0. \tag{15.5.58}$$

Consequently, we also have the relation

$$\begin{aligned}&[(1/\rho)(\partial/\partial\rho)(\rho\partial/\partial\rho) - m^2/\rho^2 + (\partial/\partial z)^2]d_{m,\alpha}(\rho,z)\\ = &\sum_{\ell=0}^{\infty}\{[(2\ell+m+2)^2 - m^2]D^{2\ell}_{m,\alpha}(z) + [(\partial/\partial z)^2 D^{2\ell-2}_{m,\alpha}(z)]\}\rho^{2\ell+m}.\end{aligned} \tag{15.5.59}$$

We have found an expansion in powers of ρ for the left side of (5.52). From (3.38) and (3.39) we already have such an expansion for the right side of (5.52), which we write in the form

$$(-2/m)(\partial/\partial z)\Psi_{m,\alpha}(\rho,z) = \sum_{\ell=0}^{\infty} r(\ell,m)C^{[2\ell+1]}_{m,\alpha}(z)\rho^{2\ell+m} \tag{15.5.60}$$

where

$$r(\ell,m) = -2(-1)^{\ell}(m!)/[m2^{2\ell}\ell!(l+m)!]. \tag{15.5.61}$$

Equating like powers of ρ on both sides of (5.52) gives the relation

$$[(2\ell+m+2)^2 - m^2]D^{2\ell}_{m,\alpha}(z) + (\partial/\partial z)^2 D^{2\ell-2}_{m,\alpha}(z) = r(\ell,m)C^{[2\ell+1]}_{m,\alpha}(z), \tag{15.5.62}$$

which can be rewritten as the recursion relation

$$D^{2\ell}_{m,\alpha}(z) = s(\ell,m)C^{[2\ell+1]}_{m,\alpha}(z) + t(\ell,m)(\partial/\partial z)^2 D^{2\ell-2}_{m,\alpha}(z) \tag{15.5.63}$$

where $s(\ell,m)$ and $t(\ell,m)$ are the coefficients

$$s(\ell,m) = r(\ell,m)/[(2\ell+m+2)^2 - m^2], \tag{15.5.64}$$

$$t(\ell,m) = -1/[(2\ell+m+2)^2 - m^2]. \tag{15.5.65}$$

We find, for the first few terms, the results

$$D^0_{m,\alpha}(z) = s(0,m)C^{[1]}_{m,\alpha}(z) = -\{1/[2m(m+1)]\}C^{[1]}_{m,\alpha}(z), \tag{15.5.66}$$

$$\begin{aligned}D^2_{m,\alpha}(z) &= s(1,m)C^{[3]}_{m,\alpha}(z) + t(1,m)(\partial/\partial z)^2 D^0_{m,\alpha}(z)\\ &= s(1,m)C^{[3]}_{m,\alpha}(z) + t(1,m)s(0,m)C^{[3]}_{m,\alpha}(z)\\ &= [s(1,m) + t(1,m)s(0,m)]C^{[3]}_{m,\alpha}(z)\\ &= \{1/[8m(m+1)(m+2)]\}C^{[3]}_{m,\alpha}(z),\end{aligned} \tag{15.5.67}$$

$$\begin{aligned}D^4_{m,\alpha}(z) &= s(2,m)C^{[5]}_{m,\alpha}(z) + t(2,m)(\partial/\partial z)^2 D^2_{m,\alpha}(z)\\ &= \{s(2,m) + t(2,m)[s(1,m) + t(1,m)s(0,m)]\}C^{[5]}_{m,\alpha}(z),\\ &= -\{1/[64m(m+1)(m+2)(m+3)]\}C^{[5]}_{m,\alpha}(z), \text{ etc.}\end{aligned} \tag{15.5.68}$$

We conclude that the $D^{2\ell}_{m,\alpha}(z)$ are completely specified in terms of the $C^{[2\ell+1]}_{m,\alpha}(z)$. Indeed, we have by induction the relation

$$D^{2\ell}_{m,\alpha}(z) = u(\ell, m)C^{[2\ell+1]}_{m,\alpha}(z) \tag{15.5.69}$$

where the coefficients $u(\ell, m)$ are given by the recursion relation

$$u(\ell, m) = s(\ell, m) + t(\ell, m)u(l - 1, m) \tag{15.5.70}$$

with

$$u(-1, m) = 0. \tag{15.5.71}$$

This recursion relation has the solution

$$u(\ell, m) = -(-1)^\ell c_\ell\{[(m - 1)!]/[(m + \ell + 1)!]\} \tag{15.5.72}$$

where

$$c_\ell = 1/[(2)(2^{2\ell})(\ell!)]. \tag{15.5.73}$$

We are now able to write explicit series expansions for the $d_{m,\alpha}(\rho, z)$ and the $\chi_{m,\alpha}$ in terms of the $C^{[n]}_{m,\alpha}(z)$. Upon combining (5.53), (5.69), (5.72), and (5.73), we find the result

$$d_{m,\alpha}(\rho, z) = -(1/2)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{(m-1)!}{2^{2\ell}\ell!(\ell+m+1)!}C^{[2\ell+1]}_{m,\alpha}(z)\rho^{2\ell+m+2}. \tag{15.5.74}$$

And, with the use of (5.48) and (5.49), we find for the $\chi_{m,\alpha}$ the expansions

$$\chi_{m,c} = (1/2)(\sin m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{(m-1)!}{2^{2\ell}\ell!(\ell+m+1)!}C^{[2\ell+1]}_{m,c}(z)\rho^{2\ell+m+2}, \tag{15.5.75}$$

$$\chi_{m,s} = -(1/2)(\cos m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{(m-1)!}{2^{2\ell}\ell!(\ell+m+1)!}C^{[2\ell+1]}_{m,s}(z)\rho^{2\ell+m+2}. \tag{15.5.76}$$

Next, we can compute series expansions for the $\nabla\chi_{m,\alpha}$. It is convenient to work out the gradients in cylindrical coordinates. We find from (5.75) and (5.76) the results

$$\begin{aligned}(\nabla\chi_{m,c})_\rho &= \partial_\rho\chi_{m,c}\\ &= (1/2)(\sin m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{(m-1)!(2\ell+m+2)}{2^{2\ell}\ell!(\ell+m+1)!}C^{[2\ell+1]}_{m,c}(z)\rho^{2\ell+m+1},\end{aligned} \tag{15.5.77}$$

$$\begin{aligned}(\nabla\chi_{m,c})_\phi &= (1/\rho)\partial_\phi\chi_{m,c}\\ &= (1/2)(\cos m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{(m)(m-1)!}{2^{2\ell}\ell!(\ell+m+1)!}C^{[2\ell+1]}_{m,c}(z)\rho^{2\ell+m+1}\\ &= (1/2)(\cos m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C^{[2\ell+1]}_{m,c}(z)\rho^{2\ell+m+1},\end{aligned} \tag{15.5.78}$$

$$\begin{aligned}(\nabla\chi_{m,c})_z &= \partial_z\chi_{m,c}\\ &= (1/2)(\sin m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{(m-1)!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+2]}(z)\rho^{2\ell+m+2};\end{aligned} \tag{15.5.79}$$

$$\begin{aligned}(\nabla\chi_{m,s})_\rho &= \partial_\rho\chi_{m,s}\\ &= -(1/2)(\cos m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{(m-1)!(2\ell+m+2)}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+1]}(z)\rho^{2\ell+m+1},\end{aligned} \tag{15.5.80}$$

$$\begin{aligned}(\nabla\chi_{m,s})_\phi &= (1/\rho)\partial_\phi\chi_{m,s}\\ &= (1/2)(\sin m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{(m)(m-1)!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+1]}(z)\rho^{2\ell+m+1}\\ &= (1/2)(\sin m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+1]}(z)\rho^{2\ell+m+1},\end{aligned} \tag{15.5.81}$$

$$\begin{aligned}(\nabla\chi_{m,s})_z &= \partial_z\chi_{m,s}\\ &= -(1/2)(\cos m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{(m-1)!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+2]}(z)\rho^{2\ell+m+2}.\end{aligned} \tag{15.5.82}$$

We have all the ingredients at hand to compute the $\hat{\boldsymbol{A}}^{m,\alpha}$. We find, in cylindrical coordinates and using (4.15) through (4.20), (5.43), and (5.77) through (5.82), the results

$$\hat{A}_\rho^{m,c} = -(1/2)(\sin m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+1]}(z)\rho^{2\ell+m+1}, \tag{15.5.83}$$

$$\hat{A}_\phi^{m,c} = (1/2)(\cos m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+1]}(z)\rho^{2\ell+m+1}, \tag{15.5.84}$$

$$\hat{A}_z^{m,c} = (\sin m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell+m}; \tag{15.5.85}$$

$$\hat{A}_\rho^{m,s} = (1/2)(\cos m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+1]}(z)\rho^{2\ell+m+1}, \tag{15.5.86}$$

$$\hat{A}_\phi^{m,s} = (1/2)(\sin m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+1]}(z)\rho^{2\ell+m+1}, \tag{15.5.87}$$

$$\hat{A}_z^{m,s} = -(\cos m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell+m}. \tag{15.5.88}$$

Correspondingly, using (2.24) and (2.25), we find for the Cartesian components of the $\hat{\boldsymbol{A}}^{m,\alpha}$ the results

$$\begin{aligned}\hat{A}_x^{m,c} &= -(1/2)[(\cos\phi)(\sin m\phi)+(\sin\phi)(\cos m\phi)]\times \\ &\quad \sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+1]}(z)\rho^{2\ell+m+1} \\ &= -(1/2)[\sin(m+1)\phi]\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+1]}(z)\rho^{2\ell+m+1}, \\ &= -(1/2)\Im[(x+iy)^{m+1}]\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+1]}(z)(x^2+y^2)^\ell,\end{aligned} \tag{15.5.89}$$

$$\begin{aligned}\hat{A}_y^{m,c} &= (1/2)[-(\sin\phi)(\sin m\phi)+(\cos\phi)(\cos m\phi)]\times \\ &\quad \sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+1]}(z)\rho^{2\ell+m+1} \\ &= (1/2)[\cos(m+1)\phi]\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+1]}(z)\rho^{2\ell+m+1}, \\ &= (1/2)\Re[(x+iy)^{m+1}]\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+1]}(z)(x^2+y^2)^\ell,\end{aligned} \tag{15.5.90}$$

$$\begin{aligned}\hat{A}_z^{m,c} &= (\sin m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell+m} \\ &= \Im[(x+iy)^m]\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)(x^2+y^2)^\ell;\end{aligned} \tag{15.5.91}$$

$$\begin{aligned}
\hat{A}_x^{m,s} &= (1/2)[(\cos\phi)(\cos m\phi) - (\sin\phi)(\sin m\phi)] \times \\
&\quad \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+1]}(z)\rho^{2\ell+m+1} \\
&= (1/2)[\cos(m+1)\phi]\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+1]}(z)\rho^{2\ell+m+1} \\
&= (1/2)\Re[(x+iy)^{m+1}]\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+1]}(z)(x^2+y^2)^{\ell},
\end{aligned} \tag{15.5.92}$$

$$\begin{aligned}
\hat{A}_y^{m,s} &= (1/2)[(\sin\phi)(\cos m\phi) + (\cos\phi)(\sin m\phi)] \times \\
&\quad \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+1]}(z)\rho^{2\ell+m+1} \\
&= (1/2)[\sin(m+1)\phi]\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+1]}(z)\rho^{2\ell+m+1} \\
&= (1/2)\Im[(x+iy)^{m+1}]\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+1]}(z)(x^2+y^2)^{\ell},
\end{aligned} \tag{15.5.93}$$

$$\begin{aligned}
\hat{A}_z^{m,s} &= -(\cos m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell+m} \\
&= -\Re[(x+iy)^{m}]\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)(x^2+y^2)^{\ell}.
\end{aligned} \tag{15.5.94}$$

Note four important facts: First, we see that the relations (5.83) through (5.88) for $\hat{\boldsymbol{A}}^{m,\alpha}$ are also defined for $m = 0$. When so evaluated they produce, in the case $\alpha = c$, a result that agrees with the $\hat{\boldsymbol{A}}^{0}$ given by (5.37) through (5.39); and [recalling (3.36)] they produce, in the case $\alpha = s$, the zero vector. That is, we may make the definitions

$$\hat{\boldsymbol{A}}^{0,c} = \hat{\boldsymbol{A}}^{0}, \tag{15.5.95}$$

$$\hat{\boldsymbol{A}}^{0,s} = 0. \tag{15.5.96}$$

Second, from their form in (5.89) through (5.94), it is evident that the Cartesian components of $\hat{\boldsymbol{A}}^{m,\alpha}$ are all harmonic functions, as expected. Third, we observe that $\hat{A}_x^{m,c}$ and $\hat{A}_y^{m,c}$ as given by (5.89) and (5.90) involve complementary trigonometric functions multiplying the same master function. The same is true of $\hat{A}_x^{m,s}$ and $\hat{A}_y^{m,s}$ as given by (5.92) and (5.93). See Exercise 5.6. For this reason we refer to the Coulomb gauge vector potential we have found as being the *symmetric* Coulomb gauge vector potential. Finally, note again that (5.89) through

(5.94) provide expansions of the vector potential in terms of homogeneous polynomials in the variables x, y with z-dependent coefficients $C^{[n]}_{m,\alpha}(z)$, and that the *minimum* degree of these polynomials is m. In summary, we have found formulas for the $\chi_{m,\alpha}$ and the $\hat{\boldsymbol{A}}^{m,\alpha}$ in terms of the $C^{[n]}_{m,\alpha}(z)$.

Symmetric Coulomb Gauge Examples for $m = 1, 2$

As an example of the use of these relations, let us compute $\hat{\boldsymbol{A}}^{1,s}$ for the dipole case $m = 1$. As before, suppose all terms in (3.33) vanish save for the dipole terms $C^{[n]}_{1,s}(z)$. Using (5.92) through (5.94) we then find, through terms of degree five, that $\hat{\boldsymbol{A}}^{1,s}$ has the expansion

$$\hat{A}^{1,s}_x = (1/4)(x^2 - y^2)C^{[1]}_{1,s}(z) - (1/48)(x^4 - y^4)C^{[3]}_{1,s}(z) + \cdots , \quad (15.5.97)$$

$$\hat{A}^{1,s}_y = (1/2)xyC^{[1]}_{1,s}(z) - (1/24)(x^3y + xy^3)C^{[3]}_{1,s}(z) + \cdots , \quad (15.5.98)$$

$$\hat{A}^{1,s}_z = -xC^{[0]}_{1,s}(z)+(1/8)(x^3+xy^2)C^{[2]}_{1,s}(z)-(1/192)(x^5+2x^3y^2+xy^4)C^{[4]}_{1,s}(z)+\cdots . \quad (15.5.99)$$

This expansion should be compared with the azimuthal-free gauge expansion given by (4.27) through (4.29). Direct calculation verifies that $\hat{\boldsymbol{A}}^{1,s}$ satisfies (5.1) and (5.4) through terms of degree four, which is what is expected based on the order of the terms that have been retained in the expansion.

As a second example of the use of these relations, let us compute $\hat{\boldsymbol{A}}^{2,s}$ for the (normal) quadrupole case $m = 2$. As before, suppose all terms in (3.33) vanish save for the quadrupole terms $C^{[n]}_{2,s}(z)$. Then, again using (5.92) through (5.94), we find, through terms of degree four, that $\hat{\boldsymbol{A}}^{2,s}$ has the expansion

$$\hat{A}^{2,s}_x = (1/6)(x^3 - 3xy^2)C^{[1]}_{2,s}(z) + \cdots , \quad (15.5.100)$$

$$\hat{A}^{2,s}_y = -(1/6)(y^3 - 3x^2y)C^{[1]}_{2,s}(z) + \cdots , \quad (15.5.101)$$

$$\hat{A}^{2,s}_z = -(x^2 - y^2)C^{[0]}_{2,s}(z) + (1/12)(x^4 - y^4)C^{[2]}_{2,s}(z) + \cdots . \quad (15.5.102)$$

This expansion should be compared with the azimuthal-free gauge expansion given by (4.30) through (4.32). Direct calculation again verifies that (5.1) and (5.4) are satisfied by $\hat{\boldsymbol{A}}^{2,s}$ through the order of the terms that have been retained in the expansion. Finally, we remark that analogous results can be found for the skew case $\hat{\boldsymbol{A}}^{2,c}$.

Exercises

15.5.1. For $\hat{\boldsymbol{A}}^{0}$ given by (5.37) through (5.39), compute the curl and divergence of $\hat{\boldsymbol{A}}^{0}$ in cylindrical coordinates.

15.5.2. Verify that (5.72) and (5.73) satisfy the recursion relation (5.70) with the initial condition (5.71).

15.5.3. Verify the expansions (5.83) through (5.88) and verify that $\nabla \cdot \hat{\boldsymbol{A}}^{m,\alpha} = 0$

15.5.4. Consider the case of straight beam-line elements, such as solenoids, quadrupoles, sextupoles, octupoles, etc., for which the design orbit lies on the z axis. Suppose we wish to retain, in the expansion of the Hamiltonian H appearing in (1.3), homogeneous polynomials in x and y through degree 4. This would be required if we wished to make a Lie factorization of $\mathcal{M}$ that retained all Lie generators of degree 4 and lower,

$$\mathcal{M} = \mathcal{R}\exp(:f_3:)\exp(:f_4:). \tag{15.5.103}$$

Assuming no particular field symmetries, and working in the Coulomb gauge of this section, show that the following generalized gradients and their derivatives would then be required:

$$\begin{aligned}
&C_0^{[0]}(z),\ C_0^{[1]}(z),\ C_0^{[2]}(z),\ C_0^{[3]}(z);\\
&C_{1,\alpha}^{[0]}(z),\ C_{1,\alpha}^{[1]}(z),\ C_{1,\alpha}^{[2]}(z);\\
&C_{2,\alpha}^{[0]}(z),\ C_{2,\alpha}^{[1]}(z),\ C_{2,\alpha}^{[2]}(z);\\
&C_{3,\alpha}^{[0]}(z);\\
&C_{4,\alpha}^{[0]}(z).
\end{aligned} \tag{15.5.104}$$

Verify that in the $m = 0$ case the $C_m^{[n]}$ with n even are actually not needed. See Subsection 5.1. Also, strictly speaking, the dipole terms, the terms in the second row of (5.104), should actually vanish in order for the design orbit to lie on the z axis. A possible exception could be the case of a wiggler/undulator where the $C_{1,\alpha}^{[n]}(z)$ oscillate in z and nearly average to zero in such a way that the design orbit does not depart significantly from the z axis.

Suppose, instead, we wish to retain homogeneous polynomials through degree 8. This would be required if we wished to make a Lie factorization of $\mathcal{M}$ that retained all Lie generators of degree 8 and lower,

$$\mathcal{M} = \mathcal{R}\exp(:f_3:)\exp(:f_4:)\exp(:f_5:)\exp(:f_6:)\exp(:f_7:)\exp(:f_8:). \tag{15.5.105}$$

Assuming no particular field symmetries, and working in the Coulomb gauge of this section, show that the following generalized gradients and their derivatives would then be required:

$$\begin{aligned}
&C_0^{[0]}(z),\ C_0^{[1]}(z),\ C_0^{[2]}(z),\ C_0^{[3]}(z),\ C_0^{[4]}(z),\ C_0^{[5]}(z),\ C_0^{[6]}(z),\ C_0^{[7]}(z);\\
&C_{1,\alpha}^{[0]}(z),\ C_{1,\alpha}^{[1]}(z),\ C_{1,\alpha}^{[2]}(z),\ C_{1,\alpha}^{[3]}(z),\ C_{1,\alpha}^{[4]}(z),\ C_{1,\alpha}^{[5]}(z),\ C_{1,\alpha}^{[6]}(z);\\
&C_{2,\alpha}^{[0]}(z),\ C_{2,\alpha}^{[1]}(z),\ C_{2,\alpha}^{[2]}(z),\ C_{2,\alpha}^{[3]}(z),\ C_{2,\alpha}^{[4]}(z),\ C_{2,\alpha}^{[5]}(z),\ C_{2,\alpha}^{[6]}(z);\\
&C_{3,\alpha}^{[0]}(z),\ C_{3,\alpha}^{[1]}(z),\ C_{3,\alpha}^{[2]}(z),\ C_{3,\alpha}^{[3]}(z),\ C_{3,\alpha}^{[4]}(z);\\
&C_{4,\alpha}^{[0]}(z),\ C_{4,\alpha}^{[1]}(z),\ C_{4,\alpha}^{[2]}(z),\ C_{4,\alpha}^{[3]}(z),\ C_{4,\alpha}^{[4]}(z);\\
&C_{5,\alpha}^{[0]}(z),\ C_{5,\alpha}^{[1]}(z),\ C_{5,\alpha}^{[2]}(z);\\
&C_{6,\alpha}^{[0]}(z),\ C_{6,\alpha}^{[1]}(z),\ C_{6,\alpha}^{[2]}(z);\\
&C_{7,\alpha}^{[0]}(z);\\
&C_{8,\alpha}^{[0]}(z).
\end{aligned} \tag{15.5.106}$$

Again verify that in the $m = 0$ case the $C_m^{[n]}$ with n even are actually not needed. And again, with the possible exception of a wiggler/undulator, the dipole terms, the terms in the second row of (5.106), should actually vanish in order for the design orbit to lie on the z axis.

15.5.5. Assume that (3.77) through (3.83) hold in the *body* of a pure multipole (with $m > 0$). Let

$$\boldsymbol{B}^m = \nabla \times \hat{\boldsymbol{A}}^m \tag{15.5.107}$$

with

$$\hat{\boldsymbol{A}}^m = \hat{\boldsymbol{A}}^{m,c} + \hat{\boldsymbol{A}}^{m,s}. \tag{15.5.108}$$

Show from (5.89) through (5.94) that in this case (the symmetric Coulomb gauge case) there are the relations

$$\hat{A}_x^m = \hat{A}_y^m = 0, \tag{15.5.109}$$

and

$$\begin{aligned} \hat{A}_z^m &= C_{m,c}^{[0]}\Im[(x+iy)^m] - C_{m,s}^{[0]}\Re[(x+iy)^m] \\ &= -\Re[(C_{m,s}^{[0]} + iC_{m,c}^{[0]})(x+iy)^m] = \Im[(C_{m,c}^{[0]} - iC_{m,s}^{[0]})(x+iy)^m]. \end{aligned} \tag{15.5.110}$$

Here the quantities $C_{m,\alpha}^{[0]}$ are assumed to be *constant* (z independent).

Note that the results (5.109) and (5.110) agree with (4.66) and (4.67). The azimuthal-free and symmetric Coulomb gauges give the *same* result in the body for all terms with $m > 0$. The difference between the two gauges occurs only in the fringe-field regions.

15.5.6. As a consequence of the symmetry present in the symmetric Coulomb gauge, verify that the two real relations (5.89) and (5.90) can be combined to produce the single complex relation

$$\hat{A}_x^{m,c} + i\hat{A}_y^{m,c} = (i/2)(x+iy)^{m+1}\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+1]}(z)(x^2+y^2)^\ell. \tag{15.5.111}$$

Also verify that the two real relations (5.92) and (5.93) can be combined to produce the single complex relation

$$\hat{A}_x^{m,s} + i\hat{A}_y^{m,s} = (1/2)(x+iy)^{m+1}\sum_{\ell=0}^{\infty}(-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+1]}(z)(x^2+y^2)^\ell. \tag{15.5.112}$$

15.6 Nonuniqueness of Coulomb Gauge

There still remains the question of uniqueness. We will see that there are other Coulomb gauge vector potentials beyond the symmetric one already found.

15.6.1 The General Case

Suppose $\lambda(x, y, z)$ is any *harmonic* function,

$$\nabla^2 \lambda = 0. \tag{15.6.1}$$

If we add $\nabla\lambda$ to $\hat{\boldsymbol{A}}$ to produce a vector potential $\tilde{\boldsymbol{A}}$, it is easily verified that the result

$$\tilde{\boldsymbol{A}} = \hat{\boldsymbol{A}} + \nabla\lambda \tag{15.6.2}$$

also satisfies the desired relations

$$\nabla \times \tilde{\boldsymbol{A}} = \boldsymbol{B} \tag{15.6.3}$$

and

$$\nabla \cdot \tilde{\boldsymbol{A}} = 0. \tag{15.6.4}$$

Conversely, if we require that the Ansatz (6.2) also yield a vector potential in the Coulomb gauge, then λ must be harmonic.

We next observe that, by construction, $\hat{\boldsymbol{A}}$ falls to zero as $|z| \to \infty$ so that we should require the same of λ in order for $\tilde{\boldsymbol{A}}$ to have the same asymptotic behavior.[9] Thanks to the work already done, it easy to describe all λ satisfying (6.1) that have this property. Namely, by repeating the arguments leading to the representation of ψ in terms of generalized gradients, we may write

$$\lambda = \sum_{m=0}^{\infty} \Lambda_{m,c}(\rho, z) \cos m\phi + \sum_{m=1}^{\infty} \Lambda_{m,s}(\rho, z) \sin m\phi \tag{15.6.5}$$

and set

$$\Lambda_{m,\alpha}(\rho, z) = \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell + m)!} L_{m,\alpha}^{[2\ell]}(z) \rho^{2\ell+m}. \tag{15.6.6}$$

Here the functions $L_{m,\alpha}^{[0]}(z)$ may be specified at will save for the condition that they fall to zero for large $|z|$.

We know that $\hat{A}_x$ and $\hat{A}_y$ are harmonic functions because we may write

$$\hat{A}_x = \sum_{m=0}^{\infty} \hat{A}_x^{m,c} + \sum_{m=1}^{\infty} \hat{A}_x^{m,s}, \text{ etc.} \tag{15.6.7}$$

and we have already seen that each term in the above sums is harmonic. It is a remarkable fact that if $\sigma(x, y, z)$ is a harmonic function that falls off for large $|z|$ and thus can be written in a form analogous to (6.5), then there is another harmonic function λ with the same properties such that

$$\partial_y \lambda = \sigma. \tag{15.6.8}$$

[9] As stated earlier, This asymptotic behavior is desirable in order that the canonical and mechanical momenta be asymptotically the same. See (1.5.30). If we are working with z as the independent variable, in which case p_z is not a dynamical variable and does not appear in the Hamiltonian, we will at least want the x and y components of $\tilde{\boldsymbol{A}}$ to vanish for large $|z|$.

Or, if one prefers, there is a λ' such that

$$\partial_x \lambda' = \sigma. \tag{15.6.9}$$

See Appendix H. Let us apply this result, for example, to the case

$$\sigma = -\hat{A}_y \tag{15.6.10}$$

and then use λ to make the gauge transformation (6.2). Doing so, we find the results

$$\tilde{A}_x = \hat{A}_x + \partial_x \lambda, \tag{15.6.11}$$

$$\tilde{A}_y = \hat{A}_y + \partial_y \lambda = \hat{A}_y + \sigma = 0, \tag{15.6.12}$$

$$\tilde{A}_z = \hat{A}_z + \partial_z \lambda. \tag{15.6.13}$$

That is, we have found a gauge which is both Coulomb and for which the y component of the vector potential is zero. We call this the *vertical-free* Coulomb gauge. Similarly, by using λ', one can find a *horizontal-free* Coulomb gauge for which the x component of the vector potential is zero.

Even a bit more can be accomplished. Suppose $\tilde{\boldsymbol{A}}$ is a vector potential in the vertical-free Coulomb gauge so that $\tilde{A}_y = 0$. Let $\tau(x, z)$ be a harmonic function that depends *only* on x and z. Such a function can be written in the form

$$\begin{aligned}\tau(x,z) &= \sum_{n=0}^{\infty}(-1)^n[1/(2n+1)!]x^{2n+1}O^{[2n]}(z) + \sum_{n=0}^{\infty}(-1)^n[1/(2n)!]x^{2n}E^{[2n]}(z) \\ &= [xO^{[0]}(z) - (1/6)x^3O^{[2]}(z) + (1/120)x^5O^{[4]}(z) + \cdots] \\ &\quad + [E^{[0]}(z) - (1/2)x^2E^{[2]}(z) + (1/24)x^4E^{[4]}(z) + \cdots]\end{aligned} \tag{15.6.14}$$

where $O^{[0]}(z)$ and $E^{[0]}(z)$ are arbitrary functions of z save that they fall off for large $|z|$. See Appendix H. Now use τ to make a further gauge transformation that sends $\tilde{\boldsymbol{A}}$ to $\check{\boldsymbol{A}}$,

$$\check{\boldsymbol{A}} = \tilde{\boldsymbol{A}} + \nabla \tau. \tag{15.6.15}$$

By construction,

$$\partial_y \tau = 0 \tag{15.6.16}$$

so that $\check{\boldsymbol{A}}$ is also vertical free,

$$\check{A}_y = 0. \tag{15.6.17}$$

And for the x and z components of $\check{\boldsymbol{A}}$ we find the results

$$\begin{aligned}\check{A}_x &= \hat{A}_x + [O^{[0]}(z) - (1/2)x^2O^{[2]}(z) + (1/24)x^4O^{[4]}(z) + \cdots] \\ &\quad + [-xE^{[2]}(z) + (1/6)x^3E^{[4]}(z) + \cdots],\end{aligned} \tag{15.6.18}$$

$$\begin{aligned}\check{A}_z &= \hat{A}_z + [xO^{[1]}(z) - (1/6)x^3O^{[3]}(z) + (1/120)x^5O^{[5]}(z) + \cdots] \\ &\quad + [E^{[1]}(z) - (1/2)x^2E^{[3]}(z) + (1/24)x^4E^{[5]}(z) + \cdots].\end{aligned} \tag{15.6.19}$$

Evidently, by a suitable choice of $O^{[0]}(z)$ and $E^{[0]}(z)$, we are able to make some further adjustments to $\check{A}_x$ and $\check{A}_z$ while keeping the gauge Coulomb and vertical free.

15.6.2 Normal Dipole Example

As an example of the further gauge freedom just described, let us consider the case of a normal dipole whose Coulomb-gauge vector potential $\hat{\boldsymbol{A}}^{1,s}$ is given by (5.95) through (5.97). We will first perform a succession of gauge transformations to make the vector potential vertical free while maintaining its Coulomb nature. Then we will adjust its x component.

To begin, use a λ, which we will denote by the symbols λ^3, for which all the $L_{m,\alpha}^{[0]}(z)$ are zero save for $L_{3,c}^{[0]}(z)$. Then, from (6.6), we find through terms of degree six the result

$$\Lambda_{3,c}(\rho,z) = L_{3,c}^{[0]}(z)\rho^3 - (1/16)L_{3,c}^{[2]}(z)\rho^5 + \cdots , \tag{15.6.20}$$

from which it follows, using (6.5), that

$$\begin{aligned} \lambda^3 &= (\cos 3\phi)\Lambda_{3,c}(\rho,z) = (\cos 3\phi)[L_{3,c}^{[0]}(z)\rho^3 - (1/16)L_{3,c}^{[2]}(z)\rho^5 + \cdots] \\ &= (x^3 - 3xy^2)L_{3,c}^{[0]}(z) - (1/16)(x^5 - 2x^3y^2 - 3xy^4)L_{3,c}^{[2]}(z) + \cdots . \end{aligned} \tag{15.6.21}$$

This λ^3 has the gradients

$$(\nabla\lambda^3)_x = 3(x^2 - y^2)L_{3,c}^{[0]}(z) - (1/16)(5x^4 - 6x^2y^2 - 3y^4)L_{3,c}^{[2]}(z) + \cdots , \tag{15.6.22}$$

$$(\nabla\lambda^3)_y = -6xyL_{3,c}^{[0]}(z) + (1/16)(4x^3y + 12xy^3)L_{3,c}^{[2]}(z) + \cdots , \tag{15.6.23}$$

$$(\nabla\lambda^3)_z = (x^3 - 3xy^2)L_{3,c}^{[1]}(z) - (1/16)(x^5 - 2x^3y^2 - 3xy^4)L_{3,c}^{[3]}(z) + \cdots . \tag{15.6.24}$$

Let use employ λ^3 to produce a transformed vector potential $\boldsymbol{A}'$ using the relation

$$\boldsymbol{A}' = \hat{\boldsymbol{A}}^{1,s} + \nabla\lambda^3. \tag{15.6.25}$$

Then, from (5.95) through (5.97) and (6.22) through (6.25), we find the results

$$\begin{aligned} A'_x &= (x^2 - y^2)[(1/4)C_{1,s}^{[1]}(z) + 3L_{3,c}^{[0]}(z)] \\ &- (1/48)(x^4 - y^4)C_{1,s}^{[3]}(z) - (1/16)(5x^4 - 6x^2y^2 - 3y^4)L_{3,c}^{[2]}(z) + \cdots , \end{aligned} \tag{15.6.26}$$

$$\begin{aligned} A'_y &= xy[(1/2)C_{1,s}^{[1]}(z) - 6L_{3,c}^{[0]}(z)] \\ &- (1/24)(x^3y + xy^3)C_{1,s}^{[3]}(z) + (1/16)(4x^3y + 12xy^3)L_{3,c}^{[2]}(z) + \cdots , \end{aligned} \tag{15.6.27}$$

$$\begin{aligned} A'_z &= -xC_{1,s}^{[0]}(z) + (1/8)(x^3 + xy^2)C_{1,s}^{[2]}(z) - (1/192)(x^5 + 2x^3y^2 + xy^4)C_{1,s}^{[4]}(z) \\ &+ (x^3 - 3xy^2)L_{3,c}^{[1]}(z) - (1/16)(x^5 - 2x^3y^2 - 3xy^4)L_{3,c}^{[3]}(z) + \cdots . \end{aligned} \tag{15.6.28}$$

Observe that we can make the leading term of A'_x vanish by setting

$$[(1/4)C_{1,s}^{[1]}(z) + 3L_{3,c}^{[0]}(z)] = 0. \tag{15.6.29}$$

Or, we can make the leading term of A'_y vanish by setting

$$[(1/2)C_{1,s}^{[1]}(z) - 6L_{3,c}^{[0]}(z)] = 0. \tag{15.6.30}$$

Suppose we decide to make the leading term of A'_y vanish. Then we have the relation

$$L_{3,c}^{[0]}(z) = (1/12)C_{1,s}^{[1]}(z), \tag{15.6.31}$$

from which it follows that

$$L_{3,c}^{[1]}(z) = (1/12)C_{1,s}^{[2]}(z), \tag{15.6.32}$$

$$L_{3,c}^{[2]}(z) = (1/12)C_{1,s}^{[3]}(z), \tag{15.6.33}$$

$$L_{3,c}^{[3]}(z) = (1/12)C_{1,s}^{[4]}(z), \text{ etc.} \tag{15.6.34}$$

When this is done, $\boldsymbol{A}'$ takes the form

$$A'_x = (1/2)(x^2 - y^2)C_{1,s}^{[1]}(z) - (1/192)(9x^4 - 6x^2y^2 - 7y^4)C_{1,s}^{[3]}(z) + \cdots , \tag{15.6.35}$$

$$A'_y = -(1/48)(x^3y - xy^3)C_{1,s}^{[3]}(z) + \cdots , \tag{15.6.36}$$

$$A'_z = \; -xC_{1,s}^{[0]}(z) + (1/24)(5x^3 - 3xy^2)C_{1,s}^{[2]}(z) - (1/96)(x^5 - xy^4)C_{1,s}^{[4]}(z) + \cdots . \tag{15.6.37}$$

At this point, as a sanity check on the algebra used to yield (6.35) through (6.37), the reader should verify (through the order of the terms retained) that $\boldsymbol{A}'$ is still Coulombic and its components are still harmonic.

Let us see if we can make the next term in A'_y vanish by performing an additional gauge transformation. Suppose we make a further gauge transformation using a λ, which we will call λ^5, for which all the $L_{m,\alpha}^{[0]}(z)$ are zero save for $L_{5,c}^{[0]}(z)$. Then we find through terms of degree six the result

$$\Lambda_{5,c}(\rho, z) = L_{5,c}^{[0]}(z)\rho^5 + \cdots , \tag{15.6.38}$$

from which it follows that

$$\begin{aligned} \lambda^5 &= (\cos 5\phi)\Lambda_{5,c}(\rho, z) = (\cos 5\phi)[L_{5,c}^{[0]}(z)\rho^5 + \cdots] \\ &= (x^5 - 10x^3y^2 + 5xy^4)L_{5,c}^{[0]}(z) + \cdots . \end{aligned} \tag{15.6.39}$$

This λ has the gradients

$$(\nabla\lambda^5)_x = (5x^4 - 30x^2y^2 + 5y^4)L_{5,c}^{[0]}(z) + \cdots , \tag{15.6.40}$$

$$(\nabla\lambda^5)_y = -20(x^3y - xy^3)L_{5,c}^{[0]}(z) + \cdots , \tag{15.6.41}$$

$$(\nabla\lambda^5)_z = (x^5 - 10x^3y^2 + 5xy^4)L^{[1]}_{5,c}(z). \tag{15.6.42}$$

Correspondingly, we will define a further transformed vector potential $\boldsymbol{A}''$ by writing

$$\boldsymbol{A}'' = \boldsymbol{A}' + \nabla\lambda^5. \tag{15.6.43}$$

Then, using (6.35) through (6.37) and (6.40) through (6.43), $\boldsymbol{A}''$ takes the form

$$\begin{aligned} A''_x &= (1/2)(x^2 - y^2)C^{[1]}_{1,s}(z) - (1/192)(9x^4 - 6x^2y^2 - 7y^4)C^{[3]}_{1,s}(z) \\ &+ (5x^4 - 30x^2y^2 + 5y^4)L^{[0]}_{5,c}(z) + \cdots , \end{aligned} \tag{15.6.44}$$

$$A''_y = -(1/48)(x^3y - xy^3)C^{[3]}_{1,s}(z) - 20(x^3y - xy^3)L^{[0]}_{5,c}(z) + \cdots , \tag{15.6.45}$$

$$\begin{aligned} A''_z &= -xC^{[0]}_{1,s}(z) + (1/24)(5x^3 - 3xy^2)C^{[2]}_{1,s}(z) - (1/96)(x^5 - xy^4)C^{[4]}_{1,s}(z) \\ &+ (x^5 - 10x^3y^2 + 5xy^4)L^{[1]}_{5,c}(z) + \cdots . \end{aligned} \tag{15.6.46}$$

We see that A''_y will vanish through terms of degree four provided $L^{[0]}_{5,c}(z)$ is selected to satisfy the relation

$$L^{[0]}_{5,c}(z) = -(1/960)C^{[3]}_{1,s}(z), \tag{15.6.47}$$

from which it follows that

$$L^{[1]}_{5,c}(z) = -(1/960)C^{[4]}_{1,s}(z), \text{ etc.} \tag{15.6.48}$$

When this condition is met, $\boldsymbol{A}''$ takes the form

$$A''_x = (1/2)(x^2 - y^2)C^{[1]}_{1,s}(z) - (1/96)(5x^4 - 6x^2y^2 - 3y^4)C^{[3]}_{1,s}(z) + \cdots , \tag{15.6.49}$$

$$A''_y = 0 + \cdots , \tag{15.6.50}$$

$$\begin{aligned} A''_z &= -xC^{[0]}_{1,s}(z) + (1/24)(5x^3 - 3xy^2)C^{[2]}_{1,s}(z) \\ &- (1/960)(11x^5 - 10x^3y^2 - 5xy^4)C^{[4]}_{1,s}(z) + \cdots . \end{aligned} \tag{15.6.51}$$

(Here the reader should again perform Coulombic and harmonic sanity checks.) We have achieved, through terms of degree four, a vertical-free Coulomb gauge vector potential for the normal dipole.

There is still the possibility of adjusting the x (and correspondingly the z component) of $\boldsymbol{A}''$ by making yet another gauge transformation using the harmonic function $\tau(x, z)$ given by (6.14). We define a still further transformed vector potential $\boldsymbol{A}'''$ by writing

$$\boldsymbol{A}''' = \boldsymbol{A}'' + \nabla\tau. \tag{15.6.52}$$

So doing gives the result

$$\begin{aligned} A_x''' &= O^{[0]}(z) - xE^{[2]}(z) + (1/2)(x^2-y^2)C^{[1]}_{1,s}(z) - (1/2)x^2O^{[2]}(z) \\ &+ (1/6)x^3E^{[4]}(z) - (1/96)(5x^4-6x^2y^2-3y^4)C^{[3]}_{1,s}(z) + (1/24)x^4O^{[4]}(z) + \cdots , \end{aligned} \tag{15.6.53}$$

$$A_y''' = 0 + \cdots , \tag{15.6.54}$$

$$\begin{aligned} A_z''' &= E^{[1]}(z) - xC^{[0]}_{1,s}(z) + xO^{[1]}(z) - (1/2)x^2E^{[3]}(z) \\ &+ (1/24)(5x^3-3xy^2)C^{[2]}_{1,s}(z) - (1/6)x^3O^{[3]}(z) + (1/24)x^4E^{[5]}(z) \\ &- (1/960)(11x^5-10x^3y^2-5xy^4)C^{[4]}_{1,s}(z) + (1/120)x^5O^{[5]}(z) + \cdots . \end{aligned} \tag{15.6.55}$$

Here the functions $O^{[0]}(z)$ and $E^{[1]}(z)$ are arbitrary except that they must vanish as $|z| \to \infty$.

Exercises

15.6.1. Verify that the vector potential $\boldsymbol{A}''$ given by (6.49) through (6.51) yields the magnetic field $\boldsymbol{B}$ given by (3.58) through (3.60), and is Coulombic and harmonic.

15.6.2. Review Exercise 4.4 and, in particular, the vector potential given by (4.58) through (4.60). Show that this vector potential is in neither the azimuthal-free nor the Coulomb gauge. Let χ be the function

$$\begin{aligned} \chi(x,z) &= -\sum_{n=1}^{\infty}(-1)^n[1/(2n+1)!]x^{2n+1}O^{[2n-1]}(z) \\ &= (1/6)x^3O^{[1]}(z) - (1/120)x^5O^{[3]}(z) + \cdots . \end{aligned} \tag{15.6.56}$$

Define a vector potential $\hat{\boldsymbol{A}}^{\rm iwd}$ by the making the gauge transformation

$$\hat{\boldsymbol{A}}^{\rm iwd} = \boldsymbol{A}^{\rm iwd} + \nabla\chi. \tag{15.6.57}$$

Show that

$$\begin{aligned} \hat{A}_x^{\rm iwd} &= \sum_{n=1}^{\infty}(-1)^n[1/(2n)!]y^{2n}O^{[2n-1]}(z) - \sum_{n=1}^{\infty}(-1)^n[1/(2n)!]x^{2n}O^{[2n-1]}(z) \\ &= -(1/2)(y^2-x^2)O^{[1]}(z) + (1/24)(y^4-x^4)O^{[3]}(z) + \cdots , \end{aligned} \tag{15.6.58}$$

$$\hat{A}_y^{\rm iwd} = 0, \tag{15.6.59}$$

$$\begin{aligned} \hat{A}_z^{\rm iwd} &= -\sum_{n=0}^{\infty}(-1)^n[1/(2n+1)!]x^{2n+1}O^{[2n]}(z) \\ &= -xO^{[0]}(z) + (1/6)x^3O^{[2]}(z) - (1/120)x^5O^{[4]}(z) + \cdots . \end{aligned} \tag{15.6.60}$$

Thus, the vector potential $\hat{\boldsymbol{A}}^{\rm iwd}$ is vertical free. Show that $\hat{\boldsymbol{A}}^{\rm iwd}$ is also in the Coulomb gauge, and that all its Cartesian components are harmonic functions.

Equations (6.49) through (6.51) give the vector potential in the vertical-free Coulomb gauge corresponding to the $C_{1,s}^{[n]}$. Change notation to call this result $\bar{\boldsymbol{A}}^{1,s}$,

$$\bar{A}_x^{1,s} = (1/2)(x^2 - y^2)C_{1,s}^{[1]}(z) - (1/96)(5x^4 - 6x^2y^2 - 3y^4)C_{1,s}^{[3]}(z) + \cdots , \tag{15.6.61}$$

$$\bar{A}_y^{1,s} = 0 + \cdots , \tag{15.6.62}$$

$$\begin{aligned}\bar{A}_z^{1,s} &= -xC_{1,s}^{[0]}(z) + (1/24)(5x^3 - 3xy^2)C_{1,s}^{[2]}(z) \\ &- (1/960)(11x^5 - 10x^3y^2 - 5xy^4)C_{1,s}^{[4]}(z) + \cdots .\end{aligned} \tag{15.6.63}$$

Show, by analogous calculations, that the vector potential in the vertical-free Coulomb gauge corresponding to the $C_{3,s}^{[n]}$ is given by the relations

$$\bar{A}_x^{3,s} = (1/4)(x^4 - 6x^2y^2 + y^4)C_{3,s}^{[1]}(z) + \cdots , \tag{15.6.64}$$

$$\bar{A}_y^{3,s} = 0, \tag{15.6.65}$$

$$\bar{A}_z^{3,s} = -(x^3 - 3xy^2)C_{3,s}^{[0]}(z) + (1/80)(7x^5 - 30x^3y^2 - 5xy^4)C_{3,s}^{[2]}(z) + \cdots . \tag{15.6.66}$$

Use these results to find, through terms of degree four, the Coulombic and vertical-free vector potential for the infinite-width dipole, and show that your results agree with (6.58) through (6.60).

Note that again, as was the case for the vector potentials found in Exercise 3.4, that the vector potential is primarily in the z direction.

Study Appendix H.3.3, which finds a Coulombic and horizontal-free vector potential for the infinite-width dipole.

15.6.3. Review Exercise 4.7. Show that, under the same assumptions, (4.66) and (4.67) also hold in the Coulomb gauge.

15.6.4. The relations (5.32) through (5.34) provide expansions for the components of $\hat{\boldsymbol{A}}^0$, the $m = 0$ vector potential in the symmetric Coulomb gauge. Find the first few terms in the expansions for the components of $\tilde{\boldsymbol{A}}^0$, the vector potential for the $m = 0$ case in the vertical-free Coulomb gauge. Recall (6.11) through (6.13).

15.7 Determination of the Vector Potential: Poincaré-Coulomb Gauge

Assuming $m \neq 0$, the relations (4.15) through (4.26) provide, through all orders, formulas for the azimuthal-free gauge vector potential in terms of the on-axis gradients. And, for all values of m, the relations (5.83) through (5.94) provide, again through all orders, formulas for the symmetric Coulomb-gauge vector potential in terms of the on-axis gradients. For general m and through all orders, are there relations that provide formulas for the Poincaré-Coulomb gauge vector potential in terms of the on-axis gradients? The purpose of this section is to explore this question.

At this point it is necessary to be precise about what we wish to accomplish. Recall the vectors $\boldsymbol{R}$, $\boldsymbol{R}_0$, and $\boldsymbol{r}$ introduced in Subsection 2.1 by writing (2.3). Since the axis of any of the straight beam-line elements we are considering is supposed to lie along the z axis, we stipulate that $\boldsymbol{R}_0$ be of the form

$$\boldsymbol{R}_0 = (0, 0, Z_0). \tag{15.7.1}$$

Correspondingly, $\boldsymbol{R}$ then takes the form

$$\boldsymbol{R} = (x, y, Z_0 + z) \tag{15.7.2}$$

where x and y are assumed to be small, and z may or may not also be small depending on the choice of Z_0.

Let $\hat{\boldsymbol{A}}^{m,\alpha}$ be the symmetric Coulomb gauge vector potential of Section 5. In view of (7.2) we may write

$$\hat{\boldsymbol{A}}^{m,\alpha}(\boldsymbol{R}) = \hat{\boldsymbol{A}}^{m,\alpha}(x, y, Z_0 + z). \tag{15.7.3}$$

Assuming it exists, let us introduce the symbols ${}^P\!\boldsymbol{A}^{m,\alpha}(x, y, z; Z_0)$ to denote the Poincaré-Coulomb gauge counterpart to the vector potential $\hat{\boldsymbol{A}}^{m,\alpha}$. Since we require that these vector potentials produce the same magnetic field,

$$\nabla \times \hat{\boldsymbol{A}}^{m,\alpha}(x, y, Z_0 + z) = \nabla \times \ {}^P\!\boldsymbol{A}^{m,\alpha}(x, y, z; Z_0), \tag{15.7.4}$$

they must be related by a gauge transformation of the form

$${}^P\!\boldsymbol{A}^{m,\alpha}(x, y, z; Z_0) = \hat{\boldsymbol{A}}^{m,\alpha}(x, y, Z_0 + z) + \nabla \hat{\chi}^P_{m,\alpha} \tag{15.7.5}$$

described by the gauge function $\hat{\chi}^P_{m,\alpha}(x, y, z; Z_0)$. Here

$$\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z), \tag{15.7.6}$$

and

$$\nabla \hat{\chi}^P_{m,\alpha} = \nabla \hat{\chi}^P_{m,\alpha}(x, y, z; Z_0). \tag{15.7.7}$$

Since both ${}^P\!\boldsymbol{A}^{m,\alpha}$ and $\hat{\boldsymbol{A}}^{m,\alpha}$ are supposed to be in the Coulomb gauge, i.e. divergence free, we see from (7.5) that the function $\hat{\chi}^P_{m,\alpha}$ must be harmonic,

$$\nabla^2 \hat{\chi}^P_{m,\alpha}(x, y, z; Z_0) = 0. \tag{15.7.8}$$

Finally if ${}^P\!\boldsymbol{A}^{m,\alpha}$ is to be in the Poincaré-Coulomb gauge, then from (7.5) we see that $\hat{\chi}^P_{m,\alpha}$ must satisfy the further condition

$$\boldsymbol{r} \cdot \nabla \hat{\chi}^P_{m,\alpha} = -\boldsymbol{r} \cdot \hat{\boldsymbol{A}}^{m,\alpha}. \tag{15.7.9}$$

Our task now is to find the functions $\hat{\chi}^P_{m,\alpha}(x, y, z; Z_0)$ in terms of the on-axis gradients.

15.7.1 The $m = 0$ Case

Observe that the $m = 0$ symmetric Coulomb gauge vector potential $\hat{\boldsymbol{A}}^0$ given by (5.37) through (5.39) has only a ϕ component. It follows, from the fact that the $\boldsymbol{e}_\rho$, $\boldsymbol{e}_\phi$, $\boldsymbol{e}_z$ form an orthonormal triad and (2.20), that there is the relation

$$\boldsymbol{r} \cdot \hat{\boldsymbol{A}}^0(\boldsymbol{R}) = 0. \tag{15.7.10}$$

Alternatively, this same relation follows from (5.32) through (5.34) and (2.5). Either way, we conclude that $\hat{\boldsymbol{A}}^0$, when evaluated with respect to any origin on the z axis, is in the Poincaré-Coulomb gauge.

Also, $\hat{\boldsymbol{A}}^0$ can be expressed in terms of on-axis gradients. Indeed, in terms of the variables employed in this section, the relations (5.32) through (5.34) and (5.36) take the form

$$\begin{aligned} \hat{A}^0_x(x, y, Z_0 + z) &= -(y/2)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{1}{2^{2\ell}\ell!(\ell+1)!} C_0^{[2\ell+1]}(Z_0 + z)(x^2 + y^2)^\ell \\ &= -(y/2)[C_0^{[1]}(Z_0 + z) - (1/8)C_0^{[3]}(Z_0 + z)(x^2 + y^2) + \cdots], \end{aligned} \tag{15.7.11}$$

$$\begin{aligned} \hat{A}^0_y(x, y, Z_0 + z) &= (x/2)\sum_{\ell=0}^{\infty}(-1)^\ell \frac{1}{2^{2\ell}\ell!(\ell+1)!} C_0^{[2\ell+1]}(Z_0 + z)(x^2 + y^2)^\ell \\ &= (x/2)[C_0^{[1]}(Z_0 + z) - (1/8)C_0^{[3]}(Z_0 + z)(x^2 + y^2) + \cdots], \end{aligned} \tag{15.7.12}$$

$$\hat{A}^0_z(x, y, Z_0 + z) = 0, \tag{15.7.13}$$

where

$$C_0^{[1]}(Z_0 + z) = B_z(0, 0, Z_0 + z). \tag{15.7.14}$$

15.7.2 The $m \geq 1$ Cases

We have seen that for the case $m = 0$ the symmetric Coulomb gauge vector potential $\hat{\boldsymbol{A}}^0$ is in the Poincaré-Coulomb gauge. What can be said about the cases $m \geq 1$? Can we construct, from the on-axis gradients, a vector potential in the Poincaré-Coulomb gauge for these cases? Here we assume that the expansion point $\boldsymbol{R}_0$ has been selected such that $\boldsymbol{r}$ is small, at least initially. Subsequently we will require that x and y remain small, but may allow z to become large.

In cylindrical coordinates the gradient operator takes the form

$$\nabla = \boldsymbol{e}_\rho(\partial/\partial\rho) + \boldsymbol{e}_\phi(1/\rho)(\partial/\partial\phi) + \boldsymbol{e}_z(\partial/\partial z). \tag{15.7.15}$$

It follows from (2.20) and (7.15) that

$$\boldsymbol{r} \cdot \nabla = \rho(\partial/\partial\rho) + z(\partial/\partial z). \tag{15.7.16}$$

Also, we conclude from (2.20) and (2.21) that

$$\boldsymbol{r} \cdot \hat{\boldsymbol{A}}^{m,\alpha} = \rho \hat{A}_\rho^{m,\alpha} + z\hat{A}_z^{m,\alpha}. \tag{15.7.17}$$

Upon combining (7.9), (7.16), and (7.17) we see that $\hat{\chi}_{m,\alpha}^P$ must satisfy the equation

$$[\rho(\partial/\partial\rho) + z(\partial/\partial z)]\hat{\chi}_{m,\alpha}^P = -(\rho\hat{A}_\rho^{m,\alpha} + z\hat{A}_z^{m,\alpha}). \tag{15.7.18}$$

Since the $\hat{\chi}_{m,\alpha}^P$ are known to be harmonic, and in view of (3.33), let us make the Ansätze

$$\hat{\chi}_{m,c}^P(x, y, z; Z_0) = -\sin(m\phi)\sum_{k=0}^{\infty}(-1)^k\frac{m!}{2^{2k}k!(k+m)!}D_{m,c}^{[2k]}(z; Z_0)\rho^{2k+m}, \tag{15.7.19}$$

$$\hat{\chi}_{m,s}^P(x, y, z; Z_0) = \cos(m\phi)\sum_{k=0}^{\infty}(-1)^k\frac{m!}{2^{2k}k!(k+m)!}D_{m,s}^{[2k]}(z; Z_0)\rho^{2k+m}, \tag{15.7.20}$$

where the functions $D_{m,\alpha}^{[0]}(z; Z_0)$ are yet to be determined. [Note that these D functions are not to be confused with those appearing in (5.53).] Recall that in Subsection 15.4.2 we found the Coulomb gauge results

$$\hat{A}_\rho^{m,c} = -(1/2)(\sin m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+1]}(Z_0+z)\rho^{2\ell+m+1}, \tag{15.7.21}$$

$$\hat{A}_z^{m,c} = (\sin m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(Z_0+z)\rho^{2\ell+m}; \tag{15.7.22}$$

$$\hat{A}_\rho^{m,s} = (1/2)(\cos m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+1]}(Z_0+z)\rho^{2\ell+m+1}, \tag{15.7.23}$$

$$\hat{A}_z^{m,s} = -(\cos m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(Z_0+z)\rho^{2\ell+m}. \tag{15.7.24}$$

[Here we have again employed the variables of this section in writing (7.21) through (7.24).] We observe that the operator appearing on the left side of (7.18) does not involve the variable ϕ. Therefore, we may cancel like trigonometric factors appearing on the right and left sides of (7.18) to find, in the case $\alpha = c$, the requirement

$$\begin{aligned}&[\rho(\partial/\partial\rho) + z(\partial/\partial z)][\sum_{k=0}^{\infty}(-1)^k\frac{m!}{2^{2k}k!(k+m)!}D_{m,c}^{[2k]}(z; Z_0)\rho^{2k+m}] = \\ &-(\rho/2)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+1]}(Z_0+z)\rho^{2\ell+m+1} \\ &+z\sum_{\ell=0}^{\infty}(-1)^\ell\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(Z_0+z)\rho^{2\ell+m};\end{aligned} \tag{15.7.25}$$

and, in the case $\alpha = s$, the requirement

$$[\rho(\partial/\partial\rho) + z(\partial/\partial z)][\sum_{k=0}^{\infty}(-1)^k \frac{m!}{2^{2k}k!(k+m)!} D_{m,s}^{[2k]}(z;Z_0)\rho^{2k+m}] =$$
$$-(\rho/2)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+1]}(Z_0+z)\rho^{2\ell+m+1}$$
$$+z\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(Z_0+z)\rho^{2\ell+m}. \tag{15.7.26}$$

Note the pleasant fact that the requirements (7.25) and (7.26) are identical in form. The operator on the left sides of (7.25) and (7.26) may be moved under the summation sign and allowed to work its will. For example, there is the general result

$$[\rho(\partial/\partial\rho) + z(\partial/\partial z)][D_{m,\alpha}^{[2k]}(z;Z_0)\rho^{2k+m}] = [(2k+m)D_{m,\alpha}^{[2k]}(z;Z_0) + zD_{m,\alpha}^{[2k+1]}(z;Z_0)]\rho^{2k+m}. \tag{15.7.27}$$

Also, the indicated multiplications on the right sides of (7.25) and (7.26) can be carried out. The net result of these two manipulations is the requirement

$$\sum_{k=0}^{\infty}[(-1)^k\frac{m!}{2^{2k}k!(k+m)!}][(2k+m)D_{m,\alpha}^{[2k]}(z;Z_0) + zD_{m,\alpha}^{[2k+1]}(z;Z_0)]\rho^{2k+m} =$$
$$\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!}{2^{2\ell+1}\ell!(\ell+m+1)!}C_{m,\alpha}^{[2\ell+1]}(Z_0+z)\rho^{2\ell+m+2}$$
$$+\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}zC_{m,\alpha}^{[2\ell]}(Z_0+z)\rho^{2\ell+m}. \tag{15.7.28}$$

Let us equate the coefficients of powers of ρ on both sides of (7.28) to obtain, we hope, relations that will specify the $D_{m,\alpha}^{[2k]}(z;Z_0)$ in terms of the $C_{m,\alpha}^{[2\ell]}(Z_0+z)$. The lowest power of ρ on the left side of (7.28) occurs for $k = 0$, and is ρ^m. Its coefficient is

$$\text{Coefficient of } \rho^m \text{ on left side } = mD_{m,\alpha}^{[0]}(z;Z_0) + zD_{m,\alpha}^{[1]}(z;Z_0). \tag{15.7.29}$$

The lowest power of ρ on the right side of (7.28) occurs for $\ell = 0$, and is also ρ^m. Its coefficient is

$$\text{Coefficient of } \rho^m \text{ on right side } = zC_{m,\alpha}^{[0]}(Z_0+z). \tag{15.7.30}$$

We conclude, so far, that there is the requirement that $D_{m,\alpha}^{[0]}(z;Z_0)$ must satisfy the differential equation

$$zD_{m,\alpha}^{[1]}(z;Z_0) + mD_{m,\alpha}^{[0]}(z;Z_0) = zC_{m,\alpha}^{[0]}(Z_0+z). \tag{15.7.31}$$

We now seek to solve (7.31). Begin by multiplying both sides of (7.31) by z^{m-1} to yield the result

$$z^m D_{m,\alpha}^{[1]}(z;Z_0) + z^{m-1}mD_{m,\alpha}^{[0]}(z;Z_0) = z^m C_{m,\alpha}^{[0]}(Z_0+z). \tag{15.7.32}$$

Observe that

$$z^m D^{[1]}_{m,\alpha}(z;Z_0) + z^{m-1} m D^{[0]}_{m,\alpha}(z;Z_0) = (d/dz)[z^m D^{[0]}_{m,\alpha}(z;Z_0)], \tag{15.7.33}$$

and therefore z^{m-1} is an integrating factor for (7.31). It follows that (7.31) can be rewritten in the form

$$(d/dz)[z^m D^{[0]}_{m,\alpha}(z;Z_0)] = z^m C^{[0]}_{m,\alpha}(Z_0+z) \tag{15.7.34}$$

with the immediate general solution

$$z^m D^{[0]}_{m,\alpha}(z;Z_0) = \text{constant } + \int_0^z dz'\ (z')^m C^{[0]}_{m,\alpha}(Z_0+z'), \tag{15.7.35}$$

or

$$D^{[0]}_{m,\alpha}(z;Z_0) = \text{constant} \times z^{-m} + \int_0^z dz'\ (z'/z)^m C^{[0]}_{m,\alpha}(Z_0+z'). \tag{15.7.36}$$

If we seek a particular solution that is analytic in z, then we must set the constant term to zero. Also, suppose we introduce a new variable of integration λ by writing

$$\lambda = z'/z. \tag{15.7.37}$$

When these steps are made, (7.36) takes the final form

$$D^{[0]}_{m,\alpha}(z;Z_0) = z\int_0^1 d\lambda\ \lambda^m C^{[0]}_{m,\alpha}(Z_0+\lambda z). \tag{15.7.38}$$

We arrived at the requirement (7.31) by equating the coefficients of the lowest power of ρ^m in (7.28). What happens if we equate the coefficients of the higher powers? We hope *nothing* new since $D^{[0]}_{m,\alpha}(z;Z_0)$ is already specified by (7.38). The next highest power of ρ appearing on the left side of (7.28) is ρ^{m+2}, and occurs for $k=1$. Its coefficient is

$$\text{Coefficient of } \rho^{m+2} \text{ on left side } = -[\frac{m!}{4(m+1)!}][(m+2)D^{[2]}_{m,\alpha}(z;Z_0) + zD^{[3]}_{m,\alpha}(z;Z_0)]. \tag{15.7.39}$$

The next highest power of ρ appearing on the right side of (7.28) is also ρ^{m+2}, and occurs for $\ell=0$ in the first term on the right and $\ell=1$ in the second term. Its coefficient is

$$\text{Coefficient of } \rho^{m+2} \text{ on right side } = -[\frac{m!}{2(m+1)!}]C^{[1]}_{m,\alpha}(Z_0+z) - [\frac{m!}{4(m+1)!}]zC^{[2]}_{m,\alpha}(Z_0+z). \tag{15.7.40}$$

Equating these two coefficients yields the result

$$(m+2)D^{[2]}_{m,\alpha}(z;Z_0) + zD^{[3]}_{m,\alpha}(z;Z_0) = 2C^{[1]}_{m,\alpha}(Z_0+z) + zC^{[2]}_{m,\alpha}(Z_0+z). \tag{15.7.41}$$

Is this result new? It is not. Differentiating both sides of the previous result (7.31) yields the relation

$$\partial_z[zD^{[1]}_{m,\alpha}(z;Z_0) + mD^{[0]}_{m,\alpha}(z;Z_0)] = \partial_z[zC^{[0]}_{m,\alpha}(Z_0+z)] \tag{15.7.42}$$

which, upon expansion, yields the result

$$zD^{[2]}_{m,\alpha}(z;Z_0) + (1+m)D^{[1]}_{m,\alpha}(z;Z_0) = zC^{[1]}_{m,\alpha}(Z_0+z) + C^{[0]}_{m,\alpha}(Z_0+z). \tag{15.7.43}$$

Next, differentiating both sides of (7.43) yields the relation

$$\partial_z[zD^{[2]}_{m,\alpha}(z;Z_0) + (1+m)D^{[1]}_{m,\alpha}(z;Z_0)] = \partial_z[zC^{[1]}_{m,\alpha}(Z_0+z) + C^{[0]}_{m,\alpha}(Z_0+z)] \qquad (15.7.44)$$

which, upon expansion, yields the result

$$zD^{[3]}_{m,\alpha}(z;Z_0) + (2+m)D^{[2]}_{m,\alpha}(z;Z_0)] = zC^{[2]}_{m,\alpha}(Z_0+z) + 2C^{[1]}_{m,\alpha}(Z_0+z)]. \qquad (15.7.45)$$

We see that this result agrees with (7.41). Further calculation shows that all the results obtained by equating the coefficients of like powers of ρ on both sides of (7.28) are consistent with the relation (7.31) and are identical to results that flow from it upon differentiation.

In summary, the $\hat{\chi}^P_{m,\alpha}$ are specified by (7.19) and (7.20) in terms of the $D^{[0]}_{m,\alpha}(z;Z_0)$, and the $D^{[0]}_{m,\alpha}(z;Z_0)$ are specified by (7.38) in terms of the $C^{[0]}_{m,\alpha}(Z_0+z)$. What remains, according to (7.5), is to compute $\nabla\hat{\chi}^P_{m,\alpha}$. Introduce the notation

$$\Delta\boldsymbol{A}^{m,\alpha} = {}^P\!\boldsymbol{A}^{m,\alpha} - \hat{\boldsymbol{A}}^{m,\alpha} = \nabla\hat{\chi}^P_{m,\alpha}. \qquad (15.7.46)$$

Then, from (7.15), (7.19), and (7.20), we have the relations

$$\Delta A^{m,c}_\rho = (\partial/\partial\rho)\hat{\chi}^P_{m,c} = -\sin(m\phi)\sum_{k=0}^\infty(-1)^k\frac{m!(2k+m)}{2^{2k}k!(k+m)!}D^{[2k]}_{m,c}(z;Z_0)\rho^{2k+m-1}, \qquad (15.7.47)$$

$$\Delta A^{m,c}_\phi = (1/\rho)(\partial/\partial\phi)\hat{\chi}^P_{m,c} = -m\cos(m\phi)\sum_{k=0}^\infty(-1)^k\frac{m!}{2^{2k}k!(k+m)!}D^{[2k]}_{m,c}(z;Z_0)\rho^{2k+m-1}, \qquad (15.7.48)$$

$$\Delta A^{m,c}_z = (\partial/\partial z)\hat{\chi}^P_{m,c} = -\sin(m\phi)\sum_{k=0}^\infty(-1)^k\frac{m!}{2^{2k}k!(k+m)!}D^{[2k+1]}_{m,c}(z;Z_0)\rho^{2k+m}; \qquad (15.7.49)$$

$$\Delta A^{m,s}_\rho = (\partial/\partial\rho)\hat{\chi}^P_{m,s} = \cos(m\phi)\sum_{k=0}^\infty(-1)^k\frac{m!(2k+m)}{2^{2k}k!(k+m)!}D^{[2k]}_{m,s}(z;Z_0)\rho^{2k+m-1}, \qquad (15.7.50)$$

$$\Delta A^{m,s}_\phi = (1/\rho)(\partial/\partial\phi)\hat{\chi}^P_{m,s} = -m\sin(m\phi)\sum_{k=0}^\infty(-1)^k\frac{m!}{2^{2k}k!(k+m)!}D^{[2k]}_{m,s}(z;Z_0)\rho^{2k+m-1}, \qquad (15.7.51)$$

$$\Delta A^{m,s}_z = (\partial/\partial z)\hat{\chi}^P_{m,s} = \cos(m\phi)\sum_{k=0}^\infty(-1)^k\frac{m!}{2^{2k}k!(k+m)!}D^{[2k+1]}_{m,s}(z;Z_0)\rho^{2k+m}. \qquad (15.7.52)$$

These are the results in cylindrical coordinates. Suppose results in Cartesian coordinates are desired. The relations (7.49) and (7.52) are already in Cartesian form. The remaining Cartesian-form results may be found using the relations

$$\Delta A^{m,\alpha}_x = \cos\phi\,\Delta A^{m,\alpha}_\rho - \sin\phi\,\Delta A^{m,\alpha}_\phi, \qquad (15.7.53)$$

$$\Delta A^{m,\alpha}_y = \sin\phi\,\Delta A^{m,\alpha}_\rho + \cos\phi\,\Delta A^{m,\alpha}_\phi. \qquad (15.7.54)$$

Recall (2.24) and (2.25).

There are two other points that require discussion. The first, a matter of consistency, is this: We have found formulas for the $\Delta\boldsymbol{A}^{m,\alpha}$ in the case $m > 0$. What happens when these formulas are evaluated for $m = 0$? From (7.47) through (7.49) we immediately see that

$$\Delta\boldsymbol{A}^{0,c} = 0, \tag{15.7.55}$$

and therefore

$$^{P}\!\boldsymbol{A}^{0,c} = \hat{\boldsymbol{A}}^{0,c}. \tag{15.7.56}$$

This result is consistent with previous results because we know that $\hat{\boldsymbol{A}}^{0,c} = \hat{\boldsymbol{A}}^{0}$ is already in the Poincaré-Coulomb gauge, and this gauge is unique. What about $\Delta\boldsymbol{A}^{0,s}$? From (3.36) we recall that $C^{[0]}_{0,s}(Z_0 + z)$ vanishes, and therefore according to (7.38) all the $D^{[n]}_{0,s}$ vanish. Consequently, according to (7.50) through (7.52), $\Delta\boldsymbol{A}^{0,s}$ also vanishes,

$$\Delta\boldsymbol{A}^{0,s} = 0. \tag{15.7.57}$$

Therefore, in view of (5.96), there is the result

$$^{P}\!\boldsymbol{A}^{0,s} = 0. \tag{15.7.58}$$

The second point has to do with the nature of the relation between the Poincaré-Coulomb vector potential we have found (which we know is unique from the work of Subsection 2.6) and the on-axis gradients. The relations (4.15) through (4.26) and (5.83) through (5.94) provide, at any point z, formulas for the vector potential in the azimuthal-free and Coulomb gauges in terms of the on-axis gradients $C^{[0]}_{m,\alpha}(z)$ and their first few derivatives at the *same* point z. In particular, if expansions in powers of x and y are required only through some finite order (as is the case), then only a finite number of derivatives of the $C^{[0]}_{m,\alpha}(z)$ are required. In this sense, we may say that these vector potentials depend *locally* on the $C^{[0]}_{m,\alpha}(z)$. By contrast, according to (7.38), it appears that computation of the $D^{[0]}_{m,\alpha}(z; Z_0)$, and therefore of the vector potential in the Poincaré-Coulomb gauge at this value of z, requires a knowledge of the $C^{[0]}_{m,\alpha}(Z_0 + z')$ over the full interval $z' \in [0, z]$. Thus the z dependence of the vector potential in the Poincaré-Coulomb gauge appears to be *nonlocal* in the $C^{[0]}_{m,\alpha}(Z_0 + z')$. This conclusion is correct if z is large, as it may well be. However if z is small, which will be the case in the vicinity of the expansion point $\boldsymbol{R}_0$, and if we are content with a polynomial expansion in powers of z, which is all that is required to find a polynomial expansion of the Poincaré-Coulomb gauge vector potential in the vicinity of $\boldsymbol{R}_0$, then we can do better. By Taylor's/Maclaurin's theorem we may write

$$\begin{aligned} C^{[0]}_{m,\alpha}(Z_0 + \lambda z) \quad = \quad & C^{[0]}_{m,\alpha}(Z_0) + C^{[1]}_{m,\alpha}(Z_0)(\lambda z) + (1/2!)C^{[2]}_{m,\alpha}(Z_0)(\lambda z)^2 \\ & + (1/3!)C^{[3]}_{m,\alpha}(Z_0)(\lambda z)^3 + \cdots . \end{aligned} \tag{15.7.59}$$

It follows from (7.38) that

$$\begin{aligned} D^{[0]}_{m,\alpha}(z; Z_0) \quad = \quad & \{[1/(m+1)]C^{[0]}_{m,\alpha}(Z_0)\}z + \{[1/(m+2)]C^{[1]}_{m,\alpha}(Z_0)\}z^2 \\ + \quad & \{[1/(m+3)](1/2!)C^{[2]}_{m,\alpha}(Z_0)\}z^3 + \{[1/(m+4)](1/3!)C^{[3]}_{m,\alpha}(Z_0)\}z^4 + \cdots . \end{aligned} \tag{15.7.60}$$

Correspondingly we conclude that, in order to obtain a polynomial expansion of the Poincaré-Coulomb gauge vector potential in the vicinity of $\boldsymbol{R}_0$, only a finite number of the $C^{[n]}_{m,\alpha}(Z_0)$ are required.

The fact that the relation between the Poincaré-Coulomb gauge vector potential and the $C^{[0]}_{m,\alpha}(Z_0)$ depends on the expansion point $\boldsymbol{R}_0$ should not be entirely surprising. The defining requirements (4.2) and (5.1) for the azimuthal-free and Coulomb-gauge vector potentials are required to hold for all z, and are thus *z independent.* By contrast, the defining requirement (2.72) for the Poincaré gauge involves $\boldsymbol{r}$, which in turn involves the expansion point.

15.8 Relations Between Gauges and Associated Symplectic Maps

We have found three general vector potentials specified in terms of on-axis gradients, namely the azimuthal free, symmetric Coulomb, and Poincaré-Coulomb gauge vector potentials. The purpose of this section is to find, in terms of on-axis gradients, the gauge transformation functions that interrelate these vector potentials. And, once these gauge transformation functions are known, there are associated symplectic maps given by relations of the form (6.2.79). See Exercises 6.2.8 and 6.5.3. These results will be of subsequent use. See, for example, Subsection 16.1.3.

15.8.1 Transformation Between Azimuthal Free Gauge and Symmetric Coulomb Gauge

The desired results for this subsection have already been found. The azimuthal free gauge and symmetric Coulomb gauge vector potentials are related by the gauge transformations (5.43) employing the gauge functions $\chi_{m,\alpha}$. And the gauge functions $\chi_{m,\alpha}$ are given in terms of on-axis gradients by the relations (5.75) and (5.76).

15.8.2 Transformation Between Symmetric Coulomb Gauge and Poincaré-Coulomb Gauge

The desired results for this subsection have also already been found. The symmetric Coulomb gauge vector potentials and the Poincaré-Coulomb gauge vector potentials are related by the gauge transformations (7.5) employing the gauge functions $\hat{\chi}^P_{m,\alpha}$. In turn, the gauge functions $\hat{\chi}^P_{m,\alpha}$ are given in terms of the functions $D^{[2k]}_{m,\alpha}$ by the relations (7.19) and (7.20). Finally, the functions $D^{[0]}_{m,\alpha}$ are given in terms of the on-axis gradients by the relations (7.38) or, equivalently, (7.60).

15.8.3 Transformation Between Azimuthal Free Gauge and Poincaré-Coulomb Gauge

Suppose (5.43) is added to (7.5). Doing so gives the result

$$ {}^{P}\boldsymbol{A}^{m,\alpha}(x,y,z;Z_0) = \boldsymbol{A}^{m,\alpha}(x,y,Z_0+z) + \nabla\chi_{m,\alpha} + \nabla\hat{\chi}^{P}_{m,\alpha}. \tag{15.8.1}$$

We see that there is the gauge transformation relation

$$ {}^{P}\boldsymbol{A}^{m,\alpha}(x,y,z;Z_0) = \boldsymbol{A}^{m,\alpha}(x,y,Z_0+z) + \nabla\chi^{P}_{m,\alpha} \tag{15.8.2}$$

where $\chi^{P}_{m,\alpha}$ is the gauge function

$$\chi^{P}_{m,\alpha}(x,y,z;Z_0) = \chi_{m,\alpha}(x,y,Z_0+z) + \hat{\chi}^{P}_{m,\alpha}(x,y,z;Z_0). \tag{15.8.3}$$

15.9 Magnetic Monopole Doublet Example

To validate the numerical methods to be presented in Chapters 17 through 20 and Chapters 23 and 24, it will be useful to have a test problem. One such problem is that of the field of a *monopole doublet*. This field has rapid spatial field variations, thereby posing a challenge to numerical methods, and is also exactly computable in analytic form. The monopole-doublet problem will also illustrate the methods of Sections 2 and 3 of this chapter.

15.9.1 Magnetic Scalar Potential and Magnetic Field

Specifically, suppose two magnetic monopoles having strengths $\pm 4\pi g$ are placed at the (x,y,z) locations

$$\boldsymbol{r}^{+} = (0,a,0), \tag{15.9.1}$$

$$\boldsymbol{r}^{-} = (0,-a,0). \tag{15.9.2}$$

See Figure 9.1, which also shows a circular cylinder with radius R (the surface $\rho = R$). These monopoles generate a scalar potential $\psi(x,y,z)$ described by the relation

$$\begin{aligned}\psi(x,y,z) &= -g[x^2+(y-a)^2+z^2]^{-1/2} + g[x^2+(y+a)^2+z^2]^{-1/2} \\ &= \psi^{+}(x,y,z) + \psi^{-}(x,y,z).\end{aligned} \tag{15.9.3}$$

[Here the notation is such that ψ^{+} is singular at $\boldsymbol{r}^{+}$ and ψ^{-} is singular at $\boldsymbol{r}^{-}$. We have also introduced a factor of 4π in the specification of the monopole strengths so that subsequent formulas such as (9.3) will be free of 4π factors.] Correspondingly, they produce a magnetic field $\boldsymbol{B} = \nabla\psi$ having the components

$$B_x = gx[x^2+(y-a)^2+z^2]^{-3/2} - gx[x^2+(y+a)^2+z^2]^{-3/2}, \tag{15.9.4}$$

$$B_y = g(y-a)[x^2+(y-a)^2+z^2]^{-3/2} - g(y+a)[x^2+(y+a)^2+z^2]^{-3/2}, \tag{15.9.5}$$

$$B_z = gz[x^2+(y-a)^2+z^2]^{-3/2} - gz[x^2+(y+a)^2+z^2]^{-3/2}. \tag{15.9.6}$$

This field is sketched in Figure 9.2. To provide further insight, Figure 9.3 shows the on-axis field component $B_y(x = 0, y = 0, z)$, and Figures 9.4 and 9.5 show the off-axis field components $B_x(\rho = 1/2, \phi = \pi/4, z)$ and $B_z(\rho = 1/2, \phi = \pi/4, z)$. In Cartesian coordinates, the field components B_x and B_z are shown along the line $x = y = \sqrt{2}/4$ cm $\simeq$ .353 cm, $z \in [-\infty, \infty]$. Note that the on-axis field component $B_y(x = 0, y = 0, z)$ falls off as $1/|z|^3$ for large $|z|$, as expected for a doublet of opposite strengths.

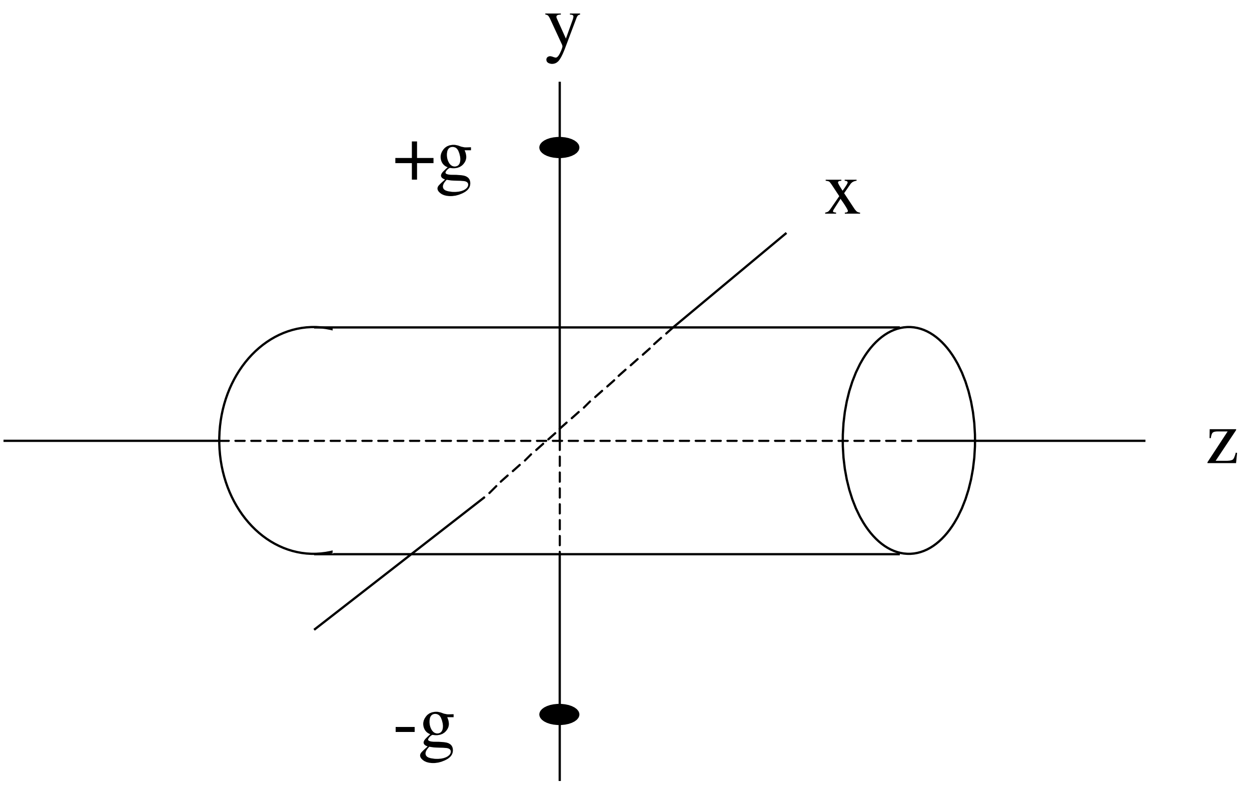


Figure 15.9.1: A monopole doublet consisting of two magnetic monopoles of equal and opposite sign placed on the y axis and centered on the origin. Also shown, for future reference, is a cylinder with circular cross section placed in the interior field.

At this point the reader might object that this field is unphysical since to this time no magnetic monopoles are known to exist. However, as far as an observer inside an interior cylinder of the kind shown in Figure 9.1 is concerned, the field he/she sees is perfectly possible because within the cylinder it obeys $\nabla \cdot \boldsymbol{B} = 0$ and $\nabla \times \boldsymbol{B} = 0$.

The radial component $B_\rho(\rho, \phi, z)$ of $\boldsymbol{B}$ is defined by the relation

$$B_\rho(\rho, \phi, z) = (\cos\phi)B_x + (\sin\phi)B_y. \tag{15.9.7}$$

Recall (2.22). Consequently, using (9.4) and (9.5), we find on the surface $\rho = R$ the result

$$\begin{aligned} B_\rho(R, \phi, z) \;=\;& gR\{[z^2 + R^2 + a^2 - 2aR\sin\phi]^{-3/2} - [z^2 + R^2 + a^2 + 2aR\sin\phi]^{-3/2}\} \\ & -ga\sin\phi\{[z^2 + R^2 + a^2 - 2aR\sin\phi]^{-3/2} + [z^2 + R^2 + a^2 + 2aR\sin\phi]^{-3/2}\}. \end{aligned} \tag{15.9.8}$$

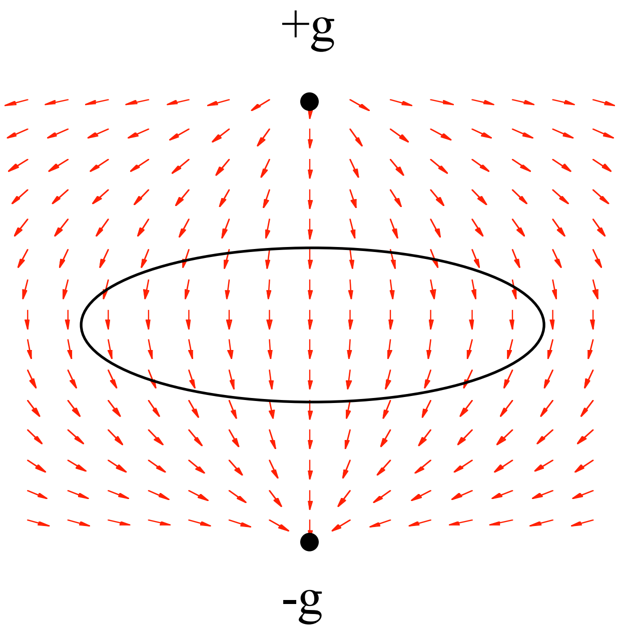


Figure 15.9.2: The interior field of a monopole doublet in the $z = 0$ plane. Also shown is an ellipse whose purpose will become clear in Sections 17.4 and 19.2.

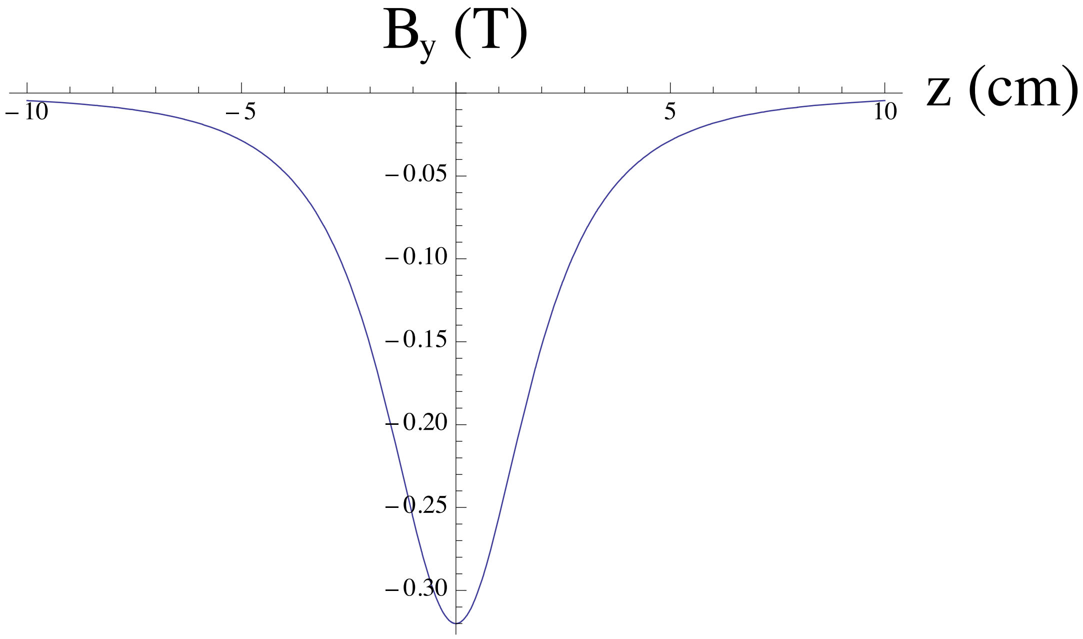


Figure 15.9.3: The on-axis field component $B_y(x = 0, y = 0, z)$ for the monopole doublet in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2. The coordinate z is given in centimeters.

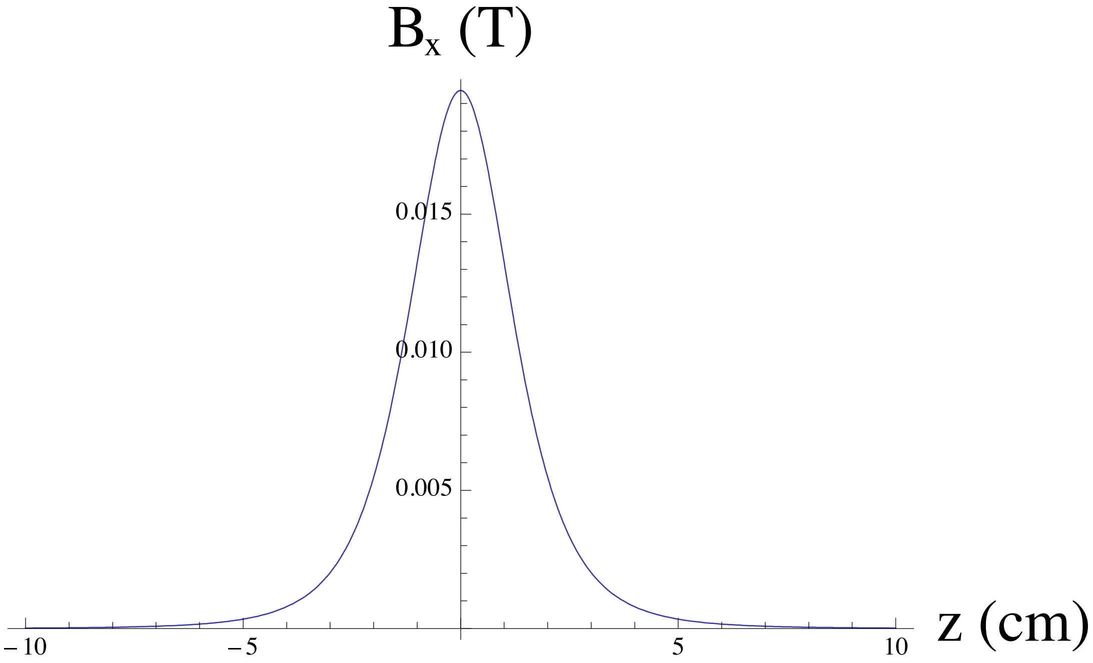


Figure 15.9.4: The field component B_x on the line $\rho = 1/2$ cm, $\phi = \pi/4$, $z \in [-\infty, \infty]$ for the monopole doublet in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2. In Cartesian coordinates, this is the line $x = y \simeq .353$ cm, $z \in [-\infty, \infty]$. The coordinate z is given in centimeters.

To provide a feel for the behavior of $B_\rho(R, \phi, z)$, Figure 9.6 displays $B_\rho(R = 2, \phi, z = 0)$ as a function of ϕ, and Figure 9.7 shows $B_\rho(R = 2, \phi = \pi/2, z)$ as a function of z. We see that the surface field is rather singular. By contrast the fields shown in Figures 9.3 through 9.5, which are those at locations interior to this surface, are less singular. This is to be expected because harmonic functions take their extrema on boundaries.

15.9.2 Analytic On-Axis Gradients for Monopole Doublet

In this subsection we will find analytic expressions for the on-axis gradients for the monopole doublet. In Chapter 19, numerical results for these gradients will be compared against these analytic results.

In view of the form of the expansion (3.33) for ψ, let us seek power series expansions for $\psi^{\pm}(x, y, z)$ in the variable ρ. In the case of ψ^+, for example, we may write

$$\begin{aligned} \psi^+(x, y, z) &= -g[x^2 + (y - a)^2 + z^2]^{-1/2} = -g[(\rho\cos\phi)^2 + (\rho\sin\phi - a)^2 + z^2]^{-1/2} \\ &= -g[a^2 + z^2 - 2a\rho\sin\phi + \rho^2]^{-1/2} \\ &= -g[a^2 + z^2]^{-1/2}\{1 - [2a\rho/(a^2 + z^2)]\sin\phi + \rho^2/(a^2 + z^2)\} \\ &= -g[a^2 + z^2]^{-1/2}[1 - 2wh + h^2]^{-1/2} \end{aligned} \tag{15.9.9}$$

where

$$h = \rho/(a^2 + z^2)^{1/2} \tag{15.9.10}$$

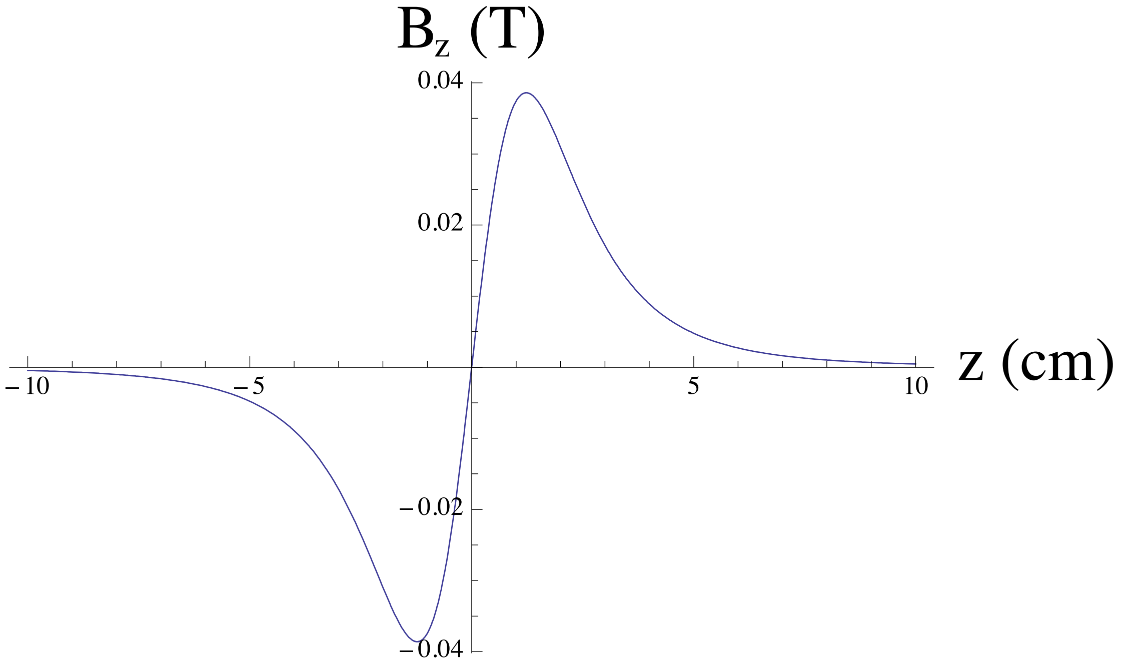


Figure 15.9.5: The field component B_z on the line $\rho = 1/2$ cm, $\phi = \pi/4$, $z \in [-\infty, \infty]$ for the monopole doublet in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2. In Cartesian coordinates, this is the line $x = y \simeq .353$ cm, $z \in [-\infty, \infty]$. The coordinate z is given in centimeters.

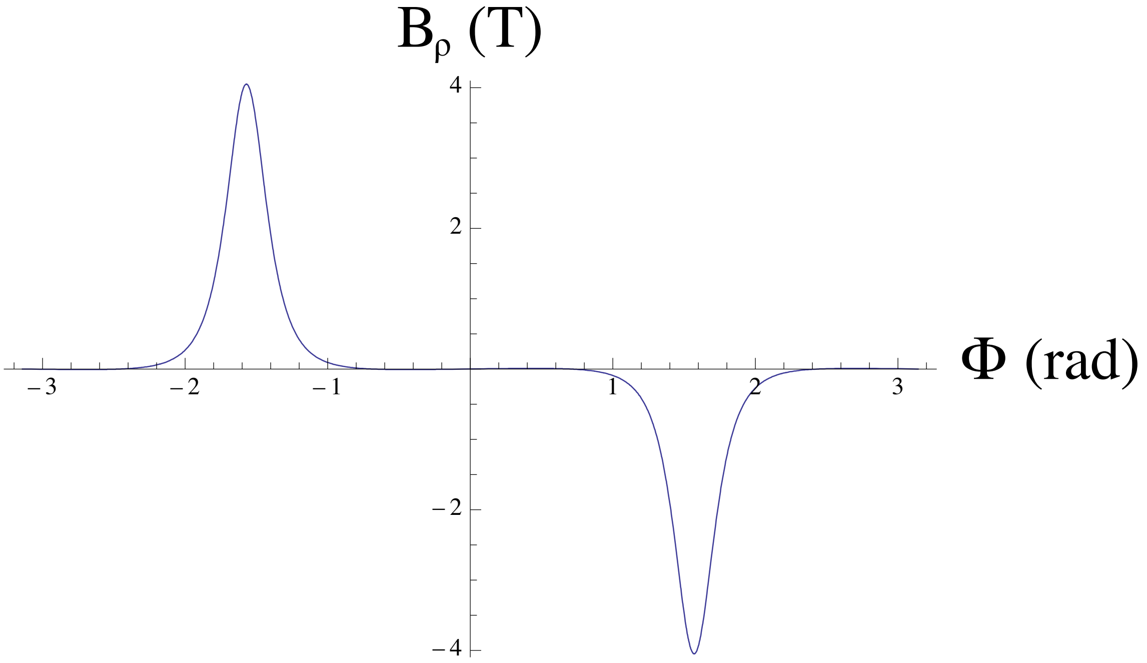


Figure 15.9.6: The quantity $B_\rho(R, \phi, z = 0)$ for the monopole doublet in the case that $R = 2$ cm, $a = 2.5$ cm, and $g = 1$ Tesla-(cm)^2.

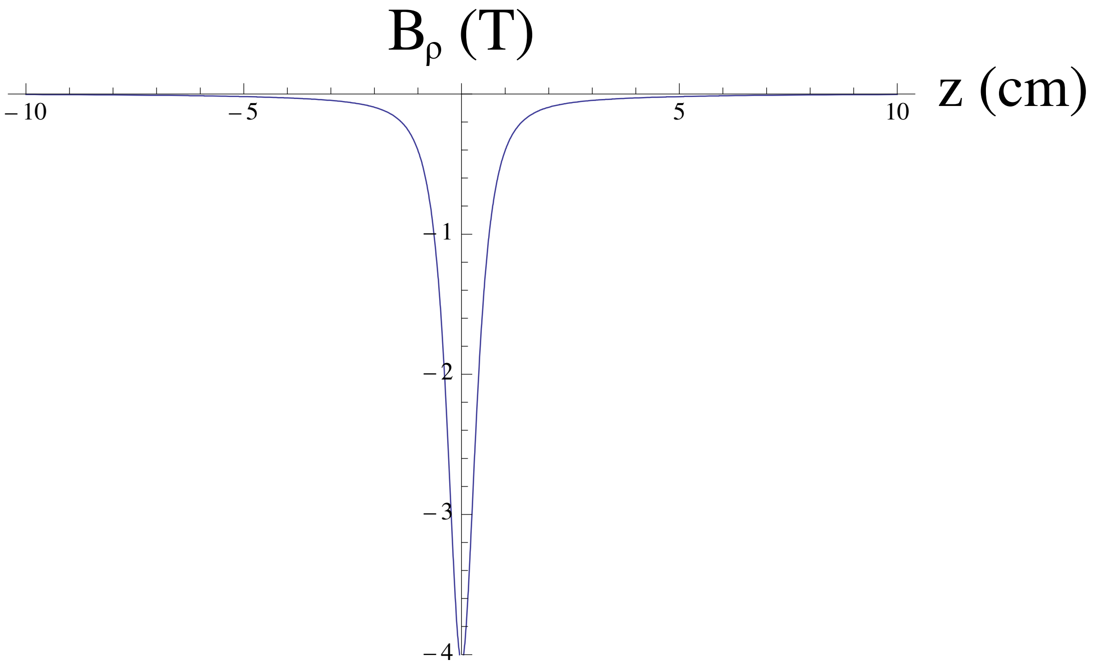


Figure 15.9.7: The quantity $B_\rho(R, \phi = \pi/2, z)$ for the monopole doublet in the case that $R = 2$ cm, $a = 2.5$ cm, and $g = 1$ Tesla-(cm)^2. The coordinate z is given in centimeters.

and

$$w = [a/(a^2 + z^2)^{1/2}] \sin\phi. \tag{15.9.11}$$

Next recall the Legendre polynomial generating function expansion

$$[1 - 2wh + h^2]^{-1/2} = \sum_{m=0}^{\infty} h^m P_m(w). \tag{15.9.12}$$

Combining (9.9) and (9.12) gives the result

$$\begin{aligned}\psi^+(x, y, z) &= -g[x^2 + (y - a)^2 + z^2]^{-1/2} \\ &= -g[a^2 + z^2]^{-1/2} \sum_{m=0}^{\infty} [\rho/(a^2 + z^2)^{1/2}]^m P_m(w).\end{aligned} \tag{15.9.13}$$

Similarly, there is the result

$$\begin{aligned}\psi^-(x, y, z) &= g[x^2 + (y + a)^2 + z^2]^{-1/2} \\ &= g[a^2 + z^2]^{-1/2} \sum_{m=0}^{\infty} [\rho/(a^2 + z^2)^{1/2}]^m P_m(-w).\end{aligned} \tag{15.9.14}$$

It follows, taking into account the parity of the the Legendre polynomials, that ψ has the expansion

$$\psi_(x, y, z) = -2g[a^2 + z^2]^{-1/2} \sum_{n=0}^{\infty} [\rho/(a^2 + z^2)^{1/2}]^{2n+1} P_{2n+1}(w). \tag{15.9.15}$$

From (3.37) through (3.43) we know there are the relations

$$\Psi_0(\rho,z) = [1/(2\pi)]\int_0^{2\pi} d\phi\, \psi(x,y,x), \tag{15.9.16}$$

$$\Psi_{m,c}(\rho,z) = (1/\pi)\int_0^{2\pi} d\phi\, \psi(x,y,x)\cos m\phi, \tag{15.9.17}$$

$$\Psi_{m,s}(\rho,z) = (1/\pi)\int_0^{2\pi} d\phi\, \psi(x,y,x)\sin m\phi. \tag{15.9.18}$$

For the case of the monopole doublet we see from (9.11) and (9.15) that ψ is an odd function of ϕ. Therefore, for the monopole doublet, we conclude that $\Psi_0(\rho,z)=0$ and $\Psi_{m,c}(\rho,z)=0$, and hence

$$C_0^{[0]}(z) = 0,$$
$$C_{m,c}^{[0]}(z) = 0. \tag{15.9.19}$$

And for $\Psi_{m,s}(\rho,z)$ we find the result

$$\Psi_{m,s}(\rho,z) = -(2g/\pi)[a^2+z^2]^{-1/2}\sum_{n=0}^{\infty}[\rho/(a^2+z^2)^{1/2}]^{2n+1}\int_0^{2\pi} d\phi\, P_{2n+1}(w)\sin m\phi. \tag{15.9.20}$$

To analyze the integral that occurs on the right side of (9.20), introduce the notation

$$\beta(z) = a/(a^2+z^2)^{1/2} \tag{15.9.21}$$

so that

$$w = \beta\sin\phi. \tag{15.9.22}$$

With this notation, we must study integrals of the form

$$c_{m',m} = \int_0^{2\pi} d\phi\, P_{m'}(\beta\sin\phi)\sin m\phi \tag{15.9.23}$$

with m' odd.

To begin, we know from the Taylor expansion for the Legendre polynomials that (for odd m')

$$P_{m'}(w) = \{[(2m')!]/[2^{m'}(m'!)^2]\}w^{m'} + \text{lower odd powers of } w. \tag{15.9.24}$$

We also know (again for odd m') that

$$(\sin\phi)^{m'} = (-1)^{(m'-1)/2}(1/2)^{m'-1}\sin m'\phi + \text{lower odd frequency sinusoidal terms.} \tag{15.9.25}$$

It follows that

$$c_{m',m} = 0 \text{ for } m \text{ even}, \tag{15.9.26}$$

$$c_{m',m} = 0 \text{ for } m' < m, \tag{15.9.27}$$

and (with m odd)

$$\begin{aligned} c_{m,m} &= (-1)^{(m-1)/2}\{[(2m)!]/[2^m(m!)^2]\}(1/2)^{m-1}\beta^m\int_0^{2\pi} d\phi\ \sin^2(m\phi) \\ &= (-1)^{(m-1)/2}\pi\{[(2m)!]/[2^{2m-1}(m!)^2]\}\beta^m. \end{aligned} \tag{15.9.28}$$

An immediate conclusion (consistent with symmetry considerations, see Subsection 3.5) is that $\Psi_{m,s}(\rho,z)=0$ for even m, and hence

$$C^{[0]}_{m,s}(z) = 0 \text{ for } m \text{ even}. \tag{15.9.29}$$

Let us insert the results obtained so far into (9.20). Doing so yields the relation

$$\begin{aligned} \Psi_{m,s}(\rho,z) &= -(2g/\pi)[a^2+z^2]^{-1/2}[\rho/(a^2+z^2)^{1/2}]^m c_{m,m} \\ &\quad -(2g/\pi)[a^2+z^2]^{-1/2}\sum_{n>n'}^{\infty}[\rho/(a^2+z^2)^{1/2}]^{2n+1}c_{2n+1,m} \end{aligned} \tag{15.9.30}$$

where (for m odd)

$$2n'+1 = m. \tag{15.9.31}$$

Also, we know from (3.39) that

$$C^{[0]}_{m,s}(z) = \lim_{\rho\to 0}(1/\rho^m)\Psi_{m,s}(\rho,z). \tag{15.9.32}$$

We conclude that (for m odd)

$$\begin{aligned} C^{[0]}_{m,s}(z) &= -(2g/\pi)(a^2+z^2)^{-1/2}(a^2+z^2)^{-m/2}c_{m,m} \\ &= -g(-1)^{(m-1)/2}\{[(2m)!]/[2^{2m-2}(m!)^2a^{m+1}]\}\beta^{2m+1}(z). \end{aligned} \tag{15.9.33}$$

Note that the $C^{[0]}_{m,s}(z)$ have the asymptotic fall off

$$|C^{[0]}_{m,s}(z)| \sim 1/|z|^{2m+1} \tag{15.9.34}$$

for large $|z|$.

Suppose we wish to retain, in the expansion of the Hamiltonian H appearing in (1.3), homogeneous polynomials through degree 8. Then, as we see from (1.4), we must retain homogeneous polynomials in the variables x,y through degree 7 in the expansions of A_x and A_y, and homogeneous polynomials in the variables x,y through degree 8 in the expansion of A_z. Inspection of (4.21) through (4.26) or (5.89) through (5.94) shows that for the cases $m=0$ or m odd we then need the $C^{[n]}_{m,\alpha}(z)$ with $(m+n)\le 7$. And for the cases of even m we need the $C^{[n]}_{m,\alpha}(z)$ with $(m+n)\le 8$. In particular, for the case of the monopole doublet, (for which only the generalized gradients with $\alpha=s$ and m odd are nonzero) we need the following functions:

$$\begin{aligned} &C^{[0]}_{1,s}(z),\ C^{[1]}_{1,s}(z),\ C^{[2]}_{1,s}(z),\ C^{[3]}_{1,s}(z),\ C^{[4]}_{1,s}(z),\ C^{[5]}_{1,s}(z),\ C^{[6]}_{1,s}(z); \\ &C^{[0]}_{3,s}(z),\ C^{[1]}_{3,s}(z),\ C^{[2]}_{3,s}(z),\ C^{[3]}_{3,s}(z),\ C^{[4]}_{3,s}(z); \\ &C^{[0]}_{5,s}(z),\ C^{[1]}_{5,s}(z),\ C^{[2]}_{5,s}(z); \\ &C^{[0]}_{7,s}(z). \end{aligned} \tag{15.9.35}$$

(See Exercise 7.1.) Graphs of a selected few of these functions, for the monopole doublet in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2, are shown in Figures 9.8 through 9.15. In these plots z has units of centimeters. Evidently the $C_{m,s}^{[0]}$ become ever more highly peaked with increasing m. Fortunately, when working through some fixed degree, we need fewer derivatives with increasing m. Note that we expect that the function $C_{m,s}^{[n]}(z)$ should have n zeroes. This is indeed the case, but some of these zeroes can be hidden in the tails. Figure 9.10 is an enlargement of Figure 9.9 showing a hidden zero for the case of $C_{1,s}^{[6]}(z)$.

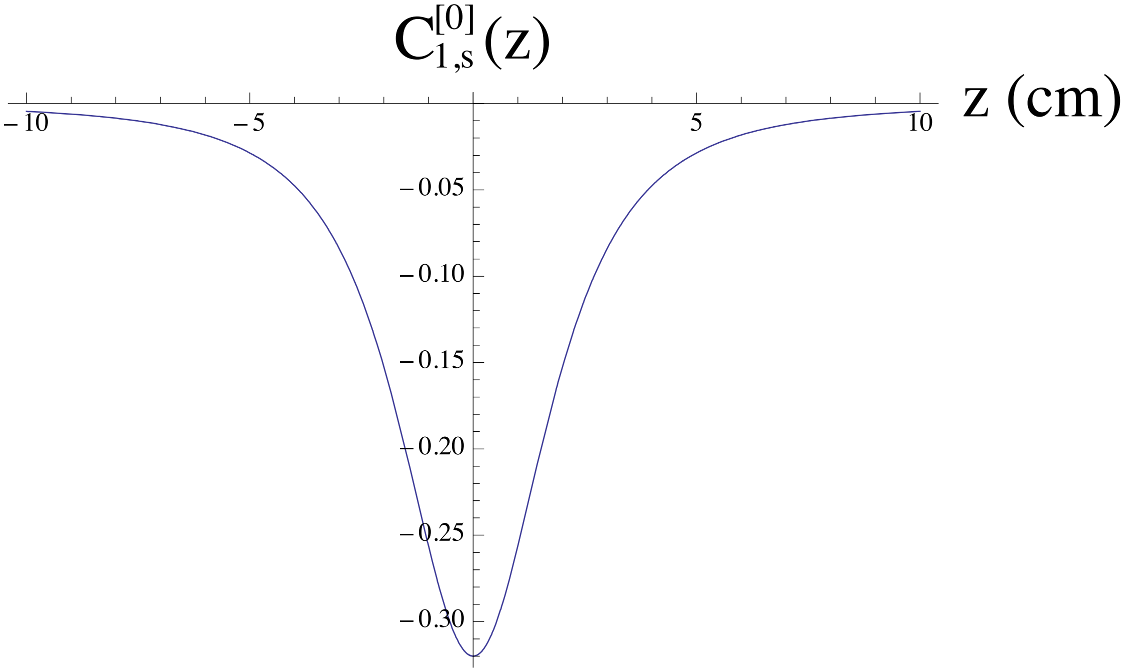


Figure 15.9.8: The on-axis gradient function $C_{1,s}^{[0]}$ for the monopole doublet in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2.

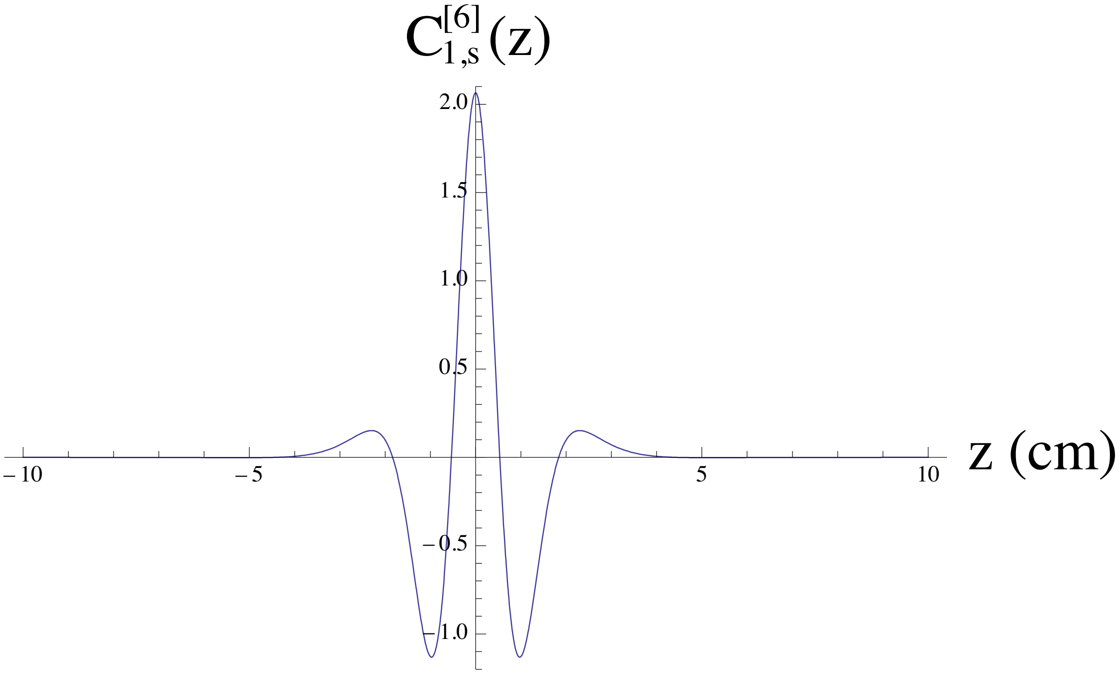


Figure 15.9.9: The on-axis gradient function $C_{1,s}^{[6]}$ for the monopole doublet in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2.

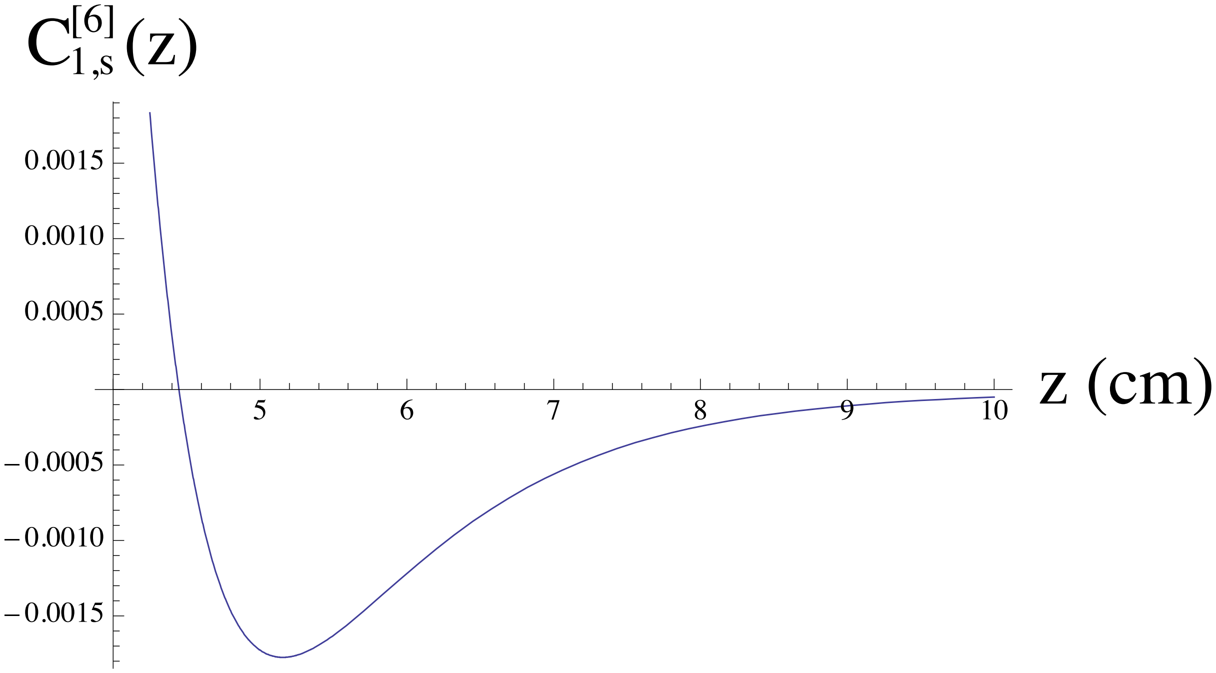


Figure 15.9.10: An enlargement of a portion of Figure 9.9 showing a zero hidden in a tail.

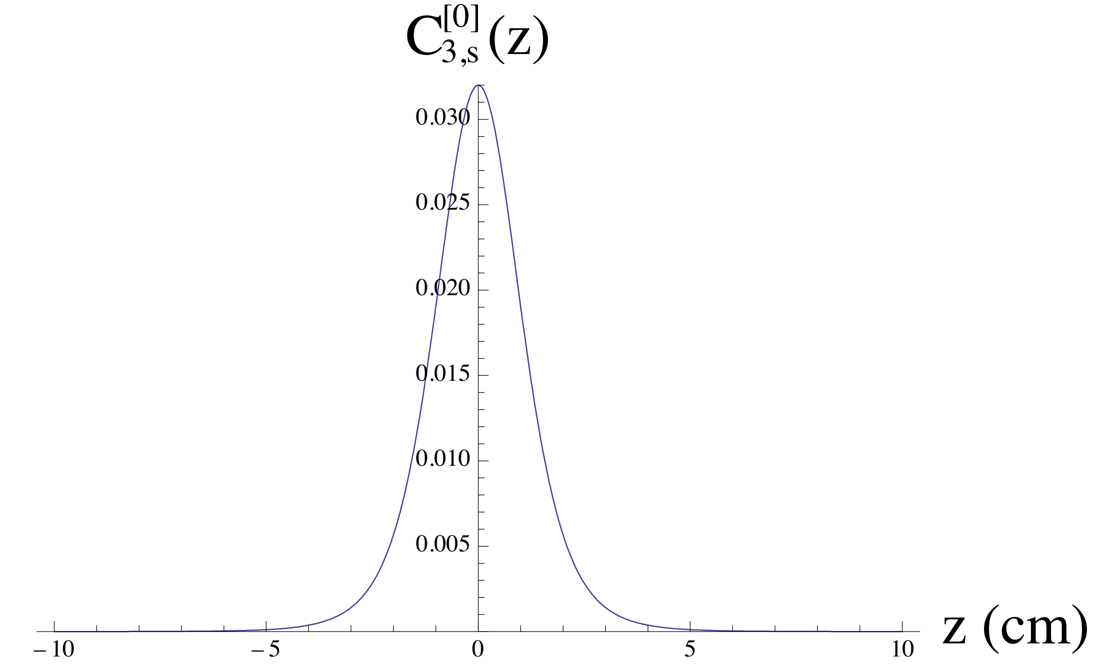


Figure 15.9.11: The on-axis gradient function $C_{3,s}^{[0]}$ for the monopole doublet in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2.

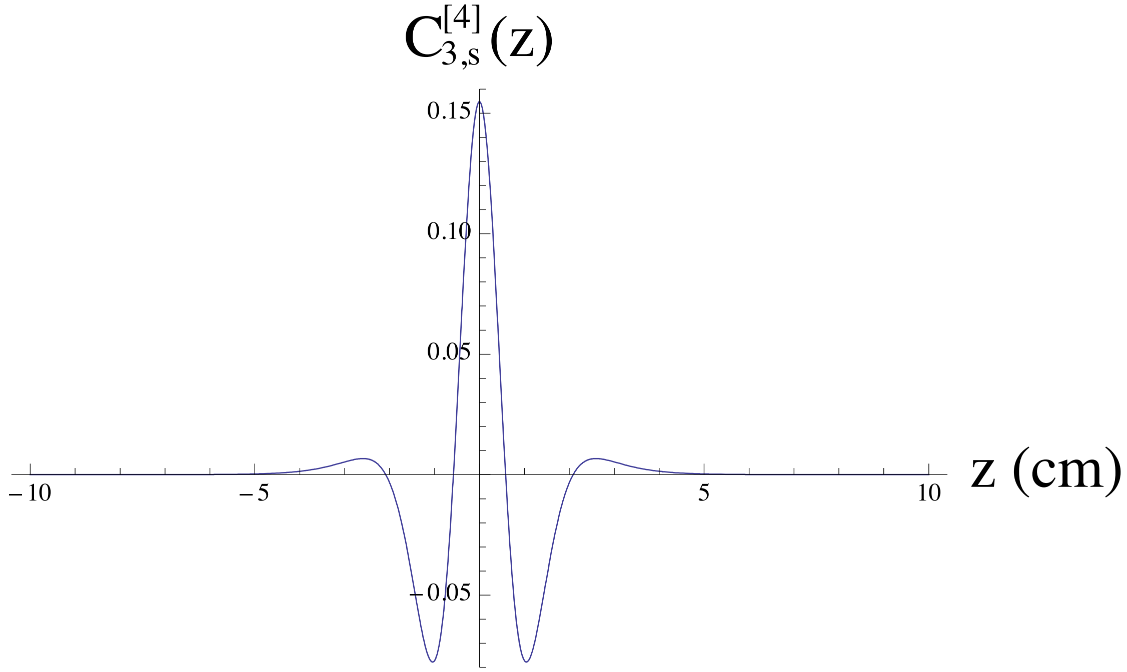


Figure 15.9.12: The on-axis gradient function $C_{3,s}^{[4]}$ for the monopole doublet in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2.

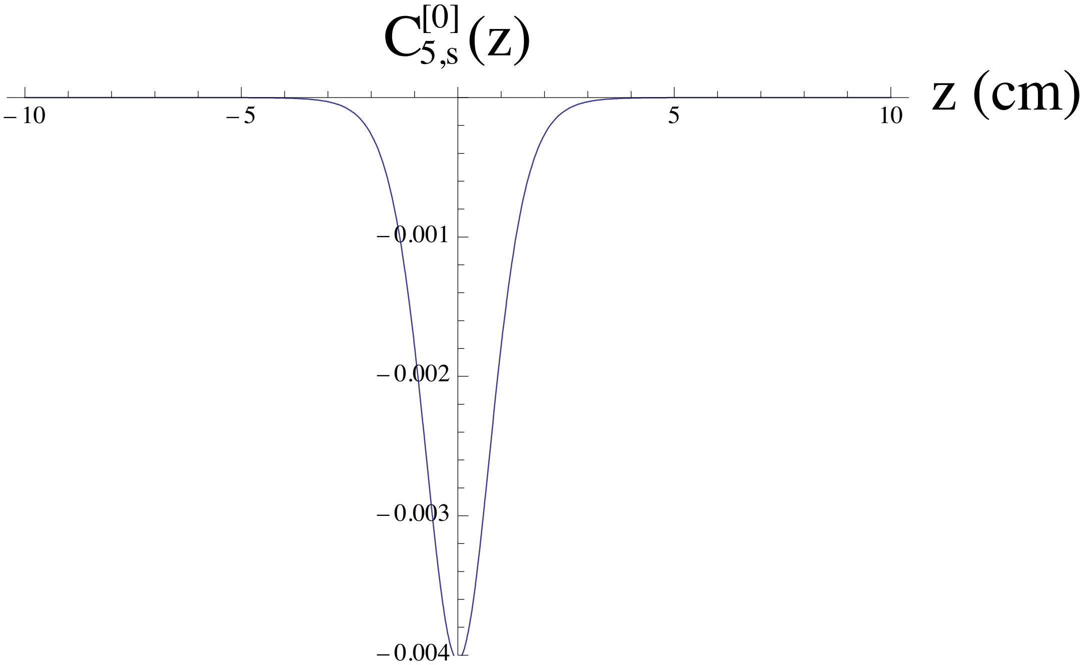


Figure 15.9.13: The on-axis gradient function $C^{[0]}_{5,s}$ for the monopole doublet in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2.

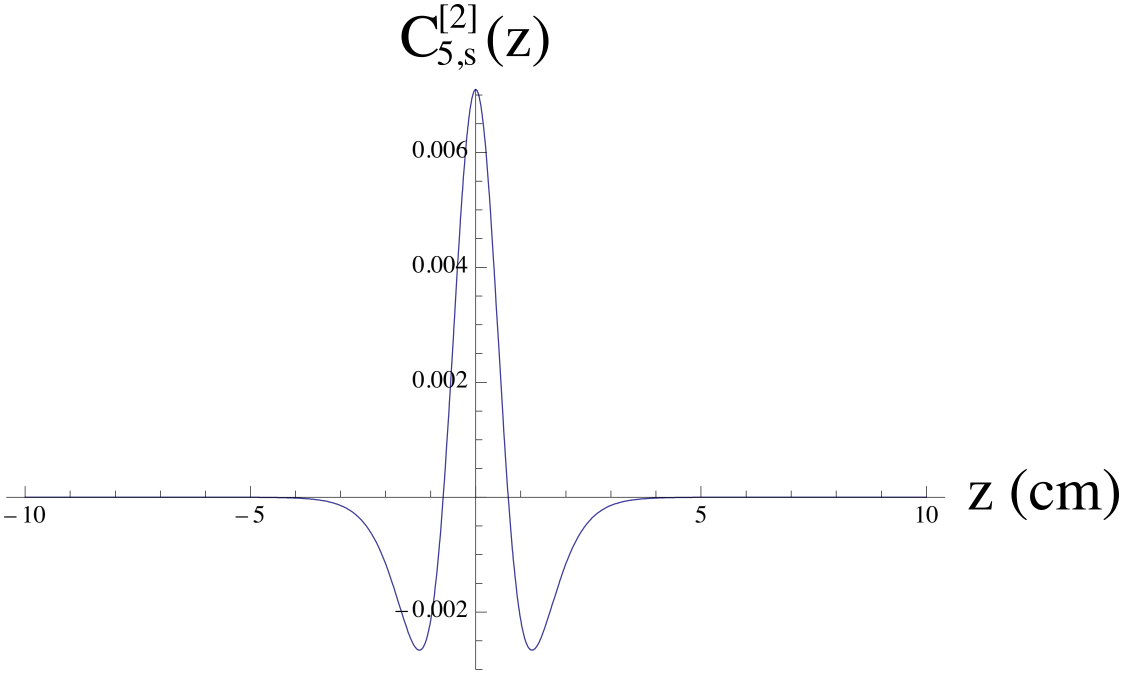


Figure 15.9.14: The on-axis gradient function $C^{[2]}_{5,s}$ for the monopole doublet in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2.

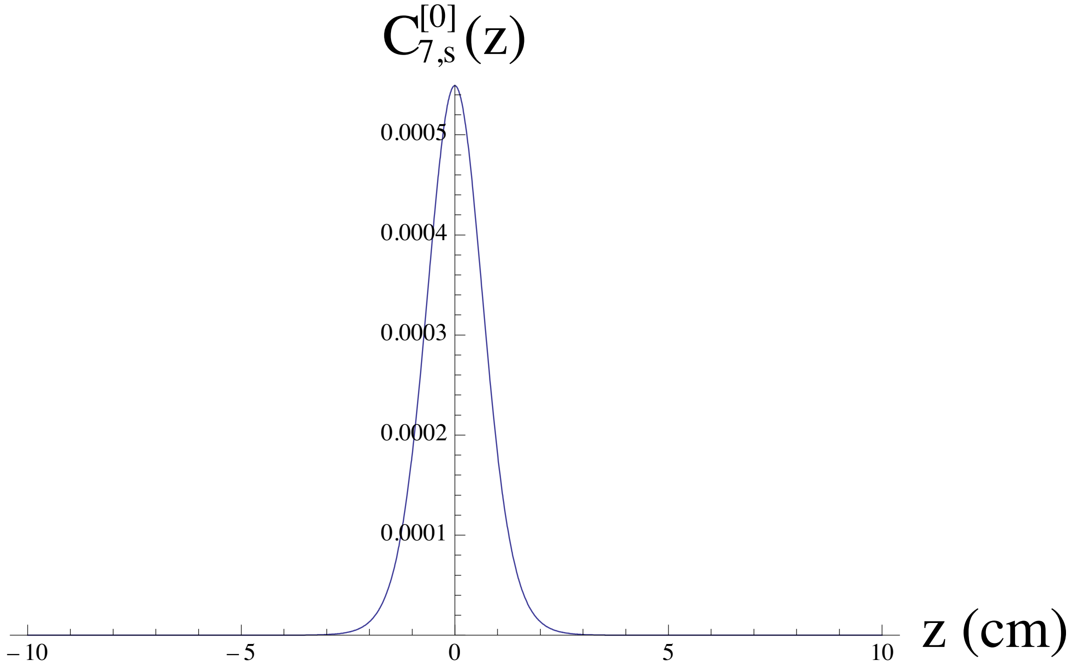


Figure 15.9.15: The on-axis gradient function $C^{[0]}_{7,s}$ for the monopole doublet in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2.

Exercises

15.9.1. Explain why Figures 9.3 and 9.8 should be the same.

15.10 Minimum Vector Potential for Magnetic Monopole Doublet

The purpose of this section is to find the first few terms in the expansion of the minimum (Poincaré-Coulomb gauge) vector potential for a magnetic monopole doublet. We will first find the minimum vector potential in terms of the scalar potential and its associated magnetic field using the results of Subsection 2.7. Then we will find the minimum vector potential in terms of the on-axis gradients using the results of Section 7.

In particular, we will be interested in expansions for the fringe-field regions and in the midplane. Suppose the doublet is located at the origin $\boldsymbol{R} = (0, 0, 0)$ as in Subsection 9.1, and we seek an expansion about the mid-plane point $\boldsymbol{R}_0 = (X_0, 0, Z_0)$. If $Z_0 \ll 0$, we will obtain an expansion in the leading fringe-field region, and if $Z_0 \gg 0$, we will obtain an expansion in the trailing fringe-field region. Moreover, if $X_0 = 0$, the expansion will be on axis; and setting $X_0 \neq 0$ allows for expansion about a point on the (curved) design orbit. See Section 23.3 and Figures 23.3.1 and 24.1.1.

15.10.1 Computation from the Scalar Potential and Associated Magnetic Field

We begin by specifying the scalar potential $\Psi(\boldsymbol{R})$. According to (9.3), it is given by the relation

$$\Psi(X,Y,Z) = -g[X^2+(Y-a)^2+Z^2]^{-1/2} + g[X^2+(Y+a)^2+Z^2]^{-1/2}. \qquad (15.10.1)$$

Next, according to (2.4), $\psi(\boldsymbol{r})$ is given by the relation

$$\begin{aligned}\psi(x,y,z) &= \Psi(\boldsymbol{R}_0+\boldsymbol{r})\\ &= -g[(X_0+x)^2+(y-a)^2+(Z_0+z)^2]^{-1/2}\\ &\quad +g[(X_0+x)^2+(y+a)^2+(Z_0+z)^2]^{-1/2}.\end{aligned} \qquad (15.10.2)$$

The right side of (10.2) can now be expanded in powers of the components of $\boldsymbol{r}$. Doing so yields, for the first few terms, the result

$$\begin{aligned}\psi(\boldsymbol{r};X_0,Z_0) &= [-2ga/(X_0^2+Z_0^2+a^2)^{3/2}]y\\ &\quad +[6ga/(X_0^2+Z_0^2+a^2)^{5/2}][y(X_0x+Z_0z)]\\ &\quad + \text{terms of order 3 and higher.}\end{aligned} \qquad (15.10.3)$$

Note that $\Psi(X_0,0,Z_0)$ vanishes so that there is no constant term in the expansion (10.3). We observe that the first term in (10.3) falls off as $(1/|X_0|)^3$ or $(1/|Z_0|)^3$ for large $|X_0|$ or $|Z_0|$, and the second falls off as $(1/|X_0|)^4$ or $(1/|Z_0|)^4$. In general, successive terms fall off with ever increasing powers of $(1/|X_0|)$ or $(1/|Z_0|)$.

Let us compute the magnetic field $\boldsymbol{B}$ associated with the first two terms in (10.3). We find the result

$$\begin{aligned}\boldsymbol{B}(\boldsymbol{r};X_0,Z_0) &= -[2ga/(X_0^2+Z_0^2+a^2)^{3/2}]\boldsymbol{e}_y\\ &\quad +[6ga/(X_0^2+Z_0^2+a^2)^{5/2}](X_0x+Z_0z)\boldsymbol{e}_y\\ &\quad +[6ga/(X_0^2+Z_0^2+a^2)^{5/2}][y(X_0\boldsymbol{e}_x+Z_0\boldsymbol{e}_z)].\end{aligned} \qquad (15.10.4)$$

Next let us find the minimum vector potential $\boldsymbol{A}^{\min}$ associated with the first two terms in (10.3). Begin by decomposing $\boldsymbol{B}$ into homogeneous polynomials by rewriting (10.4) in the form (2.109) with

$$\boldsymbol{B}^0(\boldsymbol{r};X_0,Z_0) = -[2ga/(X_0^2+Z_0^2+a^2)^{3/2}]\boldsymbol{e}_y \qquad (15.10.5)$$

and

$$\boldsymbol{B}^1(\boldsymbol{r};X_0,Z_0) = [6ga/(X_0^2+Z_0^2+a^2)^{5/2}][(X_0x+Z_0z)\boldsymbol{e}_y+y(X_0\boldsymbol{e}_x+Z_0\boldsymbol{e}_z)]. \quad (15.10.6)$$

The minimum vector potential associated with this magnetic field is given by the relations (2.109) through (2.111). Working out the indicated cross products yields the results

$$\boldsymbol{A}^{\min 1}(\boldsymbol{r};X_0,Z_0) = [ga/(X_0^2+Z_0^2+a^2)^{3/2}](-z\boldsymbol{e}_x+x\boldsymbol{e}_z), \qquad (15.10.7)$$

$$\boldsymbol{A}^{\min 2}(\boldsymbol{r}; X_0, Z_0) = [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times$$
$$[(Z_0y^2 - Z_0z^2 - X_0xz)\boldsymbol{e}_x + (X_0yz - Z_0xy)\boldsymbol{e}_y + (X_0x^2 + Z_0xz - X_0y^2)\boldsymbol{e}_z]. \tag{15.10.8}$$

Simple calculation verifies that there are the relations

$$\nabla \times \boldsymbol{A}^{\min 1}(\boldsymbol{r}; X_0, Z_0) = \boldsymbol{B}^0(\boldsymbol{r}; X_0, Z_0), \tag{15.10.9}$$

$$\nabla \times \boldsymbol{A}^{\min 2}(\boldsymbol{r}; X_0, Z_0) = \boldsymbol{B}^1(\boldsymbol{r}; X_0, Z_0), \tag{15.10.10}$$

as desired. We note that $\boldsymbol{A}^{\min 1}$ falls off as $(1/|X_0|)^3$ or $(1/|Z_0|)^3$ for large $|X_0|$ or $|Z_0|$, and $\boldsymbol{A}^{\min 2}$ falls off as $(1/|X_0|)^4$ or $(1/|Z_0|)^4$. In general, successive $\boldsymbol{A}^{\min n}$ fall off with ever increasing powers of $(1/|X_0|)$ or $(1/|Z_0|)$.

15.10.2 Computation from the On-Axis Gradients

The $m = 1$ and $\alpha = s$ vector potential in the symmetric Coulomb gauge of Section 5 is given in terms of on-axis gradients by the relation

$$\hat{A}_x^{1,s}(x, y, Z_0+z) = (1/4)(x^2-y^2)C_{1,s}^{[1]}(Z_0+z)-(1/48)(x^4-y^4)C_{1,s}^{[3]}(Z_0+z)+\cdots, \tag{15.10.11}$$

$$\hat{A}_y^{1,s}(x, y, Z_0+z) = (1/2)xyC_{1,s}^{[1]}(Z_0+z) - (1/24)(x^3y+xy^3)C_{1,s}^{[3]}(Z_0+z)+\cdots, \tag{15.10.12}$$

$$\begin{aligned}\hat{A}_z^{1,s}(x, y, Z_0+z) &= -xC_{1,s}^{[0]}(Z_0+z) + (1/8)(x^3+xy^2)C_{1,s}^{[2]}(Z_0+z)\\ &\quad -(1/192)(x^5+2x^3y^2+xy^4)C_{1,s}^{[4]}(Z_0+z)+\cdots.\end{aligned} \tag{15.10.13}$$

See (5.97) through (5.99). If we expand $\hat{\boldsymbol{A}}^{1,s}(x, y, Z_0+z)$ in powers of z, organize the results into homogeneous polynomials, and retain only terms of degree less than 3, we find the results

$$\hat{A}_x^{1,s}(x, y, Z_0+z) = (1/4)(x^2-y^2)C_{1,s}^{[1]}(Z_0)+\cdots, \tag{15.10.14}$$

$$\hat{A}_y^{1,s}(x, y, Z_0+z) = (1/2)xyC_{1,s}^{[1]}(Z_0)+\cdots, \tag{15.10.15}$$

$$\hat{A}_z^{1,s}(x, y, Z_0+z) = -xC_{1,s}^{[0]}(Z_0) - xzC_{1,s}^{[1]}(Z_0)+\cdots. \tag{15.10.16}$$

The gauge function $\hat{\chi}_{1,s}^P$ that relates the Coulomb-gauge vector potential $\hat{\boldsymbol{A}}^{1,s}$ and the Poincaré-Coulomb gauge vector potential ${}^P\!\boldsymbol{A}^{1,s}$ is given by the relation

$$\begin{aligned}\hat{\chi}_{1,s}^P(x, y, z; Z_0) &= \cos(\phi)\sum_{k=0}^{\infty}(-1)^k\frac{1!}{2^{2k}k!(k+1)!}D_{1,s}^{[2k]}(z; Z_0)\rho^{2k+1}\\ &= \rho\cos(\phi)[D_{1,s}^{[0]}(z; Z_0) - (1/8)\rho^2D_{1,s}^{[2]}(z; Z_0) + (*)\rho^4D_{1,s}^{[4]}(z; Z_0)+\cdots]\\ &= x[D_{1,s}^{[0]}(z; Z_0) - (1/8)\rho^2D_{1,s}^{[2]}(z; Z_0) + (*)\rho^4D_{1,s}^{[4]}(z; Z_0)+\cdots].\end{aligned} \tag{15.10.17}$$

See (7.5) and (7.20). We have also found, see (7.60), the relation

$$\begin{aligned} D^{[0]}_{1,s}(z;Z_0) &= \{[1/(1+1)]C^{[0]}_{1,s}(Z_0)\}z + \{[1/(1+2)]C^{[1]}_{1,s}(Z_0)\}z^2 \\ &\quad + \{[1/(1+3)](1/2!)C^{[2]}_{1,s}(0)\}z^3 + \{[1/(1+4)](1/3!)C^{[3]}_{1,s}(Z_0)\}z^4 + \cdots \\ &= (1/2)C^{[0]}_{1,s}(Z_0)z + (1/3)C^{[1]}_{1,s}(Z_0)z^2 + (1/8)C^{[2]}_{1,s}(Z_0)z^3 + (1/30)C^{[3]}_{1,s}(Z_0)z^4 + \cdots , \end{aligned} \tag{15.10.18}$$

from which it follows that

$$D^{[2]}_{1,s}(z;Z_0) = (2/3)C^{[1]}_{1,s}(Z_0) + (3/4)C^{[2]}_{1,s}(Z_0)z + (2/5)C^{[3]}_{1,s}(Z_0)z^2 + \cdots , \tag{15.10.19}$$

$$D^{[4]}_{1,s}(z;Z_0) = (4/5)C^{[3]}_{1,s}(Z_0) + \cdots . \tag{15.10.20}$$

Inserting these results into (10.17) and collecting terms of like degree give the final homogeneous polynomial expansion

$$\begin{aligned} \hat{\chi}^P_{1,s}(x,y,z;Z_0) &= x[(1/2)C^{[0]}_{1,s}(Z_0)z + (1/3)C^{[1]}_{1,s}(Z_0)z^2 + (1/8)C^{[2]}_{1,s}(Z_0)z^3 + \cdots] \\ &\quad - (1/8)x\rho^2[(2/3)C^{[1]}_{1,s}(Z_0) + (3/4)C^{[2]}_{1,s}(Z_0)z + (2/5)C^{[3]}_{1,s}(Z_0)z^2 + \cdots] \\ &\quad + (*)x\rho^4[(4/5)C^{[3]}_{1,s}(Z_0) + \cdots] + \cdots \\ &= \{[xz(1/2)]C^{[0]}_{1,s}(Z_0)\} + \{[xz^2(1/3) - \ (1/8)x\rho^2(2/3)]C^{[1]}_{1,s}(Z_0)\} + \cdots \\ &= \{[(1/2)xz]C^{[0]}_{1,s}(Z_0)\} + \{[(1/3)xz^2 - (1/12)x\rho^2]C^{[1]}_{1,s}(Z_0)\} + \cdots . \end{aligned} \tag{15.10.21}$$

The next step is to compute the $\Delta\boldsymbol{A}^{1,s}$ defined by (7.46). From (10.21) we find the results

$$\begin{aligned} \Delta A^{1,s}_x &= (\partial/\partial x)\hat{\chi}^P_{1,s}(x,y,z;Z_0) \\ &= \{[(1/2)z]C^{[0]}_{1,s}(Z_0)\} + \{[(1/3)z^2 - (1/12)(3x^2+y^2)]C^{[1]}_{1,s}(Z_0)\} + \cdots , \end{aligned} \tag{15.10.22}$$

$$\begin{aligned} \Delta A^{1,s}_y &= (\partial/\partial y)\hat{\chi}^P_{1,s}(x,y,z;Z_0) \\ &= +\{[-(1/6)xy]C^{[1]}_{1,s}(Z_0)\} + \cdots , \end{aligned} \tag{15.10.23}$$

$$\begin{aligned} \Delta A^{1,s}_z &= (\partial/\partial z)\hat{\chi}^P_{1,s}(x,y,z;Z_0) \\ &= \{[(1/2)x]C^{[0]}_{1,s}(Z_0)\} + \{[(2/3)xz]C^{[1]}_{1,s}(Z_0)\} + \cdots . \end{aligned} \tag{15.10.24}$$

Finally, we may obtain ${}^P\!\boldsymbol{A}^{1,s}$ with the aid of (7.46) and (10.14) through (10.16). Doing so gives the results

$$\begin{aligned} {}^P\!A^{1,s}_x(x,y,z;Z_0) &= (1/4)(x^2-y^2)C^{[1]}_{1,s}(Z_0) + \{[(1/2)z]C^{[0]}_{1,s}(Z_0)\} \\ &\quad + \{[(1/3)z^2 - (1/12)(3x^2+y^2)C^{[1]}_{1,s}(Z_0)\} + \cdots \\ &= \{[(1/2)z]C^{[0]}_{1,s}(Z_0)\} + \{[(1/3)z^2 - (1/3)y^2]C^{[1]}_{1,s}(Z_0)\} + \cdots , \end{aligned} \tag{15.10.25}$$

$$\begin{aligned}{}^{P}\!A_y^{1,s}(x,y,z;Z_0) &= (1/2)xyC_{1,s}^{[1]}(Z_0) + \{[-(1/6)xy]C_{1,s}^{[1]}(Z_0)\} + \cdots \\ &= \{[(1/3)xy]C_{1,s}^{[1]}(Z_0)\} + \cdots , \end{aligned} \qquad (15.10.26)$$

$$\begin{aligned}{}^{P}\!A_z^{1,s}(x,y,z;Z_0) &= -xC_{1,s}^{[0]}(Z_0) - xzC_{1,s}^{[1]}(Z_0) + \{[(1/2)x]C_{1,s}^{[0]}(Z_0)\} \\ &\quad + \{[(2/3)xz]C_{1,s}^{[1]}(Z_0)\} \cdots \\ &= \{[-(1/2)x]C_{1,s}^{[0]}(Z_0)\} - \{[(1/3)xz]C_{1,s}^{[1]}(Z_0)\} + \cdots . \end{aligned} \qquad (15.10.27)$$

How do the results (10.25) through (10.27) compare with the results (10.7) and (10.8) found in the previous subsection? Suppose $\boldsymbol{A}^{\min 1}(\boldsymbol{r}; X_0, Z_0)$ and $\boldsymbol{A}^{\min 2}(\boldsymbol{r}; X_0, Z_0)$ as given by (10.7) and (10.8) are evaluated at $X_0 = 0$. So doing gives the results

$$\boldsymbol{A}^{\min 1}(\boldsymbol{r}; X_0, Z_0)|_{X_0=0} = [ga/(a^2 + Z_0^2)^{3/2}](-z\boldsymbol{e}_x + x\boldsymbol{e}_z), \qquad (15.10.28)$$

$$\begin{aligned}\boldsymbol{A}^{\min 2}(\boldsymbol{r}; X_0, Z_0)|_{X_0=0} = [-2ga/(a^2 + Z_0^2)^{5/2}] \times \\ [(Z_0y^2 - Z_0z^2)\boldsymbol{e}_x + (-Z_0xy)\boldsymbol{e}_y + (Z_0xz)\boldsymbol{e}_z].\end{aligned} \qquad (15.10.29)$$

Also, from (8.21) and (8.33), we find the results

$$C_{1,s}^{[0]}(z) = -g\{[2!]/[a^2]\}\beta^3(z) = -2ga/(a^2 + z^2)^{3/2}, \qquad (15.10.30)$$

from which it follows that

$$C_{1,s}^{[0]}(Z_0) = -2ga/(a^2 + Z_0^2)^{3/2}, \qquad (15.10.31)$$

$$C_{1,s}^{[1]}(Z_0) = 6gaZ_0/(a^2 + Z_0^2)^{5/2}. \qquad (15.10.32)$$

Consequently, (10.28) and (10.29) can be rewritten in the form

$$\boldsymbol{A}^{\min 1}(\boldsymbol{r}; X_0, Z_0)|_{X_0=0} = -(1/2)C_{1,s}^{[0]}(Z_0)(-z\boldsymbol{e}_x + x\boldsymbol{e}_z), \qquad (15.10.33)$$

$$\boldsymbol{A}^{\min 2}(\boldsymbol{r}; X_0, Z_0)|_{X_0=0} = (-1/3)C_{1,s}^{[1]}(Z_0)[(y^2 - z^2)\boldsymbol{e}_x + (-xy)\boldsymbol{e}_y + (xz)\boldsymbol{e}_z]. \qquad (15.10.34)$$

Comparison of (10.33) and (10.34) with (10.25) through (10.27) reveals that (10.33) agrees with the first-degree terms in (10.25) through (10.27), and (10.34) agrees with the second-degree terms in (10.25) through (10.27). Therefore the on-axis minimum vector potential expansion computed from the scalar potential and associated magnetic field agrees with the on-axis minimum vector potential expansion computed from the on-axis gradients, as desired and required.

Finally we remind the reader that, although we have been considering the case of a monopole doublet field, the relations (10.11) through (10.13) and (10.21) through (10.27) hold for *any* $m = 1$ and $\alpha = s$ magnetic field no matter what its source. The same is true for the relation

$$C_{1.s}^{[0]}(z) = B_y(0, 0, z). \qquad (15.10.35)$$

Recall (3.59).

15.11 Calculation of Scalar and Vector Potentials from Current Data

The previous sections in this chapter have studied how magnetic fields and vector potentials may be described in terms of terms of scalar potentials and their various expansions. In this section we will explore the relation between vector and scalar potentials and the currents that produce them.

15.11.1 Calculation of Vector Potential from Current Data

15.11.1.1 Preliminary Steps

The vector potential $\boldsymbol{A}$ is a vector field with the property

$$\boldsymbol{B} = \nabla \times \boldsymbol{A} \tag{15.11.1}$$

where $\boldsymbol{B}$ is the underlying magnetic field of physical interest. In the static (no time dependence) case $\boldsymbol{B}$ is given in terms of the current density $\boldsymbol{j}(\boldsymbol{r})$ by the *Biot-Savart* law

$$\boldsymbol{B}(\boldsymbol{r}) = [\mu_0/(4\pi)] \int d^3\boldsymbol{r}'\, \boldsymbol{j}(\boldsymbol{r}') \times \{(\boldsymbol{r}-\boldsymbol{r}')/[||(\boldsymbol{r}-\boldsymbol{r}')||^3]\}. \tag{15.11.2}$$

If we define $\boldsymbol{A}$ by the rule

$$\boldsymbol{A}(\boldsymbol{r}) = [\mu_0/(4\pi)] \int d^3\boldsymbol{r}'\, \boldsymbol{j}(\boldsymbol{r}')\{1/[||(\boldsymbol{r}-\boldsymbol{r}')||]\}, \tag{15.11.3}$$

then direct computation shows that this $\boldsymbol{A}$ satisfies (11.1). Indeed, we find

$$\begin{aligned}\nabla \times \boldsymbol{A} &= -[\mu_0/(4\pi)] \int d^3\boldsymbol{r}'\, \boldsymbol{j}(\boldsymbol{r}') \times \nabla\{1/[||(\boldsymbol{r}-\boldsymbol{r}')||]\} \\ &= -[\mu_0/(4\pi)] \int d^3\boldsymbol{r}'\, \boldsymbol{j}(\boldsymbol{r}') \times \{-(\boldsymbol{r}-\boldsymbol{r}')/[||(\boldsymbol{r}-\boldsymbol{r}')||^3]\} \\ &= \boldsymbol{B}(\boldsymbol{r}).\end{aligned} \tag{15.11.4}$$

We can also verify by direct computation that this $\boldsymbol{A}$ is divergence free,

$$\begin{aligned}\nabla \cdot \boldsymbol{A} &= [\mu_0/(4\pi)] \int d^3\boldsymbol{r}'\, \boldsymbol{j}(\boldsymbol{r}') \cdot \nabla\{1/[||(\boldsymbol{r}-\boldsymbol{r}')||]\} \\ &= [\mu_0/(4\pi)] \int d^3\boldsymbol{r}'\, \boldsymbol{j}(\boldsymbol{r}') \cdot \nabla'\{1/[||(\boldsymbol{r}-\boldsymbol{r}')||]\} \\ &= -[\mu_0/(4\pi)] \int d^3\boldsymbol{r}'\, [\nabla' \cdot \boldsymbol{j}(\boldsymbol{r}')]\{1/[||(\boldsymbol{r}-\boldsymbol{r}')||]\} \\ &= 0.\end{aligned} \tag{15.11.5}$$

Here we have used integration by parts and the current conservation relation

$$\nabla' \cdot \boldsymbol{j}(\boldsymbol{r}') = 0. \tag{15.11.6}$$

See Exercise 11.2. Therefore the vector potential given by (11.3) is in a Coulomb gauge.

From (11.2) it follows that there is (in the static case) the differential relation

$$\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{j}. \tag{15.11.7}$$

Upon combining (11.1) and (11.7) we see that there is the general result

$$\nabla \times (\nabla \times \boldsymbol{A}) = -\nabla^2 \boldsymbol{A} + \nabla(\nabla \cdot \boldsymbol{A}) = \mu_0 \boldsymbol{j}. \tag{15.11.8}$$

And, when (11.5) is taken into account, we see that for the $\boldsymbol{A}$ given by (11.3) there is the result

$$-\nabla^2 \boldsymbol{A} = \mu_0 \boldsymbol{j}. \tag{15.11.9}$$

In particular, in current-free regions, the Cartesian components of this $\boldsymbol{A}$ are harmonic functions. Recall the discussion at the beginning of Section 5. We note that the result (11.9) can also be found directly from the definition (11.3),

$$\begin{aligned} -\nabla^2 \boldsymbol{A}(\boldsymbol{r}) &= -[\mu_0/(4\pi)] \int d^3\boldsymbol{r}'\, \boldsymbol{j}(\boldsymbol{r}') \nabla^2 \{1/[||(\boldsymbol{r} - \boldsymbol{r}')||]\} \\ &= -[\mu_0/(4\pi)] \int d^3\boldsymbol{r}'\, \boldsymbol{j}(\boldsymbol{r}')(-4\pi)\delta_3(\boldsymbol{r} - \boldsymbol{r}') \\ &= \mu_0 \boldsymbol{j}(\boldsymbol{r}). \end{aligned} \tag{15.11.10}$$

15.11.1.2 Use of Green Function in Cylindrical Coordinates

Suppose we attempt to compute, in cylindrical coordinates, the vector potential $\boldsymbol{A}$ in terms of $\boldsymbol{j}$ with the aid of (11.3). To do so we will need the volume element $d^3\boldsymbol{r}'$ in cylindrical coordinates,

$$d^3\boldsymbol{r}' = dz' d\phi' \rho' d\rho'. \tag{15.11.11}$$

We will also need the Green function $1/[||(\boldsymbol{r} - \boldsymbol{r}')||]$ in cylindrical coordinates. It can be shown that it is given by the relation

$$1/[||(\boldsymbol{r} - \boldsymbol{r}')||] = (2/\pi) \sum_{m=-\infty}^{\infty} \exp(im\phi) \exp(-im\phi') \int_0^\infty dk\ \cos[k(z - z')] I_m(k\rho_<) K_m(k\rho_>). \tag{15.11.12}$$

Here

$$\rho_< = \text{ the lesser of } \rho \text{ and } \rho' \tag{15.11.13}$$

and

$$\rho_> = \text{ the greater of } \rho \text{ and } \rho'. \tag{15.11.14}$$

See the books of Jackson and Arfken listed in the bibliography at the end of this chapter.

As it stands, (11.12) is not exactly in the form we need. It can be verified that there is the relation

$$\begin{aligned} &\int_0^\infty dk\ \cos[k(z - z')] I_m(k\rho_<) K_m(k\rho_>) = \\ &(1/2) \int_{-\infty}^{\infty} dk\ \exp(ikz) \exp(-ikz') I_m(|k|\rho_<) K_m(|k|\rho_>). \end{aligned} \tag{15.11.15}$$

See exercise 11.3. This relation can be employed in (11.12) to bring it to the form

$$1/[||(\boldsymbol{r}-\boldsymbol{r}')||] =$$
$$(1/\pi)\sum_{m=-\infty}^{\infty}\exp(im\phi)\exp(-im\phi')\int_{-\infty}^{\infty}dk\ \exp(ikz)\exp(-ikz')I_m(|k|\rho_<)K_m(|k|\rho_>). \tag{15.11.16}$$

We remark that (11.16) is known to appear in the literature without the absolute value signs $|*|$ about k in the arguments of I_m and K_m. Such appearances are incorrect. They are also ill defined because $K_m(w)$ is not well defined for negative values of w due to logarithmic terms at the origin.

Let us now employ (11.11) and (11.16) in (11.3) to compute $\boldsymbol{A}$ in terms of $\boldsymbol{j}$. So doing gives the result

$$\boldsymbol{A}(\boldsymbol{r}) = (1/\pi)\int dz'd\phi'\rho'd\rho'\ \boldsymbol{j}(\rho',\phi',z')\ \times$$
$$\sum_{m=-\infty}^{\infty}\exp(im\phi)\exp(-im\phi')\int_{-\infty}^{\infty}dk\ \exp(ikz)\exp(-ikz')I_m(|k|\rho_<)K_m(|k|\rho_>). \tag{15.11.17}$$

This result can be rearranged to take the form

$$\boldsymbol{A}(\boldsymbol{r}) = (1/\pi)\sum_{m=-\infty}^{\infty}\exp(im\phi)\ \times$$
$$\int_{-\infty}^{\infty}dk\ \exp(ikz)I_m(|k|\rho)\int dz'd\phi'\rho'd\rho'\ \boldsymbol{j}(\rho',\phi',z')\exp(-im\phi')\exp(-ikz')K_m(|k|\rho'). \tag{15.11.18}$$

Here we have assumed that the current $\boldsymbol{j}$ lies outside (vanishes inside) a cylinder of radius a and we are interested in the vector potential inside the cylinder. Then we have the relations $\rho \in (0,a)$ and $\rho' \in (a,\infty)$ so that $\rho_< = \rho$ and $\rho_> = \rho'$. Finally, (11.18) can be written in the more compact form

$$\boldsymbol{A}(\boldsymbol{r}) = (1/\pi)\sum_{m=-\infty}^{\infty}\exp(im\phi)\int_{-\infty}^{\infty}dk\ \exp(ikz)I_m(|k|\rho)\tilde{\boldsymbol{j}}(m,k) \tag{15.11.19}$$

where

$$\tilde{\boldsymbol{j}}(m,k) = \int dz'd\phi'\rho'd\rho'\ \boldsymbol{j}(\rho',\phi',z')\exp(-im\phi')\exp(-ikz')K_m(|k|\rho'). \tag{15.11.20}$$

15.11.1.3 Complex Cylindrical Harmonic Expansion

We observe that (11.19) is beginning to take on the appearance of a (complex) cylindrical harmonic expansion. Following the pattern of Section 15.3.1, let us work to enhance the appearance of a cylindrical harmonic expansion. Begin by rewriting (11.19) in the form

$$\boldsymbol{A} = \sum_{m=-\infty}^{\infty} \exp(im\phi)\tilde{\boldsymbol{A}}(m,\rho,z) \tag{15.11.21}$$

where

$$\tilde{\boldsymbol{A}}(m,\rho,z) = (1/\pi)\int_{-\infty}^{\infty} dk\ \exp(ikz)I_m(|k|\rho)\tilde{\boldsymbol{j}}(m,k). \tag{15.11.22}$$

Next employ the Taylor expansion (15.3.11) in (11.22). As a first step we see that

$$\begin{aligned} I_m(|k|\rho) &= (1/2)^{|m|}|k|^{|m|}\rho^{|m|}\sum_{\ell=0}^{\infty}(|k|\rho)^{2\ell}/[2^{2\ell}\ell!(\ell+|m|)!] \\ &= (1/2)^{|m|}|k|^{|m|}\rho^{|m|}\sum_{\ell=0}^{\infty}(k\rho)^{2\ell}/[2^{2\ell}\ell!(\ell+|m|)!]. \end{aligned} \tag{15.11.23}$$

Consequently, we may rewrite (11.22) in the form

$$\begin{aligned} &\tilde{\boldsymbol{A}}(m,\rho,z) = (1/\pi)\int_{-\infty}^{\infty} dk\ \tilde{\boldsymbol{j}}(m,k)\exp(ikz)I_m(|k|\rho) = \\ &(1/\pi)\int_{-\infty}^{\infty} dk\ \tilde{\boldsymbol{j}}(m,k)\exp(ikz)(1/2)^{|m|}|k|^{|m|}\rho^{|m|}\sum_{\ell=0}^{\infty}(k\rho)^{2\ell}/[2^{2\ell}\ell!(\ell+|m|)!] = \\ &\sum_{\ell=0}^{\infty}\{1/[2^{2\ell}\ell!(\ell+|m|)!]\}\rho^{2\ell+|m|}(1/\pi)(1/2)^{|m|}\int_{-\infty}^{\infty} dk\ |k|^{|m|}k^{2\ell}\tilde{\boldsymbol{j}}(m,k)\exp(ikz). \end{aligned} \tag{15.11.24}$$

Define (vector) functions $\boldsymbol{C}^{[0]}(m,z)$ by writing

$$\boldsymbol{C}^{[0]}(m,z) = (1/\pi)(1/2)^{|m|}(1/|m|!)\int_{-\infty}^{\infty} dk\ |k|^{|m|}\tilde{\boldsymbol{j}}(m,k)\exp(ikz). \tag{15.11.25}$$

Also define functions $\boldsymbol{C}^{[n]}(m,z)$ by writing

$$\boldsymbol{C}^{[n]}(m,z) = (\partial_z)^n\boldsymbol{C}^{[0]}(m,z). \tag{15.11.26}$$

Then, by differentiating under the integral sign, we have the result

$$\boldsymbol{C}^{[n]}(m,z) = i^n(1/\pi)(1/2)^{|m|}(1/|m|!)\int_{-\infty}^{\infty} dk\ |k|^{|m|}k^n\tilde{\boldsymbol{j}}(m,k)\exp(ikz) \tag{15.11.27}$$

and, in particular,

$$\boldsymbol{C}^{[2\ell]}(m,z) = (-1)^\ell(1/\pi)(1/2)^{|m|}(1/|m|!)\int_{-\infty}^{\infty} dk\ k^{2\ell}|k|^{|m|}\tilde{\boldsymbol{j}}(m,k)\exp(ikz). \tag{15.11.28}$$

Therefore we also write the relation

$$(1/\pi)(1/2)^{|m|}(1/|m|!)\int_{-\infty}^{\infty} dk\ k^{2\ell}|k|^{|m|}\tilde{\boldsymbol{j}}(m,k)\exp(ikz) = (-1)^\ell|m|!\boldsymbol{C}^{[2\ell]}(m,z). \tag{15.11.29}$$

Combining (11.24) and (11.29) gives the results

$$\tilde{\boldsymbol{A}}(m,\rho,z)=\sum_{\ell=0}^{\infty}(-1)^{\ell}|m|!\{1/[2^{2\ell}\ell!(\ell+|m|)!]\}\rho^{2\ell+|m|}\boldsymbol{C}^{[2\ell]}(m,z), \tag{15.11.30}$$

And then using (11.21) gives the final result

$$\boldsymbol{A}(\rho,\phi,z)=\sum_{m=-\infty}^{\infty}\exp(im\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}|m|!\{1/[2^{2\ell}\ell!(\ell+|m|)!]\}\rho^{2\ell+|m|}\boldsymbol{C}^{[2\ell]}(m,z). \tag{15.11.31}$$

15.11.1.4 Real Cylindrical Harmonic Expansion

Section 15.3.2

15.11.2 Calculation of Scalar Potential from Current Data

Exercises

15.11.1. Verify (11.1) through (11.4).

15.11.2. Integration by parts.

15.11.3. The aim of this exercise is to arrive at the relation (11.15). Begin by verifying the equations listed below:

$$\begin{aligned}&\int_0^{\infty}dk\ \cos[k(z-z')]I_m(k\rho_<)K_m(k\rho_>)=\\&\int_0^{\infty}dk\ (1/2)\exp[k(z-z')]I_m(k\rho_<)K_m(k\rho_>)\\&+\int_0^{\infty}dk\ (1/2)\exp[-k(z-z')]I_m(k\rho_<)K_m(k\rho_>);\end{aligned} \tag{15.11.32}$$

$$\begin{aligned}&\int_0^{\infty}dk\ \cos[k(z-z')]I_m(k\rho_<)K_m(k\rho_>)=\\&\int_0^{\infty}dk\ (1/2)\exp[k(z-z')]I_m(|k|\rho_<)K_m(|k|\rho_>)\\&+\int_0^{\infty}dk\ (1/2)\exp[-k(z-z')]I_m(|k|\rho_<)K_m(|k|\rho_>);\end{aligned} \tag{15.11.33}$$

$$\begin{aligned}&\int_0^{\infty}dk\ \cos[k(z-z')]I_m(k\rho_<)K_m(k\rho_>)=\\&\int_0^{\infty}dk\ (1/2)\exp[ik(z-z')]I_m(|k|\rho_<)K_m(|k|\rho_>)\\&+\int_{-\infty}^{0}dk\ (1/2)\exp[ik(z-z')]I_m(|k|\rho_<)K_m(|k|\rho_>);\end{aligned} \tag{15.11.34}$$

$$\int_0^\infty dk\ \cos[k(z-z')]I_m(k\rho_<)K_m(k\rho_>) =$$
$$\int_{-\infty}^\infty dk\ (1/2)\exp[ik(z-z')]I_m(|k|\rho_<)K_m(|k|\rho_>); \tag{15.11.35}$$

$$\int_0^\infty dk\ \cos[k(z-z')]I_m(k\rho_<)K_m(k\rho_>) =$$
$$\int_{-\infty}^\infty dk\ (1/2)\exp(ikz)\exp(-ikz')I_m(|k|\rho_<)K_m(|k|\rho_>). \tag{15.11.36}$$

Finally, verify that (11.15) follows from (11.28).

15.12 Closing Remarks

15.12.1 Caveat about Significance of Integrated Multipoles

Suppose the relations (3.65) and (3.66) are used to compute the integrals of the transverse field components $B_x(x,y,z)$ and $B_y(x,y,z)$ over the range $z=-\infty$ to $z=+\infty$. We observe that for $n \geq 0$ there are the relations

$$\int_{-\infty}^\infty dz\ C_0^{[n+2]}(z) = C_0^{[n+1]}(z)|_{-\infty}^\infty, \tag{15.12.1}$$

$$\int_{-\infty}^\infty dz\ C_{m,\alpha}^{[n+1]}(z) = C_{m,\alpha}^{[n]}(z)|_{-\infty}^\infty. \tag{15.12.2}$$

Also we know that the $C_0^{[n+1]}(z)$ and the $C_{m,\alpha}^{[n]}(z)$ vanish at $z=\pm\infty$. Consequently, all of the terms in the sums (3.65) and (3.66) integrate to zero save for those that involve $C_0^{[1]}(z)$ and the $C_{m,\alpha}^{[0]}(z)$. We these observations in mind, we find the results

$$\begin{aligned}\int_{-\infty}^\infty dz\ B_x(x,y,z) &= \sum_{m=0}^\infty (m+1)\rho^m\cos(m\phi)\int_{-\infty}^\infty dz\ C_{m+1,c}^{[0]}(z) \\ &\quad + \sum_{m=0}^\infty (m+1)\rho^m\sin(m\phi)\int_{-\infty}^\infty dz\ C_{m+1,s}^{[0]}(z) \\ &= \sum_{m'=1}^\infty m'\rho^{m'-1}\cos[(m'-1)\phi]\int_{-\infty}^\infty dz\ C_{m',c}^{[0]}(z) \\ &\quad + \sum_{m'=1}^\infty m'\rho^{m'-1}\sin[(m'-1)\phi]\int_{-\infty}^\infty dz\ C_{m',s}^{[0]}(z),\end{aligned} \tag{15.12.3}$$

$$\begin{aligned}\int_{-\infty}^{\infty} dz\ B_y(x,y,z) &= \sum_{m=0}^{\infty}(m+1)\rho^m\cos(m\phi)\int_{-\infty}^{\infty} dz\ C^{[0]}_{m+1,s}(z)\\ &\quad -\sum_{m=0}^{\infty}(m+1)\rho^m\sin(m\phi)\int_{-\infty}^{\infty} dz\ C^{[0]}_{m+1,c}(z)\\ &= \sum_{m'=1}^{\infty} m'\rho^{m'-1}\cos[(m'-1)\phi]\int_{-\infty}^{\infty} dz\ C^{[0]}_{m',s}(z)\\ &\quad -\sum_{m'=1}^{\infty} m'\rho^{m'-1}\sin[(m'-1)\phi]\int_{-\infty}^{\infty} dz\ C^{[0]}_{m',c}(z).\end{aligned} \tag{15.12.4}$$

Note that (11.3) and (11.4) are consistent with (3.77), (3.78), (3.80), and (3.81).

What are we to conclude from these results? The "multipole" content of a magnet is often specified, in effect, in terms of the *integrated multipole* quantities $\int_{-\infty}^{\infty} dz\ C^{[0]}_{m',\alpha}(z)$ for $m' = 1, 2, \cdots$. This is because magnet measurements are frequently made using spinning coils whose length is such that they extend beyond the ends of the magnets to include the fringe-field regions. (See Appendix K.) Hence, the use of such coils measures $\int_{-\infty}^{\infty} dz\ B_x(x,y,z)$ and $\int_{-\infty}^{\infty} dz\ B_y(x,y,z)$ which, according to (11.3) and (11.4), is equivalent to measuring the integrated multipoles. Moreover, the size of the integrated multipoles is often taken as a figure of merit for any given magnet.

Is this reasonable? We know that some terms of the form $\exp(: f_3 :)\exp(: f_4 :)\cdots$ in the transfer map can have deleterious effects on the dynamic aperture. Recall, for example, Section 1.2.3 which illustrated the effect of the term $\exp(: q^3 :)$ in the simplest nonlinear case. We also know that the generators $f_3, f_4, \cdots$ arise from $H_3, H_4, \cdots$ terms in the Hamiltonian. Finally, we know that nonzero on-axis gradients of the form $C^{[0]}_{3,\alpha}(z), C^{[0]}_{4,\alpha}(z), \cdots$ produce nonzero terms of the form $H_3, H_4, \cdots$ in the Hamiltonian.[10] Therefore, if the integrated multipoles are large for $m' = 3, 4, \cdots$, we expect that nonlinear terms in the map will be important and the dynamic aperture is likely be small. Consequently, a good rule of thumb would appear to be that the integrated $m' = 3, 4, \cdots$ multipole terms should be small to minimize possibly deleterious nonlinear terms in the transfer map.

But, while minimizing the integrated multipoles would seem to be a possible way of minimizing the nonlinear terms in the transfer map, so doing is *not necessarily sufficient.* Consider, for example, the transfer map for a composite system consisting of two identical back-to back sextuples save that they are oppositely powered. All integrated multipoles for such a system would be exactly zero. However, the transfer map for this system could still have large nonlinear terms including those with $f_3 \neq 0$.[11] Observe also that fringe-field terms

[10] However, nonzero on-axis gradients of the form $C^{[0]}_{3,\alpha}(z), C^{[0]}_{4,\alpha}(z), \cdots$ are not the only source of $H_3, H_4, \cdots$ terms in the Hamiltonian. Such terms also arise, for example, from the expansion of the square root in (1.4) and occur even if $A_x = A_y = 0$.

[11] There are at least two other instances of this apparently malevolent principle: The first involves superconducting dipoles. They are often equipped with multipole-corrector coil packages at each end. Even if these coils are powered so that the composite system (dipole plus correctors) has net integrated multipole values of zero for the first few m' values (with $m' \geq 3$), so doing does not guarantee that the net transfer map is free of f_n terms for the first few values of $n \geq 3$. The second concerns room-temperature quadrupoles.

contribute to $H_3, H_4, \cdots$, and therefore can produce nonlinear terms in the transfer map.[12] However, as noted earlier, all of the terms in the sums (3.65) and (3.66) integrate to zero save for those that involve $C_0^{[1]}(z)$ and the $C_{m,\alpha}^{[0]}(z)$. Consequently all fringe-field terms integrate to zero, and their presence is therefore undetectable solely from an examination of the values of the integrated multipoles. Note also that $m = 0$ (solenoid) terms make no contribution to the integrated multipoles. However, it can be shown that solenoid fringe-fields make nonlinear contributions to the transfer map.[13]

We conclude and emphasize that what is needed for a realistic calculation of transfer maps are the functions $C_0^{[1]}(z)$ and the $C_{m,\alpha}^{[0]}(z)$ *themselves*, and *not* just their integrals.

15.12.2 Need for Generalized Gradients and the Use of Surface Data

From the work of the previous sections, we have learned that the dynamics of a charged particle passing through a region of space occupied by a magnetic field described by the scalar potential (3.33), or the azimuthal-free vector potential $\boldsymbol{A}$ given by (4.21) through (4.26), or the symmetric Coulomb gauge vector potential $\hat{\boldsymbol{A}}$ given by (5.89) through (5.94), or their vertical-free and possibly further adjusted variants as illustrated for the normal dipole, are completely determined by a knowledge of the generalized on-axis gradient functions $C_0^{[1]}(z)$ and $C_{m,\alpha}^{[0]}(z)$ and their derivatives. In Chapter 16 we will treat cases for which the generalized gradients can be computed analytically. In Chapters 17 through 21 we will describe several general methods for computing the generalized gradients and their derivatives numerically based on the use of numerical field data on a *surface*. The surfaces employed will be those of cylinders with circular, elliptical, or rectangular cross sections. These methods are *smoothing*. That is, they have the virtue of being relatively insensitive to errors in the input data. Consequently, they are ideally suited for numerical use.

15.12.3 Limitations Imposed by Symmetry and Hamilton and Maxwell

In the introduction to this chapter we noted that there are possible limitations on what transfer maps can be achieved. The first limitation is that the transfer map must be symplectic. The second arises from the fact that, in many instances, the electric and magnetic fields within beam-line elements must arise from fields that satisfy the source-free Maxwell equations. These limitations, combined with symmetry assumptions, may place restrictions upon what can actually be achieved. For example there is a remarkable theorem, due to Scherzer, which states that any imaging system having cylindrical symmetry must have negative spherical aberration. Consequently it is impossible to design, using only electric and

Sometimes they are hand-fitted during manufacture with end shims so that the net integrated $m' = 6$ multipole (which, according to Subsection 3.5, is allowed) is in fact zero. So doing does not guarantee that the net transfer map is free of f_6 terms. For a further discussion of correction methods, see Section 12.11.

[12] For example, dipole fringe-field effects in the hard-edge limit produce an f_4 some of whose terms are infinite.

[13] For example, solenoid fringe-field effects in the hard-edge limit also produce an f_4 some of whose terms are infinite.

magnetic elements with cylindrical symmetry, an electron microscope that is free of spherical aberration. As a practical consequence, the resolution of such microscopes is limited to a few Angstroms. To achieve zero spherical aberration it is necessary to break cylindrical symmetry with the careful use of nonlinear elements such as sextuples or octupoles. This is now done in the highest resolution electron microscopes with the result that it is now possible to achieve resolution at the atomic and subatomic level. For a discussion of Scherzer's theorem, and the possible correction of spherical aberration, see the reference at the end of the bibliography for this chapter.

Exercises

15.12.1. Verify that the integrated transverse fields satisfy the transverse Laplace equation,

$$\nabla_\perp^2 \int_{-\infty}^{\infty} dz\; B_x(x,y,z) = \nabla_\perp^2 \int_{-\infty}^{\infty} dz\; B_y(x,y,z) = 0. \tag{15.12.5}$$

Hint: See Exercise 3.4.

Bibliography

The Poincaré-Coulomb Gauge

[1] W. Brittin, W. Smythe, and W. Wyss, “Poincaré gauge in electrodynamics”, *Am. J. Phys.* **50**, p.693 (1982).

[2] F. Cornish, “The Poincaré and related gauges in electromagnetic theory”, *Am. J. Phys.* **52**, p.460 (1984).

[3] J. D. Jackson, “From Lorenz to Coulomb and other explicit gauge transformations”, LBNL-50079 (2002) and *Am. J. Phys.* **70**, 917 (2002). Alternatively, see the Web site http://arxiv.org/ftp/physics/papers/0204/0204034.pdf.

Bessel and Other Special Functions

[4] G.N. Watson, *A Treatise on the Theory of Bessel Functions*, Cambridge (1922).

[5] E.T. Whittaker and G.N. Watson, *A Course of Modern Analysis*, Cambridge (1952).

[6] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover (1972). Also available on the Web by Googling “abramowitz and stegun 1972”.

[7] F. Olver, D. Lozier, R. Boisvert, and C. Clark, Editors, *NIST Handbook of Mathematical Functions*, Cambridge (2010). See also the Web site http://dlmf.nist.gov/.

[8] H. Bateman, A. Erdelyi, W. Magnus, F. Oberhettinger, F.G. Tricomi, *Higher Transcendental Functions*, 3 Volumes, McGraw-Hill (1953). Chapter 7 in Volume 2 treats Bessel functions.

[9] H. Bateman, *Tables of Integral Transforms*, Volumes I, Ii, and II, McGraw-Hill (1954).

Electromagnetic Fields

[10] J. D. Jackson, *Classical Electrodynamics*, John Wiley (1999).

[11] G. Arfken, *Mathematical Methods for Physicists*, Second Edition, Academic Press (1970).

[12] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, Vols. 1 and 2, McGraw-Hill (1953).

[13] S. Russenschuck, *Field Computation for Accelerator Magnets: Analytical and Numerical Methods for Electromagnetic Design and Optimization*, Wiley (2010).

Magnetic Optics and Electron Microscopes

[14] A. Dragt, "Numerical third-order transfer map for solenoid", *Nuclear Instruments and Methods in Physics Research* **A298**, p. 441 (1990).

[15] M. Bassetti and C. Biscari, "Analytic Formulae for Magnetic Multipoles", *Particle Accelerators*, **52**, 221-250 (1996).

[16] M. Bassetti and C. Biscari, "Cylinder Model of Multipoles", Section 2.2.2 of *Handbook of Accelerator Physics and Engineering*, A. Chao and M. Tigner, Eds., World Scientific (2002).

[17] P. W. Hawkes and E. Kasper, *Principles of Electron Optics*, Vols. 1 through 3, Academic Press (1996).

[18] A. B. El-Kareh and J. C. J. El-Kareh, *Electron Beams, Lenses, and Optics*, Vols. 1 and 2, Academic Press (1970).

[19] P. A. Sturrock, "The Aberrations of Magnetic Electron Lenses due to Asymmetries", *Philosophical Transactions of the Royal Society of London, Series A, Mathematical and Physical Sciences*, Vol. 243, No. 868, pp. 387-429 (1951).

[20] P. W. Hawkes, Edit., *Advances in Imaging and Electron Physics*, Academic Press, Elsevier Inc. This continuing journal merges the long-running journals *Advances in Electronics and Electron Physics* and *Advances in Optical and Electron Microscopy.*

Scherzer's Theorem

[21] A. Dragt and E. Forest, "Lie Algebraic Theory of Charged-Particle Optics and Electron Microscopes", *Advances in Electronics and Electron Physics*, Volume 67, pp. 65-117, P. W. Hawkes, Edit. (1986). See the reference above to this journal.

Chapter 16

Realistic Transfer Maps for Straight Iron-Free Beam-Line Elements

Chapter 15 described cylindrical harmonic expansions for straight elements, and showed how the scalar potential, magnetic field, and various vector potentials can be described in terms of generalized on-axis gradients. In most cases these on-axis gradients must be computed numerically. How this may be done in general for straight elements is described in Chapters 17 through 21. However, in some iron-free cases the on-axis gradients can be computed analytically, including their fringe-field behavior. This chapter treats several of these cases. We remark that although our discussion is limited to magnetic beam-line elements, electrostatic beam-line elements can be treated in an analogous way.

In principle, the fringe field of any individual beam-line element at either end of the element has *infinite* extent. (This is particularly true of iron-free elements, but is also a consideration even in the case of some iron-dominated elements.) However in practice in many instances we may wish to regard a beam line as a collection of separated/isolated elements. To do this it is necessary to make an approximation in which leading and trailing end fields are "terminated" in some way. The crucial problem is how to relate canonical coordinates in the absence of a magnetic field with canonical coordinates in the presence of a magnetic field. The first part of this chapter is devoted to describing how this problem may be treated in general for straight beam-line elements. The remaining part of the chapter treats various specific straight iron-free beam-line elements.

16.1 Terminating End Fields

16.1.1 Preliminary Observations

We begin with some preliminary observations. In Cartesian coordinates the Hamiltonian describing charged-particle motion with z as the independent variable is given by the relation

$$K = -[(p_t^{\rm can})^2/c^2 - m^2c^2 - (p_x^{\rm can} - qA_x)^2 - (p_y^{\rm can} - qA_y)^2]^{1/2} - qA_z. \tag{16.1.1}$$

Here we have assumed that the electric scalar potential ψ vanishes and $\boldsymbol{A}$ is static so that there is no electric field. Also, we have used the notation $p_x^{\rm can}$, $p_y^{\rm can}$, and $p_t^{\rm can}$ to indicate that it is the components of the *canonical* momenta that are involved in a Hamiltonian description of motion. See (1.6.16).

According to Hamilton's equations of motion, the change of a coordinate, say $x(z)$, with z is given by

$$\begin{aligned} dx/dz &= \partial K/\partial p_x^{\rm can} \\ &= (p_x^{\rm can}-qA_x)/[(p_t^{\rm can})^2/c^2-m^2c^2-(p_x^{\rm can}-qA_x)^2-(p_y^{\rm can}-qA_y)^2]^{1/2} \\ &= (p_x^{\rm can}-qA_x)/[-(K+qA_z)]. \end{aligned} \tag{16.1.2}$$

Let us verify that this result agrees with what we already know. Recall that

$$K=-p_z^{\rm can}. \tag{16.1.3}$$

See (1.6.6). It follows that (1.2) can be rewritten in the form

$$dx/dz=(p_x^{\rm can}-qA_x)/(p_z^{\rm can}-qA_z). \tag{16.1.4}$$

According to (1.5.27) through (1.5.30) there is the relation

$$\boldsymbol{p}^{\rm can}-q\boldsymbol{A}=\boldsymbol{p}^{\rm mech} \tag{16.1.5}$$

where $\boldsymbol{p}^{\rm mech}$ is the *mechanical* momentum given by

$$\boldsymbol{p}^{\rm mech}=\gamma m\boldsymbol{v}. \tag{16.1.6}$$

Consequently, (1.4) can be rewritten in the form

$$dx/dz=p_x^{\rm mech}/p_z^{\rm mech}=\gamma m v_x/(\gamma m v_z)=v_x/v_z=\frac{dx/dt}{dz/dt}. \tag{16.1.7}$$

Evidently, the far left and far right sides of (1.7) agree. It is also easy to see that results analogous to those just found also hold for $y(z)$.

To complete the story we need to examine also the equation of motion for $t(z)$. In this case application of the standard Hamiltonian rules gives the result

$$\begin{aligned} dt/dz &= \partial K/\partial p_t^{\rm can} \\ &= (-p_t^{\rm can}/c^2)/[(p_t^{\rm can})^2/c^2-m^2c^2-(p_x^{\rm can}-qA_x)^2-(p_y^{\rm can}-qA_y)^2]^{1/2} \\ &= (-p_t^{\rm can}/c^2)/[-(K+qA_z)]. \end{aligned} \tag{16.1.8}$$

Now use of (1.3), (1.5), (1.6), and (1.6.17) yields the relation

$$dt/dz=(-p_t^{\rm can}/c^2)/p_z^{\rm mech}=\gamma m/(\gamma m v_z)=\frac{1}{dz/dt} \tag{16.1.9}$$

so that the far left and far right sides of (1.9) also agree.

16.1.2 Matching Conditions

We now consider what matching conditions should be imposed upon entry and exit of fringe-field regions. To proceed further it is useful to introduce some notation. Let $z^{\rm en}$ denote the z value where a transition is to be made from a region where the magnetic field is *taken* to vanish to the beginning of the leading fringe-field region. That is, any charged particle in question *enters* the leading fringe-field region when $z = z^{\rm en}$. Similarly, let $z^{\rm ex}$ denote the z value where a transition is to be made from the end of a trailing fringe-field region to a region where the magnetic field is again taken to vanish. That is, any charged particle in question *exits* the trailing fringe-field region when $z = z^{\rm ex}$. Our task is to find matching relations at $z^{\rm en}$ and $z^{\rm ex}$.

16.1.2.1 Entering a Leading Fringe-Field Region

Let us begin with a consideration of the transition between a field-free region and a leading fringe-field region. Let $K^{\rm ben}$ be the Hamiltonian *before entry* into the fringe-field region, and let $K^{\rm aen}$ be the Hamiltonian *after entry* into the fringe-field region. Then, since the magnetic field and its associated vector potential are assumed to vanish before entry, we have the relation

$$K^{\rm ben} = -[(p_t^{\rm canben})^2/c^2 - m^2c^2 - (p_x^{\rm canben})^2 - (p_y^{\rm canben})^2]^{1/2}. \tag{16.1.10}$$

And, since the magnetic field (and therefore also the vector potential) is nonzero after entry, we have the relation

$$K^{\rm aen} = -[(p_t^{\rm canaen})^2/c^2 - m^2c^2 - (p_x^{\rm canaen} - qA_x)^2 - (p_y^{\rm canean} - qA_y)^2]^{1/2} - qA_z. \tag{16.1.11}$$

Here we have added the suffixes *ben* and *aen* to the phase-space coordinates to denote their values *before entry* and *after entry.* Our task is to relate these phase-space coordinates.

As a first step, we naturally require that the coordinates be continous at $z^{\rm en}$,

$$x^{\rm aen} = x^{\rm ben}, \tag{16.1.12}$$

$$y^{\rm aen} = y^{\rm ben}, \tag{16.1.13}$$

$$t^{\rm aen} = t^{\rm ben}, \tag{16.1.14}$$

when $z = z^{\rm en}$. The next step is specify what is to be done with the momenta.

One possibility is to require that the slopes/"velocities" dx/dz, dy/dz, and dt/dz be continuous at $z^{\rm en}$. Let us work out the consequences of such a requirement. Before entry we have the result

$$\begin{aligned} & dx/dz = \partial K^{\rm ben}/\partial p_x^{\rm canben} = \\ & p_x^{\rm canben}/[(p_t^{\rm canben})^2/c^2 - m^2c^2 - (p_x^{\rm canben})^2 - (p_y^{\rm canben})^2]^{1/2}, \end{aligned} \tag{16.1.15}$$

and after entry there is the result

$$\begin{aligned} & dx/dz = \partial K^{\rm aen}/\partial p_x^{\rm canaen} = \\ & (p_x^{\rm canaen} - qA_x)/[(p_t^{\rm canaen})^2/c^2 - m^2c^2 - (p_x^{\rm canaen} - qA_x)^2 - (p_y^{\rm canaen} - qA_y)^2]^{1/2}. \end{aligned} \tag{16.1.16}$$

See (1.2). An analogous result holds for dy/dz. Finally, for dt/dz there is the before entry result

$$dt/dz = \partial K^{\rm ben}/\partial p_t^{\rm canben} = \\ (-p_t^{\rm canben}/c^2)/[(p_t^{\rm canben})^2/c^2 - m^2c^2 - (p_x^{\rm canben})^2 - (p_y^{\rm canben})^2]^{1/2}, \tag{16.1.17}$$

and the after entry result

$$dt/dz = \partial K^{\rm aen}/\partial p_t^{\rm canaen} = \\ (-p_t^{\rm canaen}/c^2)/[(p_t^{\rm canaen})^2/c^2 - m^2c^2 - (p_x^{\rm canaen} - qA_x)^2 - (p_y^{\rm canaen} - qA_y)^2]^{1/2}. \tag{16.1.18}$$

See (1.8). Now equate the far right sides of (1.15) and (1.16), the far right sides of their dy/dz counterparts, and the far right sides of (1.17) and (1.18). So doing yields the transition matching relations

$$p_x^{\rm canaen} - qA_x = p_x^{\rm canben}, \tag{16.1.19}$$

$$p_y^{\rm canaen} - qA_y = p_y^{\rm canben}, \tag{16.1.20}$$

$$p_t^{\rm canaen} = p_t^{\rm canben}. \tag{16.1.21}$$

In view of (1.5) the relations (1.19) and (1.20) can also be written in the form

$$p_x^{\rm mechaen} = p_x^{\rm mechben}, \tag{16.1.22}$$

$$p_y^{\rm mechaen} = p_y^{\rm mechben}. \tag{16.1.23}$$

Moreover, under our assumption that the electrical potential $\psi = 0$, (1.21) and (1.6.17) yield the relation

$$\boldsymbol{p}^{\rm mechaen} \cdot \boldsymbol{p}^{\rm mechaen} = \boldsymbol{p}^{\rm mechben} \cdot \boldsymbol{p}^{\rm mechben}. \tag{16.1.24}$$

This relation, when combined with (1.22) and (1.23), yields the further result

$$p_z^{\rm mechaen} = p_z^{\rm mechben}. \tag{16.1.25}$$

We conclude that imposition of the requirement that the slopes/"velocities" be continuous entails that the mechanical momenta be continuous.

The relation (1.21) is satisfactory because we would hope that the energy would not change upon entry into the leading fringe-field region. Again recall (1.6.17). However, we also desire that the phase-space transition matching relations be a symplectic transformation. Calculation shows that the transformation given by (1.12) through (1.14) and (1.19) through (1.21) is *not* symplectic. Compute the Poisson bracket of the left sides of (1.19) and (1.20) to find the result

$$\begin{aligned}[p_x^{\rm canaen} - qA_x, p_y^{\rm canaen} - qA_y] &= [p_x^{\rm canaen}, -qA_y] + [-qA_x, p_y^{\rm canaen}] \\ &= q\{\partial A_y/\partial x^{\rm aen} - \partial A_x/\partial y^{\rm aen}\} = qB_z. \end{aligned} \tag{16.1.26}$$

[Recall (1.7.40).] While hopefully small, generally $B_z(x, y, z^{\rm en})$ differs from zero at the beginning of the leading fringe-field region. On the other hand, the Poisson bracket of the right sides of (1.19) and (1.20) must vanish since $p_x^{\rm canben}$ and $p_y^{\rm canben}$ are supposed to be canonical momenta. Therefore the phase-space transformation given by (1.12) through (1.14) and (1.19) through (1.21) is generally *not* symplectic. Review, at this point, Exercise 6.4.11.

We expect that neglect of the magnetic field in the region $z < z^{\rm en}$ will lead to some error in trajectories. However, we do not want this error to violate the symplectic condition. The simplest way to maintain the symplectic condition is to retain the relations (1.12) through (1.14) and replace the relations (1.19) through (1.21) by the relations

$$p_x^{\rm canaen} = p_x^{\rm canben}, \tag{16.1.27}$$

$$p_y^{\rm canaen} = p_y^{\rm canben}, \tag{16.1.28}$$

$$p_t^{\rm canaen} = p_t^{\rm canben}. \tag{16.1.29}$$

In this case the transition matching relations (1.12) through (1.14) and (1.27) through (1.29) amount to the identity map $\mathcal{I}$, and the symplectic condition is trivially satisfied. Now, however, the error in trajectories manifests itself in that the slopes/"velocities" dx/dz, dy/dz, and dt/dz may be expected to be discontinuous at at $z^{\rm en}$. Inspection of (1.15) and (1.16), their dy/dz counterparts, and (1.17) and (1.18) shows that, in lowest approximation, these discontinuities are proportional to components of $\boldsymbol{A}(x, y, z^{\rm en})$. Indeed, again in view of (1.3) and (1.5), the transition relations (1.27) and (1.28) can be written in the form

$$\Delta p_x^{\rm mech} = p_x^{\rm mechaen} - p_x^{\rm mechben} = qA_x(x, y, z^{\rm en}), \tag{16.1.30}$$

$$\Delta p_y^{\rm mech} = p_y^{\rm mechaen} - p_y^{\rm mechben} = qA_y(x, y, z^{\rm en}). \tag{16.1.31}$$

Also, in view of (1.29), (1.24) continues to hold. Therefore, upon combining (1.29) through (1.31), we see that $p_z^{\rm mechaen}$ is given by the relation

$$\begin{aligned} p_z^{\rm mechaen} &= [(p_z^{\rm mechben})^2 + (p_x^{\rm mechben})^2 - (p_x^{\rm mechaen})^2 + (p_y^{\rm mechben})^2 - (p_y^{\rm mechaen})^2]^{1/2} \\ &= [(p_z^{\rm mechben})^2 - (\Delta p_x^{\rm mech})(\Sigma p_x^{\rm mech}) - (\Delta p_y^{\rm mech})(\Sigma p_y^{\rm mech})]^{1/2} \\ &= [(p_z^{\rm mechben})^2 - qA_x(x, y, z^{\rm en})(\Sigma p_x^{\rm mech}) - qA_y(x, y, z^{\rm en})(\Sigma p_y^{\rm mech})]^{1/2} \end{aligned} \tag{16.1.32}$$

where

$$\Sigma p_x^{\rm mech} = p_x^{\rm mechaen} + p_x^{\rm mechben} = 2p_x^{\rm mechben} + qA_x(x, y, z^{\rm en}), \tag{16.1.33}$$

$$\Sigma p_y^{\rm mech} = p_y^{\rm mechaen} + p_y^{\rm mechben} = 2p_y^{\rm mechben} + qA_y(x, y, z^{\rm en}). \tag{16.1.34}$$

We conclude that imposition of continuity in the canonical momenta as expressed by (1.27) through (1.29) entails a discontinuity in the mechanical momenta, and this discontinuity depends on the size of $\boldsymbol{A}(x, y, z^{\rm en})$. It is therefore desirable to work in a gauge where $\boldsymbol{A}(x, y, z^{\rm en})$ is as *small* as possible. Subsequently, we will explore the use of the minimum vector potential of Section 15.2.5 for this purpose.

16.1.2.2 Exiting a Trailing Fringe-Field Region

The transition between a trailing fringe-field region and a subsequent field-free region may be considered in an analogous way. We again require continuity in the coordinates and canonical momenta. As described earlier, let $z^{\rm ex}$ denote the z value where a transition is to be made from the end of a trailing fringe-field region to a region where the magnetic field is again taken to vanish. That is, any charged particle in question *exits* the trailing fringe-field region when $z = z^{\rm ex}$. We will also add the suffixes *aex* and *bex* to the phase-space coordinates to denote their values *after* and *before exit.* In terms of this notation we impose the matching conditions

$$x^{\rm aex} = x^{\rm bex}, \tag{16.1.35}$$

$$y^{\rm aex} = y^{\rm bex}, \tag{16.1.36}$$

$$t^{\rm aex} = t^{\rm bex}, \tag{16.1.37}$$

$$p_x^{\rm canaex} = p_x^{\rm canbex}, \tag{16.1.38}$$

$$p_y^{\rm canaex} = p_y^{\rm canbex}, \tag{16.1.39}$$

$$p_t^{\rm canaex} = p_t^{\rm canbex} \tag{16.1.40}$$

when $z = z^{\rm ex}$. So so doing entails discontinuities in the mechanical momenta given by the relations

$$\Delta p_x^{\rm mech} = p_x^{\rm mechaex} - p_x^{\rm mechbex} = qA_x(x, y, z^{\rm ex}), \tag{16.1.41}$$

$$\Delta p_y^{\rm mech} = p_y^{\rm mechaex} - p_y^{\rm mechbex} = qA_y(x, y, z^{\rm ex}), \tag{16.1.42}$$

$$\begin{aligned} p_z^{\rm mechaex} &= [(p_z^{\rm mechbex})^2 + (p_x^{\rm mechbex})^2 - (p_x^{\rm mechaex})^2 + (p_y^{\rm mechbex})^2 - (p_y^{\rm mechaex})^2]^{1/2} \\ &= [(p_z^{\rm mechbex})^2 - (\Delta p_x^{\rm mech})(\Sigma p_x^{\rm mech}) - (\Delta p_y^{\rm mech})(\Sigma p_y^{\rm mech})]^{1/2} \\ &= [(p_z^{\rm mechbex})^2 - qA_x(x, y, z^{\rm ex})(\Sigma p_x^{\rm mech}) - qA_y(x, y, z^{\rm ex})(\Sigma p_y^{\rm mech})]^{1/2} \end{aligned} \tag{16.1.43}$$

where

$$\Sigma p_x^{\rm mech} = p_x^{\rm mechaex} + p_x^{\rm mechbex} = 2p_x^{\rm mechbex} + qA_x(x, y, z^{\rm ex}), \tag{16.1.44}$$

$$\Sigma p_y^{\rm mech} = p_y^{\rm mechaex} + p_y^{\rm mechbex} = 2p_y^{\rm mechbex} + qA_y(x, y, z^{\rm ex}). \tag{16.1.45}$$

That is, imposition of continuity in the canonical momenta as expressed by (1.38) through (1.40) again entails discontinuities in the associated mechanical momenta. It is therefore also desirable to work in a gauge where $\boldsymbol{A}(x, y, z^{\rm ex})$ is as small as possible.

16.1.2.3 Modified Hamiltonian, Vector Potential, Magnetic Field, and Current

One way to view the symplectic matching relations (1.12) through (1.14), (1.27) through (1.29), and (1.35) through (1.40) is to replace the Hamiltonian (1.1) by a modified Hamiltonian $K^{\rm mod}$ given by

$$K^{\rm mod} = -[(p_t^{\rm can})^2/c^2 - m^2c^2 - (p_x^{\rm can} - qA_x^{\rm mod})^2 - (p_y^{\rm can} - qA_y^{\rm mod})^2]^{1/2} - qA_z^{\rm mod} \tag{16.1.46}$$

where

$$\boldsymbol{A}^{\rm mod}(x,y,z)=\theta(z-z^{\rm en})\theta(z^{\rm ex}-z)\boldsymbol{A}(x,y,z). \tag{16.1.47}$$

That is, the vector potential is taken to vanish for $z<z^{\rm en}$, turns on at $z=z^{\rm en}$, and again turns off for $z>z^{\rm ex}$. A little thought shows that integrating the equations of motion associated with this modified Hamiltonian automatically produces the matching relations (1.12) through (1.14), (1.27) through (1.29), and (1.35) through (1.40).

What is the modified magnetic field $\boldsymbol{B}^{\rm mod}$ associated with this modified vector potential? Evaluation of

$$\boldsymbol{B}^{\rm mod}=\nabla\times\boldsymbol{A}^{\rm mod} \tag{16.1.48}$$

gives the relations

$$\begin{aligned} B_x^{\rm mod}(x,y,z) &= \partial_y A_z^{\rm mod}-\partial_z A_y^{\rm mod} \\ &= \theta(z-z^{\rm en})\theta(z^{\rm ex}-z)B_x(x,y,z) \\ &\quad -[\delta(z-z^{\rm en})-\delta(z^{\rm ex}-z)]A_y(x,y,z), \end{aligned} \tag{16.1.49}$$

$$\begin{aligned} B_y^{\rm mod}(x,y,z) &= \partial_z A_x^{\rm mod}-\partial_x A_z^{\rm mod} \\ &= \theta(z-z^{\rm en})\theta(z^{\rm ex}-z)B_y(x,y,z) \\ &\quad +[\delta(z-z^{\rm en})-\delta(z^{\rm ex}-z)]A_x(x,y,z), \end{aligned} \tag{16.1.50}$$

$$\begin{aligned} B_z^{\rm mod}(x,y,z) &= \partial_x A_y^{\rm mod}-\partial_y A_x^{\rm mod} \\ &= \theta(z-z^{\rm en})\theta(z^{\rm ex}-z)B_z(x,y,z). \end{aligned} \tag{16.1.51}$$

By the construction (1.48), the modified magnetic field is divergence free,

$$\nabla\cdot\boldsymbol{B}^{\rm mod}=0, \tag{16.1.52}$$

as required. [Note that making the simple Ansatz $\boldsymbol{B}^{\rm mod}=\theta(z-z^{\rm en})\theta(z^{\rm ex}-z)\boldsymbol{B}$ violates the requirement (1.52). It is this Ansatz that would arise naturally if one were integrating the non-canonical Lorentz-force equations given in Exercise 1.6.16.]

What current $\boldsymbol{j}^{\rm mod}$ produces this modified magnetic field? It is specified by employing $\boldsymbol{B}^{\rm mod}$ as given by (1.49) through (1.51) in the relation

$$\mu_0\boldsymbol{j}^{\rm mod}=\nabla\times\boldsymbol{B}^{\rm mod}. \tag{16.1.53}$$

Doing so directly leads to considerable algebra, which can be bypassed with the use of suitable vector identities. Proceed as follows: Combining (1.48) and (1.53) gives the relation

$$\mu_0\boldsymbol{j}^{\rm mod}=\nabla\times(\nabla\times\boldsymbol{A}^{\rm mod})=\nabla(\nabla\cdot\boldsymbol{A}^{\rm mod})-\nabla^2\boldsymbol{A}^{\rm mod}. \tag{16.1.54}$$

Let us work on the first term on the right side of (1.54). From the definition (1.47) there is the result

$$\nabla\cdot\boldsymbol{A}^{\rm mod}=\boldsymbol{A}\cdot\nabla[\theta(z-z^{\rm en})\theta(z^{\rm ex}-z)]+\theta(z-z^{\rm en})\theta(z^{\rm ex}-z)\nabla\cdot\boldsymbol{A}. \tag{16.1.55}$$

Evaluation of the first term on the right in (1.55) gives the result

$$\begin{aligned} \boldsymbol{A}\cdot\nabla[\theta(z-z^{\rm en})\theta(z^{\rm ex}-z)] &= A_z\partial_z[\theta(z-z^{\rm en})\theta(z^{\rm ex}-z)] \\ &= [\delta(z-z^{\rm en})-\delta(z^{\rm ex}-z)]A_z. \end{aligned} \tag{16.1.56}$$

We also assume that $\boldsymbol{A}$ is in a Coulomb gauge, $\nabla\cdot\boldsymbol{A}=0$, so that (1.55) becomes

$$\nabla\cdot\boldsymbol{A}^{\rm mod}=[\delta(z-z^{\rm en})-\delta(z^{\rm ex}-z)]A_z, \tag{16.1.57}$$

and therefore

$$\begin{aligned} \nabla(\nabla\cdot\boldsymbol{A}^{\rm mod}) &= \boldsymbol{e}_x[\delta(z-z^{\rm en})-\delta(z^{\rm ex}-z)]\partial_x A_z \\ &\quad +\boldsymbol{e}_y[\delta(z-z^{\rm en})-\delta(z^{\rm ex}-z)]\partial_y A_z \\ &\quad +\boldsymbol{e}_z[\delta(z-z^{\rm en})-\delta(z^{\rm ex}-z)]\partial_z A_z \\ &\quad +\boldsymbol{e}_z[\delta'(z-z^{\rm en})+\delta'(z^{\rm ex}-z)]A_z. \end{aligned} \tag{16.1.58}$$

Next we turn to working out $-\nabla^2\boldsymbol{A}^{\rm mod}$, the second term on the right side of (1.54). For the x component we have the intermediate result

$$\begin{aligned} -\nabla^2 A_x^{\rm mod} &= -\nabla^2[\theta(z-z^{\rm en})\theta(z^{\rm ex}-z)A_x] \\ &= -\theta(z-z^{\rm en})\theta(z^{\rm ex}-z)(\partial_x^2+\partial_y^2)A_x-\partial_z^2[\theta(z-z^{\rm en})\theta(z^{\rm ex}-z)A_x]. \end{aligned} \tag{16.1.59}$$

By the product rule there is the relation

$$\begin{aligned} \partial_z[\theta(z-z^{\rm en})\theta(z^{\rm ex}-z)A_x] &= [\delta(z-z^{\rm en})-\delta(z^{\rm ex}-z)]A_x \\ &\quad +\theta(z-z^{\rm en})\theta(z^{\rm ex}-z)\partial_z A_x, \end{aligned} \tag{16.1.60}$$

from which it follows that

$$\begin{aligned} \partial_z^2[\theta(z-z^{\rm en})\theta(z^{\rm ex}-z)A_x] &= [\delta'(z-z^{\rm en})+\delta'(z^{\rm ex}-z)]A_x \\ &\quad +2[\delta(z-z^{\rm en})-\delta(z^{\rm ex}-z)]\partial_z A_x \\ &\quad +\theta(z-z^{\rm en})\theta(z^{\rm ex}-z)\partial_z^2 A_x. \end{aligned} \tag{16.1.61}$$

Combining (1.59) and (1.61) yields the next intermediate result

$$-\nabla^2 A_x^{\rm mod}=-2[\delta(z-z^{\rm en})-\delta(z^{\rm ex}-z)]\partial_z A_x-[\delta'(z-z^{\rm en})+\delta'(z^{\rm ex}-z)]A_x, \tag{16.1.62}$$

Here we have used the fact that $\boldsymbol{A}$ is harmonic. See (15.5.4). Similarly, there are the analogous next intermediate results

$$-\nabla^2 A_y^{\rm mod}=-2[\delta(z-z^{\rm en})-\delta(z^{\rm ex}-z)]\partial_z A_y-[\delta'(z-z^{\rm en})+\delta'(z^{\rm ex}-z)]A_y, \tag{16.1.63}$$

$$-\nabla^2 A_z^{\rm mod}=-2[\delta(z-z^{\rm en})-\delta(z^{\rm ex}-z)]\partial_z A_z-[\delta'(z-z^{\rm en})+\delta'(z^{\rm ex}-z)]A_z. \tag{16.1.64}$$

We are now able to combine the two terms on the right side of (1.54), using (1.58) and (1.62) through (1.64), to yield the desired final results

$$\begin{aligned}\mu_0 j_x^{\rm mod} &= [\delta(z-z^{\rm en})-\delta(z^{\rm ex}-z)][\partial_x A_z - 2\partial_z A_x] \\ &\quad -[\delta'(z-z^{\rm en})+\delta'(z^{\rm ex}-z)]A_x, \end{aligned}\tag{16.1.65}$$

$$\begin{aligned}\mu_0 j_y^{\rm mod} &= [\delta(z-z^{\rm en})-\delta(z^{\rm ex}-z)][\partial_y A_z - 2\partial_z A_y] \\ &\quad -[\delta'(z-z^{\rm en})+\delta'(z^{\rm ex}-z)]A_y, \end{aligned}\tag{16.1.66}$$

$$\mu_0 j_z^{\rm mod} = -[\delta(z-z^{\rm en})-\delta(z^{\rm ex}-z)]\partial_z A_z. \tag{16.1.67}$$

Evidently requiring the vector potential to vanish for $z < z^{\rm en}$, turn on at $z = z^{\rm en}$, and again turn off for $z > z^{\rm ex}$ is equivalent to introducing sheet (corresponding to the δ functions) and double-sheet (corresponding to the δ' functions) currents at $z = z^{\rm en}$ and $z = z^{\rm ex}$. And the strengths of these currents are proportional to the values of $\boldsymbol{A}$ and its first derivatives at $z = z^{\rm en}$ and $z = z^{\rm ex}$.

16.1.3 Changing Gauge

It may be useful to change gauges at various points during the course of integrating a trajectory and computing an associated transfer map. For example, to minimize end-field termination effects, it is desirable to change to minimum vector potentials at $z = z^{\rm en}$ and $z = z^{\rm ex}$. Suppose the gauge is to be *changed* at the point $z = z^c$. Let x^b, y^b, and t^b denote coordinate functions *before* the change, and let x^a, y^a, and t^a denote coordinate functions *after* the change. Also, let $\boldsymbol{A}^b(x^b, y^b; z)$ and $\boldsymbol{A}^a(x^a, y^a; z)$ be the vector potentials before ($z < z^c$) and after ($z > z^c$) the change point z^c. Finally, let $p_x^{\rm canb}$, $p_y^{\rm canb}$, $p_t^{\rm canb}$ be the canonical momentum functions before the change, and let $p_x^{\rm cana}$, $p_y^{\rm cana}$, $p_t^{\rm cana}$ be the canonical momentum functions after the change. In terms of these quantities, the before and after Hamiltonians K^b and K^a are given by the relations

$$K^b = -[(p_t^{\rm canb})^2/c^2 - m^2c^2 - (p_x^{\rm canb} - qA_x^b)^2 - (p_y^{\rm canb} - qA_y^b)^2]^{1/2} - qA_z^b \text{ for } z < z^c, \tag{16.1.68}$$

$$K^a = -[(p_t^{\rm cana})^2/c^2 - m^2c^2 - (p_x^{\rm cana} - qA_x^a)^2 - (p_y^{\rm cana} - qA_y^a)^2]^{1/2} - qA_z^a \text{ for } z > z^c. \tag{16.1.69}$$

What should be the matching relations between the phase-space quantities before and after? Since the choice of gauge should have no physical effect, there is the immediate requirement that the coordinate functions be continuous:

$$\begin{aligned} &x^a(z) = x^b(z) \text{ when } z = z^c, \\ &y^a(z) = y^b(z) \text{ when } z = z^c, \\ &t^a(z) = t^b(z) \text{ when } z = z^c. \end{aligned}\tag{16.1.70}$$

For the same reason, we require that the velocities, and hence the mechanical momenta, be continuous. From (1.5) and (1.6) we see that this requirement is equivalent to the relations

$$\boldsymbol{p}^{\rm cana} - q\boldsymbol{A^a} = \boldsymbol{p}^{\rm canb} - q\boldsymbol{A^b} \text{ when } z = z^c. \tag{16.1.71}$$

In terms of components, (1.71) yields the matching relations

$$\begin{aligned} p_x^{\text{cana}} &= p_x^{\text{canb}} + q(A_x^a - A_x^b) \text{ when } z = z^c, \\ p_y^{\text{cana}} &= p_y^{\text{canb}} + q(A_y^a - A_y^b) \text{ when } z = z^c. \end{aligned} \tag{16.1.72}$$

Finally, the total energy cannot change under a gauge transformation and therefore, since we have assumed that the electric scalar potential ψ vanishes, there is the matching relation

$$p_t^{\text{cana}} = p_t^{\text{canb}} \text{ when } z = z^c. \tag{16.1.73}$$

We note that this relation also follows from (1.6.17).

We assume there is some common overlap region where both $\boldsymbol{A}^b$ and $\boldsymbol{A}^a$ are defined. Since they both give rise to the same magnetic field, there is the relation

$$\nabla \times (\boldsymbol{A}^a - \boldsymbol{A}^b) = 0. \tag{16.1.74}$$

It follows that there is a function χ such that

$$\boldsymbol{A}^a - \boldsymbol{A}^b = \nabla\chi. \tag{16.1.75}$$

Consequently, the relations (1.72) can be rewritten in the form

$$\begin{aligned} p_x^{\text{cana}} &= p_x^{\text{canb}} + q(\partial/\partial x)\chi \text{ when } z = z^c, \\ p_y^{\text{cana}} &= p_y^{\text{canb}} + q(\partial/\partial y)\chi \text{ when } z = z^c. \end{aligned} \tag{16.1.76}$$

There is one last step. Let $\mathcal{T}^c$ be the symplectic *transformation* map defined by the relation

$$\mathcal{T}^c = \exp(q :\chi:). \tag{16.1.77}$$

With aid of this map it is easily verified that the relations (1.70), (1.72), and (1.73) can be rewritten in the form

$$\begin{aligned} x^a(z) &= \exp(q :\chi:)x^b(z) \text{ with } z = z^c, \\ y^a(z) &= \exp(q :\chi:)y^b(z) \text{ with } z = z^c, \\ t^a(z) &= \exp(q :\chi:)t^b(z) \text{ with } z = z^c; \end{aligned} \tag{16.1.78}$$

$$\begin{aligned} p_x^{\text{cana}}(z) &= \exp(q :\chi:)p_x^{\text{canb}}(z) \text{ with } z = z^c, \\ p_y^{\text{cana}}(z) &= \exp(q :\chi:)p_y^{\text{canb}}(z) \text{ with } z = z^c, \\ p_t^{\text{cana}}(z) &= \exp(q :\chi:)p_t^{\text{canb}}(z) \text{ with } z = z^c. \end{aligned} \tag{16.1.79}$$

We have determined that a change in gauge amounts to making a symplectic transformation. Review Exercises 6.2.8 and 6.5.3.

16.2 Solenoids

The remainder of this chapter is devoted to the treatment of various specific straight beam-line elements for which the on-axis gradients can be found analytically. We begin with the case of solenoids.

16.2.1 Preliminaries

A solenoid is a straight beam-line element whose field is described by a cylindrical harmonic expansion that contains (ideally) only an $m = 0$ term. Figure 2.1 illustrates a Cartesian coordinate system for the treatment of a solenoid. We recall from Section 15.3.3 that in this case the magnetic scalar potential ψ has the expansion

$$\begin{aligned}\psi(x,y,z) &= \psi_0(x,y,z) = \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_0^{[2\ell]}(z)\rho^{2\ell} \\ &= C_0^{[0]}(z) - (1/4)(x^2+y^2)C_0^{[2]}(z) + \cdots \qquad (16.2.1)\end{aligned}$$

with

$$\rho^2 = x^2 + y^2. \qquad (16.2.2)$$

See (15.3.53) and (15.5.5). Correspondingly, the associated magnetic field has the expansion

$$B_x = \partial_x\psi_0 = -(1/2)xC_0^{[2]}(z) + \cdots, \qquad (16.2.3)$$

$$B_y = \partial_y\psi_0 = -(1/2)yC_0^{[2]}(z) + \cdots, \qquad (16.2.4)$$

$$\begin{aligned}B_z &= \partial_z\psi_0 = \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_0^{[2\ell+1]}(z)\rho^{2\ell} \\ &= C_0^{[1]}(z) - (1/4)(x^2+y^2)C_0^{[3]}(z) + \cdots. \qquad (16.2.5)\end{aligned}$$

In particular, there is the result

$$\boldsymbol{B}(0,0,z) = C_0^{[1]}(z)\boldsymbol{e}_z. \qquad (16.2.6)$$

Also, according to Section 15.5.1, there is a suitable associated vector potential $\hat{\boldsymbol{A}}^0$ (in the symmetric Coulomb gauge which, in the case of a solenoid, is also the Poincaré-Coulomb gauge) given by the relation

$$\hat{A}_x^0 = -yU, \qquad (16.2.7)$$

$$\hat{A}_y^0 = xU, \qquad (16.2.8)$$

$$\hat{A}_z^0 = 0, \qquad (16.2.9)$$

where U is defined to be

$$U(\rho,z) = (1/2)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!(\ell+1)!}C_0^{[2\ell+1]}(z)\rho^{2\ell}. \qquad (16.2.10)$$

Correspondingly, the vector potential will have an expansion in x and y of the form

$$\hat{A}_x^0 = -yU = -y(1/2)[C_0^{[1]}(z) - (1/8)C_0^{[3]}(z)(x^2+y^2) + \cdots], \qquad (16.2.11)$$

$$\hat{A}_y^0 = xU = x(1/2)[C_0^{[1]}(z) - (1/8)C_0^{[3]}(z)(x^2+y^2) + \cdots], \qquad (16.2.12)$$

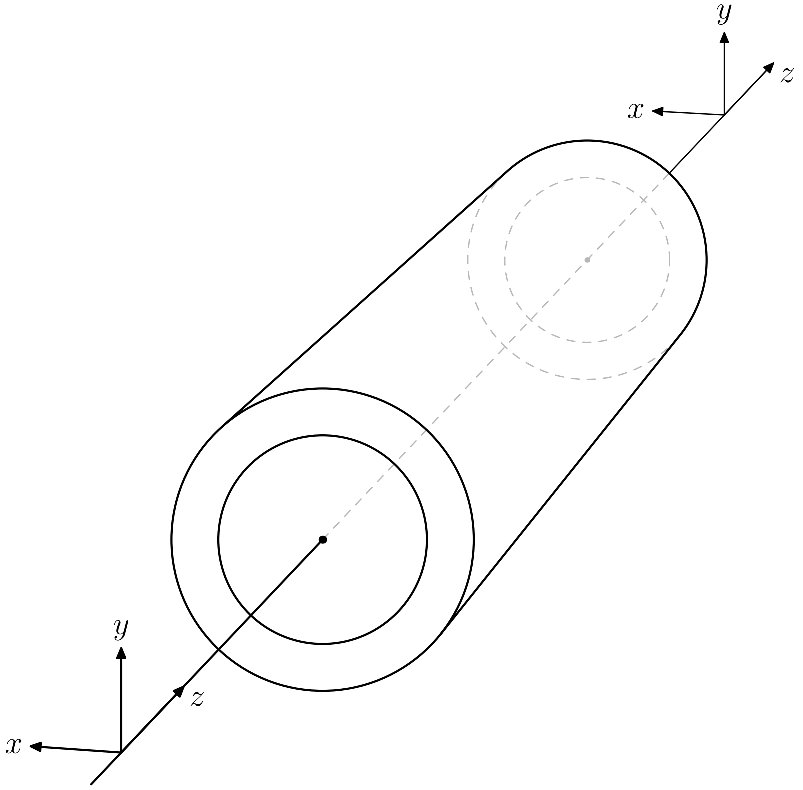


Figure 16.2.1: Coordinate system for a solenoid.

$$\hat{A}^0_z = 0, \tag{16.2.13}$$

which can be written in the vector form

$$\hat{\boldsymbol{A}}^0(\boldsymbol{r}) = -(1/2)\boldsymbol{r} \times \boldsymbol{B}(0,0,z) + \text{higher order terms.} \tag{16.2.14}$$

Note the resemblance between (2.14) and (15.2.61). This resemblance should not surprise us because we know from Section 15.7.1 that $\hat{\boldsymbol{A}}^0$ is in the Poincaré-Coulomb gauge with respect to any origin on the z axis.

From (2.3) through (2.5) and (2.7) through (2.13) we see that both $\boldsymbol{B}$ and $\hat{\boldsymbol{A}}^0$ are completely specified in terms of a single "master" function $C_0^{[1]}(z)$ and its derivatives. Moreover, according to (2.6), the function $C_0^{[1]}(z)$ is given in terms of the longitudinal on-axis field by the relation

$$C_0^{[1]}(z) = B_z(0,0,z). \tag{16.2.15}$$

We observe that for a long uniform solenoid the on-axis field $B_z(0,0,z)$ will be nearly constant in the body of the solenoid, and therefore the quantities $C_0^{[n]}(z)$ will be small in this region for $n > 1$. However, these derivatives may be large in fringe-field regions. We will next see what can be said more specifically about the master function $C_0^{[1]}(z)$ in various cases.

16.2.2 Simple Air-Core Solenoid

16.2.2.1 General Properties of the On-Axis Field

For simplicity, consider initially the case of a simple air-core solenoid consisting of a single-layer circular cylindrical winding of length L and radius $\rho = a$, and powered so that the interior field (in the infinite-length limit) is B. For such a solenoid it can be shown that the on-axis field is given by the relation

$$B_z(0,0,z) = B\{z/[z^2 + a^2]^{1/2} - (z - L)/[(z - L)^2 + a^2]^{1/2}\}/2 \tag{16.2.16}$$

where the cylinder axis is the z axis and the winding extends from $z = 0$ to $z = L$. The on-axis fields of more general air-core solenoids can be found from (2.16) by superposition. See Subsection 2.4. Here we assume that the effect of a solenoidal winding is well approximated by a uniform current sheet (or a collection of uniform current sheets) in the $\boldsymbol{e}_\phi$ direction. For a discussion of helical effects, see the book by W. Smythe cited in the references at the end of this chapter.

Suppose we define a soft-edge "bump" function bump(z, a, L) by the rule

$$\text{bump}(z, a, L) = \{z/[z^2 + a^2]^{1/2} - (z - L)/[(z - L)^2 + a^2]^{1/2}\}/2 \tag{16.2.17}$$

so that (2.16) can be written in the form

$$B_z(0,0,z) = B \ \text{bump}(z, a, L). \tag{16.2.18}$$

Then there is also the result

$$C_0^{[1]}(z) = B_z(0,0,z) = B \ \text{bump}(z, a, L). \tag{16.2.19}$$

It can be verified that the soft-edge bump function has the properties

$$\text{bump}(z, a, L) \simeq 1 \text{ for } z \in [0, L], \tag{16.2.20}$$

$$\text{bump}(z, a, L) \simeq 0 \text{ elsewhere}, \tag{16.2.21}$$

$$\text{bump}(L/2 + w, a, L) = \text{bump}(L/2 - w, a, L), \tag{16.2.22}$$

It can also be shown that

$$\int_{-\infty}^{\infty} \text{bump}(z, a, L) dz = L. \tag{16.2.23}$$

See Exercise 2.3. In particular, from the results above, it follows that for a simple air-core solenoid there is the relation

$$\int_{-\infty}^{\infty} C_0^{[1]}(z) dz = \int_{-\infty}^{\infty} B_z(0,0,z) dz = BL. \tag{16.2.24}$$

At this point two remarks are in order: The first is that we have been using the term *bump function* in a slightly different way from that often employed in mathematics. In mathematics a bump function is generally a smooth (C^∞) function with exact value 1 over some region and exact value 0 slightly outside this region. By contrast, in (2.20) and (2.21), we require only that this be approximately true. The second is that the relation (2.24) also holds in the case of a *thick* air-core solenoid. See Subsection 2.4 and Exercise 2.9.

We also observe that, according to (2.17), the soft-edge bump function can be written in the form

$$\mathrm{bump}(z,a,L) = [\mathrm{sgn}(z,a) - \mathrm{sgn}(z-L,a)]/2 \tag{16.2.25}$$

where $\mathrm{sgn}(z,a)$ is an approximating "signum" function given by the relation

$$\mathrm{sgn}(z,a) = z/[z^2+a^2]^{1/2}. \tag{16.2.26}$$

Figures 2.2 and 2.3 illustrate the behavior of the approximating signum function for two different values of a. Evidently the approximating signum function becomes the true signum function in the limit that a goes to zero,

$$\lim_{a\to 0} \mathrm{sgn}(z,a) = \mathrm{sgn}(z). \tag{16.2.27}$$

Recall that the true signum function has the definition

$$\begin{aligned}
&\mathrm{sgn}(z) = 1 \text{ if } z>0,\\
&\mathrm{sgn}(z) = 0 \text{ if } z=0,\\
&\mathrm{sgn}(z) = -1 \text{ if } z<0.
\end{aligned} \tag{16.2.28}$$

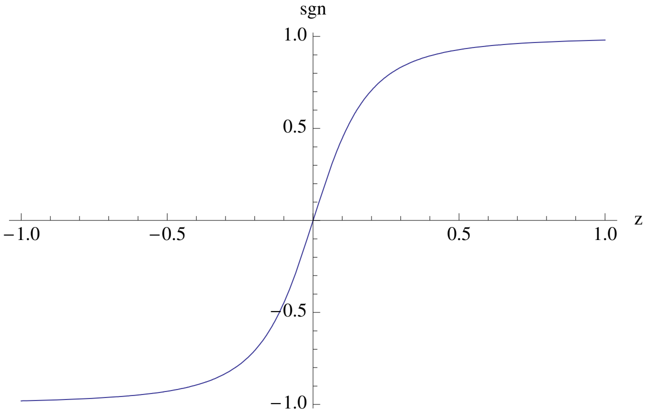


Figure 16.2.2: The approximating signum function (2.26) when $a = .2$.

In this same limit the soft-edge bump function becomes the hard-edge bump function,

$$\lim_{a\to 0} \mathrm{bump}(z,a,L) = \mathrm{bump}(z,L). \tag{16.2.29}$$

The hard-edge bump function has the properties,

$$\begin{aligned}
&\mathrm{bump}(z,L) = 1 \text{ for } z\in(0,L),\\
&\mathrm{bump}(0,L) = \mathrm{bump}(L,L) = 1/2,\\
&\mathrm{bump}(z,L) = 0 \text{ elsewhere.}
\end{aligned} \tag{16.2.30}$$

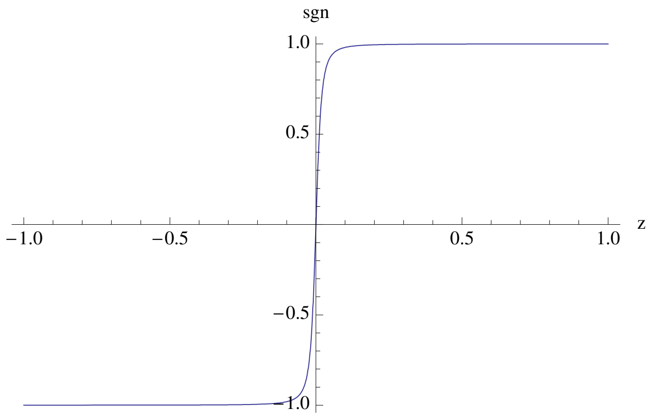


Figure 16.2.3: The approximating signum function (2.26) when $a = .02$.

Figures 2.4 and 2.5 illustrate the bump-function properties (2.20) through (2.22) for a fixed value of L and two different values of a. As expected, the simple air-core solenoid soft-edge bump function approaches a hard-edge bump function in the limit $a \to 0$. We also see that the quantity a plays the role of a *characteristic length* that controls the rate of fall off. The fringe-field region is large if a is large, and vanishes as a goes to zero. Finally, we note that $\mathrm{sgn}(z, a)$ is *analytic* as a function of z save for branch points at $z = \pm ia$. Correspondingly, $\mathrm{bump}(z, a, L)$ and hence also all the $C_0^{[n]}(z)$ are analytic in z save for branch points at $z = \pm ia$ and $z = L \pm ia$. Therefore approximating $C_0^{[1]}(z)$ by a series of straight-line segments, as is sometimes done in the literature, violates its fundamental analytic properties.

Note also that use of the term *fringe field* to describe what is going on here can be a bit misleading. It is true that the field does extend beyond/outside the solenoid boundaries $z = 0$ and $z = L$, but it is also affected/diminished *inside* the boundaries, particularly noticeably in the vicinity of the boundaries. Finally note that (2.24) holds for all a and does not depend on a. Thus, so to speak, whatever on-axis field "disappears" from inside the boundaries of the solenoid due to fringing behavior is in fact found outside the boundaries.

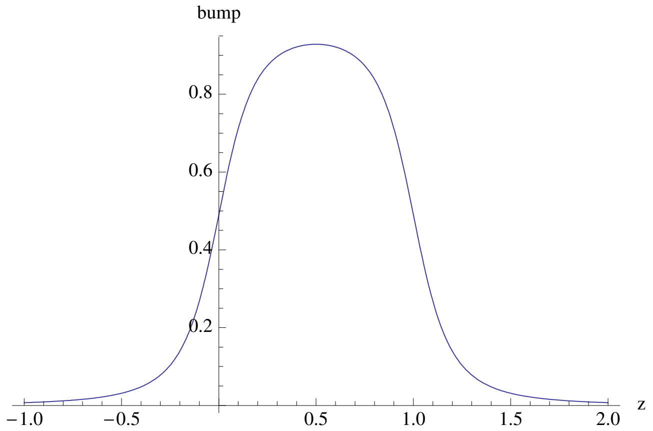


Figure 16.2.4: The soft-edge bump function (2.17) when $a = .2$ and $L = 1$.

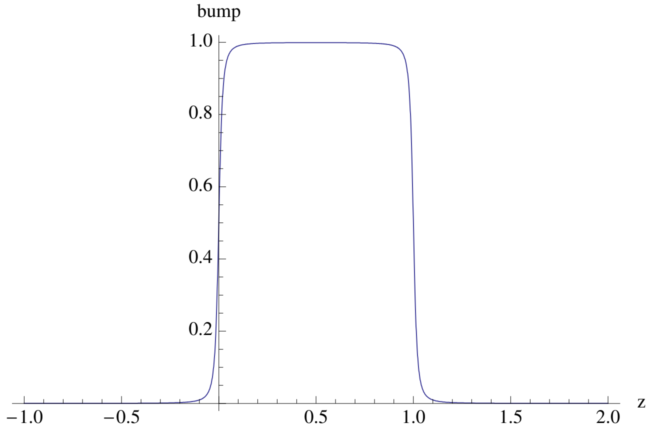


Figure 16.2.5: The soft-edge bump function (2.17) when $a = .02$ and $L = 1$.

16.2.2.2 Asymptotic Behavior of the On-Axis Field

Let us examine the behaviors of the approximating signum function and soft-edge bump function in more detail. Evidently the approximating signum function is an odd function of z,

$$\mathrm{sgn}(-z,a) = \mathrm{sgn}(z,a), \tag{16.2.31}$$

and, correspondingly, satisfies the relation

$$\mathrm{sgn}(0,a) = 0. \tag{16.2.32}$$

It is also easy to verify that

$$\mathrm{sgn}(\pm a,a) = \pm 1/\sqrt{2} = \pm .707\cdots\ . \tag{16.2.33}$$

Finally, It can be verified from (2.26) that, when $(a/z)^2 < 1$, there are the asymptotic behaviors

$$\mathrm{sgn}(z,a) = 1 - (1/2)(a/z)^2 + (3/8)(a/z)^4 - (15/48)(a/z)^6 + \cdots \text{ as } z \to +\infty, \tag{16.2.34}$$

$$\mathrm{sgn}(z,a) = -1 + (1/2)(a/z)^2 - (3/8)(a/z)^4 + (15/48)(a/z)^6 - \cdots \text{ as } z \to -\infty. \quad (16.2.35)$$

The behavior of the soft-edge bump function is a bit more complicated but, according to (2.25), follows from that of the approximating signum function. We begin with two simple observations. From (2.25) and (2.32) we see that there are the "end" ($z = 0$ and $z = L$) values

$$\mathrm{bump}(0,a,L) = \mathrm{bump}(L,a,L) = (1/2)L/(L^2 + a^2)^{1/2} \to (1/2) \text{ as } (a/L) \to 0, \quad (16.2.36)$$

and the "center" ($z = L/2$) value

$$\mathrm{bump}(L/2,a,L) = L/(L^2 + 4a^2)^{1/2} \to 1 \text{ as } (a/L) \to 0. \quad (16.2.37)$$

We will next study the soft-edge bump function's *near* leading/entering end behavior, its behavior when $z \in (-L, 0)$. [Its near trailing/exiting end behavior, its behavior when $z \in (L, 2L)$, then follows by symmetry.] For this purpose, according to (2.25), we need to know the behavior of both $\mathrm{sgn}(z,a)$ and $\mathrm{sgn}(z-L,a)$ when $z < 0$. The behavior of $\mathrm{sgn}(z,a)$ for $z < 0$ is given by (2.35). For $\mathrm{sgn}(z - L, a)$ we find (when $z \approx 0$) the expansion

$$\begin{aligned}\mathrm{sgn}(z-L,a) &= (z-L)/[(z-L)^2 + a^2]^{1/2} = -[1 + a^2/(z-L)^2]^{1/2} \\ &= -1 + (1/2)[a/(z-L)]^2 - (3/8)[a/(z-L)]^4 + \cdots \\ &= -1 + (1/2)(a/L)^2 + (a^2 z/L^3) + \cdots . \quad (16.2.38)\end{aligned}$$

(Here we assume that both z/L and a/L are small.) Consequently, for the range $z \approx 0$ but $z < -a$ so that (2.35) holds, we conclude from (2.25) there is the expansion

$$\begin{aligned}\mathrm{bump}(z,a,L) &= [\mathrm{sgn}(z,a) - \mathrm{sgn}(z-L,a)]/2 \\ &= -1/2 + (1/4)(a/z)^2 - (3/16)(a/z)^4 + \cdots \\ &\quad +1/2 - (1/4)(a/L)^2 - (1/2)(a^2 z/L^3) - \cdots \\ &= +(1/4)(a/z)^2 - (1/4)(a/L)^2 - (1/2)(a/L)^2(z/L) - \cdots . \\ &\qquad (16.2.39)\end{aligned}$$

We see that the $(a/z)^2$ term dominates for small z and therefore $\mathrm{bump}(z,a,L)$ decreases as $1/z^2$ as z becomes more negative. Upon reflection, this result should be expected. Close by the end of a solenoid the external field looks like a *monopole* field; and the field of a monopole falls off with distance as the inverse square.

Complete asymptotic behavior does not set in until $z < -L$. In that case, using (2.35), we see that $\mathrm{sgn}(z - L, a)$ has the expansion

$$\begin{aligned}\mathrm{sgn}(z-L,a) &= -1 + (1/2)[a/(z-L)]^2 - (3/8)[a/(z-L)]^4 + \cdots \\ &= -1 + (1/2)(a/z)^2 + (La^2/z^3) \\ &\quad +(3/8)(4a^2L^2 - a^4)/z^4 + \cdots \text{ when } z < -L. \quad (16.2.40)\end{aligned}$$

We now find from (2.25) the result

$$\begin{aligned}\mathrm{bump}(z,a,L) &= [\mathrm{sgn}(z,a) - \mathrm{sgn}(z-L,a)]/2 \\ &= -1/2 + (1/4)(a/z)^2 - (3/16)(a/z)^4 + \cdots \\ &\quad +1/2 - (1/4)(a/z)^2 - (1/2)(La^2/z^3) + \cdots \\ &= -(1/2)La^2/z^3 + O(1/z^4) \text{ when } z \to -\infty. \quad (16.2.41)\end{aligned}$$

Correspondingly, the on-axis gradient $C_0^{[1]}(z)$ and on-axis field $B_z(0,0,z)$ fall off for very large ($z < -L$) distances as

$$C_0^{[1]}(z) = B_z(0,0,z) = -(1/2)BLa^2/z^3 + \cdots \text{ when } z \to -\infty. \tag{16.2.42}$$

Analogous fall off occurs when $z > 2L$. This result is also to be expected. From far enough away, the end fields of a solenoid of length L look like those of two monopoles of opposite signs a distance L apart, and therefore combine to appear as the field of a *dipole* once one is more than a distance L away. Finally, at large distances, the field of a dipole falls off as $1/|z|^3$.

We close this subsection with the injunction that, although the asymptotic expansions we have examined are illuminating, there is no substitute for computing $B_z(0,0,z)$ exactly using (2.17) and (2.18).

16.2.2.3 Properties of the Vector Potential

The computation of orbits in and transfer maps for solenoids, using a Hamiltonian formulation, requires the use of a vector potential. We will employ the vector potential given by (2.11) through (2.13). Evidently, depending on the order to which we wish to work, we need the functions $C_0^{[1]}(z)$, $C_0^{[3]}(z)$, $C_0^{[5]}(z)$,$\cdots$. To get a feel for what is involved, let us examine, for example, the function $C_0^{[3]}(z)$. From (2.15) and (2.19) we see that

$$C_0^{[3]}(z) = (\partial/\partial z)^2 B_z(0,0,z) = B\ (\partial/\partial z)^2 \text{bump}(z,a,L) = B\ \text{bump}''(z,a,L). \tag{16.2.43}$$

Figures (2.6) and (2.7) illustrate the function bump$''$ for a fixed value of L and two different values of a. Evidently the function bump$''$ becomes quite singular at the ends of the solenoid in the limit $a \to 0$. Indeed it approaches the function $\delta'(z)$ at the leading end, and the function $-\delta'(z-L)$ at the trailing end. Moreover, it falls off quite rapidly beyond the fringe-field regions. From (2.39) and (2.41) we conclude that there is the near-by asymptotic behavior

$$\text{bump}''(z,a,L) = (3/2)a^2/z^4 + \cdots \text{ as } z \to -\infty, \tag{16.2.44}$$

and the far asymptotic behavior

$$\text{bump}''(z,a,L) = -6La^2/z^5 + \cdots \text{ as } z \to -\infty, \tag{16.2.45}$$

The still higher derivatives of the bump function, needed to compute the $C_0^{[n]}(z)$ for still larger values of n, are even more singular in the limit $a \to 0$, and fall off ever more rapidly as $z \to -\infty$. Analogous results hold for $z > L$ and $z > 2L$.

16.2.3 Opposing Simple Solenoid Doublet

We have seen that the on-axis field of a single simple solenoid has the *far* asymptotic behavior (2.42). A sequence of solenoids, all having the same "sign", will have the same far fall-off behavior. In this subsection we will study the far fall-off behavior for what we call an *opposing solenoid doublet*. By this term we mean a pair of solenoids, each of length L,

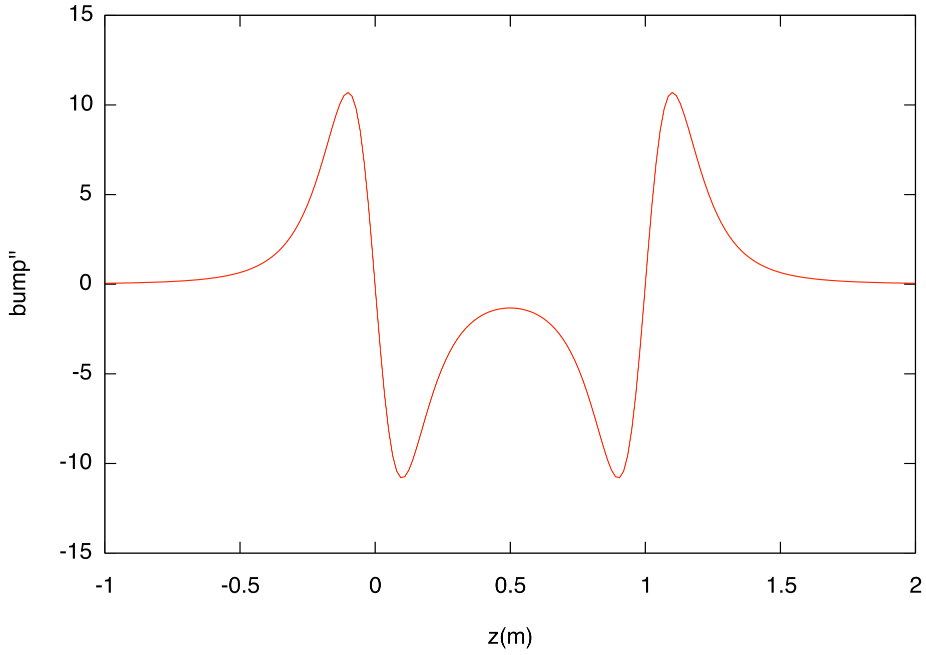


Figure 16.2.6: The function bump$''$ when $a = .2$ and $L = 1$.

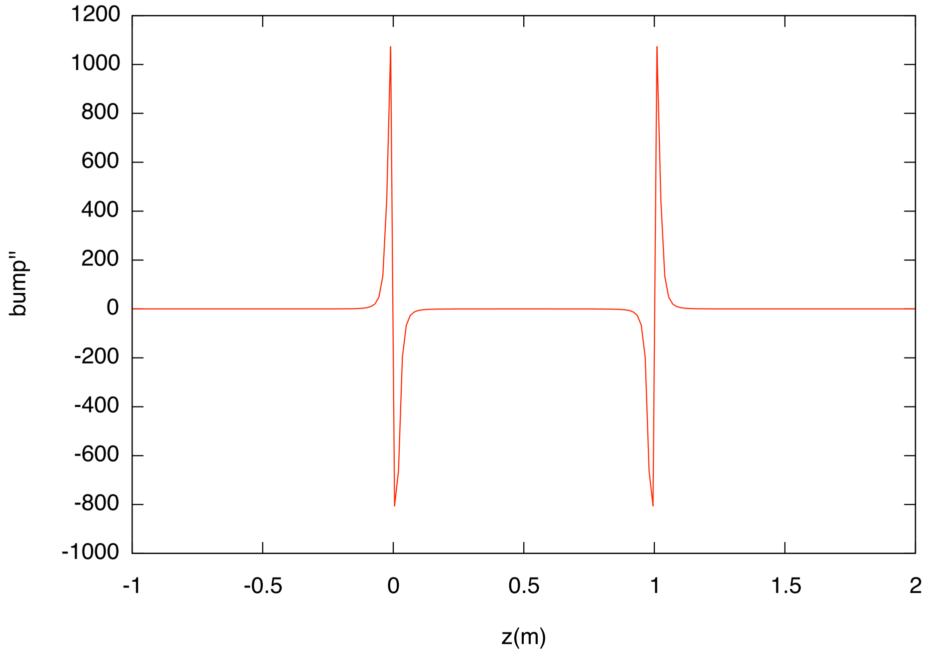


Figure 16.2.7: The function bump$''$ when $a = .02$ and $L = 1$.

separated by a space D, and having opposite strengths. By superposition, the on-axis field for such a pair of solenoids is given by the relation

$$\begin{aligned}
&B_z^{\text{osd}}(0,0,z) = B\ \text{bump}(z,a,L) - B\ \text{bump}(z-L-D,a,L)\\
&= B\{z/[z^2+a^2]^{1/2} - (z-L)/[(z-L)^2+a^2]^{1/2}\}/2\\
&-B\{(z-L-D)/[(z-L-D)^2+a^2]^{1/2} - (z-2L-D)/[(z-2L-D)^2+a^2]^{1/2}\}/2.\\
&= B[\text{sgn}(z,a) - \text{sgn}(z-L,a) - \text{sgn}(z-L-D,a) + \text{sgn}(z-2L-D,a)].
\end{aligned} \tag{16.2.46}$$

Computation using expansions such as (2.42) shows that for this opposing solenoid doublet there is the far fall-off behavior

$$C_0^{[1]}(z) = B_z^{\text{osd}}(0,0,z) = 3Ba^2L(L+D)/z^4 + O(1/z^5) \text{ when } z \to -\infty. \tag{16.2.47}$$

Comparison with (2.42) shows that this $1/z^4$ far fall-off behavior is one order higher in $1/z$ than that for a single solenoid. This result is to be expected because the end of each solenoid looks like a monopole, at a far distance the four ends of the two solenoids in the opposing solenoid doublet look like an in-line quadrupole, and the field of a quadrupole falls off as $1/z^4$.

16.2.4 More Complicated Air-Core Solenoids

The fields for more complicated air-core solenoids can be found from those of simple single-layer air-core solenoids by superposition. Consider, for example, the on-axis field of an air-core solenoid that has a multi-layer winding with inner radius a_1 and outer radius a_2. We will call such a solenoid a *thick* solenoid. We observe that there is the integral relation

$$\int_{a_1}^{a_2} da\ \{1/[z^2+a^2]^{1/2}\} = \log\Big(\{[z^2+a_2^2]^{1/2}+a_2\}/\{[z^2+a_1^2]^{1/2}+a_1\}\Big). \tag{16.2.48}$$

Correspondingly, the on-axis field of such a thick solenoid is given by the relation

$$\begin{aligned}
B_z(0,0,z) \ &= \ (B/2)[1/(a_2-a_1)]\Big[z\log\Big(\{[z^2+a_2^2]^{1/2}+a_2\}/\{[z^2+a_1^2]^{1/2}+a_1\}\Big)\\
&- \ (z-L)\log\Big(\{[(z-L)^2+a_2^2]^{1/2}+a_2\}/\{[(z-L)^2+a_1^2]^{1/2}+a_1\}\Big)\Big].
\end{aligned} \tag{16.2.49}$$

Here again the winding extends from $z=0$ to $z=L$ and the interior field (in the infinite length limit) is B.

Evidently, from this result and by superposition, analytic on-axis results can be obtained for any combination of concentric coils of various lengths, thicknesses, locations, and powerings. Note, because we have assumed cylindrical symmetry in all cases, only the $m=0$ terms are present in the expansion (15.3.33) so that (2.1) continues to hold.

16.2.5 Computation of Transfer Map

In this subsection we will compute the transfer map for a solenoid (or a collection of solenoids) including fringe-field effects. To do so we begin with the Hamiltonian (1.1) and employ the vector potential given by (2.7) through (2.10). We then introduce dimensionless scaled deviation variables and the associated scaled deviation-variable Hamiltonian. Finally, we expand the scaled deviation-variable Hamiltonian in a Taylor series, and employ this Taylor series to compute the transfer map.

16.2.5.1 Dimensionless Scaled Deviation Variables and Scaled Deviation-Variable Hamiltonian

A solenoid is an example of a straight beam-line element for which the design orbit may be taken to be the z axis (a straight line) traversed with constant velocity. According to the results of Section 13.1.5, the scaled deviation variable Hamiltonian $H(X,Y,\tau,P_x,P_y,P_\tau;z)$ for any such element is given by

$$\begin{aligned}&H(X,Y,\tau,P_x,P_y,P_\tau;z)=\\&\quad-(1/\ell)\{[1-(2P_\tau/\beta_0)+P_\tau^2-(P_x-A_x^s)^2-(P_y-A_y^s)^2]^{1/2}+(P_\tau/\beta_0)-(1/\beta_0^2)\}.\end{aligned}\tag{16.2.50}$$

Here the dimensionless scaled deviation variables $(X,Y,\tau,P_x,P_y,P_\tau)$ are defined in terms of the original variables (x,y,t,p_x,p_y,p_t) by the relations (13.1.21) through (13.1.26), and the *scaled* vector potential $\boldsymbol{A}^s$ is defined in terms of the original vector potential $\hat{\boldsymbol{A}}^0$ by the relation

$$\boldsymbol{A}^s(X,Y,z)=(q/p^0)\hat{\boldsymbol{A}}^0(\ell X,\ell Y,z).\tag{16.2.51}$$

We have also used (2.9).

The relations (2.7) through (2.10) and (2.51) can be used to find $\boldsymbol{A}^s$ for solenoids. To do so it is convenient to rewrite (2.10) in the form

$$U(\rho,z)=(1/2)\sum_{n=0}^{\infty}(-1)^n\frac{1}{2^{2n}n!(n+1)!}B^{[2n]}(z)\rho^{2n}.\tag{16.2.52}$$

Here we have introduced, in accord with (2.15), the notation

$$B^{[0]}(z)=C_0^{[1]}(z)=B_z(0,0,z),\tag{16.2.53}$$

$$B^{[2n]}(z)=(\partial/\partial z)^{2n}B^{[0]}(z)=(\partial/\partial z)^{2n}C_0^{[1]}(z)=C_0^{[2n+1]}(z).\tag{16.2.54}$$

Combining these relations yields the results

$$A_x^s=-YU^s(X,Y,z)\tag{16.2.55}$$

$$A_y^s=XU^s(X,Y,z),\tag{16.2.56}$$

$$A_z^s=0,\tag{16.2.57}$$

where

$$U^s(X,Y,z) = \ell(q/p^0)U(\ell X,\ell Y,z) = (1/2)\sum_{n=0}^{\infty}(-1)^n \frac{1}{2^{2n}n!(n+1)!}b^{[2n]}(z)(X^2+Y^2)^n. \tag{16.2.58}$$

Here we have introduced the notation

$$b^{[2n]}(z) = (q/p^0)\ell^{2n+1}B^{[2n]}(z), \tag{16.2.59}$$

and observe, in view of (1.5.81) and (2.54), that the quantities $b^{[2n]}(z)$ are dimensionless. Also, from (2.50), we conclude that the scaled deviation variable Hamiltonina H has dimensions of 1/length.

16.2.5.2 Symmetry of Scaled Deviation-Variable Hamiltonian

The Hamiltonian H given by (2.50) has a symmetry that is worth noting. Define two two-dimensional vectors $\boldsymbol{Q}$ and $\boldsymbol{P}$ by the rules

$$\boldsymbol{Q} = X\boldsymbol{e}_x + Y\boldsymbol{e}_y, \tag{16.2.60}$$

$$\boldsymbol{P} = P_x\boldsymbol{e}_x + P_y\boldsymbol{e}_y. \tag{16.2.61}$$

Also make the definitions

$$Q^2 = \boldsymbol{Q}\cdot\boldsymbol{Q} = X^2+Y^2, \tag{16.2.62}$$

$$P^2 = \boldsymbol{P}\cdot\boldsymbol{P} = P_x^2+P_y^2, \tag{16.2.63}$$

$$J_z = (\boldsymbol{Q}\times\boldsymbol{P})\cdot\boldsymbol{e}_z = XP_y - YP_x. \tag{16.2.64}$$

With the aid of these definitions we may write

$$\begin{aligned}(P_x-A_x^s)^2+(P_y-A_y^s)^2 &= (P_x)^2+(P_y)^2+(A_x^s)^2+(A_y^s)^2-2P_xA_x^s-2P_yA_y^s\\ &= \boldsymbol{P}\cdot\boldsymbol{P}+\boldsymbol{A}^s\cdot\boldsymbol{A}^s-2\boldsymbol{P}\cdot\boldsymbol{A}^s.\end{aligned} \tag{16.2.65}$$

We also observe with the aid of (2.55) and (2.56) that there are the relations

$$\boldsymbol{P}\cdot\boldsymbol{A}^s = (XP_y-YP_x)U^s = J_zU^s, \tag{16.2.66}$$

$$\boldsymbol{Q}\cdot\boldsymbol{A}^s = 0, \tag{16.2.67}$$

and

$$\boldsymbol{A}^s\cdot\boldsymbol{A}^s = Q^2(U^s)^2. \tag{16.2.68}$$

The relation (2.67) is a consequence of our decision to employ the Poincaré-Coulomb gauge. Also note that, according to (2.58), U^s depends only on Q^2 and z. With the aid of these relations we may also write

$$(P_x-A_x^s)^2+(P_y-A_y^s)^2 = P^2+Q^2(U^s)^2-2J_zU^s. \tag{16.2.69}$$

It is easily verified that J_z has the properties

$$: J_z : X = [J_z, X] = [(XP_y - YP_x), X] = Y, \tag{16.2.70}$$

$$: J_z : Y = -X, \tag{16.2.71}$$

$$: J_z : P_x = P_y, \tag{16.2.72}$$

$$: J_z : P_y = -P_x, \tag{16.2.73}$$

$$: J_z : \tau =: J_z : P_\tau = 0. \tag{16.2.74}$$

Consequently, as the notation is meant to suggest, the Lie operator $: J_z :$ is the generator of rotations about the z axis. It follows that there are the relations

$$: J_z : Q^2 =: J_z : U^s =: J_z : P^2 =: J_z : (\boldsymbol{P} \cdot \boldsymbol{Q}) = 0. \tag{16.2.75}$$

We remark that the last relation in (2.75) is consistent with the identity

$$(\boldsymbol{P} \cdot \boldsymbol{Q})^2 = Q^2 P^2 - J_z^2. \tag{16.2.76}$$

We also see from (2.55) and (2.56) that there are the relations

$$: J_z : A_x^s == [J_z, -YU^s] = -[J_z, Y]U^s = XU^s = A_y^s \tag{16.2.77}$$

and

$$: J_z : A_y^s == [J_z, XU^s] = [J_z, X]U^s = YU^s = -A_x^s. \tag{16.2.78}$$

It follows that

$$: J_z : (\boldsymbol{A}^s \cdot \boldsymbol{A}^s) =: J_z : (\boldsymbol{Q} \cdot \boldsymbol{A}^s) =: J_z : (\boldsymbol{P} \cdot \boldsymbol{A}^s) = 0. \tag{16.2.79}$$

We note that these last results can be viewed as a consequence of (2.68), (2.67), and (2.66). But they can also be viewed as consequence of the relations (2.70) through (2.73) and (2.77) and (2.78).

Based on the work so far H as given by (2.50) can be rewritten in either of the forms

$$\begin{aligned} &H(X, Y, \tau, P_x, P_y, P_\tau; z) = \\ &-(1/\ell)\{[1 - (2P_\tau/\beta_0) + P_\tau^2 - \boldsymbol{P} \cdot \boldsymbol{P} - \boldsymbol{A}^s \cdot \boldsymbol{A}^s + 2\boldsymbol{P} \cdot \boldsymbol{A}^s]^{1/2} + (P_\tau/\beta_0) - (1/\beta_0^2)\}. \end{aligned} \tag{16.2.80}$$

and

$$\begin{aligned} &H(X, Y, \tau, P_x, P_y, P_\tau; z) = \\ &-(1/\ell)\{[1 - (2P_\tau/\beta_0) + P_\tau^2 - P^2 - Q^2(U^s)^2 + 2J_z U^s]^{1/2} + (P_\tau/\beta_0) - (1/\beta_0^2)\}. \end{aligned} \tag{16.2.81}$$

From either of these forms it is evident, using (2.75) and (2.79), that

$$: J_z : H = [J_z, H] = 0. \tag{16.2.82}$$

That is, H is invariant under rotations about the z axis and, conversely, J_z is an integral of motion. (We also say that H and J_z are in involution.) Note that this invariance stems from the fact that we are dealing with the the magnetic field case described by an $m = 0$ scalar potential ψ. This scalar potential has rotational symmetry about the z axis, and the gauge for the associated vector potential has been judiciously chosen to maintain this symmetry. See Exercise 2.10.

16.2.5.3 Properties of Transfer Map and Factorization of Linear Part

Because of the J_zU^s term in (2.81) [or the $\boldsymbol{P}\cdot\boldsymbol{A}^s$ term in (2.80)] the map $\mathcal{M}$ generated by H produces rotations about the z axis as well as other effects. Consequently $\mathcal{M}$ does not preserve the X, P_x and Y, P_y planes. When performing fitting operations it is easier to understand what is happening when motion in the X, P_x and Y, P_y planes is uncoupled. Uncoupling can be accomplished by a trick. Define a Hamiltonian with no rotational parts, call it H^{nonrot}, by removing the J_zU^s term in (2.81):

$$\begin{aligned}&H^{\text{nonrot}}(X,Y,\tau,P_x,P_y,P_\tau;z) =\\ &\quad -(1/\ell)\{[1-(2P_\tau/\beta_0)+P_\tau^2-P^2-Q^2(U^s)^2]^{1/2}+(P_\tau/\beta_0)-(1/\beta_0^2)\}.\end{aligned} \tag{16.2.83}$$

The map generated by H^{nonrot}, call it $\mathcal{M}^{\text{nonrot}}$, will preserve the X, P_x and Y, P_y planes. See Exercise 2.17. To proceed, first carry out the desired fitting operation using $\mathcal{M}^{\text{nonrot}}$ in place of $\mathcal{M}$. After a fit has been achieved using $\mathcal{M}^{\text{nonrot}}$ in place of $\mathcal{M}$, continue on using the associated full $\mathcal{M}$ in subsequent calculations.

In general the maps $\mathcal{M}$ and $\mathcal{M}^{\text{nonrot}}$ do not commute. However, as shown in Exercise 2.15, the matrices M and M^{nonrot} associated with their linear parts do commute,

$$M^{\text{nonrot}}M = MM^{\text{nonrot}}. \tag{16.2.84}$$

In this case it is possible to define a matrix M^{rot} by the rule

$$M^{\text{rot}} = (M^{\text{nonrot}})^{-1}M = M(M^{\text{nonrot}})^{-1} \tag{16.2.85}$$

so that

$$M = M^{\text{rot}}M^{\text{nonrot}} = M^{\text{nonrot}}M^{\text{rot}}. \tag{16.2.86}$$

As the notation suggests, M^{rot} describes rotations about the z axis.

For example, for the matrix M^{body} displayed in Exhibit 2.11 in Subsection 2.6.2, Exhibits 2.1 and 2.2 below display the factors M^{nonrot} and M^{rot}.

Exhibit 16.2.1: The matrix M^{nonrot} factor of the M^{body} displayed in Exhibit 2.11

```
 9.94943E-01  9.98265E-01  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
-1.01064E-02  9.94943E-01  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  9.94943E-01  9.98265E-01  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00 -1.01064E-02  9.94943E-01  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  4.11143E-01
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
```

Exhibit 16.2.2: The matrix M^{rot} factor of the M^{body} displayed in Exhibit 2.11

```
 9.94963E-01  0.00000E+00  1.00240E-01  0.00000E+00  0.00000E+00  0.00000E+00
 0.00000E+00  9.94963E-01  0.00000E+00  1.00240E-01  0.00000E+00  0.00000E+00
-1.00240E-01  0.00000E+00  9.94963E-01  0.00000E+00  0.00000E+00  0.00000E+00
 0.00000E+00 -1.00240E-01  0.00000E+00  9.94963E-01  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
```

As can be seen, the effect of $M^{\rm nonrot}$ is to produce (equal) *focussing* in both planes. (Note also that $M^{\rm nonrot}$ does not introduce any coupling between planes.) And $M^{\rm rot}$ produces a rotation about the z axis by an angle $\theta_{\rm rot}$. In the case that the entrance and exit planes are *well outside* the fringe-field regions [which, as can be seen from Figure 2.5 and (2.42), is not quite true for this example], $\theta_{\rm rot}$ is given (in radians) by the relation

$$\theta_{\rm rot} = [1/(2\ {\rm brho})]\int_{-\infty}^{+\infty} B_z(0,0,z)\,dz = BL/(2\ {\rm brho}). \tag{16.2.87}$$

Here brho is the magnetic rigidity. See Exercise 1.5.9. Note that the result (2.87) does not depend on a in the case of a simple solenoid or on a_1 and a_2 in the case of a thick solenoid.

16.2.5.4 Expansion of Scaled Deviation-Variable Hamiltonian

To find the transfer map $\mathcal{M}$ about the design orbit it is necessary to express the scaled deviation-variable Hamiltonian H as a sum of homogeneous polynomials,

$$H = \sum_{m=0}^{\infty} H_m. \tag{16.2.88}$$

Doing so for the Hamiltonian (2.50) gives, for the first few terms, the results

$$H_0 = 1/(\beta_0^2\gamma_0^2\ell), \tag{16.2.89}$$

$$H_1 = 0, \tag{16.2.90}$$

$$\begin{aligned} H_2 = {} & [1/(2\ell)](P_x^2 + P_y^2) - [b^{[0]}/(2\ell)](XP_y - YP_x) \\ & + [(b^{[0]})^2/(8\ell)](X^2 + Y^2) + [1/(2\beta_0^2\gamma_0^2\ell)]P_\tau^2, \end{aligned} \tag{16.2.91}$$

$$\begin{aligned} H_3 = {} & [1/(2\beta_0\ell)]P_\tau(P_x^2 + P_y^2) - [b^{[0]}/(2\beta_0\ell)]P_\tau(XP_y - YP_x) \\ & + [(b^{[0]})^2/(8\beta_0\ell)]P_\tau(X^2 + Y^2) + [1/(\beta_0^3\gamma_0^2\ell)]P_\tau^3, \end{aligned} \tag{16.2.92}$$

$$\begin{aligned} H_4 = {} & (1/8\ell)(P_x^4 + 2P_x^2P_y^2 + P_y^4) - [b^{[0]}/(4\ell)](P_x^2 + P_y^2)(XP_y - YP_x) \\ & + [(b^{[0]})^2/(16\ell)](X^2P_x^2 + Y^2P_y^2) + [3(b^{[0]})^2/(16\ell)](X^2P_y^2 + Y^2P_x^2) \\ & - [(b^{[0]})^2/(4\ell)](XP_xYP_y) + \{[b^{[2]} - (b^{[0]})^3]/(16\ell)\}(X^2 + Y^2)(XP_y - YP_x) \\ & + \{[(b^{[0]})^4 - 4b^{[0]}b^{[2]}]/(128\ell)\}(X^4 + 2X^2Y^2 + Y^4) - [(3 - \beta_0^3)/(4\beta_0^2\ell)]P_\tau^2(P_x^2 + P_y^2) \\ & - [b^{[0]}(3 - \beta_0^3)/(4\beta_0^2\ell)]P_\tau^2(XP_y - YP_x) + [(b^{[0]})^2(3 - \beta_0^3)/(16\beta_0^2\ell)]P_\tau^2(X^2 + Y^2) \\ & + [(5 - \beta_0)^2)/(8\beta_0^4\gamma_0^2\ell)]P_\tau^4. \end{aligned} \tag{16.2.93}$$

These are the terms required to compute $\mathcal{M}$ through third order. Note that H_1 vanishes as it should. If it did not, the phase-space path obtained by setting all deviation variables to zero (the design orbit) would not be a solution of the equations of motion. We also remark that the constant piece H_0 is irrelevant to the actual motion, and does not enter into the calculation of $\mathcal{M}$. It is presented only as an aid for those who wish to check the expansion (2.88) through (2.93).

In view of the symmetry (2.82), it is also instructive to have an expansion of H in terms of the variables P_τ, P^2, J_z, and Q^2. In terms of these variables there are the results

$$H_2 = [1/(2\ell)]P^2 - [b^{[0]}/(2\ell)]J_z + [(b^{[0]})^2/(8\ell)]Q^2 + [1/(2\beta_0^2\gamma_0^2\ell)]P_\tau^2, \tag{16.2.94}$$

$$H_3 = (1/\beta_0)P_\tau H_2, \tag{16.2.95}$$

$$\begin{aligned} H_4 = {} & [1/(8\ell)](P^2)^2 - [b^{[0]}/4\ell)]P^2 J_z + [b^{[0]}/(8\ell)]J_z^2 + [-b^{[0]}/(8\ell) + 3(b^{[0]})^2/(16\ell)]P^2 Q^2 \\ & + \{[b^{[2]} - (b^{[0]})^3]/(16\ell)\}Q^2 J_z + \{[(b^{[0]})^4 - 4b^{[0]}b^{[2]}]/(128\ell)\}(Q^2)^2 \\ & + [(3-\beta_0^2)/(4\beta_0^2\ell)]P_\tau^2 P^2 - [(3-\beta_0^2)b^{[0]}/(4\beta_0^2\ell)]P_\tau^2 J_z \\ & + [(3-\beta_0^2)(b^{[0]})^2/(16\beta_0^2\ell)]P_\tau^2 Q^2 + [(5-\beta_0)^2)/(8\beta_0^4\gamma_0^2\ell)]P_\tau^4. \end{aligned} \tag{16.2.96}$$

Note that, because H_3 as given by (2.95) is proportional to P_τ, all second-order aberrations for any solenoid transfer map are purely chromatic.

We close this subsection with the remark that the result given by (2.94) through (2.96) holds for any solenoid. For simplicity, in subsequent sections we will apply them to the case of a simple air-core solenoid, in which case (2.19) holds. But they are also applicable to more complicated air-core solenoids as described in Subsection 2.4 as well as solenoids containing iron.

16.2.6 Solenoidal Fringe-Field Effects: Attempts to Hard-Edge Model Them

16.2.6.1 Convergence and Divergence

Suppose we wish to make a *simple* model of fringe-field effects. The hope would be to find a model whose fields are not too different from those that can be attained by feasible magnet construction and for which analytic calculations can be made using simple approximations and without too much difficulty, thereby bypassing the need for detailed numerical calculation involving a detailed knowledge of the functions $b^{[n]}(z)$. One idea for doing so is to consider a model in which the bump function in (2.18) and (2.19) is replaced by a bump function having the properties (2.30). This so-called *hard-edge* model, for which the on-axis field begins and ends abruptly, has only limited utility. Here are several objections to this approach:

- Real solenoids, and in particular multi-layer solenoids as described in Subsection 2.4, have extended fringe fields. From (2.49) one sees that the on-axis field involves both the inner radius a_1 and the outer radius a_2. For a realistic/*thick* multi-layer solenoid a_2 is relatively large. Correspondingly, the fringe fields falls off only slowly. Therefore beginning and terminating the on-axis field abruptly is a poor approximation for real solenoids.

- Suppose we restrict our attention to single-layer solenoids as described in Subsection 2.2. In this case (which we have called the *simple* solenoid case), as examination of Figures 2.4 and 2.5 illustrates, it might be useful to attempt a hard-edge model. That is, we might attempt to compute the transfer map $\mathcal{M}$ when $a = 0$ because then, according to (2.29), the soft-edge bump function becomes the hard-edge bump function. In this case however, as described in Subsection 2.2.3, the on-axis gradient $C_0^{[3]}$ involves bump$''(z, a, L)$ which takes on the appearance of $\delta'(z)$ and $-\delta'(z - L)$ in the hard-edge limit. See Figure (2.7). We note that the appearance of the δ' functions is a consequence of the representation (2.1) which itself is a consequence of the Maxwell equations for $\boldsymbol{B}$. We also note that H_4 as given by (2.96) involves $b^{[2]}(z)$ which in turn, according to (2.43) and (2.59), can involve the pesky δ' functions. Therefore the differential equation (10.5.61) for f_4 is ill defined in the hard-edge case. One might hope to deal with this complication by making calculations for $a \neq 0$, including all fringe-field effects, and then taking the limit $a \to 0$. When this is done it can be shown that some of the the third- and higher-order aberrations (described by the f_n with $n \geq 4$) of the transfer map $\mathcal{M}$ for a solenoid become *infinite* in the hard-edge ($a \to 0$) limit! Thus, the hard-edge limit is unphysical for a solenoid if these aberrations are important.[1] Correspondingly, third-order solenoid aberrations can be reduced by making the fringe-field regions large. It also helps to make the solenoid weak since aberrations are proportional to B. If this is done, the solenoid must also be made long (to compensate for the small B) in order to maintain the desired paraxial properties.

- Finally we must acknowlege the obvious but irritating fact that the aperture of the simple solenoid, which must contain the beam, shrinks to zero as $a \to 0$.

To illustrate some of these points, let us examine the transfer map for the specific simple air-core solenoid we have been discussing. To do so we will employ the Lie-algebraic charged particle beam transport code MaryLie. Among the beam-line elements it treats is the simple air-core solenoid. Exhibits 2.3 through 2.5 below show (through third order) the transfer map $\mathcal{M}$ for the three cases $a = 0.2$, $a = .02$, and $a = .002$. (Here we use the indexing scheme of Table 39.2.1.) In all cases the solenoid has length $L = 1$, and the entry and exit planes are taken to be at $z = z^{\text{en}} = -1$ and $z = z^{\text{ex}} = 2$, respectively. All lengths are in meters, and we have used the terminology of Subsection 1.2. See Figures 2.4 and 2.5. The quantity B has the value $B = 1$ Tesla and the magnetic rigidity (brho) is that for 800 MeV protons. Finally, the scale length is taken to be $\ell = 1$ meter. Numerical integration of the differential equations (described in Section 10.5.2) to compute $\mathcal{M}$ was carried out employing the Adams10 routine described in Appendix B.8. Both these differential equations and the Adams10 routine are incorporated into MaryLie. The number of integrations steps was 5000 for the cases $a = 0.2$ and $a = .02$, and 10,000 for the case $a = .002$. Results are accurate to

[1] There is confusion/error on this point in the literature. Some authors give aberration results through third order and in the hard-edge limit for many common beam-line elements, but give results for simple solenoids only through second order. And their accompanying discussion can be read to imply that no difficulty is expected in extending the simple solenoid results through third order. Other authors propose formulas for the third-order aberrations of a simple solenoid in the hard-edge limit, and these formulas are independent of a and thus yield *finite* results when $a = 0$. Yet other authors correctly recognize that attempting to make the fringe-field region very small, say by adding extra coils at solenoid ends, leads to some very large third-order aberrations.

at least 10 significant figures.

Evidently the matrices for the three cases are not very different. Moreover, the f_3 Lie generators (which describe second-order aberrations), those with indices 28 through 83, are comparable for the three cases. Both these behaviors are consistent with some sort of convergence occurring for the matrix and f_3 entries as $a \to 0$. See Exercise 2.16. Examination of the equations of motion for these quantities when H_2 and H_3 are given by (2.91) and (2.92), see (10.5.32) and (10.5.60), shows that convergence is to be expected. However, the matter is delicate because the function $b^{[0]}(z)$ that appears in H_2 and hence also in H_3 [see (2.94) and (2.95)] is discontinuous at $z = 0$ and $z = L$ in the limit $a \to 0$. (See also Exercise 2.2.) Therefore the assumptions of Theorems 1.3.1 and 1.3.2 are violated in the limit $a \to 0$. (Note that in this context a is a parameter.)

```
 Exhibit 16.2.3: Transfer map for the case a = 0.2

 matrix for map is :

 9.83483E-01  2.96811E+00  1.00099E-01  3.02094E-01  0.00000E+00  0.00000E+00
-7.58348E-03  9.83483E-01 -7.71845E-04  1.00099E-01  0.00000E+00  0.00000E+00
-1.00099E-01 -3.02094E-01  9.83483E-01  2.96811E+00  0.00000E+00  0.00000E+00
 7.71845E-04 -1.00099E-01 -7.58348E-03  9.83483E-01  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  1.23343E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00

 nonzero elements in generating polynomial are :

 f( 33)=f( 20 00 01 )=-4.56819202124767E-03
 f( 38)=f( 11 00 01 )= 1.16503508346131E-04
 f( 45)=f( 10 01 01 )= 0.12049092663579
 f( 53)=f( 02 00 01 )= -1.7728464321596
 f( 57)=f( 01 10 01 )=-0.12049092663579
 f( 67)=f( 00 20 01 )=-4.56819202124767E-03
 f( 70)=f( 00 11 01 )= 1.16503508346112E-04
 f( 76)=f( 00 02 01 )= -1.7728464321596
 f( 83)=f( 00 00 03 )=-0.73260547246490
 f( 84)=f( 40 00 00 )=-3.74253654402362E-03
 f( 85)=f( 31 00 00 )= 2.24430345842169E-02
 f( 87)=f( 30 01 00 )= 8.95982012441282E-04
 f( 90)=f( 22 00 00 )=-5.26850707023760E-02
 f( 91)=f( 21 10 00 )=-8.95982012441342E-04
 f( 92)=f( 21 01 00 )=-2.48332996215296E-03
 f( 95)=f( 20 20 00 )=-7.48507308804724E-03
 f( 96)=f( 20 11 00 )= 2.24430345842169E-02
 f( 99)=f( 20 02 00 )=-2.26497144701727E-02
 f(104)=f( 20 00 02 )=-6.25873997450046E-03
 f(105)=f( 13 00 00 )= 5.13684894940268E-02
 f(106)=f( 12 10 00 )= 2.48332996215340E-03
 f(107)=f( 12 01 00 )= 2.86233975277233E-02
 f(110)=f( 11 20 00 )= 2.24430345842170E-02
 f(111)=f( 11 11 00 )=-6.00707124644067E-02
 f(114)=f( 11 02 00 )= 5.13684894940265E-02
 f(119)=f( 11 00 02 )= 3.65870158733148E-04
 f(121)=f( 10 21 00 )= 8.95982012441354E-04
 f(124)=f( 10 12 00 )=-2.48332996215361E-03
 f(130)=f( 10 03 00 )= 2.86233975277249E-02
 f(135)=f( 10 01 02 )= 0.16398429796274
 f(140)=f( 04 00 00 )=-0.39062543855982
 f(141)=f( 03 10 00 )=-2.86233975277245E-02
 f(145)=f( 02 20 00 )=-2.26497144701727E-02
 f(146)=f( 02 11 00 )= 5.13684894940265E-02
 f(149)=f( 02 02 00 )=-0.78125087711963
 f(154)=f( 02 00 02 )= -2.4021742345697
 f(155)=f( 01 30 00 )=-8.95982012441327E-04
 f(156)=f( 01 21 00 )= 2.48332996215335E-03
 f(159)=f( 01 12 00 )=-2.86233975277243E-02
 f(164)=f( 01 10 02 )=-0.16398429796274
 f(175)=f( 00 40 00 )=-3.74253654402362E-03
 f(176)=f( 00 31 00 )= 2.24430345842170E-02
 f(179)=f( 00 22 00 )=-5.26850707023762E-02
 f(184)=f( 00 20 02 )=-6.25873997450046E-03
 f(185)=f( 00 13 00 )= 5.13684894940276E-02
 f(190)=f( 00 11 02 )= 3.65870158733149E-04
 f(195)=f( 00 04 00 )=-0.39062543855982
 f(200)=f( 00 02 02 )= -2.4021742345697
 f(209)=f( 00 00 04 )=-0.93366303371890
```

Exhibit 16.2.4: Transfer map for the case a=0.02

```
 matrix for map is :

 9.79608E-01  2.96235E+00  1.00691E-01  3.04490E-01  0.00000E+00  0.00000E+00
-1.00980E-02  9.79608E-01 -1.03794E-03  1.00691E-01  0.00000E+00  0.00000E+00
-1.00691E-01 -3.04490E-01  9.79608E-01  2.96235E+00  0.00000E+00  0.00000E+00
 1.03794E-03 -1.00691E-01 -1.00980E-02  9.79608E-01  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  1.23343E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00

 nonzero elements in generating polynomial are :

 f( 33)=f( 20 00 01 )=-6.10129602475707E-03
 f( 38)=f( 11 00 01 )= 2.06122460889397E-04
 f( 45)=f( 10 01 01 )= 0.12167479425840
 f( 53)=f( 02 00 01 )= -1.7698863189223
 f( 57)=f( 01 10 01 )=-0.12167479425840
 f( 67)=f( 00 20 01 )=-6.10129602475707E-03
 f( 70)=f( 00 11 01 )= 2.06122460889378E-04
 f( 76)=f( 00 02 01 )= -1.7698863189223
 f( 83)=f( 00 00 03 )=-0.73260547246490
 f( 84)=f( 40 00 00 )=-3.82799704943991E-02
 f( 85)=f( 31 00 00 )= 0.22920563229564
 f( 87)=f( 30 01 00 )= 7.71671926080641E-04
 f( 90)=f( 22 00 00 )=-0.56748370148935
 f( 91)=f( 21 10 00 )=-7.71671926080904E-04
 f( 92)=f( 21 01 00 )=-2.05451760035706E-03
 f( 95)=f( 20 20 00 )=-7.65599409887983E-02
 f( 96)=f( 20 11 00 )= 0.22920563229564
 f( 99)=f( 20 02 00 )=-0.19593987204417
 f(104)=f( 20 00 02 )=-8.37702532799553E-03
 f(105)=f( 13 00 00 )= 0.66447262222389
 f(106)=f( 12 10 00 )= 2.05451760035958E-03
 f(107)=f( 12 01 00 )= 2.70686638900989E-02
 f(110)=f( 11 20 00 )= 0.22920563229564
 f(111)=f( 11 11 00 )=-0.74308765889035
 f(114)=f( 11 02 00 )= 0.66447262222389
 f(119)=f( 11 00 02 )= 6.47127530769082E-04
 f(121)=f( 10 21 00 )= 7.71671926081047E-04
 f(124)=f( 10 12 00 )=-2.05451760035932E-03
 f(130)=f( 10 03 00 )= 2.70686638901053E-02
 f(135)=f( 10 01 02 )= 0.16559550393811
 f(140)=f( 04 00 00 )=-0.67707437134121
 f(141)=f( 03 10 00 )=-2.70686638901064E-02
 f(145)=f( 02 20 00 )=-0.19593987204417
 f(146)=f( 02 11 00 )= 0.66447262222389
 f(149)=f( 02 02 00 )= -1.3541487426824
 f(154)=f( 02 00 02 )= -2.3947144417881
 f(155)=f( 01 30 00 )=-7.71671926080822E-04
 f(156)=f( 01 21 00 )= 2.05451760035805E-03
 f(159)=f( 01 12 00 )=-2.70686638901048E-02
 f(164)=f( 01 10 02 )=-0.16559550393811
 f(175)=f( 00 40 00 )=-3.82799704943991E-02
 f(176)=f( 00 31 00 )= 0.22920563229564
 f(179)=f( 00 22 00 )=-0.56748370148935
 f(184)=f( 00 20 02 )=-8.37702532799553E-03
 f(185)=f( 00 13 00 )= 0.66447262222389
 f(190)=f( 00 11 02 )= 6.47127530769100E-04
 f(195)=f( 00 04 00 )=-0.67707437134121
 f(200)=f( 00 02 02 )= -2.3947144417881
 f(209)=f( 00 00 04 )=-0.93366303371890
```

```
 Exhibit 16.2.5: Transfer map for the case a=0.002

 matrix for map is :

 9.79172E-01  2.96175E+00  1.00656E-01  3.04459E-01  0.00000E+00  0.00000E+00
-1.03876E-02  9.79172E-01 -1.06781E-03  1.00656E-01  0.00000E+00  0.00000E+00
-1.00656E-01 -3.04459E-01  9.79172E-01  2.96175E+00  0.00000E+00  0.00000E+00
 1.06781E-03 -1.00656E-01 -1.03876E-02  9.79172E-01  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  1.23343E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00

 nonzero elements in generating polynomial are :

 f( 33)=f( 20 00 01 )=-6.27795556809269E-03
 f( 38)=f( 11 00 01 )= 2.16268873229983E-04
 f( 45)=f( 10 01 01 )= 0.12168683913260
 f( 53)=f( 02 00 01 )= -1.7696108701855
```

```
f( 57)=f( 01 10 01 )=-0.12168683913260
f( 67)=f( 00 20 01 )=-6.27795556809270E-03
f( 70)=f( 00 11 01 )= 2.16268873229994E-04
f( 76)=f( 00 02 01 )= -1.7696108701855
f( 83)=f( 00 00 03 )=-0.73260547246506
f( 84)=f( 40 00 00 )=-0.38235532452970
f( 85)=f( 31 00 00 )=  2.2888185879143
f( 87)=f( 30 01 00 )= 8.00590462204597E-04
f( 90)=f( 22 00 00 )= -5.7102701093391
f( 91)=f( 21 10 00 )=-8.00590462205045E-04
f( 92)=f( 21 01 00 )=-2.13388593660145E-03
f( 95)=f( 20 20 00 )=-0.76471064905939
f( 96)=f( 20 11 00 )=  2.2888185879143
f( 99)=f( 20 02 00 )= -1.9103970318647
f(104)=f( 20 00 02 )=-8.62088196355748E-03
f(105)=f( 13 00 00 )=  6.8352404267418
f(106)=f( 12 10 00 )= 2.13388593640862E-03
f(107)=f( 12 01 00 )= 2.71197052999494E-02
f(110)=f( 11 20 00 )=  2.2888185879143
f(111)=f( 11 11 00 )= -7.5997461549496
f(114)=f( 11 02 00 )=  6.8352404267445
f(119)=f( 11 00 02 )= 6.78947912878182E-04
f(121)=f( 10 21 00 )= 8.00590462204979E-04
f(124)=f( 10 12 00 )=-2.13388593639684E-03
f(130)=f( 10 03 00 )= 2.71197052987057E-02
f(135)=f( 10 01 02 )= 0.16561189662672
f(140)=f( 04 00 00 )= -3.5930682684087
f(141)=f( 03 10 00 )=-2.71197052987419E-02
f(145)=f( 02 20 00 )= -1.9103970318647
f(146)=f( 02 11 00 )=  6.8352404267445
f(149)=f( 02 02 00 )= -7.1861365368258
f(154)=f( 02 00 02 )= -2.3940220818360
f(155)=f( 01 30 00 )=-8.00590462202088E-04
f(156)=f( 01 21 00 )= 2.13388593658432E-03
f(159)=f( 01 12 00 )=-2.71197052999558E-02
f(164)=f( 01 10 02 )=-0.16561189662672
f(175)=f( 00 40 00 )=-0.38235532452970
f(176)=f( 00 31 00 )=  2.2888185879143
f(179)=f( 00 22 00 )= -5.7102701093391
f(184)=f( 00 20 02 )=-8.62088196355748E-03
f(185)=f( 00 13 00 )=  6.8352404267418
f(190)=f( 00 11 02 )= 6.78947912878175E-04
f(195)=f( 00 04 00 )= -3.5930682684087
f(200)=f( 00 02 02 )= -2.3940220818360
f(209)=f( 00 00 04 )=-0.93366303371922
```

What can be said about the f_4 Lie generators (which describe third-order aberrations), those with indices 84 through 209? Some of them are quite different for the three values of a. For example, the values of $f(84)$, which are the coefficients of X^4 in the f_4 Lie generators for the three cases, are quite different. Examination shows that these values are $f(84) = -3.74253654402362E - 03$, $f(84) = -3.82799704943991E - 02$, and $f(84) = -0.38235532452970$ for the cases $a = 0.2$, $a = .02$, and $a = .002$, respectively. This behavior suggests the coefficient of X^4 is *diverging* (in magnitude) to ∞ as $a \to 0$. The same is true of some of the other f_4 entries. Indeed, it can be illustrated numerically and demonstrated analytically that the divergent f_4 entries behave as $1/a$ as $a \to 0$ so that, for example, the product $af(84)$ approaches a *constant* as $a \to 0$. Table 2.1 below illustrates this divergence/behavior for the case of $f(84)$.

Table 16.2.1: Numerical behavior of $f(84)$ for small values of a.

a	$f(84)$	$af(84)$
.2	-3.7425E-3	-7.4850E-4
.02	-3.8279E-2	-7.6558E-4
.002	-3.8235E-1	-7.6470E-4

We have made a preliminary study of the $a \to 0$ behavior of the transfer map $\mathcal{M}$ for a simple solenoid. In the rest of this subsection we will examine the matter in greater detail.

16.2.6.2 Behavior of Linear Part

16.2.6.2.1 Factorization into Three a Dependent Maps/Matrices

Something more can be said about M, the matrix for the *linear* part of $\mathcal{M}$, if one attempts to form hard-edge limits/approximations. Let $\mathcal{M}_{-1\to 2}$ be the transfer map between the planes $z = -1$ and $z = 2$, respectively. It is the map displayed in Exhibits 2.3 through 2.5 for three different values of a. Also, employing analogous notation, consider the maps $\mathcal{M}_{-1\to 0}$, $\mathcal{M}_{0\to 1}$, and $\mathcal{M}_{1\to 2}$. Then we have the relation

$$\mathcal{M}_{-1\to 2} = \mathcal{M}_{-1\to 0}\mathcal{M}_{0\to 1}\mathcal{M}_{1\to 2}. \tag{16.2.97}$$

Next, let $\mathcal{D}$ be the map for a drift of length 1 meter. Employ this map to define implicitly two other maps $\mathcal{M}^{\rm lff}$ and $\mathcal{M}^{\rm tff}$ by writing

$$\mathcal{M}_{-1\to 0} = \mathcal{D}\mathcal{M}^{\rm lff} \tag{16.2.98}$$

and

$$\mathcal{M}_{1\to 2} = \mathcal{M}^{\rm tff}\mathcal{D}. \tag{16.2.99}$$

Then, particularly when a is small, we may view $\mathcal{M}^{\rm lff}$ and $\mathcal{M}^{\rm tff}$ as *leading* and *trailing fringe-field* maps. Of course, (2.98) and (2.99) can be solved to give the explicit definitions

$$\mathcal{M}^{\rm lff} = \mathcal{D}^{-1}\mathcal{M}_{-1\to 0} \tag{16.2.100}$$

$$\mathcal{M}^{\rm tff} = \mathcal{M}_{1\to 2}\mathcal{D}^{-1}. \tag{16.2.101}$$

Also, make the definition

$$\mathcal{M}^{\rm body} = \mathcal{M}_{0\to 1}. \tag{16.2.102}$$

Then we have the factorization

$$\mathcal{M}_{-1\to 2} = \mathcal{D}\mathcal{M}^{\rm lff}\mathcal{M}^{\rm body}\mathcal{M}^{\rm tff}\mathcal{D}. \tag{16.2.103}$$

Note that all these maps are symplectic. Note also that, as illustrated in Figure 2.5, when a is sufficiently small, the map $\mathcal{M}_{-1\to 2}$ essentially describes transport through a 1 meter drift followed by transport through a 1 meter solenoid followed by transport through a final 1 meter drift. We therefore expect, when a is sufficiently small, that $\mathcal{M}^{\rm lff}$ and $\mathcal{M}^{\rm tff}$ will describe leading and trailing fringe-field effects *outside* the solenoid, and $\mathcal{M}^{\rm body}$ will describe all effects occurring *within* the solenoid itself.

Finally, in view of (2.103), we make the definition

$$\mathcal{M}_{\rm solenoid} = \mathcal{M}^{\rm lff}\mathcal{M}^{\rm body}\mathcal{M}^{\rm tff}. \tag{16.2.104}$$

Note that $\mathcal{M}_{\rm solenoid}$ has been factored into *three* a dependent maps. Correspondingly for the associated linear parts there will be the relation

$$M_{\rm solenoid} = M^{\rm tff}M^{\rm body}M^{\rm lff}. \tag{16.2.105}$$

The matrix $M_{\rm solenoid}$ has also been factorized into three a dependent matrices.

From the previous discussion we expect that some of the third- and higher-order aberration parts of the maps $\mathcal{M}^{\rm lff}$, $\mathcal{M}^{\rm tff}$, and $\mathcal{M}^{\rm body}$ may diverge as $a \to 0$. But for now let us examine the linear/matrix parts of the maps $\mathcal{M}^{\rm lff}$, $\mathcal{M}^{\rm body}$, and $\mathcal{M}^{\rm tff}$ in the hard-edge limit $a \to 0$. We begin with the maps $\mathcal{M}^{\rm lff}$ and $\mathcal{M}^{\rm tff}$. Exhibits 2.6 through 2.9 show the matrices associated with these maps for the values $a = 0.2$ and $a = 0.02$.

Exhibit 16.2.6: The matrix $M^{\rm lff}$ for the case $a = 0.2$

```
 9.99944E-01  3.31737E-06  8.72573E-03  2.89481E-08  0.00000E+00  0.00000E+00
-2.11420E-04  9.99980E-01 -1.84490E-06  8.72604E-03  0.00000E+00  0.00000E+00
-8.72573E-03 -2.89481E-08  9.99944E-01  3.31737E-06  0.00000E+00  0.00000E+00
 1.84490E-06 -8.72604E-03 -2.11420E-04  9.99980E-01  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
```

Exhibit 16.2.7: The matrix $M^{\rm lff}$ for the case $a = 0.02$

```
 9.99999E-01  4.78327E-09  1.00901E-03  0.00000E+00  0.00000E+00  0.00000E+00
-2.24996E-05  1.00000E+00 -2.27024E-08  1.00901E-03  0.00000E+00  0.00000E+00
-1.00901E-03  0.00000E+00  9.99999E-01  4.78327E-09  0.00000E+00  0.00000E+00
 2.27024E-08 -1.00901E-03 -2.24996E-05  1.00000E+00  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
```

Exhibit 16.2.8: The matrix $M^{\rm tff}$ for the case $a = 0.2$

```
 9.99980E-01  3.31737E-06  8.72604E-03  2.89481E-08  0.00000E+00  0.00000E+00
-2.11420E-04  9.99944E-01 -1.84490E-06  8.72573E-03  0.00000E+00  0.00000E+00
-8.72604E-03 -2.89481E-08  9.99980E-01  3.31737E-06  0.00000E+00  0.00000E+00
 1.84490E-06 -8.72573E-03 -2.11420E-04  9.99944E-01  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
```

Exhibit 16.2.9: The matrix M^{tff} for the case $a = 0.02$

```
  1.00000E+00  4.78328E-09  1.00901E-03  0.00000E+00  0.00000E+00  0.00000E+00
 -2.24996E-05  9.99999E-01 -2.27024E-08  1.00901E-03  0.00000E+00  0.00000E+00
 -1.00901E-03  0.00000E+00  1.00000E+00  4.78328E-09  0.00000E+00  0.00000E+00
  2.27024E-08 -1.00901E-03 -2.24996E-05  9.99999E-01  0.00000E+00  0.00000E+00
  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  0.00000E+00
  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
```

Upon comparing Exhibits 2.6 and 2.7 we see that there appears to be the limiting behavior

$$\lim_{a\to 0} M^{\mathrm{lff}} = I. \tag{16.2.106}$$

And, upon comparing Exhibits 2.8 and 2.9, we see that there appears to be the limiting behavior

$$\lim_{a\to 0} M^{\mathrm{tff}} = I. \tag{16.2.107}$$

Thus it appears that, for a solenoid in the hard-edge limit, there are *no* effects on the *linear* part of the transfer map due to fringe fields outside the solenoid. These results can be proved analytically using the H_2 given by (2.91) to compute $M_{-1\to 0}$ and $M_{1\to 2}$ since in the hard-edge limit $b^{[0]}(z)$ vanishes for $z < 0$ and $z > L = 1$, and therefore the resulting M for such computations will simply be that for a 1 meter drift.

What can be said about the linear part of $\mathcal{M}^{\mathrm{body}}$? Exhibits 2.10 and 2.11 show the matrices associated with these maps for the values $a = 0.2$ and $a = 0.02$.

Exhibit 16.2.10: The matrix M^{body} for the case $a = 0.2$

```
  9.92886E-01  9.95137E-01  8.35777E-02  8.37672E-02  0.00000E+00  0.00000E+00
 -7.17625E-03  9.92886E-01 -6.04072E-04  8.35777E-02  0.00000E+00  0.00000E+00
 -8.35777E-02 -8.37672E-02  9.92886E-01  9.95137E-01  0.00000E+00  0.00000E+00
  6.04072E-04 -8.35777E-02 -7.17625E-03  9.92886E-01  0.00000E+00  0.00000E+00
  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  4.11143E-01
  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
```

Exhibit 16.2.11: The matrix M^{body} for the case $a = 0.02$

```
  9.89931E-01  9.93237E-01  9.97335E-02  1.00067E-01  0.00000E+00  0.00000E+00
 -1.00555E-02  9.89931E-01 -1.01307E-03  9.97335E-02  0.00000E+00  0.00000E+00
 -9.97335E-02 -1.00067E-01  9.89931E-01  9.93237E-01  0.00000E+00  0.00000E+00
  1.01307E-03 -9.97335E-02 -1.00555E-02  9.89931E-01  0.00000E+00  0.00000E+00
  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  4.11143E-01
  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
```

Comparison of the matrices in Exhibits 2.10 and 2.11 shows that some sort of limit also appears to be approached by M^{body} as $a \to 0$. But what is this limit? Exhibit 2.12 shows M^{uniform}, the matrix computed (numerically) using the Hamiltonian H_2 given by (2.91) with $b^{[0]}(z)$ having a *constant* value. (This matrix can be computed analytically as well as numerically. The analytic result is that quoted in Section 13.4. See Exercise 2.16.)

Examination of the matrices in the Exhibits 2.10 through 2.12 shows that there appears to be the limiting behavior

$$\lim_{a\to 0} M^{\text{body}} = M^{\text{uniform}}. \tag{16.2.108}$$

This result can be proved analytically using the H_2 given by (2.91) to compute $M_{0\to 1}$ since in the hard-edge limit $b^{[0]}(z)$ is constant for z in the *open* interval $z \in (0,1)$.

Exhibit 16.2.12: The matrix M^{uniform} for the case of a uniform field

```
  9.89543E-01  9.93019E-01  1.01722E-01  1.02079E-01  0.00000E+00  0.00000E+00
 -1.04201E-02  9.89543E-01 -1.07116E-03  1.01722E-01  0.00000E+00  0.00000E+00
 -1.01722E-01 -1.02079E-01  9.89543E-01  9.93019E-01  0.00000E+00  0.00000E+00
  1.07116E-03 -1.01722E-01 -1.04201E-02  9.89543E-01  0.00000E+00  0.00000E+00
  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  4.11143E-01
  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
```

From (2.106) through (2.108) we conclude that in the hard-edge limit there are *no* fringe-field contributions to the *linear* part of the transfer map for a solenoid.That is, there is the limiting result

$$\lim_{a\to 0} M_{\text{solenoid}} = M^{\text{uniform}}. \tag{16.2.109}$$

(Curiously, this is the same result as that found by neglecting fringe fields entirely!) Some other authors have also reached the same conclusion by other methods. However we hasten to emphasize, as we have already seen, that in the hard-edge limit there are disastrous fringe-field effects for some third- and higher-order aberrations. It is therefore highly desirable, in the case of a solenoid, to treat fringe-field effects with care (which must be done numerically) using realistic profiles $b^{[n]}(z)$.

Yet other authors provide formulas for matrices M_{fringe} and $M_{\text{longitudinal}}$ which are meant to play roles analogous to M^{lff}, M^{tff}, and M^{body}. These matrices differ from those given by the linear parts of (2.100) through (2.102); and their limiting values differ from those given by (2.106) through (2.108). They are also *not* symplectic. However, when $a = 0$, their product does give the symplectic result

$$M_{\text{fringe}} M_{\text{longitudinal}} M_{\text{fringe}}^{-1} = M^{\text{uniform}}. \tag{16.2.110}$$

Based on the results of these authors one might be tempted to conclude that, at least in some way, fringe fields even in the hard-edge limit do play some role in determining the linear part of the transfer map for a solenoid. [Note however that their net effect cancels out because of the result (2.110).] The explanation for this confusing circumstance is that these authors employ in essence mechanical rather than canonical momenta in their calculations. So doing is expected to yield nonsymplectic results in the presence of magnetic fields. Recall Exercise 1.7.5. But these nonsymplectic results can ultimately cancel, as in (2.110), when a full/complete calculation is made providing the magnetic field vanishes before entry into and after exit from the solenoid. See Exercise 6.4.11.

16.2.6.2.2 Factorization Involving Only Two a Dependent Maps/Matrices

There is another variation on the theme we have been exploring. Suppose instead of (2.103) we attempt an Ansatz of the form

$$\mathcal{M}_{-1\to 2} = \mathcal{D}\mathcal{M}^{\rm LFF}\mathcal{M}^{\rm uniform}\mathcal{M}^{\rm TFF}\mathcal{D}. \tag{16.2.111}$$

Here $\mathcal{M}^{\rm uniform}$ is the map computed for a length of 1 meter using the Hamiltonian H given by (2.88) through (2.93) with $b^{[0]}$ having a *constant* value and correspondingly $b^{[n]} = 0$ for $n > 0$. (For future reference, we will call this Hamiltonian $H^{\rm uniform}$.) If the Ansatz (2.111) is successful, the maps $\mathcal{M}^{\rm LFF}$ and $\mathcal{M}^{\rm TFF}$ will describe *both* the effects arising from the depletion of the field within the body of the solenoid and the effects of the fields that extend beyond the ends of the solenoid. In view of (2.111) we make the definition

$$\mathcal{M}_{\rm solenoid} = \mathcal{M}^{\rm LFF}\mathcal{M}^{\rm uniform}\mathcal{M}^{\rm TFF}. \tag{16.2.112}$$

Assuming success of the Ansatz (2.111), we see that $\mathcal{M}_{\rm solenoid}$ has been factorized in a way that involves only *two* a dependent maps, namely $\mathcal{M}^{\rm LFF}$ and $\mathcal{M}^{\rm TFF}$, and one a independent map, namely $\mathcal{M}^{\rm uniform}$. Correspondingly for the associated linear parts there will be the relation

$$M_{\rm solenoid} = M^{\rm TFF}M^{\rm uniform}M^{\rm LFF}. \tag{16.2.113}$$

The matrix $M_{\rm solenoid}$ has been factorized in a way that involves only two a dependent matrices and one a independent matrix

Let us pause momentarily at this point to compare the factorizations (2.104) and (2.112). The map $\mathcal{M}_{\rm solenoid}$ is the same in both. [Correspondingly, the matrices $M_{\rm solenoid}$ given by (2.105) and (2.113) are the same.] But (2.104) may be viewed as a kind of *local* factorization in that it treats *separately* effects that occur before, within, and after the body of the solenoid. By contrast (2.112) may be viewed as a *nonlocal/lumped* factorization in that $\mathcal{M}^{\rm LFF}$ describes effects that occur both before and after entry into the body of the solenoid and $\mathcal{M}^{\rm TFF}$ describes effects that occur both within the body of the solenoid and after exit from the body of the solenoid. No attempt is made to describe separately what occurs only within the body of the solenoid itself.[2]

To continue, how can we determine the maps $\mathcal{M}^{\rm LFF}$ and $\mathcal{M}^{\rm TFF}$? Let $\mathcal{H}$ be the map of a uniform "half" solenoid, the map computed for a uniform solenoid with a length of 1/2 meter using the Hamiltonian H given by (2.88) through (2.93) with $b^{[0]}$ having a constant value and correspondingly $b^{[n]} = 0$ for $n > 0$. Then evidently there will be the relation

$$\mathcal{M}^{\rm uniform} = \mathcal{H}\mathcal{H} \tag{16.2.114}$$

and the Ansatz (2.111) becomes

$$\mathcal{M}_{-1\to 2} = \mathcal{D}\mathcal{M}^{\rm LFF}\mathcal{H}\mathcal{H}\mathcal{M}^{\rm TFF}\mathcal{D}. \tag{16.2.115}$$

[2] Strictly speaking, the factorization (2.104) is not completely local since, according to the definition (2.98), $\mathcal{M}^{\rm lff}$ lumps together at the end of the leading drift all the fringe-field effects that have accumulated prior to the body of the solenoid; and, according to (2.99), $\mathcal{M}^{\rm tff}$ lumps together at the beginning of the trailing drift all the fringe-field effects that will accumulate after the body of the solenoid.

Observe that there is also the factorization

$$\mathcal{M}_{-1\to 2} = \mathcal{M}_{-1\to 0.5}\mathcal{M}_{0.5\to 2}. \tag{16.2.116}$$

We are therefore led to make the implicit definitions

$$\mathcal{D}\mathcal{M}^{\rm LFF}\mathcal{H} = \mathcal{M}_{-1\to 0.5} \tag{16.2.117}$$

and

$$\mathcal{H}\mathcal{M}^{\rm TFF}\mathcal{D} = \mathcal{M}_{0.5\to 2}. \tag{16.2.118}$$

It is easily verified using (2.114) and (2.116) through (2.118) that (2.111) is then satisfied. Thus the Ansatz (2.111) has been justified. Moreover, (2.117) and (2.118) can be solved for $\mathcal{M}^{\rm LFF}$ and $\mathcal{M}^{\rm LFF}$ to give the explicit results

$$\mathcal{M}^{\rm LFF} = \mathcal{D}^{-1}\mathcal{M}_{-1\to 0.5}\mathcal{H}^{-1} \tag{16.2.119}$$

and

$$\mathcal{M}^{\rm TFF} = \mathcal{H}^{-1}\mathcal{M}_{0.5\to 2}\mathcal{D}^{-1}. \tag{16.2.120}$$

Note that, according to (2.119) and (2.120), both $\mathcal{M}^{\rm LFF}$ and $\mathcal{M}^{\rm TFF}$ are symplectic maps. Consequently the matrices $M^{\rm LFF}$ and $M^{\rm TFF}$ associated with their linear parts will be symplectic matrices.

What can be said about the nature of $M^{\rm LFF}$ and $M^{\rm TFF}$ as functions of a? Exhibits 2.13 through 2.18 display these matrices for the values $a = 0.2$, $a = .02$, and $a = .002$. For clarity of presentation, in making these calculations the rotational parts of the solenoid maps have been removed. Recall Subsection 2.5.3. In principle, because of (2.87), there should be no rotational parts in $M^{\rm LFF}$ and $M^{\rm TFF}$ in the limit that the external leading fringe field is allowed to begin at $z^{\rm en} = -\infty$ and the external trailing fringe field is allowed to extend to $z^{\rm ex} = +\infty$. In the calculations described here (and without the rotational components of the solenoid maps being removed) there are negligible but numerically noticeable rotational components in the matrices $M^{\rm LFF}$ and $M^{\rm TFF}$ associated with the maps $\mathcal{M}^{\rm LFF}$ and $\mathcal{M}^{\rm TFF}$ as defined by (2.119) and (2.120). For example, in computing (2.119) the rotational part of $\mathcal{M}_{-1\to 0.5}$ does not completely cancel the rotational part of $\mathcal{H}^{-1}$ because the leading external fringe field region is taken to begin at $z = z^{\rm en} = -1$. We have verified that if the leading external fringe field region is taken to begin at a larger negative value of z, for example $z^{\rm en} = -10$, then the cancellation of rotational parts is more nearly complete. (Of course, in this case we must also employ for $\mathcal{D}$ the map of a 10 meter drift.) The cancellation also is more nearly complete the smaller the value of a, as is to be expected from comparison of Figures 2.4 and 2.5. The subtlety of these considerations arises from the relatively slow fall off of air-core solenoid fringe fields.

Exhibit 16.2.13: The matrix $M^{\rm LFF}$ for the case $a = 0.2$

```
 9.99702E-01 -7.50022E-05  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
 1.42997E-03  1.00030E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  9.99702E-01 -7.50022E-05  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  1.42997E-03  1.00030E+00  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
```

Exhibit 16.2.14: The matrix $M^{\rm LFF}$ for the case $a = 0.02$

```
 9.99992E-01 -1.22107E-06  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
 1.62696E-04  1.00001E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  9.99992E-01 -1.22107E-06  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  1.62696E-04  1.00001E+00  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
```

Exhibit 16.2.15: The matrix $M^{\rm LFF}$ for the case $a = 0.002$

```
 1.00000E+00 -1.10603E-08  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
 1.64409E-05  1.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  1.00000E+00 -1.10603E-08  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  1.64409E-05  1.00000E+00  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
```

Exhibit 16.2.16: The matrix $M^{\rm TFF}$ for the case $a = 0.2$

```
 1.00030E+00 -7.50022E-05  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
 1.42997E-03  9.99702E-01  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  1.00030E+00 -7.50022E-05  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  1.42997E-03  9.99702E-01  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
```

Exhibit 16.2.17: The matrix $M^{\rm TFF}$ for the case $a = 0.02$

```
 1.00001E+00 -1.22107E-06  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
 1.62696E-04  9.99992E-01  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  1.00001E+00 -1.22107E-06  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  1.62696E-04  9.99992E-01  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
```

Exhibit 16.2.18: The matrix $M^{\rm TFF}$ for the case $a = 0.002$

```
 1.00000E+00 -1.10596E-08  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
 1.64409E-05  1.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  1.00000E+00 -1.10596E-08  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  1.64409E-05  1.00000E+00  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
```

From Exhibits 2.13 through 2.15 we infer that there is the limiting behavior

$$\lim_{a\to 0} M^{\rm LFF} = I. \tag{16.2.121}$$

And from Exhibits 2.16 through 2.18 we infer that there is the limiting behavior

$$\lim_{a\to 0} M^{\rm TFF} = I. \tag{16.2.122}$$

These limiting behaviors can also be verified analytically, and are to be expected. Again we conclude that in the hard-edge limit there are *no* fringe-field contributions to the *linear* part of the transfer map for a solenoid. That is, (2.109) again holds.

There is some irony here. We have seen that the matrices for the linear parts of $\mathcal{M}^{\rm LFF}$ and $\mathcal{M}^{\rm TFF}$ have the benign limiting behavior (2.121) and (2.122). We also know from (2.112) that the only a dependence in $\mathcal{M}_{\rm solenoid}$ arises from that in $\mathcal{M}^{\rm LFF}$ and $\mathcal{M}^{\rm TFF}$, and we have seen that $\mathcal{M}_{\rm solenoid}$ has some divergent third-order aberrations as $a \to 0$. We conclude that while the limiting behavior of the linear parts of $\mathcal{M}^{\rm LFF}$ and $\mathcal{M}^{\rm TFF}$ is benign, that of some of the nonlinear parts of $\mathcal{M}^{\rm LFF}$ and $\mathcal{M}^{\rm TFF}$ is pathological.

What can be said about the linear parts of $\mathcal{M}^{\rm LFF}$ and $\mathcal{M}^{\rm TFF}$ when a is small but *nonzero*? From Exhibits 2.13 through 2.15 we see that, for small a, the effect of the leading fringe field is to produce *identical defocussing* in both planes. And from Exhibits 2.16 through 2.18 we see that the same is true for the effect of the trailing fringe field.[3] The modest and small a effect of fringe fields is to *decrease* the focussing effect of a solenoid compared to that predicted by $M^{\rm uniform}$. Recall (2.113).

Moreover, as illustrated in Table 2.2 below, for small a the defocussing strength behaves linearly in a. That is, for example, the quantity $M_{21}^{\rm LFF} = M_{43}^{\rm LFF}$ is proportional to a when a is sufficiently small so that the product $(1/a)M_{21}^{\rm LFF}$ approaches a *constant* as $a \to 0$.[4] Identical results hold for $M_{21}^{\rm TFF} = M_{43}^{\rm TFF}$.

Finally, we remark that these results are consistent with those obtained by some other authors using other methods.

[3] In passing we also note that, for a given value of a and within the announced numerical accuracy, the matrices $M^{\rm LFF}$ and $M^{\rm TFF}$ differ only by permutations of various *diagonal* entries. This result is a consequence of *reversal symmetry*. That is, for a given value of a, the matrices $M^{\rm LFF}$ and $M^{\rm TFF}$ are *reverses* of each other. See Chapter 36.

[4] For modest values of a such as $a = 0.2$, and still larger values of a, there are some effects on $M_{21}^{\rm LFF}$ that arise from the approximation we have made that there are no leading external fringe-field effects before $z^{\rm en} = -1$. Similarly there are some effects on $M_{21}^{\rm TFF}$ that arise from the approximation we have made that there are no trailing external fringe-field effects after $z^{\rm ex} = 2$. These effects disappear as $a \to 0$ because then external fringe fields become more and more confined to the vicinities of the entrance and exit of the solenoid. As an indication of the size of these effects, suppose the leading external fringe field region is taken to begin at a larger negative value of z, for example $z = z^{\rm en} = -10$. (Of course, in this case we must also employ for $\mathcal{D}$ the map of a 10 meter drift.) Then, for $a = 0.2$, there are the results $M_{21}^{\rm LFF} = 1.42982$E-03 and $(1/a)M_{21}^{\rm LFF} = 7.14910$E-03, which are to be compared with those in the first line of Table 2.2. Evidently in this case the effects of the approximations we have made are small.

Table 16.2.2: Numerical behavior of $M_{21}^{\rm LFF}$ for small values of a.

a	$M_{21}^{\rm LFF}$	$(1/a)M_{21}^{\rm LFF}$
.2	1.42997E-03	7.14985E-03
.02	1.62696E-04	8.13480E-03
.002	1.64409E-05	8.22045E-03

16.2.6.3 Behavior of Nonlinear Part

We have concluded that, while the limiting behavior of the linear parts of $\mathcal{M}^{\rm LFF}$ and $\mathcal{M}^{\rm TFF}$ is benign, that of some of the nonlinear parts of $\mathcal{M}^{\rm LFF}$ and $\mathcal{M}^{\rm TFF}$ is pathological. In this subsection we will examine in more detail the behavior of the nonlinear part of $\mathcal{M}^{\rm LFF}$ in the limit $a \to 0$. For brevity, we will not present the behavior of $\mathcal{M}^{\rm TFF}$. But, as expected, it is found to be analogous to that of $\mathcal{M}^{\rm LFF}$.

Exhibits 2.19 through 2.21 display (through third order) the maps $\mathcal{M}^{\rm LFF}$ given by (2.119) for the cases $a = 0.2$, $a = 0.02$, and $a = 0.002$, respectively. For these exhibits the rotational parts of the solenoidal maps have *not* been removed.

Exhibit 16.2.19: The map $\mathcal{M}^{\rm LFF}$ for the case $a = 0.2$

```
 matrix for map is :

  9.99702E-01 -7.47109E-05 -9.30614E-06  6.95479E-10  0.00000E+00  0.00000E+00
  1.42984E-03  1.00030E+00 -1.33102E-08 -9.31170E-06  0.00000E+00  0.00000E+00
  9.30614E-06 -6.95479E-10  9.99702E-01 -7.47109E-05  0.00000E+00  0.00000E+00
  1.33102E-08  9.31170E-06  1.42984E-03  1.00030E+00  0.00000E+00  0.00000E+00
  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  0.00000E+00
  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00

 nonzero elements in generating polynomial are :

  f( 33)=f( 20 00 01 )= 8.48173754616857E-04
  f( 38)=f( 11 00 01 )= 7.08661171957406E-04
  f( 45)=f( 10 01 01 )=-1.10582080947857E-05
  f( 53)=f( 02 00 01 )= 1.33079847471684E-04
  f( 57)=f( 01 10 01 )= 1.10582080948066E-05
  f( 67)=f( 00 20 01 )= 8.48173754616855E-04
  f( 70)=f( 00 11 01 )= 7.08661171957389E-04
  f( 76)=f( 00 02 01 )= 1.33079847476181E-04
  f( 84)=f( 40 00 00 )=-1.86910737357663E-03
  f( 85)=f( 31 00 00 )=-2.23275870608366E-03
  f( 87)=f( 30 01 00 )=-1.51855928877267E-04
  f( 90)=f( 22 00 00 )=-3.06010303211954E-04
  f( 91)=f( 21 10 00 )= 1.51855928877281E-04
  f( 92)=f( 21 01 00 )= 2.37154993794211E-02
  f( 95)=f( 20 20 00 )=-3.73821474715327E-03
  f( 96)=f( 20 11 00 )=-2.23275870608358E-03
  f( 99)=f( 20 02 00 )= 8.51691896034609E-04
  f(104)=f( 20 00 02 )= 1.15254677042892E-03
  f(105)=f( 13 00 00 )= 1.26460888238766E-04
  f(106)=f( 12 10 00 )=-2.37154993794227E-02
  f(107)=f( 12 01 00 )=-9.17588458782363E-04
  f(110)=f( 11 20 00 )=-2.23275870608357E-03
  f(111)=f( 11 11 00 )=-2.31540439851645E-03
  f(114)=f( 11 02 00 )= 1.26460888456775E-04
  f(119)=f( 11 00 02 )= 1.38437434049872E-03
  f(121)=f( 10 21 00 )=-1.51855928877260E-04
  f(124)=f( 10 12 00 )= 2.37154993794226E-02
  f(130)=f( 10 03 00 )=-9.17588458807038E-04
  f(135)=f( 10 01 02 )=-1.50498675709195E-05
  f(140)=f( 04 00 00 )= 4.36074169958001E-05
  f(141)=f( 03 10 00 )= 9.17588458803895E-04
  f(145)=f( 02 20 00 )= 8.51691896035069E-04
  f(146)=f( 02 11 00 )= 1.26460888454029E-04
  f(149)=f( 02 02 00 )= 8.72148317848653E-05
  f(154)=f( 02 00 02 )= 3.39093545585142E-04
  f(155)=f( 01 30 00 )= 1.51855928877314E-04
  f(156)=f( 01 21 00 )=-2.37154993794229E-02
  f(159)=f( 01 12 00 )= 9.17588458801365E-04
  f(164)=f( 01 10 02 )= 1.50498675709225E-05
  f(175)=f( 00 40 00 )=-1.86910737357663E-03
  f(176)=f( 00 31 00 )=-2.23275870608367E-03
  f(179)=f( 00 22 00 )=-3.06010303212333E-04
  f(184)=f( 00 20 02 )= 1.15254677042892E-03
  f(185)=f( 00 13 00 )= 1.26460888246612E-04
  f(190)=f( 00 11 02 )= 1.38437434049872E-03
  f(195)=f( 00 04 00 )= 4.36074169358758E-05
  f(200)=f( 00 02 02 )= 3.39093545579732E-04
```

Exhibit 16.2.20: The map $\mathcal{M}^{\rm LFF}$ for the case $a = 0.02$

```
matrix for map is :

 9.99992E-01 -1.22277E-06 -9.20989E-08  0.00000E+00  0.00000E+00  0.00000E+00
 1.62716E-04  1.00001E+00  0.00000E+00 -9.21003E-08  0.00000E+00  0.00000E+00
 9.20989E-08  0.00000E+00  9.99992E-01 -1.22277E-06  0.00000E+00  0.00000E+00
 0.00000E+00  9.21003E-08  1.62716E-04  1.00001E+00  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00

nonzero elements in generating polynomial are :

 f( 33)=f( 20 00 01 )= 9.66314417891064E-05
 f( 38)=f( 11 00 01 )= 1.85289937516709E-05
 f( 45)=f( 10 01 01 )=-1.09406551700164E-07
 f( 53)=f( 02 00 01 )= 2.17830287646548E-06
 f( 57)=f( 01 10 01 )= 1.09406551727920E-07
 f( 67)=f( 00 20 01 )= 9.66314417891068E-05
 f( 70)=f( 00 11 01 )= 1.85289937516412E-05
```

```
f( 76)=f( 00 02 01 )= 2.17830287385645E-06
f( 84)=f( 40 00 00 )=-1.93149067689109E-02
f( 85)=f( 31 00 00 )=-2.61905549343168E-03
f( 87)=f( 30 01 00 )=-1.86391130390008E-05
f( 90)=f( 22 00 00 )=-9.05018682344674E-05
f( 91)=f( 21 10 00 )= 1.86391130388983E-05
f( 92)=f( 21 01 00 )= 2.55870298344304E-02
f( 95)=f( 20 20 00 )=-3.86298135378219E-02
f( 96)=f( 20 11 00 )=-2.61905549343095E-03
f( 99)=f( 20 02 00 )= 7.83183807162944E-05
f(104)=f( 20 00 02 )= 1.31485684215657E-04
f(105)=f( 13 00 00 )=-7.58534145784010E-07
f(106)=f( 12 10 00 )=-2.55870298344239E-02
f(107)=f( 12 01 00 )=-1.03168131481354E-05
f(110)=f( 11 20 00 )=-2.61905549343097E-03
f(111)=f( 11 11 00 )=-3.37640497875350E-04
f(114)=f( 11 02 00 )=-7.58534755148649E-07
f(119)=f( 11 00 02 )= 3.62090101634249E-05
f(121)=f( 10 21 00 )=-1.86391130388087E-05
f(124)=f( 10 12 00 )= 2.55870298344245E-02
f(130)=f( 10 03 00 )=-1.03168130791351E-05
f(135)=f( 10 01 02 )=-1.48898803105739E-07
f(140)=f( 04 00 00 )= 5.20628154886127E-07
f(141)=f( 03 10 00 )= 1.03168130669920E-05
f(145)=f( 02 20 00 )= 7.83183807188067E-05
f(146)=f( 02 11 00 )=-7.58534727284429E-07
f(149)=f( 02 02 00 )= 1.04126206200428E-06
f(154)=f( 02 00 02 )= 5.55068286516206E-06
f(155)=f( 01 30 00 )= 1.86391130389137E-05
f(156)=f( 01 21 00 )=-2.55870298344221E-02
f(159)=f( 01 12 00 )= 1.03168129842943E-05
f(164)=f( 01 10 02 )= 1.48898803052262E-07
f(175)=f( 00 40 00 )=-1.93149067689109E-02
f(176)=f( 00 31 00 )=-2.61905549343156E-03
f(179)=f( 00 22 00 )=-9.05018682329818E-05
f(184)=f( 00 20 02 )= 1.31485684215654E-04
f(185)=f( 00 13 00 )=-7.58534226352400E-07
f(190)=f( 00 11 02 )= 3.62090101634544E-05
f(195)=f( 00 04 00 )= 5.20628151791380E-07
f(200)=f( 00 02 02 )= 5.55068286537210E-06
```

Exhibit 16.2.21: The map $\mathcal{M}^{\rm LFF}$ for the case $a = 0.002$

```
 matrix for map is :

  1.00000E+00 -1.27806E-08  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
  1.64617E-05  1.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00
  0.00000E+00  0.00000E+00  1.00000E+00 -1.27806E-08  0.00000E+00  0.00000E+00
  0.00000E+00  0.00000E+00  1.64617E-05  1.00000E+00  0.00000E+00  0.00000E+00
  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00  0.00000E+00
  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  0.00000E+00  1.00000E+00
 nonzero elements in generating polynomial are :

  f( 33)=f( 20 00 01 )= 9.77739063782491E-06
  f( 38)=f( 11 00 01 )= 2.99993492913329E-07
  f( 45)=f( 10 01 01 )= 1.11383166578882E-10
  f( 53)=f( 02 00 01 )= 2.27936177710220E-08
  f( 57)=f( 01 10 01 )=-1.11383180456670E-10
  f( 67)=f( 00 20 01 )= 9.77739063782491E-06
  f( 70)=f( 00 11 01 )= 2.99993492886962E-07
  f( 76)=f( 00 02 01 )= 2.27936205465795E-08
  f( 84)=f( 40 00 00 )=-0.19316037246934
  f( 85)=f( 31 00 00 )=-2.62331874601927E-03
  f( 87)=f( 30 01 00 )=-1.89599356992097E-06
  f( 90)=f( 22 00 00 )=-9.71312429641274E-06
  f( 91)=f( 21 10 00 )= 1.89599356725453E-06
  f( 92)=f( 21 01 00 )= 2.56091096783283E-02
  f( 95)=f( 20 20 00 )=-0.38632074493868
  f( 96)=f( 20 11 00 )=-2.62331874601768E-03
  f( 99)=f( 20 02 00 )= 7.73683945196578E-06
  f(104)=f( 20 00 02 )= 1.33064336534576E-05
  f(105)=f( 13 00 00 )=-3.17816277210040E-08
  f(106)=f( 12 10 00 )=-2.56091096782642E-02
  f(107)=f( 12 01 00 )=-1.02864023718979E-07
  f(110)=f( 11 20 00 )=-2.62331874601767E-03
  f(111)=f( 11 11 00 )=-3.48999279807779E-05
  f(114)=f( 11 02 00 )=-3.17767246773918E-08
  f(119)=f( 11 00 02 )= 5.86324364974507E-07
  f(121)=f( 10 21 00 )=-1.89599356821861E-06
  f(124)=f( 10 12 00 )= 2.56091096783162E-02
  f(130)=f( 10 03 00 )=-1.02863765470695E-07
  f(135)=f( 10 01 02 )= 1.51604383154817E-10
  f(140)=f( 04 00 00 )= 5.33338799513228E-09
  f(141)=f( 03 10 00 )= 1.02862986930269E-07
  f(145)=f( 02 20 00 )= 7.73683948127914E-06
  f(146)=f( 02 11 00 )=-3.17772725634192E-08
  f(149)=f( 02 02 00 )= 1.06253442294646E-08
  f(154)=f( 02 00 02 )= 5.80574061863604E-08
  f(155)=f( 01 30 00 )= 1.89599356946647E-06
  f(156)=f( 01 21 00 )=-2.56091096783008E-02
  f(159)=f( 01 12 00 )= 1.02862771318019E-07
  f(164)=f( 01 10 02 )=-1.51604352741914E-10
  f(175)=f( 00 40 00 )=-0.19316037246934
  f(176)=f( 00 31 00 )=-2.62331874601742E-03
  f(179)=f( 00 22 00 )=-9.71312426696364E-06
  f(184)=f( 00 20 02 )= 1.33064336534594E-05
  f(185)=f( 00 13 00 )=-3.17829202628548E-08
  f(190)=f( 00 11 02 )= 5.86324364877742E-07
  f(195)=f( 00 04 00 )= 5.34906990923290E-09
  f(200)=f( 00 02 02 )= 5.80574133293362E-08
```

Examination of these exhibits shows that, as expected, (2.121) continues to hold even when the rotational parts of M^{LFF} have not been removed. We reiterate that, in the hard-edge limit and when canonical coordinates are employed, there are *no* fringe-field contributions to the *linear* part of the transfer map for a solenoid.

What can be said about the exp(: f_3 :) content of $\mathcal{M}^{\rm LFF}$? Examination of the f_3 contents of these same exhibits, the generators with indices 28 through 83, shows that numerically there is the limiting behavior

$$\lim_{a\to 0} f_3^{\rm LFF} = 0. \qquad (16.2.123)$$

The same result can be obtained analytically. Consequently, in the hard-edge limit and when canonical coordinates are employed, there are also *no* fringe-field contributions to the *quadratic* part of the transfer map for a solenoid. All second-order aberrations associated with $\mathcal{M}^{\rm LFF}$ vanish in the hard-edge limit.

What can be said about the exp(: f_4 :) content of $\mathcal{M}^{\rm LFF}$? Examination of the generators with indices 84 through 209 shows that numerically some of the $f_4^{\rm LFF}$ generators *grow/diverge* in magnitude as $a \to 0$. Indeed, it can be illustrated numerically and demonstrated analytically that the divergent $f_4^{\rm LFF}$ entries behave as $1/a$ as $a \to 0$. For example, Table 2.3 below illustrates this divergence for the case of $f^{\rm LFF}(84)$.

Table 16.2.3: Numerical behavior of $f^{\rm LFF}(84)$ for small values of a.

a	$f^{\rm LFF}(84)$	$af^{\rm LFF}(84)$
.2	-1.8691E-3	-3.7382E-4
.02	-1.9315E-2	-3.8630E-4
.002	-1.9316E-1	-3.8632E-4

At this point here are two other remarks to be made. First, there are also divergent $f_n^{\rm LFF}$ generators for $n > 4$. Second, as already observed, the only a dependence in (2.112) is in the maps $\mathcal{M}^{\rm LFF}$ and $\mathcal{M}^{\rm TFF}$. We conclude that *all divergent* aberrations for a simple solenoid in the hard-edge limit arise from divergencies in the fringe-field maps.

Upon thinking in more detail, there is more that can be inferred about the a dependence of the $f_4^{\rm LFF}$ generators. We have already seen that H and J_z are in involution. Recall (2.82). Indeed, upon examination we see that all the terms in the expansion of H given in (2.94) through (2.96) are separately in involution with J_z. Therefore, since f_4 is constructed from the ingredients of H using only Lie algebraic operations, we expect that $f_4^{\rm LFF}$ and J_z will be in involution. Below we list the various possible *static* (τ independent) f_4 polynomials, call then I_j, that are in *involution* with J_z (and hence *invariant* under the action of rotations generated by : J_z :) and display their monomial content.

$$I_1 = (P^2)^2 = (P_x^2 + P_y^2)^2 = P_x^4 + 2P_x^2P_y^2 + P_y^4, \qquad (16.2.124)$$

$$\begin{aligned} I_2 &= P^2 J_z = (P_x^2 + P_y^2)(XP_y - YP_x) \\ &= XP_x^2P_y - P_x^3Y + XP_y^3 - P_xYP_y^2, \end{aligned} \qquad (16.2.125)$$

$$I_3 = J_z^2 = (XP_y - YP_x)^2 = X^2P_y^2 - 2XP_xYP_y + P_x^2Y^2, \qquad (16.2.126)$$

$$I_4 = P^2Q^2 = (P_x^2 + P_y^2)(X^2 + Y^2) = X^2P_x^2 + X^2P_y^2 + P_x^2Y^2 + Y^2P_y^2, \tag{16.2.127}$$

$$\begin{aligned} I_5 &= Q^2J_z = (X^2 + Y^2)(XP_y - YP_x) \\ &= X^3P_y - X^2P_xY + XY^2P_y - P_xY^3, \end{aligned} \tag{16.2.128}$$

$$I_6 = (Q^2)^2 = (X^2 + Y^2)^2 = X^4 + 2X^2Y^2 + Y^4, \tag{16.2.129}$$

$$I_7 = P_\tau^2P^2 = P_x^2P_\tau^2 + P_y^2P_\tau^2, \tag{16.2.130}$$

$$I_8 = P_\tau^2J_z = XP_yP_\tau^2 - P_xYP_\tau^2, \tag{16.2.131}$$

$$I_9 = P_\tau^2Q^2 = X^2P_\tau^2 + Y^2P_\tau^2, \tag{16.2.132}$$

$$I_{10} = P_\tau^4. \tag{16.2.133}$$

Note that the monomials $X^2P_y^2$ and $P_x^2Y^2$ appear in both the invariants I_3 and I_4. All other monomials appear at most once in the invariants I_j.

According to the reasoning of the previous paragraph, we expect that $f_4^{\rm LFF}$ can be expressed/expanded (in a unique way) in terms of the I_j for any value of a. This is indeed the case. The relations (2.134) through (2.136) below display this expansion for the values $a = 0.2$, $a = 0.02$, and $a = 0.002$, respectively.

$$\begin{aligned} f_4^{\rm LFF}|_{a=0.2} &= (4.36074169958001E-05)I_1 - (9.17588458782363E-04)I_2 \\ &+ (1.15770219925823E-03)I_3 - (3.06010303211954E-04)I_4 \\ &- (1.51855928877267E-04)I_5 - (1.86910737357663E-03)I_6 \\ &+ (3.39093545585142E-04)I_7 - (1.50498675709195E-05)I_8 \\ &+ (1.15254677042892E-03)I_9 + (0)I_{10}; \end{aligned} \tag{16.2.134}$$

$$\begin{aligned} f_4^{\rm LFF}|_{a=0.02} &= (5.20628154886127E-07)I_1 - (1.03168131481354E-05)I_2 \\ &+ (1.68820248937675E-04)I_3 - (9.05018682344674E-05)I_4 \\ &- (1.86391130390008E-05)I_5 - (1.93149067689109E-02)I_6 \\ &+ (5.55068286516206E-06)I_7 - (1.48898803105739E-07)I_8 \\ &+ (1.31485684215657E-04)I_9 + (0)I_{10}; \end{aligned} \tag{16.2.135}$$

$$\begin{aligned} f_4^{\rm LFF}|_{a=0.002} &= (5.33338799513228E-09)I_1 - (1.02864023718979E-07)I_2 \\ &+ (1.74499639903890E-05)I_3 - (9.71312429641274E-06)I_4 \\ &- (1.89599356992097E-06)I_5 - (0.19316037246934)I_6 \\ &+ (5.80574061863604E-08)I_7 + (1.51604383154817E-10)I_8 \\ &+ (1.33064336534576E-05)I_9 + (0)I_{10}. \end{aligned} \tag{16.2.136}$$

Inspection of (2.134) through (2.136) shows that, save for I_6, the coefficients of all the I_j *vanish* as $a \to 0$. See Exercise 2.21. By contrast, the aberrations associated with the ingredients of I_6 *grow* as $1/a$ as $a \to 0$. Recall Table 2.3 and see Exercise 2.22. Thus the behavior of f_4^{LFF} is *simple* in the hard-edge limit in that most (all but three) of its entries *vanish* in this limit. But it is also *pathological* in that the entries for the generators X^4, X^2Y^2, and Y^4, those that occur in I_6, *diverge* in this limit.[5] Finally we remark that, while we have illustrated these results numerically, they can also be proven analytically. The pathological behavior arises from the appearance of the the δ' functions that occur in the $a \to 0$ limit, and must occur in any hard-edge model.

16.2.7 Consequences of Terminating Solenoidal End Fields

Suppose we wish to find the transfer map for a solenoid with the approximation that the leading fringe-field region begins at the "entry" point $z = z^{\text{en}}$ and the trailing fringe-field region ends at the "exit" point $z = z^{\text{ex}}$. That is, we make the approximation that the vector potential is to be set to zero for $z < z^{\text{en}}$ and $z > z^{\text{ex}}$. Since $\hat{\boldsymbol{A}}^0$, the vector potential we will employ, is in the Poincaré-Coulomb gauge with respect to any origin on the z axis, we may use (following the methods of Section 1) this vector potential to terminate end fields both before entry of the leading fringe field and after exit of the trailing fringe field. That is, there is no need to make gauge transformations at these points because the vector potential is already in the minimum gauge. In this subsection we will study the consequences of terminating end fields using the Poincaré-Coulomb gauge.

Let us begin by finding the associated discontinuities in the mechanical momenta as given by (1.30), (1.31), (1.41), and (1.42). For the vector potential we use (2.7) through (2.10). So doing using (1.30) and (1.31) gives, upon entry, the results

$$\begin{aligned}
\Delta p_x^{\text{mech}} &= qA_x(x,y,z^{\text{en}}) = -qyU(\rho,z^{\text{en}}) \\
&= -qy(1/2)[C_0^{[1]}(z^{\text{en}}) - (1/8)C_0^{[3]}(z^{\text{en}})(x^2+y^2)+\cdots] \\
&= -qy(1/2)[B_z(0,0,z^{\text{en}}) - (1/8)B_z''(0,0,z^{\text{en}})(x^2+y^2)+\cdots] \\
&= -qy(B/2)[\text{bump}(z^{\text{en}},a,L) - (1/8)\text{bump}''(z^{\text{en}},a,L)(x^2+y^2)+\cdots],
\end{aligned} \tag{16.2.137}$$

$$\begin{aligned}
\Delta p_y^{\text{mech}} &= qA_y(x,y,z^{\text{en}}) = qxU(\rho,z^{\text{en}}) \\
&= qx(1/2)[C_0^{[1]}(z^{\text{en}}) - (1/8)C_0^{[3]}(z^{\text{en}})(x^2+y^2)+\cdots] \\
&= qx(1/2)[B_z(0,0,z^{\text{en}}) - (1/8)B_z''(0,0,z^{\text{en}})(x^2+y^2)+\cdots] \\
&= qx(B/2)[\text{bump}(z^{\text{en}},a,L) - (1/8)\text{bump}''(z^{\text{en}},a,L)(x^2+y^2)+\cdots].
\end{aligned} \tag{16.2.138}$$

[5] We hasten to add that the aberrations associated with $\mathcal{M}^{\text{LFF}}$ and $\mathcal{M}^{\text{TFF}}$ are not the only aberrations for $\mathcal{M}_{\text{solenoid}}$. According to (2.112) there will also be aberrations associated with $\mathcal{M}^{\text{uniform}}$. But they are always finite and, by definition, a independent. They are also relatively easy to compute because $\mathcal{M}^{\text{uniform}}$ arises from the z *independent* Hamiltonian H^{uniform}.

Here, in writing the last lines of (2.137) and (2.138), we have assumed the field profile of Subsection 2.2. Similarly, upon exit, we find from (1.41), and (1.42) the discontinuity results

$$\begin{aligned}\Delta p_x^{\text{mech}} &= qA_x(x,y,z^{\text{ex}}) = -qyU(\rho,z^{\text{ex}}) \\ &= -qy(1/2)[C_0^{[1]}(z^{\text{ex}}) - (1/8)C_0^{[3]}(z^{\text{ex}})(x^2+y^2)+\cdots] \\ &= -qy(1/2)[B_z(0,0,z^{\text{ex}}) - (1/8)B_z''(0,0,z^{\text{ex}})(x^2+y^2)+\cdots] \\ &= -qy(B/2)[\text{bump}(z^{\text{ex}},a,L) - (1/8)\text{bump}''(z^{\text{ex}},a,L)(x^2+y^2)+\cdots],\end{aligned} \tag{16.2.139}$$

$$\begin{aligned}\Delta p_y^{\text{mech}} &= qA_y(x,y,z^{\text{ex}}) = qxU(\rho,z^{\text{ex}}) \\ &= qx(1/2)[C_0^{[1]}(z^{\text{ex}}) - (1/8)C_0^{[3]}(z^{\text{ex}})(x^2+y^2)+\cdots] \\ &= qx(1/2)[B_z(0,0,z^{\text{ex}}) - (1/8)B_z''(0,0,z^{\text{ex}})(x^2+y^2)+\cdots] \\ &= qx(B/2)[\text{bump}(z^{\text{ex}},a,L) - (1/8)\text{bump}''(z^{\text{ex}},a,L)(x^2+y^2)+\cdots].\end{aligned} \tag{16.2.140}$$

We see that in all cases the discontinuities are proportional to $B_z(0,0,z)$ and its derivatives at $z = z^{\text{en}}$ or $z = z^{\text{ex}}$. See Figures 2.4 through 2.7 for examples of how these functions behave in the case of a simple air-code solenoid. Moreover, the discontinuities also vanish as the spatial deviations from the z axis (the design orbit) become small.

We close this subsection by finding, at entry and exit, the surface currents implied by our termination procedure/approximation. Since $\hat{A}_z^0 = 0$ in the Poincaré-Coulomb gauge for any solenoid or collection of solenoids, the relations (1.65) through (1.67) take the form

$$\begin{aligned}\mu_0 j_x^{\text{mod}} &= -2[\delta(z-z^{\text{en}}) - \delta(z^{\text{ex}}-z)]\partial_z\hat{A}_x^0 \\ &\quad -[\delta'(z-z^{\text{en}}) + \delta'(z^{\text{ex}}-z)]\hat{A}_x^0,\end{aligned} \tag{16.2.141}$$

$$\begin{aligned}\mu_0 j_y^{\text{mod}} &= -2[\delta(z-z^{\text{en}}) - \delta(z^{\text{ex}}-z)]\partial_z\hat{A}_y^0 \\ &\quad -[\delta'(z-z^{\text{en}}) + \delta'(z^{\text{ex}}-z)]\hat{A}_y^0,\end{aligned} \tag{16.2.142}$$

$$\mu_0 j_z^{\text{mod}} = 0. \tag{16.2.143}$$

Let us evaluate (2.141) and (2.142) using the explicit form for $\hat{\boldsymbol{A}}^0$ given by (2.7) through (2.9. Doing so gives the intermediate results

$$\begin{aligned}\mu_0 j_x^{\text{mod}} &= 2[\delta(z-z^{\text{en}}) - \delta(z^{\text{ex}}-z)]y\partial_z U \\ &\quad +[\delta'(z-z^{\text{en}}) + \delta'(z^{\text{ex}}-z)]yU,\end{aligned} \tag{16.2.144}$$

$$\begin{aligned}\mu_0 j_y^{\text{mod}} &= -2[\delta(z-z^{\text{en}}) - \delta(z^{\text{ex}}-z)]x\partial_z U \\ &\quad -[\delta'(z-z^{\text{en}}) + \delta'(z^{\text{ex}}-z)]xU.\end{aligned} \tag{16.2.145}$$

At this point it is convenient to employ cylindrical components for $\boldsymbol{j}^{\rm mod}$ using the relations

$$j_\rho^{\rm mod} = \cos\phi\, j_x^{\rm mod} + \sin\phi\, j_y^{\rm mod}, \tag{16.2.146}$$

$$j_\phi^{\rm mod} = -\sin\phi\, j_x^{\rm mod} + \cos\phi\, j_y^{\rm mod}. \tag{16.2.147}$$

Recall (15.2.22) and (15.2.23). Implementing these substitutions gives the results

$$\mu_0 j_\rho^{\rm mod} = 0, \tag{16.2.148}$$

$$\begin{aligned} \mu_0 j_\phi^{\rm mod} &= -2[\delta(z-z^{\rm en}) - \delta(z^{\rm ex}-z)]\rho\partial_z U \\ &\quad -[\delta'(z-z^{\rm en}) + \delta'(z^{\rm ex}-z)]\rho U. \end{aligned} \tag{16.2.149}$$

Here we have used the relations

$$y\cos\phi - x\sin\phi = 0, \tag{16.2.150}$$

$$y\sin\phi + x\cos\phi = \rho. \tag{16.2.151}$$

We see that $\boldsymbol{j}^{\rm mod}$, the current that is required to cancel the residual solenoidal fringe field, has only a ϕ component. This is to be expected since the current that produces the solenoidal field itself has only a ϕ component. The last step is to use the expansion (2.10) for U. With the aid of this expansion we find the final result

$$\begin{aligned} \mu_0 j_\phi^{\rm mod} &= -2[\delta(z-z^{\rm en}) - \delta(z^{\rm ex}-z)]\rho\partial_z U \\ &\quad -[\delta'(z-z^{\rm en}) + \delta'(z^{\rm ex}-z)]\rho U \\ &= [\delta(z-z^{\rm en}) - \delta(z^{\rm ex}-z)]\rho[C_0^{[2]}(z) - (1/8)C_0^{[4]}(z)(x^2+y^2) + \cdots] \\ &\quad +(1/2)[\delta'(z-z^{\rm en}) + \delta'(z^{\rm ex}-z)]\rho[C_0^{[1]}(z) - (1/8)C_0^{[3]}(z)(x^2+y^2) + \cdots] \\ &= [\delta(z-z^{\rm en}) - \delta(z^{\rm ex}-z)]\rho[B_z'(0,0,z) - (1/8)B_z'''(0,0,z)(x^2+y^2) + \cdots] \\ &\quad +(1/2)[\delta'(z-z^{\rm en}) + \delta'(z^{\rm ex}-z)]\rho[B_z(0,0,z) - (1/8)B_z''(0,0,z)(x^2+y^2) + \cdots]. \end{aligned} \tag{16.2.152}$$

Like the discontinuities in the mechanical momenta, $\boldsymbol{j}^{\rm mod}$ is also proportional to $B_z(0,0,z)$ and its derivatives at $z = z^{\rm en}$ or $z = z^{\rm ex}$, and also vanishes as the spatial deviations from the z axis (the design orbit) become small.

Exercises

16.2.1. Verify that $B_z(0,0,z)$ as given by (2.16) describes the on-axis field of a simple air-core solenoid. Verify that there is the result

$$B = \mu_0 IN/L \tag{16.2.153}$$

where I is the current in the coil, N is the number of turns in the single-layer winding, and L is the length of the coil. Hint: Use (2.24) and Ampère's law. It is of historical interest to note that the name *solenoid* was coined by Ampère.

16.2.2. Show that for the on-axis field of a simple solenoid as given by (2.16) there are, for the fields at the midpoint/center and either end, the limiting behaviors

$$\lim_{L\to\infty} B_z(0,0,L/2) = B, \tag{16.2.154}$$

$$\lim_{L\to\infty} B_z(0,0,0) = \lim_{L\to\infty} B_z(0,0,L) = B/2. \tag{16.2.155}$$

Verify that the same is true for any bump function model that is constructed from approximating signum functions.

16.2.3. The purpose of this exercise is to verify the relations (2.20) through (2.23). Show that the approximating signum function (2.26) has the properties

$$\text{sgn}(-z,a) = -\text{sgn}(z,a), \tag{16.2.156}$$

$$\lim_{z\to\pm\infty} \text{sgn}(z,a) = \pm 1. \tag{16.2.157}$$

Verify the limiting behavior (2.27). Sketch sgn(z,a), $-$sgn$(z-L,a)$, and bump(z,a,L) as given by (2.25) to verify the relations (2.20) through (2.22).

What remains is to prove the relation (2.23). Begin by writing

$$\int_{-\infty}^{\infty} \text{bump}(z,a,L)dz = \lim_{w\to\infty} \int_{-w}^{w} \text{bump}(z,a,L)dz. \tag{16.2.158}$$

Next verify from the representation (2.25) that

$$\int_{-w}^{w} \text{bump}(z,a,L)dz = (1/2)\int_{-w}^{w} \text{sgn}(z,a)dz - (1/2)\int_{-w}^{w} \text{sgn}(z-L,a)dz. \tag{16.2.159}$$

Show that the first integral on the right side of (2.159) vanishes because of (2.156). Show that by making the change of variables $x = z - L$ the second integral on the right side of (2.159) becomes

$$\begin{aligned} -(1/2)\int_{-w}^{w} \text{sgn}(z-L,a)dz &= -(1/2)\int_{-w-L}^{w-L} \text{sgn}(x,a)dx \\ &= -(1/2)\int_{-w-L}^{w+L} \text{sgn}(x,a)dx + (1/2)\int_{w-L}^{w+L} \text{sgn}(x,a)dx. \end{aligned} \tag{16.2.160}$$

Verify that the first integral in the second line of (2.160) vanishes, again because of (2.156). It follows that there is the result

$$\int_{-w}^{w} \text{bump}(z,a,L)dz = (1/2)\int_{w-L}^{w+L} \text{sgn}(x,a)dx. \tag{16.2.161}$$

Show from (2.157) that there is the result

$$\lim_{w\to\infty} (1/2)\int_{w-L}^{w+L} \text{sgn}(x,a)dx = (1/2)2L = L. \tag{16.2.162}$$

Put all your intermediate results together to obtain the final result

$$\int_{-\infty}^{\infty} \text{bump}(z,a,L)dz = L, \tag{16.2.163}$$

as desired. Note that the proof of this result has depended only on the representation (2.25) and properties (2.156) and (2.157), which are required properties of any approximating signum function.

16.2.4. Verify the fall-off relations (2.34) and (2.35) and the limiting behaviors (2.36) and (2.37).

16.2.5. Verify the expansion (2.38) and the near leading end fall-off behavior (2.39).

16.2.6. Verify the expansion (2.40) and the far fall-off behavior given by (2.41) and (2.42).

16.2.7. Verify the near-by and far asymptotic behaviors (2.44) and (2.45).

16.2.8. Verify the far fall-off behavior (2.47).

16.2.9. Verify the relation (2.49) for the on-axis field of a thick solenoid. Show that (2.24), (2.154), and (2.155) continue to hold. Show that (2.153) also holds providing that N is now the total of number of turns in the whole multilayer winding. Finally, use the notation $B_z(0,0,z;a)$ and $B_z(0,0,z;a_1,a_2)$ to denote the right sides of (2.16) and (2.49), respectively. Show that there is the limiting relation

$$\lim_{a_2 \to a} B_z(0,0,z;a,a_2) = B_z(0,0,z;a). \tag{16.2.164}$$

Hint: Use the representation (2.48).

16.2.10. Observe that (2.7) through (2.9) are the same as (15.5.7) through (15.5.9). Verify that using (15.2.22), (15.2.23), and (15.5.7) through (15.5.9) gives the results (15.5.37) through (15.5.39). That is, $\hat{\boldsymbol{A}}^0$ and hence $\boldsymbol{A}^s$ has only a ϕ component. Verify that this component manifests rotational symmetry about the z axis by having no ϕ dependence.

16.2.11. Verify that the $b^{[2n]}(z)$ as given by (2.59) are dimensionless.

16.2.12. In (2.64) the quantity J_z is defined in terms of canonical variables. What happens if mechanical momenta are employed instead? Define *scaled* mechanical momenta P_x^{mech} and P_y^{mech} by writing

$$P_x^{\text{mech}} = P_x - A_x^s, \tag{16.2.165}$$

$$P_y^{\text{mech}} = P_y - A_y^s; \tag{16.2.166}$$

from which it follows that there are the relations

$$P_x = P_x^{\text{mech}} + A_x^s, \tag{16.2.167}$$

$$P_y = P_y^{\text{mech}} + A_y^s. \tag{16.2.168}$$

[Recall the unscaled results (1.5.30).] Show that the expression for J_z in terms of mechanical momenta is given by the relation

$$J_z = XP_y^{\text{mech}} - YP_x^{\text{mech}} + XA_y^s - YA_x^s. \tag{16.2.169}$$

Verify, using (2.55), (2.56), and (2.58), that there is the result

$$\begin{aligned} XA_y^s - YA_x^s &= (X^2+Y^2)U^s(X,Y,z) \\ &= (1/2)\sum_{n=0}^{\infty}(-1)^n \frac{1}{2^{2n}n!(n+1)!} b^{[2n]}(z)(X^2+Y^2)^{n+1}. \end{aligned} \tag{16.2.170}$$

Show from (15.5.38) and (2.52) that there is the result

$$\begin{aligned} \rho\hat{A}_\phi^0 &= \rho^2 U(\rho,z) = (1/2)\sum_{n=0}^{\infty}(-1)^n \frac{1}{2^{2n}n!(n+1)!} C_0^{[2n+1]}(z)\rho^{2n+2} \\ &= (1/2)\sum_{n=0}^{\infty}(-1)^n \frac{1}{2^{2n}n!(n+1)!} B^{[2n]}(z)\rho^{2n+2}. \end{aligned} \tag{16.2.171}$$

Verify from (13.1.21) and (13.1.22) that

$$\rho^{2n+2} = \ell^{2n+2}(X^2+Y^2)^{n+1}. \tag{16.2.172}$$

Therefore (2.171) can also be written in the form

$$\rho\hat{A}_\phi^0 = (1/2)\sum_{n=0}^{\infty}(-1)^n \frac{1}{2^{2n}n!(n+1)!} B^{[2n]}(z)\ell^{2n+2}(X^2+Y^2)^{n+1}. \tag{16.2.173}$$

Verify from (2.59) that

$$B^{[2n]}(z)\ell^{2n+2} = (\ell p^0/q)b^{[2n]}(z) \tag{16.2.174}$$

so that

$$\begin{aligned} \rho\hat{A}_\phi^0 &= (\ell p^0/q)(1/2)\sum_{n=0}^{\infty}(-1)^n \frac{1}{2^{2n}n!(n+1)!} b^{[2n]}(z)(X^2+Y^2)^{n+1} \\ &= (\ell p^0/q)(X^2+Y^2)U^s(X,Y,z) = (\ell p^0/q)(XA_y^s - YA_x^s). \end{aligned} \tag{16.2.175}$$

Make the definition

$$J_z^{\text{mech}} = XP_y^{\text{mech}} - YP_x^{\text{mech}}. \tag{16.2.176}$$

Consequently, verify that (2.169) can be rewritten in the form

$$J_z = J_z^{\text{mech}} + [q/(\ell p^0)]\rho\hat{A}_\phi^0. \tag{16.2.177}$$

According to (2.82) J_z is conserved. Also, far outside a solenoid, $\hat{A}_\phi^0$ and hence $\rho\hat{A}_\phi^0$ vanish. Therefore, (2.177) shows that J_z^{mech} must have different values inside and outside a solenoid and what their difference must be. Show that if a particle enters a solenoid from a zero field region with some initial value of J_z and ultimately exits into a second zero field region, then its final value of J_z must be the same as its initial value.

16.2.13. Verify the identity (2.76). Show that (2.75) and (2.79) are a consequence of (2.70) through (2.73) and (2.77) and (2.78).

16.2.14. Compare the f_4 contents of the maps in Exhibits 2.3, 2.4, and 2.5. Which f_4 entries appear to be diverging (in magnitude) to ∞ as $a \to 0$? Why, for a given value of a, are there the relations $f(84) = f(175)$ and $f(95) = 2f(84)$? Can you find other relations of this kind? Hint: Show that the Lie generators f_n for a solenoid map must satisfy

$$: J_z : f_n = 0. \tag{16.2.178}$$

Show that the same is true for a drift map and therefore for a composite of drift and solenoid maps.

16.2.15. The purpose of this exercise is to verify the factorization of the linear part as described in Subsection 2.5.3. Begin by writing the transfer map for a solenoid in the general form

$$\mathcal{M} = \mathcal{R} \exp(: f_3 :) \exp(: f_4 :) \cdots . \tag{16.2.179}$$

Then $\mathcal{R}$ will be determined by H_2 as given by (2.94). Verify that H_2 can be written in the form

$$H_2 = H_2^{\rm nonrot} + H_2^{\rm rot} \tag{16.2.180}$$

where

$$H_2^{\rm nonrot} = [1/(2\ell)]P^2 + \{[b^{[0]}(z)]^2/(8\ell)\}Q^2 + [1/(2\beta_0^2\gamma_0^2\ell)]P_\tau^2 \tag{16.2.181}$$

and

$$H_2^{\rm rot} = -[b^{[0]}(z)/(2\ell)]J_z. \tag{16.2.182}$$

Let $\mathcal{R}$, $\mathcal{R}^{\rm nonrot}$, and $\mathcal{R}^{\rm rot}$ be the maps generated by H_2, $H_2^{\rm nonrot}$, and $H_2^{\rm rot}$, respectively. Verify that $: H_2^{\rm nonrot} :$ and $: H_2^{\rm rot} :$ commute and consequently prove, using the results of Exercise 10.2.2, that there are the relations

$$\mathcal{R} = \mathcal{R}^{\rm nonrot}\mathcal{R}^{\rm rot} = \mathcal{R}^{\rm rot}\mathcal{R}^{\rm nonrot}. \tag{16.2.183}$$

Correspondingly, there are the associated matrix relations

$$R = R^{\rm rot}R^{\rm nonrot} = R^{\rm nonrot}R^{\rm rot}. \tag{16.2.184}$$

16.2.16. Through terms of second order, the transfer map for a solenoid can be written in the general form

$$\mathcal{M} = \mathcal{R} \exp(: f_3 :). \tag{16.2.185}$$

The goal of this exercise is to compute for a simple solenoid, in the hard-edge limit $a \to 0$, both the matrix R associated with $\mathcal{R}$ and the Lie generator f_3. For a further discussion of motion in a uniform magnetic field, see Exercise 32.2.7.

Examination of H_2 and H_3 as given by (2.94) and (2.95) shows that their z dependence is given entirely in terms of $b^{[0]}(z)$. This function is bounded for all a, and in the hard-edge limit takes on a *constant* value in the open interval $z \in (0, L)$. See Figures 2.4 and 2.5. We recall that here z plays the role of the independent variable, and therefore in the hard-edge limit H_2 and H_3 do not depend on the independent variable. Consequently show that, in the hard-edge limit and through terms of second order, the transfer map for a solenoid can be written in the form

$$\mathcal{M} = \exp(-L : H_2 + H_3 :). \tag{16.2.186}$$

Recall (7.4.1). Next verify that $: H_2 :$ and $: H_3 :$ commute under the assumption that $b^{[0]}(z)$ is constant in the open interval $z \in (0, L)$. It follows that (2.186) can be rewritten in the form

$$\mathcal{M} = \exp(-L : H_2 :) \exp(-L : H_3 :). \tag{16.2.187}$$

Comparison of (2.185) and (2.187) yields the results

$$\mathcal{R} = \exp(-L : H_2 :) \tag{16.2.188}$$

and

$$f_3 = -LH_3. \tag{16.2.189}$$

What remains is to find the matrix R associated with $\mathcal{R}$. Review Exercise 2.15. The discussion there holds for all (well behaved) functions $b^{[0]}(z)$ and therefore also holds when $b^{[0]}(z)$ is constant in the open interval $z \in (0, L)$. In the case that $b^{[0]}(z)$ is constant in the open interval $z \in (0, L)$ there are the results

$$\mathcal{R}^{\text{nonrot}} = \exp(-L : H_2^{\text{nonrot}} :) \tag{16.2.190}$$

and

$$\mathcal{R}^{\text{rot}} = \exp(-L : H_2^{\text{rot}} :). \tag{16.2.191}$$

At this point we pause to observe that f_3 as given by (2.95) and (2.189) can also be decomposed into two parts whose associated Lie operators commute. We may write

$$f_3 = f_3^{\text{nonrot}} + f_3^{\text{rot}} \tag{16.2.192}$$

where

$$f_3^{\text{nonrot}} = -L(1/\beta_0)P_\tau H_2^{\text{nonrot}} \tag{16.2.193}$$

and

$$f_3^{\text{rot}} = -L(1/\beta_0)P_\tau H_2^{\text{rot}}. \tag{16.2.194}$$

To continue, your next task is to compute R^{nonrot}, the matrix associated with $\mathcal{R}^{\text{nonrot}}$. Begin by verifying that H_2^{nonrot} as given by (2.181) consists of three pieces associated with the x, P_x; y, P_y; and τ, P_τ planes, and that the Lie operators associated with different pieces all commute. Correspondingly R^{nonrot} has nonzero entries consisting only of 2×2 matrices centered on the diagonal. What are these matrices? Consider the Lie transformation

$$\exp\{: [-L/(2\ell)]P_x^2 - [L(b^{[0]})^2/(8\ell)]X^2 :\} \tag{16.2.195}$$

that is associated with the x, P_x part of $\mathcal{R}^{\text{nonrot}}$ when $b^{[0]}(z)$ is constant in the open interval $z \in (0, L)$. Let R_{XP_x} be the 2×2 matrix that describes the action of this Lie transformation on the X, P_x plane. Use the formalism and results of Section 8.7.2 to make the identifications

$$b = L/\ell, \tag{16.2.196}$$

$$a = 0, \tag{16.2.197}$$

$$c = [L/(4\ell)](b^{[0]})^2, \tag{16.2.198}$$

from which it follows that

$$\Delta = -[L/(2\ell)]^2(b^{[0]})^2 \tag{16.2.199}$$

and

$$\Delta^{1/2} = i[L/(2\ell)](b^{[0]}). \tag{16.2.200}$$

Show using (8.7.35) that there is the result

$$R_{XP_x} = \begin{pmatrix} \cosh(\Delta^{1/2}) & b[\sinh(\Delta^{1/2})]/\Delta^{1/2} \\ -c[\sinh(\Delta^{1/2})]/\Delta^{1/2} & \cosh(\Delta^{1/2}) \end{pmatrix}. \tag{16.2.201}$$

Also verify that there are the relations

$$b/\Delta^{1/2} = (L/\ell)(-i)(2\ell/L)(1/b^{[0]}) = -i(2/b^{[0]}) \tag{16.2.202}$$

and

$$c/\Delta^{1/2} = [L/(4\ell)](b^{[0]})^2(-i)(2\ell/L)(1/b^{[0]}) = -i(b^{[0]}/2). \tag{16.2.203}$$

Introduce the notation

$$k = (b^{[0]}/2) \tag{16.2.204}$$

and

$$\psi = [L/(2\ell)](b^{[0]}) = k(L/\ell) \tag{16.2.205}$$

so that (2.200), (2.202), and (2.203) take the forms

$$\Delta^{1/2} = i\psi, \tag{16.2.206}$$

$$b/\Delta^{1/2} = -i/k, \tag{16.2.207}$$

and

$$c/\Delta^{1/2} = -ik. \tag{16.2.208}$$

Verify, using this notation, that (2.201) can be written in the final form

$$R_{XP_x} = \begin{pmatrix} \cos(\psi) & (1/k)\sin(\psi) \\ -k\sin(\psi) & \cos(\psi) \end{pmatrix}. \tag{16.2.209}$$

With regard of the action of $\mathcal{R}^{\rm nonrot}$ on the Y, P_y plane, verify from symmetry considerations that

$$R_{YP_y} = R_{XP_x}. \tag{16.2.210}$$

Finally show that the the effect of the Lie transformation $\exp\{:[-L/(2\beta_0^2\gamma_0^2\ell)]P_\tau^2:\}$ on the τ, P_τ plane is given by the relations

$$\exp\{:[-L/(2\beta_0^2\gamma_0^2\ell)]P_\tau^2:\}\tau = \tau + [L/(\beta_0^2\gamma_0^2\ell)]P_\tau, \tag{16.2.211}$$

$$\exp\{:[-L/(2\beta_0^2\gamma_0^2\ell)]P_\tau^2:\}P_\tau = P_\tau. \tag{16.2.212}$$

Consequently there is the corresponding matrix

$$R_{\tau P_\tau} = \begin{pmatrix} 1 & \eta \\ 0 & 1 \end{pmatrix} \tag{16.2.213}$$

where

$$\eta = L/(\beta_0^2 \gamma_0^2 \ell). \tag{16.2.214}$$

You have shown that

$$R^{\text{nonrot}} = \begin{pmatrix} C & (1/k)S & 0 & 0 & 0 & 0 \\ -kS & C & 0 & 0 & 0 & 0 \\ 0 & 0 & C & (1/k)S & 0 & 0 \\ 0 & 0 & -kS & C & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \eta \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{16.2.215}$$

where we have used the short-hand notation

$$C = \cos(\psi) \tag{16.2.216}$$

and

$$S = \sin(\psi). \tag{16.2.217}$$

Note that the entries in R^{nonrot} are *even* functions of k and hence R^{nonrot} is *invariant* under the replacement $b^{[0]} \rightarrow -b^{[0]}$. This symmetry is also evident from the form of H_2^{nonrot} as given in (2.167).

Now turn to the calculation of R^{rot}, the matrix associated with $\mathcal{R}^{\text{rot}}$. Verify that $\mathcal{R}^{\text{rot}}$ can be written on the form

$$\mathcal{R}^{\text{rot}} = \exp(-L : H_2^{\text{rot}} :) = \exp(\psi : J_z :). \tag{16.2.218}$$

Use the results (2.70) and (2.71) to verify the relation

$$\begin{aligned} &\mathcal{R}^{\text{rot}} X = \exp(\psi : J_z :) X = \\ &X + \psi : J_z : X + \psi^2 (1/2!) : J_z :^2 X + \psi^3 (1/3!) : J_z :^3 X + \cdots = \\ &X + \psi Y - \psi^2 (1/2!) X - \psi^3 (1/3!) Y + \cdots = \\ &X[1 - \psi^2 (1/2!) + \cdots] + Y[\psi - \psi^3 (1/3!) + \cdots] = \\ &X \cos(\psi) + Y \sin(\psi). \end{aligned} \tag{16.2.219}$$

In a similar way verify that there is the relation

$$\mathcal{R}^{\text{rot}} Y = -X \sin(\psi) + Y \cos(\psi). \tag{16.2.220}$$

Use the results (2.72) and (2.73) to find the analogous relations

$$\mathcal{R}^{\text{rot}} P_x = P_x \cos(\psi) + P_y \sin(\psi) \tag{16.2.221}$$

and

$$\mathcal{R}^{\text{rot}} P_y = -P_x \sin(\psi) + P_y \cos(\psi). \tag{16.2.222}$$

Finally observe from (2.74) that $\mathcal{R}^{\rm rot}$ leaves τ and P_τ in peace. You have shown that

$$R^{\rm rot} = \begin{pmatrix} C & 0 & S & 0 & 0 & 0 \\ 0 & C & 0 & S & 0 & 0 \\ -S & 0 & C & 0 & 0 & 0 \\ 0 & -S & 0 & C & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{16.2.223}$$

Note that, according to (2.218), making the replacement $b^{[0]} \rightarrow -b^{[0]}$ entails the replacement $R^{\rm rot} \rightarrow (R^{\rm rot})^{-1}$.

We are almost done. According to (2.184) there will be the matrix relation

$$R = R^{\rm rot}R^{\rm nonrot} = R^{\rm nonrot}R^{\rm rot} = M^{\rm uniform}. \tag{16.2.224}$$

Verify that

$$R = M^{\rm uniform} = \begin{pmatrix} C^2 & (1/k)SC & SC & (1/k)S^2 & 0 & 0 \\ -kSC & C^2 & -kS^2 & SC & 0 & 0 \\ -SC & -(1/k)S^2 & C^2 & (1/k)SC & 0 & 0 \\ kS^2 & -SC & -kSC & C^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \eta \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{16.2.225}$$

Show that (2.225) agrees with (13.4.1) when use is made of (2.59) and (1.5.81).

Verify that, if we wish, we may write

$$R = (R^{\rm rot})^{1/2}R^{\rm nonrot}(R^{\rm rot})^{1/2} \tag{16.2.226}$$

where

$$(R^{\rm rot})^{1/2} = \exp[-(L/2) : H_2^{\rm rot} :] = \exp[(\psi/2) : J_z :]. \tag{16.2.227}$$

However, there does not seem to be much point in doing so except, perhaps, to exhibit the reversibility properties of R. See Section 36.2.

16.2.17. Review Section 2.5.3. This exercise explores properties of $H^{\rm nonrot}$ and $\mathcal{M}^{\rm nonrot}$. We will employ the notation

$$\zeta = (X, P_x, Y, P_y, \tau, P_\tau). \tag{16.2.228}$$

Begin by considering *initial* conditions with the property

$$\zeta_1^{\rm in} = X^{\rm in} = 0, \tag{16.2.229}$$

$$\zeta_2^{\rm in} = P_x^{\rm in} = 0, \tag{16.2.230}$$

$$\zeta_3^{\rm in} = Y^{\rm in} = \text{anything}, \tag{16.2.231}$$

$$\zeta_4^{\rm in} = P_y^{\rm in} = \text{anything}, \tag{16.2.232}$$

$$\zeta_5^{\rm in} = \tau^{\rm in} = \text{anything}, \tag{16.2.233}$$

$$\zeta_6^{\rm in} = P_\tau^{\rm in} = \text{anything}. \tag{16.2.234}$$

These are the initial conditions for motion that is, at least initially, in the vertical plane. Show that the resulting subsequent motion, when governed by $H^{\rm nonrot}$, remains in the vertical plane. In particular, show that then the associated *final* conditions under the action of $\mathcal{M}^{\rm nonrot}$ have the property

$$X^{\rm fin} = 0, \tag{16.2.235}$$

$$P_x^{\rm fin} = 0. \tag{16.2.236}$$

Next show that

$$: J_z : \mathcal{M}^{\rm nonrot} = \mathcal{M}^{\rm nonrot} : J_z :, \tag{16.2.237}$$

and consequently

$$\mathcal{R}^{\rm rot}\mathcal{M}^{\rm nonrot} = \mathcal{M}^{\rm nonrot}\mathcal{R}^{\rm rot}. \tag{16.2.238}$$

Suppose we summarize the results (2.229) through (2.236) by writing

$$\zeta^{\rm fin} = \mathcal{M}^{\rm nonrot}\zeta^{\rm fn}. \tag{16.2.239}$$

Use (2.238) and (2.239) to show that

$$\mathcal{M}^{\rm nonrot}\mathcal{R}^{\rm rot}\zeta^{\rm in} = \mathcal{R}^{\rm rot}\zeta^{\rm fin}. \tag{16.2.240}$$

You have shown that $\mathcal{M}^{\rm nonrot}$ preserves *all* planes obtained by rotating, by any angle ψ, the vertical plane about the z axis. This includes, of course, the horizontal plane.

16.2.18. Review the discussion of Section 15.11.1. Using (2.3) through (2.6), show that for a solenoid the transverse components of its magnetic field have the integral property

$$\int_{-\infty}^{\infty} dz\ B_x(x,y,z) = \int_{-\infty}^{\infty} dz\ B_y(x,y,z) = 0. \tag{16.2.241}$$

16.2.19. Review Exercise 2.1. Consider a "one-turn" solenoid consisting of a single circular loop of radius a lying in the $z=0$ plane, centered on the origin, and carrying a current I. Show that in this case

$$B_z^{\rm one\ turn}(0,0,z)) = \mu_0 I\delta(z,a) \tag{16.2.242}$$

where $\delta(z,a)$ is an approximating delta function defined by

$$\delta(z,a) = (a^2/2)/(z^2+a^2)^{3/2}. \tag{16.2.243}$$

See the discussion following (3.13). See also Section 3.3. Show that

$$\int_{-\infty}^{\infty} dz\ \delta(z,a) = 1. \tag{16.2.244}$$

Show also that there is the relation

$$\text{bump}(z,a,L) = \int_0^L dz'\ \delta(z-z',a). \tag{16.2.245}$$

Show that (2.23) follows from (2.244) and (2.245).

16.2.20. Verify that (2.134) through (2.136) reproduce the f_4 content of Exhibits 2.19 through 2.21.

16.2.21. The purpose of this exercise is to study how the coefficients of the I_i, save for I_6, in the relations (2.134) through (2.136) tend to zero as $a \to 0$. Let us write relations of this kind in the general form

$$f_4^{\text{LFF}}(a) = \sum_{i=1}^{10} c_i^{\text{LFF}}(a) I_i \tag{16.2.246}$$

where we have indicated that f_4^{LFF} is a dependent. Verify that for the case of $c_1^{\text{LFF}}(a)$ one may make the Table 2.4 below, and conclude that $c_1^{\text{LFF}}(a)$ vanishes as a^2 when a goes to zero. Verify that the same is true for $c_2^{\text{LFF}}(a)$, $c_7^{\text{LFF}}(a)$, and $c_8^{\text{LFF}}(a)$. By contrast, verify that $c_3^{\text{LFF}}(a)$, $c_4^{\text{LFF}}(a)$, $c_5^{\text{LFF}}(a)$, and $c_9^{\text{LFF}}(a)$ vanish as a^1 when a goes to zero. Evidently $c_{10}^{\text{LFF}}(a)$ vanishes for all values of a.

Table 16.2.4: Numerical behavior of $c_1^{\text{LFF}}(a)$ for small values of a.

a	c_1^{LFF}	$(1/a^2)c_1^{\text{LFF}}$
.2	4.3607E-5	1.09019E-3
.02	5.2063E-7	1.30158E-3
.002	5.3334E-9	1.33335E-3

16.2.22. By making a suitable table, illustrate that $c_6^{\text{LFF}}(a)$ diverges as $1/a$ when a goes to zero.

16.2.23. Show that H_4 as given by (2.96) can be written in the form

$$H_4 = \sum_{i=1}^{10} d_i(z) I_i, \tag{16.2.247}$$

and verify that it is d_5 and d_6 that contain the pesky δ' functions. In particular, verify that d_5 contains $b^{[2]}$ (which involves δ' functions) and d_6 contains the product $b^{[0]}b^{[2]}$. Reason that the latter is more singular than the former because $b^{[0]}$ is discontinuous in the hard-edge limit. Show that there are the results

$$\text{sgn}'(z,a) = \partial_z \text{sgn}(z,a) = 2\delta(z,a) \tag{16.2.248}$$

and

$$\begin{aligned}\int_{-\infty}^{\infty} dz\ \text{sgn}(z,a)\delta'(z,a) &= [\text{sgn}(z,a)\delta(z,a)]|_{z=-\infty}^{z=+\infty} - \int_{-\infty}^{\infty} dz\ \text{sgn}'(z,a)\delta(z,a) \\ = 0 - 2\int_{-\infty}^{\infty} dz\ \delta(z,a)\delta(z,a) &= -2\int_{-\infty}^{\infty} dz\ \delta^2(z,a) = -(3\pi/16)(1/a).\end{aligned} \tag{16.2.249}$$

Review and, if you have not already done so, perform Exercise 2.22. In agreement with the results of this exercise, make the small a Ansatz

$$c_6^{\rm LFF}(a) \simeq e^{\rm LFF}/a \tag{16.2.250}$$

where $e^{\rm LFF}$ is a coefficient to be determined. Find a formula for $e^{\rm LFF}$ in terms of beam parameters and the parameters for a simple solenoid.

16.2.24. Exercise on the vector spherical harmonic decomposition of the vector potential for the magnetic field of a simple solenoid.

16.3 Iron-Free Dipoles

16.3.1 Preliminaries

A *straight* dipole is a straight beam-line element whose field is described by a cylindrical harmonic expansion that contains primarily an $m = 1$ term.[6] We recall from Section 15.3.3 that for the $m = 1$ case the magnetic scalar potential ψ has the expansion

$$\psi(x, y, z) = \psi_{1,s}(x, y, z) = y[C_{1,s}^{[0]}(z) - (1/8)(x^2 + y^2)C_{1,s}^{[2]}(z) + \cdots]. \tag{16.3.1}$$

See (15.3.57). [Here we have retained only the *normal* $m = 1$ term, but a skew ($\psi_{1,c}$) term is also possible. See (15.3.40) and Exercise 15.4.1.] Correspondingly, the associated magnetic field has the expansion

$$B_x = \partial_x\psi_{1,s} = -(1/4)xyC_{1,s}^{[2]}(z) + \cdots\ , \tag{16.3.2}$$

$$B_y = \partial_y\psi_{1,s} = C_{1,s}^{[0]}(z) - (1/8)(x^2 + 3y^2)C_{1,s}^{[2]}(z) + \cdots\ , \tag{16.3.3}$$

$$B_z = \partial_z\psi_{1,s} = y[C_{1,s}^{[1]}(z) - (1/8)(x^2 + y^2)C_{1,s}^{[3]}(z) + \cdots]. \tag{16.3.4}$$

From (3.2) through (3.4) we see that $\boldsymbol{B}$ is completely specified in terms of a single "master" function $C_{1,s}^{[0]}(z)$ and its derivatives. Moreover, according to (3.3), the on-axis field has only a B_y component, and it is given by the relation

$$B_y(0, 0, z) = C_{1,s}^{[0]}(z). \tag{16.3.5}$$

See Exercise 1.5.7.

There is also a suitable associated vector potential $\hat{\boldsymbol{A}}^{1,s}$ given (in symmetric Coulomb gauge) by the relations

$$\hat{A}_x^{1,s} = (1/4)(x^2 - y^2)C_{1,s}^{[1]}(z) - (1/48)(x^4 - y^4)C_{1,s}^{[3]}(z) + \cdots\ , \tag{16.3.6}$$

$$\hat{A}_y^{1,s} = (1/2)xyC_{1,s}^{[1]}(z) - (1/24)(x^3y + xy^3)C_{1,s}^{[3]}(z) + \cdots\ , \tag{16.3.7}$$

[6] In practice dipoles are often *bent* because the design orbit in a dipole is bent. In this case a cylindrical harmonic expansion is of limited use. See Subsection 3.7.

$$\hat{A}_z^{1,s} = -xC_{1,s}^{[0]}(z) + (1/8)(x^3+xy^2)C_{1,s}^{[2]}(z) - (1/192)(x^5+2x^3y^2+xy^4)C_{1,s}^{[4]}(z) + \cdots . \quad (16.3.8)$$

Recall (15.5.97) through (15.5.99). We observe that $\hat{\boldsymbol{A}}^{1,s}$, like $\boldsymbol{B}$, is also completely specified in terms of $C_{1,s}^{[0]}(z)$ and its derivatives. Vector potentials in other gauges are also possible, in particular the azimuthal-free gauge, and they too are completely specified in terms of $C_{1,s}^{[0]}(z)$ and its derivatives. Recall Sections 15.4 through 15.7.

We will next examine what can be said about the master function $C_{1,s}^{[0]}(z)$ in various cases.

16.3.2 Single Monopole Doublet

The simplest way to *mathematically* model a dipole field, as the name suggests, is to properly locate and assign strengths to two monopoles. Suppose two monopoles having strengths $\mp 4\pi g$ are placed at the (x, y, z) locations $(0, a, 0)$ and $(0, -a, 0)$. Note that these locations lie on a cylinder of radius a. A pair of monopoles of opposite sign is what we have called a *monopole doublet*. Recall Section 15.9. Note, however, that in this section we have *interchanged* the strengths of the two monopoles so as to produce an on-axis $\boldsymbol{B}$ field that points in the $+\boldsymbol{e}_y$ direction.

We will verify, as expected, that this monopole doublet produces an interior field whose cylindrical multipole expansion begins with an $m = 1$ term. From the symmetry of this array we expect that, in addition to the leading $m = 1$ term, there may also be $m = 3, 5, \cdots$ terms. Recall Subsection 15.3.5 for explicit results.

Define a *unit* monopole to be a monopole for which $g = 1$. Let $\chi(\boldsymbol{r}; \boldsymbol{r}')$ be the scalar potential for a unit monopole located at the point $\boldsymbol{r}'$,

$$\begin{aligned} \chi(\boldsymbol{r}; \boldsymbol{r}') &= \chi(x, y, z; x', y', z') = -1/||\boldsymbol{r} - \boldsymbol{r}'|| \\ &= -1/[(x-x')^2 + (y-y')^2 + (z-z')^2]^{1/2}. \end{aligned} \quad (16.3.9)$$

Then the potential ψ^{doub} for the monopole doublet we are considering here is given by the relation

$$\begin{aligned} \psi^{\text{doub}}(x, y, z) &= g\chi(x, y, z; 0, a, 0) - g\chi(x, y, z; 0, -a, 0) \\ &= g[x^2 + (y-a)^2 + z^2]^{-1/2} - g[x^2 + (y+a)^2 + z^2]^{-1/2}, \end{aligned} \quad (16.3.10)$$

in agreement with (15.9.1) through (15.9.3) save for a sign.

Next expand $\psi^{\text{doub}}(x, y, z)$ as a power series in x and y. So doing yields the result

$$\psi^{\text{doub}}(x, y, z) = 2gay/(z^2 + a^2)^{3/2} + \text{higher order terms}. \quad (16.3.11)$$

Comparison of (3.11) with (3.1) reveals that, for the monopole doublet we are considering here, the $m = 1$ on-axis gradient is given by the relation

$$C_{1,s}^{[0]}(z) = 2ga/(z^2 + a^2)^{3/2} = (4g/a)\delta(z, a) \quad (16.3.12)$$

where $\delta(z, a)$ is the approximating delta function already introduced in Exercise 2.19,

$$\delta(z, a) = (a^2/2)/(z^2 + a^2)^{3/2}. \quad (16.3.13)$$

The result (3.12) agrees, save for the expected sign difference, with (15.9.21) and (15.9.33) evaluated at $m = 1$.

Calculation shows that the approximating delta function (3.13) has the properties

$$\delta(-z, a) = \delta(z, a), \tag{16.3.14}$$

$$\delta(0, a) = 1/(2a), \tag{16.3.15}$$

$$\delta(z, a) = (a^2/2)/|z|^3 + O(1/|z|^5) \text{ as } |z| \to \infty, \tag{16.3.16}$$

$$\int_{-\infty}^{\infty} dz\ \delta(z, a) = 1. \tag{16.3.17}$$

Figures 3.1 and 3.2 illustrate the behavior of this approximating delta function for two different values of a. Evidently the approximating delta function (3.13) becomes the true delta function in the limit $a \to 0$,

$$\lim_{a\to 0} \delta(z, a) = \delta(z), \tag{16.3.18}$$

and a controls the fall-off rate. From (3.12) we see that the $m = 1$ on-axis gradient for a monopole doublet has the fall-off behavior

$$C_{1,s}^{[0]}(z) = 2ga/|z|^3 + O(1/|z|^5) \text{ as } |z| \to \infty. \tag{16.3.19}$$

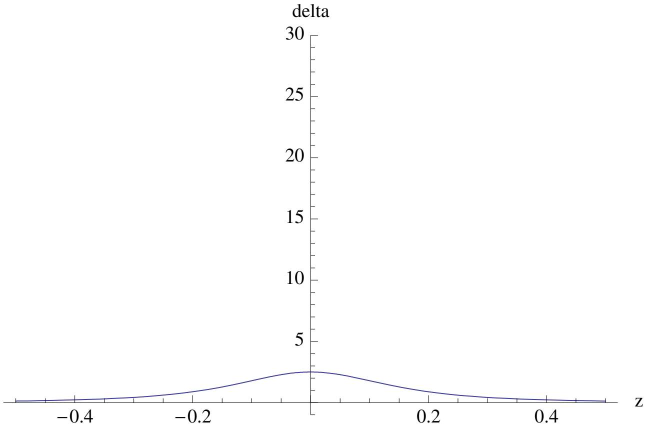


Figure 16.3.1: The approximating delta function (3.13) when $a = .2$.

16.3.3 Line of Monopole Doublets

Next consider, as a second mathematical model, the case of a *line* of monopole doublets extending from $z = 0$ to $z = L$. Associated with a line of monopole doublets will be a soft-edge bump function described by the relation

$$\text{bump}(z, a, L) = \int_0^L dz'\ \delta(z - z', a). \tag{16.3.20}$$

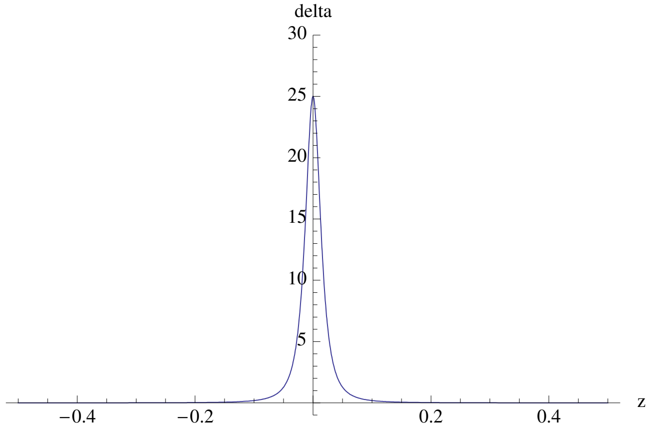


Figure 16.3.2: The approximating delta function (3.13) when $a = .02$.

In terms of this soft-edge bump function the $m = 1$ on-axis field gradient $C^{[0]}_{1,s}(z)$ for a line of monopole doublets can be written in the form

$$C^{[0]}_{1,s}(z) = B \operatorname{bump}(z, a, L) \tag{16.3.21}$$

where B is the dipole strength in the infinite length limit. Recall Exercise 1.5.7 and (3.2) through (3.5). It is given in terms of the doublet parameters by the relation

$$B = 4G/a \tag{16.3.22}$$

where $4\pi G$ is the monopole strength per unit length.

To evaluate (3.20), make the change of variables $\zeta = z - z'$ so that (3.20) becomes

$$\operatorname{bump}(z, a, L) = -\int_{z}^{z-L} d\zeta\ \delta(\zeta, a) = \int_{z-L}^{z} d\zeta\ \delta(\zeta, a). \tag{16.3.23}$$

The approximating delta function (3.13) has as an indefinite integral the result

$$\int d\zeta\ \delta(\zeta, a) = (1/2)\zeta/(\zeta^2 + a^2)^{1/2} = (1/2)\operatorname{sgn}(\zeta, a) \tag{16.3.24}$$

where $\operatorname{sgn}(\zeta, a)$ is the approximating signum function (2.26). It follows from (3.23) and (3.24) that

$$\operatorname{bump}(z, a, L) = [\operatorname{sgn}(z, a) - \operatorname{sgn}(z - L, a)]/2, \tag{16.3.25}$$

a result identical to (2.25). We conclude that the soft-edge bump function for the $m = 1$ on-axis gradient arising from a line of monopole doublets is the same as the soft-edge bump function for a simple air-core solenoid. In particular, the fall off for the $m = 1$ on-axis gradient arising from a line of monopole doublets goes as $1/|z|^3$, just as it does for a single monopole doublet.

16.3.4 Current Windings for two Air-Core Dipoles

The work of the previous Subsections 3.2 and 3.3 is deficient in at least two ways. First, according to Section 15.9.2, a monopole doublet has nonzero on-axis gradients for all odd values of m. See also Section 15.3.5. Moreover, even the integrated on-axis gradients for a monopole doublet are nonzero for all odd values of m,

$$\int_{-\infty}^{\infty} dz\, C_{m,s}^{[0]}(z) \neq 0 \text{ for } m \text{ odd.} \tag{16.3.26}$$

[See (15.9.21) and (15.9.33), and, for graphic examples, Figures 15.9.8, 15.9.11, 15.19.13, and 15.9.15.] The same is true for a line of monopole doublets.

Second, unlike the case of a solenoid, no prescription has been given of current windings/distributions that could be fabricated to produce the field of a line of multipole doublets or, better yet, the field of a reasonably "pure dipole". The aim of this subsection is to describe two commonly used (or contemplated) thin-shell windings on a straight circular cylinder of radius a and length L such that the magnetic field produced within the bore has *primarily* an $m = 1$ component.

Then, in the next two subsections, we will consider windings such that *only* the $m = 1$ on-axis gradient is nonzero for the field produced by such windings. We will call such a winding an *ideal* air-core dipole. However it is not the case, even for an ideal air-core dipole, that the field is that of a *perfect* dipole, a field, say, only in the $\boldsymbol{e}_y$ direction. According to (15.2.61) through (15.2.64) there are additional components in the fringe-field regions at the ends of the dipole where $C_{1,s}^{[0]}(z)$ is changing.

As already stated, we will consider air-core dipoles that consist of a thin-shell winding placed on a circular cylinder of radius $\rho = a$ and length L. We will also arrange the coordinate system so that the winding begins at $z = 0$ and ends at $z = L$.

There are two commonly used or considered approaches as to what the nature of this winding should be. One approach is to arrange to have most of the winding running on straight lines parallel to the cylinder axis to form what are called *saddle* coils. See Figure 3.3. Moreover, the spacing between successive straight lines is arranged so that the cross-sectional current density for the straight-line portion of the winding has (effectively on average) a $\cos(\theta)$ distribution. [Here suppose the coordinate system shown in Figure 2.1 is also employed in part c of Figure 3.3 above. Then, following customary nomenclature, θ is the angle ϕ defined by (15.2.12) through (15.2.14). From now on we will refer to a $\cos(\phi)$ distribution.] This can be accomplished by placing the winding in grooves machined into the underlying cylinder or by placing appropriate sized variable width spacers (not depicted) between successive straight wires.

It can be shown that for a $\cos(\phi)$ current distribution for the straight-line portion of the winding, and assuming the coils are long so that L is large, the field in the vicinity of $z = L/2$ is that of a reasonably pure dipole. See Exercise 3.5. That is, we already know from symmetry considerations, see Section 15.3.5, that only the $C_{m,s}^{[0]}(z)$ for odd m can be nonzero. And for a $\cos(\phi)$ current distribution for the straight-line portion of the winding it can be shown that $C_{1,s}^{[0]}(z)$ is substantial, and the remaining $C_{m,s}^{[0]}(z)$ are small, in the vicinity of $z = L/2$.

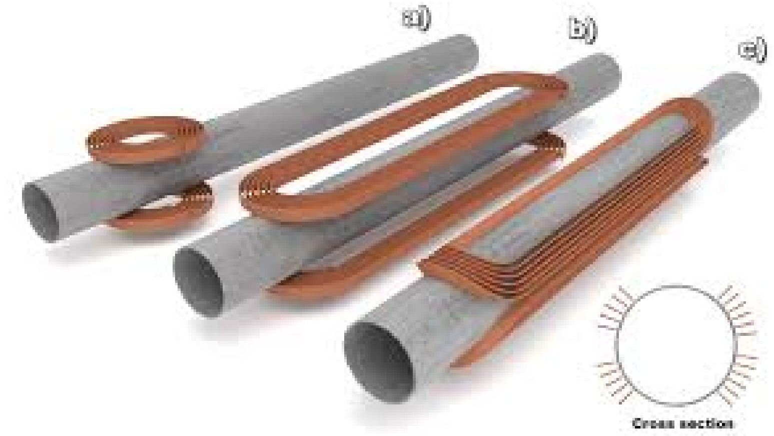


Figure 16.3.3: Coils and cylinders: Part c of this figure shows coils draped like saddles, above and below, over a circular cylinder. Apart from the coil ends, most of the winding runs along straight lines parallel to the cylinder axis.

What can be said about the field *away* from the central region $z \simeq L/2$? There the $C^{[0]}_{3,s}(z)$, $C^{[0]}_{5,s}(z)$, $\cdots$ can be substantial due to the currents flowing around the cylinder at the coil ends. Thus, an air-core dipole made with saddle $\cos(\phi)$ coils is not ideal as defined above. Moreover, in general there will be the integral results

$$\int_{-\infty}^{\infty} dz\, C^{[0]}_{m,s}(z) \neq 0 \text{ for } m \text{ odd}. \tag{16.3.27}$$

In particular the undesired $m = 3$, $m = 5$, $\cdots$ integrated multipole strengths will in general be nonzero.

But it is in principle possible to drive various undesired integrated multipole strengths to zero by placing and appropriately powering suitable multipole corrector windings at or near the ends of the main coil. Of course, when this is done, the net $C^{[0]}_{3,s}(z)$, $C^{[0]}_{5,s}(z)$, $\cdots$ remain nonzero. Only their integrals vanish.

A second approach, variously called "canted $\cos(\theta)$", "double helix", or "tilted solenoid", is based on a simple configuration where a conductor is wound around a cylinder as two oppositely *tilted* solenoids. Figure 3.4 illustrates such a winding.[7] See the Ph.D. thesis of *Brouwer* cited at the end of this chapter for an extensive treatment of canted $\cos(\theta)$ dipoles.

It can be shown that two tilted solenoids when powered as shown produce what is primarily a dipole field. Each layer produces a combination of vertical and solenoidal fields. If the currents are directed as shown, the vertical components add to produce a primarily dipole field and the solenoidal components cancel (save for end effects which are modest and also integrate to zero).[8] (And if the current in one of the layers is directed as shown and the other is reversed, the vertical components essentially cancel and the solenoidal components

[7] The use of the term "$\cos(\theta)$" in this context may seem slightly confusing since each tilted solenoid is wound *uniformly* with no variable width spacers between turns. However, as study of Figure 3.4 suggests, it can be shown that the net effect of the two tilted layers is to produce, in the overlap region and in the z direction, a $\cos(\phi)$ current distribution.

[8] Here we assume the layers have infinitesimal thickness so that they both have the same radii. If not,

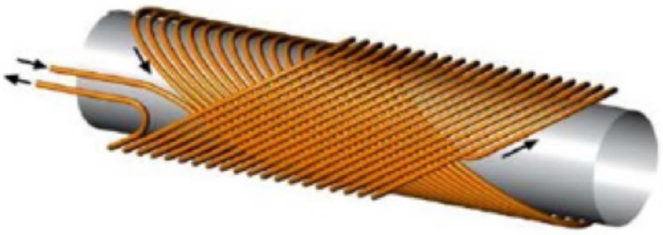


Figure 16.3.4: A winding composed of two oppositely tilted solenoids to form a canted $\cos(\theta)$ dipole.

add to produce a primarily solenoidal field.) Figure 3.5 shows, in Tesla, the solenoidal field component $B_z(0,0,z) = C_0^{[1]}(z)$ for the individual layers as well as the net $C_0^{[1]}(z)$ for a two-layer canted $\cos(\theta)$ dipole. Note that, due to the advertised cancellation, the net solenoidal field is small save for end effects. The solenoidal field at the center location $z = L/2$ is 0.003 Tesla and arises from the layers having slightly different radii.

With regard to the dipole and higher multipole fields, and again from symmetry considerations, only the $C_{m,s}^{[0]}(z)$ for odd m can be nonzero. And, if L is large, it can be shown that $C_{1,s}^{[0]}(z)$ is substantial, and the remaining $C_{m,s}^{[0]}(z)$ are small, in the vicinity of $z = L/2$. Figure 3.6 shows, for example, $C_{1,s}^{[0]}(z)$; and Figure 3.7 shows $C_{3,s}^{[0]}(z)$ and $C_{5,s}^{[0]}(z)$. But note that the $C_{3,s}^{[0]}(z)$, $C_{5,s}^{[0]}(z)$, $\cdots$ are *not* zero for all z, and therefore this air-core dipole is also not ideal.

Nevertheless, this air-core dipole does have a remarkable property: It can be shown that $\int_{-\infty}^{\infty} dz\, C_{1,s}^{[0]}(z)$ is substantial, and for all the remaining $C_{m,s}^{[0]}(z)$ there is the relation

$$\int_{-\infty}^{\infty} dz\, C_{m,s}^{[0]}(z) = 0. \tag{16.3.28}$$

That is, the *integrated* strengths of all undesired multipoles vanish for this air-core winding! Again see Figure 3.7.[9]

cancellation is not perfect so that there is a small residual solenoidal field even apart from end effects. In any event the effect of solenoidal fields can presumably be compensated if desired by the addition of other windings or the use of skew quadrupoles. For example, the x, y coupling effect of the strong solenoidal fields associated with some detectors in storage rings/colliders is routinely compensated by the use of skew quadrupoles.

[9] Windings (for air-core magnets) which have the property that all integrated multipoles vanish save for some desired multipole are sometimes called *Lambertson* windings. See the references at the end of this chapter.

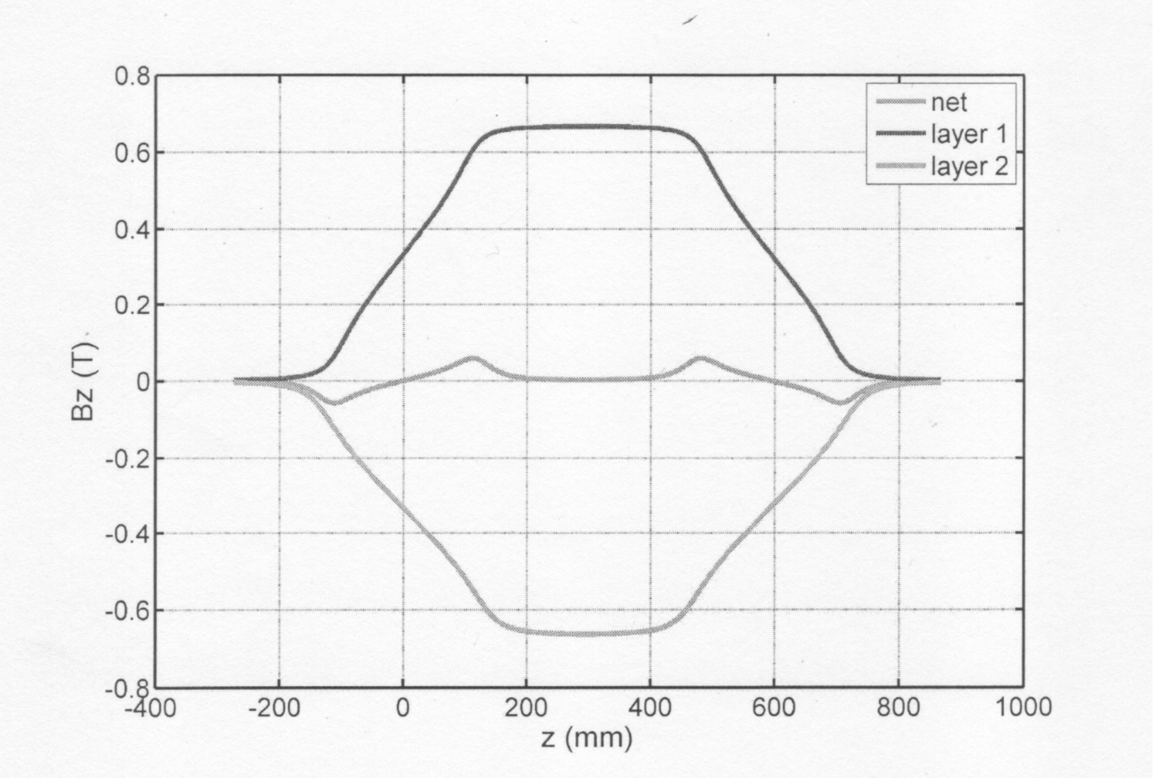


Figure 16.3.5: The individual and net $C_0^{[1]}(z)$ for a canted $\cos(\theta)$ dipole.

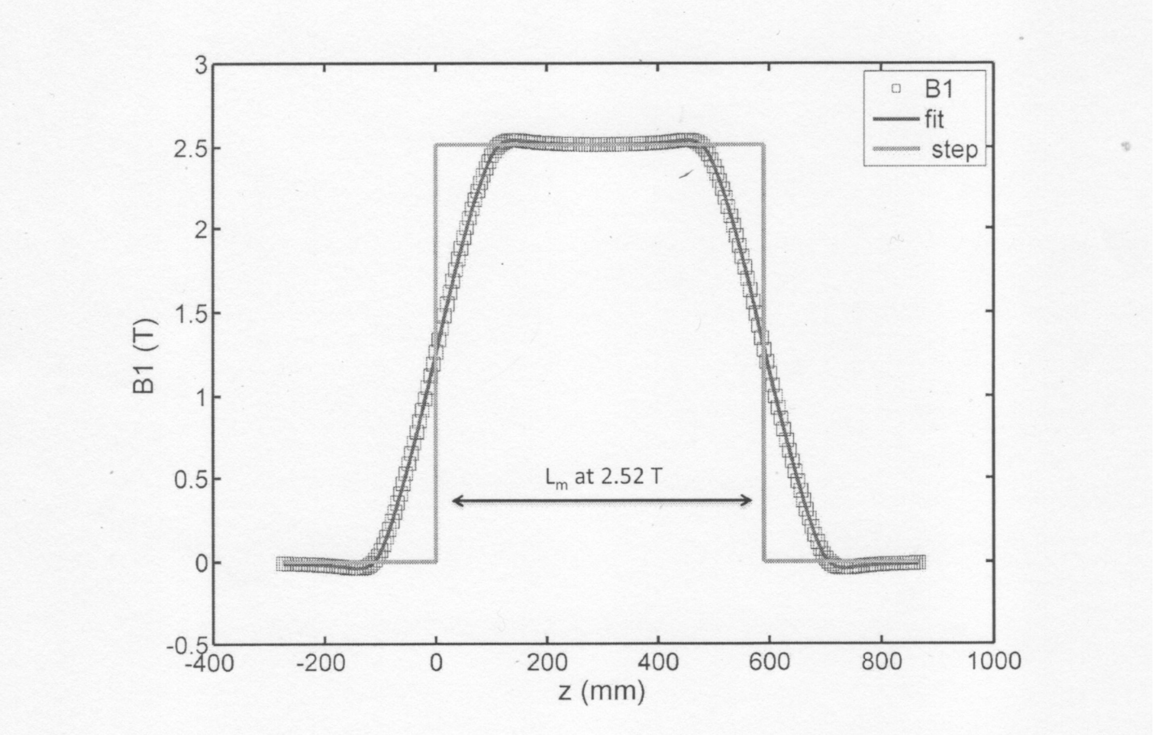


Figure 16.3.6: The on-axis gradient $C_{1,s}^{[0]}(z)$ in Tesla for a canted $\cos(\theta)$ dipole. Also shown is a hard-edge bump function approximation.

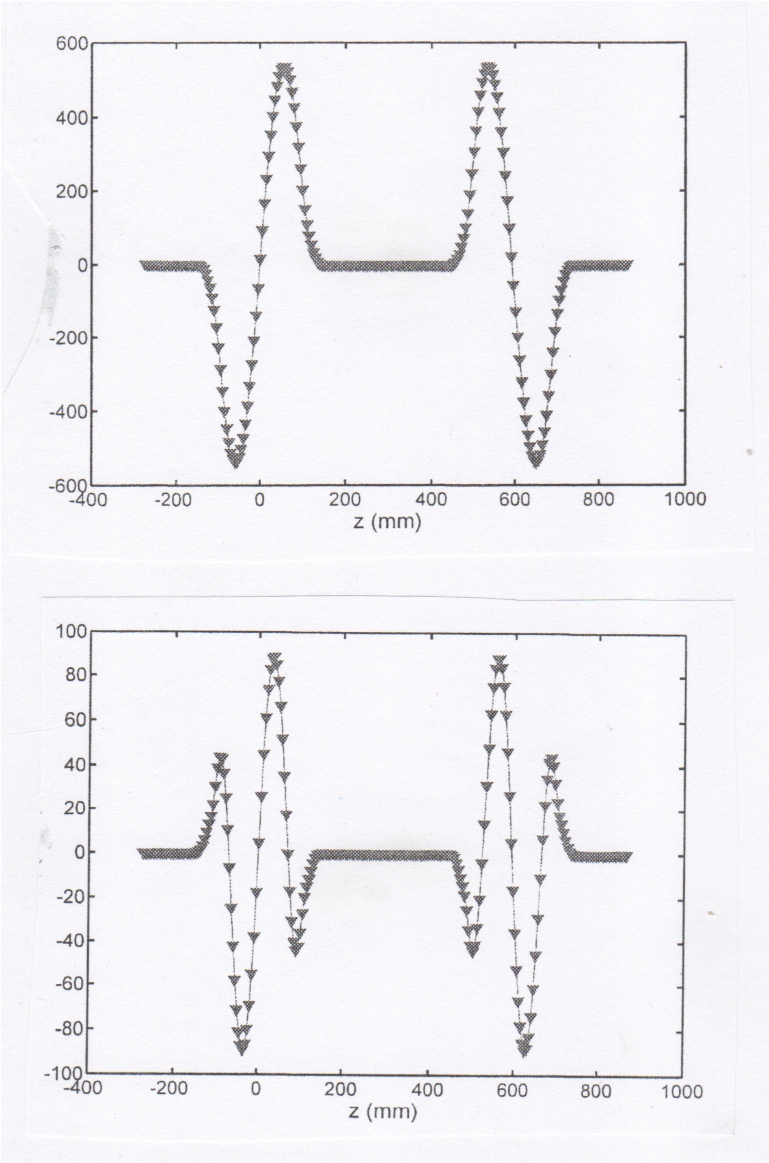


Figure 16.3.7: The on-axis gradient $C_{3,s}^{[0]}(z)$ (above) and on-axis gradient $C_{5,s}^{[0]}(z)$ (below) in dimensionless units for a canted $\cos(\theta)$ dipole. They are small for $z \simeq L/2$, but do not vanish everywhere. Nevertheless their integrated strengths do vanish.

16.3.5 Current Winding for an Ideal Air-Core Dipole

We have seen two examples of windings that fail to produce ideal air-core dipoles. Is there a winding that succeeds? Yes, there are several. This subsection will describe one that is easy to visualize and for which it is possible to compute $C_{1,s}^{[0]}(z)$ analytically.

Figure 3.8 shows a net of coils placed on a circular cylinder. Figure 3.9 shows a top view of the right end of the cylinder and illustrates the sign convention for describing the current flow in the right ($+z$) end of each coil and the currents as they flow along the long sides of each coil. A dot denotes the tip of an arrow as it comes up from below the plane of the figure, and a cross denotes the feather of an arrow as it goes down below the plane of the figure. In this figure $n = 12$ coils are displayed with the circular arc of each I_k current subtending an angle of $\Delta = 2\pi/n = 30°$. We also define angles ϕ_k by the rule

$$\phi_k = k\Delta. \tag{16.3.29}$$

Thus, at each angular location ϕ_k there is an upward and downward current pair so that, for example, there is a *net* current $(I_1 - I_n)$ along the side of the cylinder in the z direction at the location $\phi_0 = 0$.

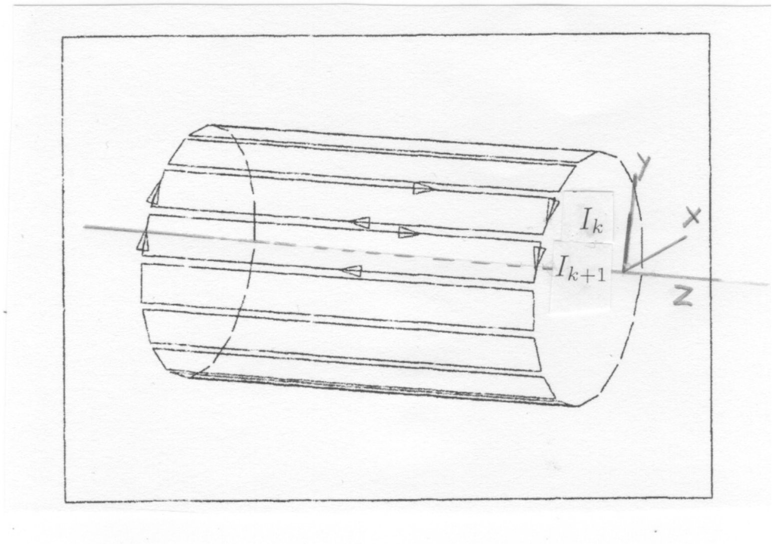


Figure 16.3.8: A net of coils draped over a cylinder. The k^{th} coil carries a current I_k.

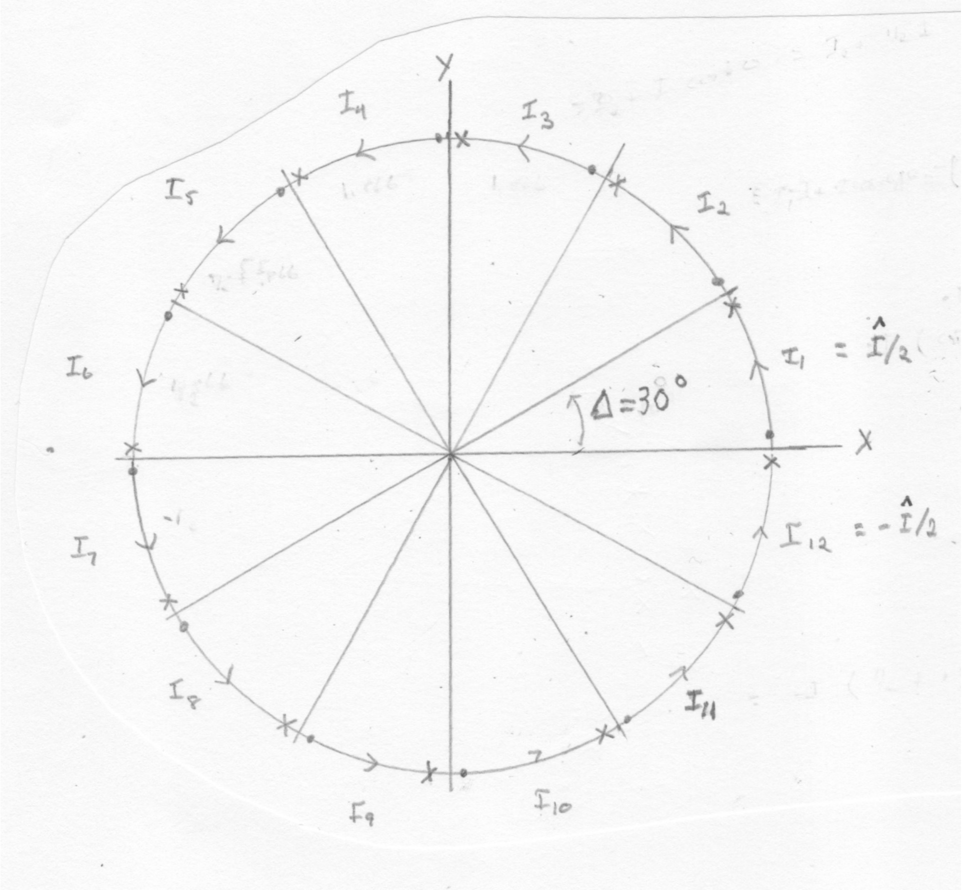


Figure 16.3.9: Top view of the right ends of the coils shown in Figure 3.8. The z axis comes out of the plane of the paper.

In order to achieve an effective $\cos\phi$ current distribution along the *length* of the cylinder, the currents in adjacent coils are required to be related by the rules

$$\begin{aligned}
&I_2 - I_1 = \hat{I}\cos\phi_1,\\
&I_3 - I_2 = \hat{I}\cos\phi_2,\\
&I_{k+1} - I_k = \hat{I}\cos\phi_k,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\ ,
\end{aligned}\qquad (16.3.30)$$

$$\begin{aligned}
&I_n - I_{n-1} = \hat{I}\cos\phi_{n-1},\\
&I_1 - I_n = \hat{I}\cos\phi_0 = \hat{I},
\end{aligned}\qquad (16.3.31)$$

where $\hat{I}$ is some amount of current yet to be determined. As a sanity check, we observe that the sum of the left sides of (3.30) through (3.31) vanishes. And for the right sides computation shows that the sum also vanishes, as desired,

$$\begin{aligned}
&\hat{I}[1+\cos\phi_1+\cos\phi_2+\cdots+\cos\phi_{n-1}] = \hat{I}\Re\{\sum_{k=0}^{n-1}\exp(ik\Delta)\}\\
&= \hat{I}\Re\{[1-\exp(in\Delta)]/[1-\exp(i\Delta)]\}\\
&= \hat{I}\Re\{[1-\exp(2\pi i)]/[1-\exp(i2\pi/n)]\} = 0. \qquad (16.3.32)
\end{aligned}$$

Observe that the left sides of (3.30) through (3.31) only involve current differences, and therefore the currents themselves still need to be determined. This can be done by specifying that

$$I_1 = \hat{I}/2. \qquad (16.3.33)$$

Note that this specification, when employed with the last equation in (3.31), produces the pleasant result

$$I_n = -\hat{I}/2. \qquad (16.3.34)$$

It can be shown that (3.30) through (3.31) and (3.33), when taken together, yield the relations

$$I_k = (\hat{I}/2)[\sin(\Delta/2)]^{-1}\sin\psi_k \qquad (16.3.35)$$

where

$$\psi_k = (k-1/2)\Delta. \qquad (16.3.36)$$

See Exercise *. Note that ψ_k is the angular location of the *midpoint* of the arc that carries the current I_k. For example, $\psi_1 = \Delta/2$.

Consider cases for which n has a value of the form $n = 4\ell$ where ℓ is an integer. Then, in the *continuum* large ℓ limit (and with $\hat{I}$ adjusted accordingly), it can be shown that the collection of coils of the kind shown schematically in Figure 3.8 and more precisely in Figure 3.9 (for the case $\ell = 3$) with the I_k described by (3.35) and (3.36) produces an ideal dipole field. That is, $C_{1,s}^{[0]}(z)$ is substantial, and the remaining $C_{m,s}^{[0]}(z)$ *all vainsh*. See Exercise *.

Moreover, in this case $C_{1,s}^{[0]}(z)$ can be computed analytically. (See the work of *Bassetti* and *Biscari* cited in the references at the end of this chapter.) If the winding is on a cylinder of radius a that begins at $z = 0$ and extends to $z = L$, then the on-axis gradient for such an air-core dipole, as is the case for the on-axis field for a simple air-core solenoid and the on-axis gradient for a line of monopole doublets, can be described in terms of a soft-edge bump function which we will again call bump(z, a, L). That is, the on-axis gradient $C_{1,s}^{[0]}(z)$ can be written in the form

$$C_{1,s}^{[0]}(z) = B\ \text{bump}(z, a, L) \qquad (16.3.37)$$

where B is again the dipole strength in the infinite length limit.

Like the soft-edge bump function for a simple air-core solenoid and a line of monopole doublets, the soft-edge bump function for this ideal air-core dipole can be written in terms of an associated approximating signum function in the form

$$\text{bump}(z, a, L) = [\text{sgn}(z, a) - \text{sgn}(z - L, a)]/2. \qquad (16.3.38)$$

It can be shown that for this ideal air-core dipole the approximating signum function sgn(z, a) is given by the relation

$$\mathrm{sgn}(z,a) = z(z^2 + 2a^2)/(z^2 + a^2)^{3/2}. \tag{16.3.39}$$

Figures 3.10 and 3.11 illustrate the behavior of this approximating signum function for two different values of a. Evidently the approximating signum function (3.39) becomes the true signum function in the limit $a \to 0$.

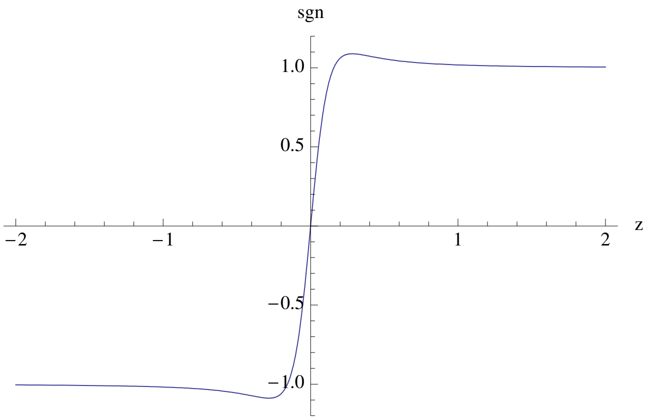


Figure 16.3.10: The approximating signum function (3.39) when $a = .2$.

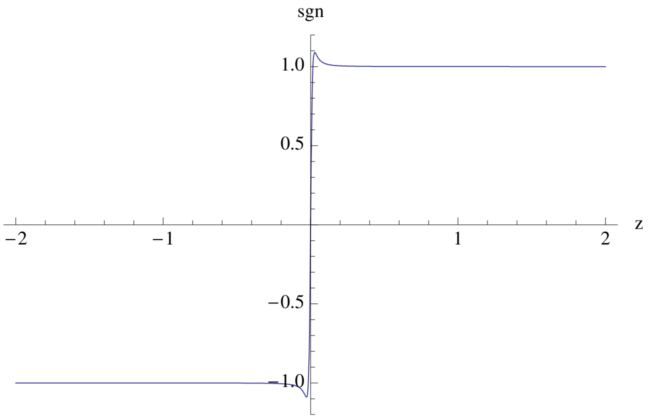


Figure 16.3.11: The approximating signum function (3.39) when $a = .02$.

It follows from (3.38) and (3.39) that, like the soft-edge bump function for a simple air-core solenoid and a line of monopole doublets, the soft-edge bump function for this ideal air-core dipole satisfies the relations (2.20 through (2.22). Recall Exercise 2.3. We see from (2.23) and (3.37) that for this ideal air-core dipole there is the relation

$$\int_{-\infty}^{\infty} C_{1,s}^{[0]}(z)dz = BL. \tag{16.3.40}$$

Figures 3.12 and 3.13 illustrate the properties (2.20) through (2.22) for a fixed value of L and two different values of the radius a. Evidently the ideal air-core dipole soft-edge bump function given by (3.38) and (3.39) becomes a hard-edge bump function in the limit $a \to 0$. The radius a plays the role of a characteristic length that controls the rate of fall off. The fringe-field region is large if a is large, and vanishes as a goes to zero.

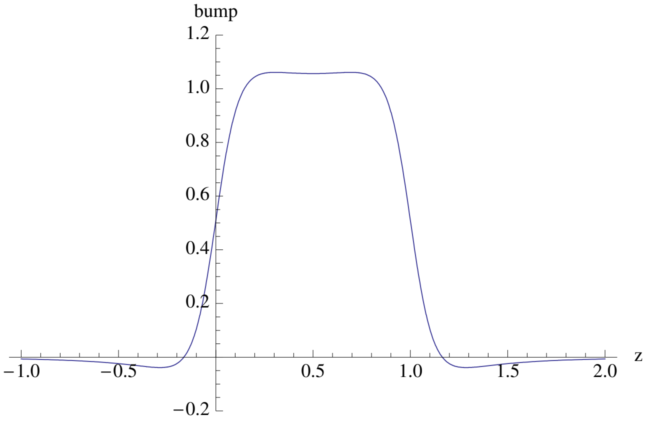


Figure 16.3.12: The soft-edge bump function bump(z, a, L) given by (3.38) and (3.39) when $a = 0.2$ and $L = 1$.

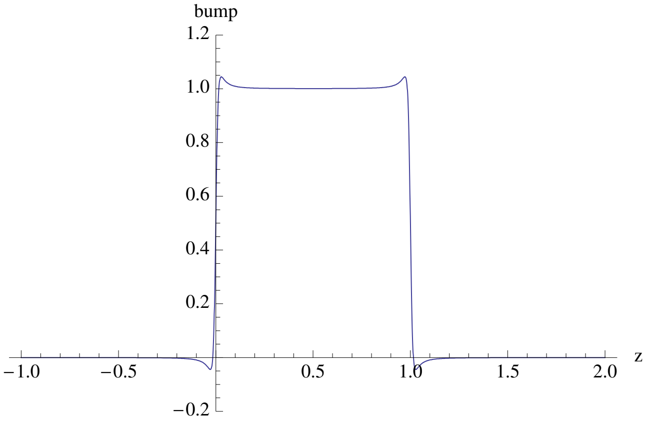


Figure 16.3.13: The soft-edge bump function bump(z, a, L) given by (3.38) and (3.39) when $a = 0.02$ and $L = 1$.

From (3.39) and (3.38) we find the asymptotic behaviors

$$\mathrm{sgn}(z, a) = 1 + (1/2)a^2/z^2 + O(1/z^4) \text{ as } z \to +\infty, \tag{16.3.41}$$

$$\mathrm{bump}(z, a, L) = -(1/2)La^2/|z|^3 + O(1/|z|^4) \text{ as } |z| \to \infty. \tag{16.3.42}$$

Consequently $C^{[0]}_{1,s}(z)$ falls off for large distances as

$$C^{[0]}_{1,s}(z) = -(1/2)BLa^2/|z|^3 + O(1/|z|^4) \text{ as } |z| \to \infty. \tag{16.3.43}$$

We see that the fall off goes as $1/|z|^3$, just as it does for the simple air-core solenoid, a single monopole doublet, and a line of monopole doublets. Compare (2.42) and (3.43). Note, however, that the approximating signum functions (2.26) and (3.39) are different. Compare, for example, Figures 2.2 and 3.10. Also, compare (2.34) with (3.41). Correspondingly, the bump functions for a line of monopole doublets and this idealized air-core dipole are different, and the relations (2.42) and (3.43) differ in sign. Finally we observe, for example from Figures 3.12 and 3.13, that for an ideal air-core dipole the bump function bump(z, a, L), while positive in the center of the dipole, can be *negative* outside the dipole. This change in sign *cannot* be modeled by an *Enge* function profile.[10]

Our discussion of dipole fringe fields so far has treated the iron-free case, and we have found a $1/|z|^3$ fall off in all cases susceptible to easy analysis. When iron is present, and the coils are buried in iron or field clamps are employed, the fall off can in principle be much faster including the possibility of essentially exponential fall off.

16.3.6 Current Windings for other Ideal Air-Core Dipoles

16.3.7 Limited Utility of Cylindrical Harmonic Expansions for Dipoles

Strictly speaking, and as already alluded to in a previous footnote, cylindrical harmonic expansions for the field of a dipole are of limited use. First, there is this observation: If it is desired that the bore be much smaller than the length of the dipole, as is frequently the case, then the dipole must be bent to accommodate the design orbit. In this case a cylindrical harmonic analysis of the field is no longer possible.[11] Second, cylindrical harmonic expansions are expected to be valid (rapidly convergent) only in the vicinity of the z axis. But the orbit in a dipole is bent and therefore cannot be confined to the vicinity of the z axis unless the bend angle is suitably small. Thus there is a conflict between the desire to have a simple model of the design orbit accompanied by a practical dipole design (an essentially circular arc reasonably closely surrounded by coil windings/iron), and the desire for a simple model (cylindrical harmonic expansion) of the dipole field. This conflict occurs both for iron-free dipoles and dipoles with iron. Chapter 22, which does not presuppose a cylindrical harmonic expansion, treats the problem of finding realistic transfer maps for curved beam-line elements with significant sagitta.

At this point we pause to note that there is one area where this conflict does not occur, of at least may be less significant: the modeling of a wiggler/undulator which may be viewed as a string of short dipoles and for which the design orbit throughout the length of the element does not differ much from a straight line. In this case a cylindrical harmonic analysis of the field is appropriate and useful providing the amplitude of the wiggles in the design orbit is modest compared to the half gap of the dipoles. See Section 4 of this chapter. Finally, wigglers/undulators may be treated using the methods of Chapter 22 without the use of

[10] Note that for a canted $\cos(\theta)$ dipole, see Figure 3.6, the field can also be negative outside the dipole.

[11] Often, in the case of large storage rings/colliders, long superconducting dipoles with small bores are initially built as straight rectangular magnets with windings (essentially $\cos\phi$) and iron designed in such a way as to produce as far as practical a pure $m = 1$ field. These magnets are then *mechanically* bent to accommodate a curved design orbit, thereby producing a sector bend with normal entry and exit. The hope, partially verified by experience, is that the transfer map for such a bent dipole will not have unacceptable nonlinearities.

cylindrical harmonic expansions.

Let us return to the main discussion: instances where the conflict must be addressed. Consider the case of a straight (unbent) rectangular dipole of length L with equal entry and exit angles. Make the approximation that the design orbit is a circular arc within the dipole and straight lines outside the dipole in the fringe-field regions. Assume the bend angle is θ. Then, by simple geometry, in order to *just accommodate* the design trajectory the dipole (without bending) must have a bore radius a given by the relation

$$a/L = (1 - \cos\theta/2)/(4\sin\theta/2) = \theta/(16) + O(\theta^3). \tag{16.3.44}$$

See Exercise *. Figure 3.14 displays the ratio a/L as a function of θ.[12] Suppose, for example, that $\theta = 9$ degrees and $L = 1$ meter. Then use of (3.44) gives the result $a/L = .0098$ and therefore $a = .98$ cm.

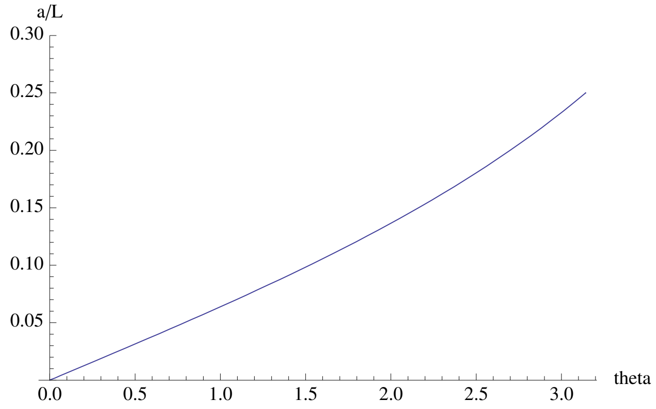


Figure 16.3.14: The ratio a/L as a function of θ.

Next, beyond the simple geometric considerations we have just explored, we must acknowledge that a cylindrical harmonic expansion is expected to be valid only in the vicinity of the z axis. Suppose we assume that the actual bore should be at least twice the geometrically needed bore in order for the cylindrical harmonic expansion to be reasonably accurate in the region traversed by the design orbit. In this case we should, say for convenience, make the Ansatz $a = 2$ cm $= .02$ meters. When $a = .02$ and $L = 1$ the field profile is that of Figure 3.13, from which we see that the fringe-field region appears to be relatively small. Specifically, from (3.38) and (3.39), we find the result

$$\text{bump}(z = -.3, a = .02, L = 1) = \text{bump}(z = 1.3, a = .02, L = 1) = -1.04 \times 10^{-3}. \tag{16.3.45}$$

Consequently, in this case one must be a distance of about 30 cm from the ends of the dipole for the on-axis field value to fall to 10^{-3} of its central on-axis value. Thus, in this case and with a 10^{-3} fall-off criterion, the fringe-field region on either end of the dipole is about 1/3 the length of the dipole.

Evidently, a more detailed analysis of this case would involve numerical integration to determine the design orbit accurately. And computation of the transfer map about this

[12] Note that, for an unbent dipole with fixed design-orbit bend angle, the bore is proportional to the length.

design orbit would require integration of the map equations of Section 10.5 using a Hamiltonian based on the vector potential given by (3.6) through (3.8). Note that in this case, according to (3.5) and (3.37) there are the relations

$$C_{1,s}^{[1]}(z) = B \text{ bump}'(z, a, L), \tag{16.3.46}$$

$$C_{1,s}^{[2]}(z) = B \text{ bump}''(z, a, L), \text{ etc.} \tag{16.3.47}$$

These functions are shown in Figures 3.15 and 3.16 below. Like its counterpart shown in Figure 2.7 for a solenoid of the same geometry, the function bump$''$ for the ideal air-core dipole of Section 3.5 is quite singular in the case $a = 0.02$ and $L = 1$. We may therefore expect that the transfer map for this ideal air-core dipole will have substantial higher-order aberrations.

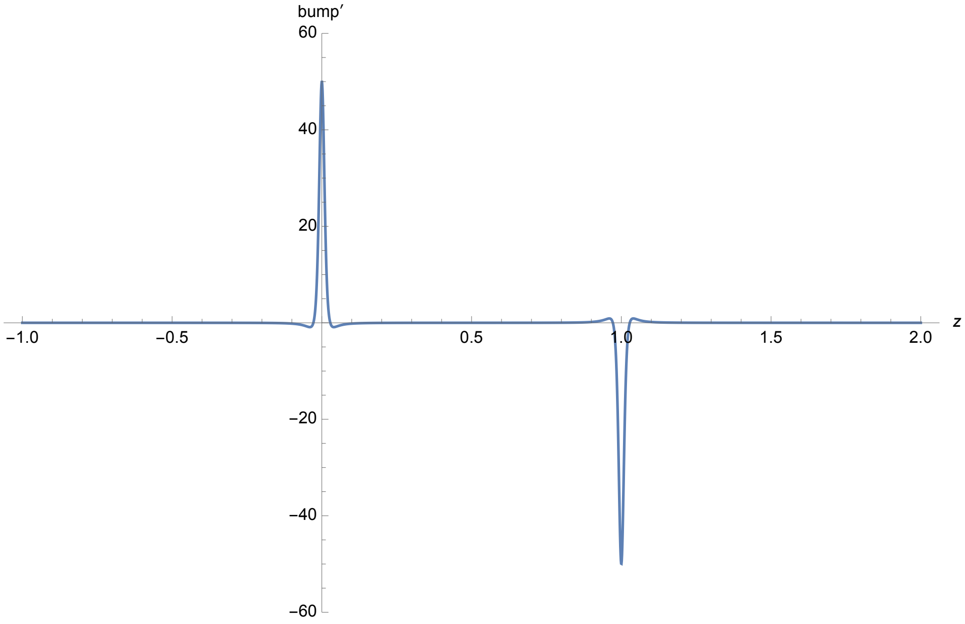


Figure 16.3.15: The function bump$'(z, a = 0.02, L = 1)$ associated with (3.38).

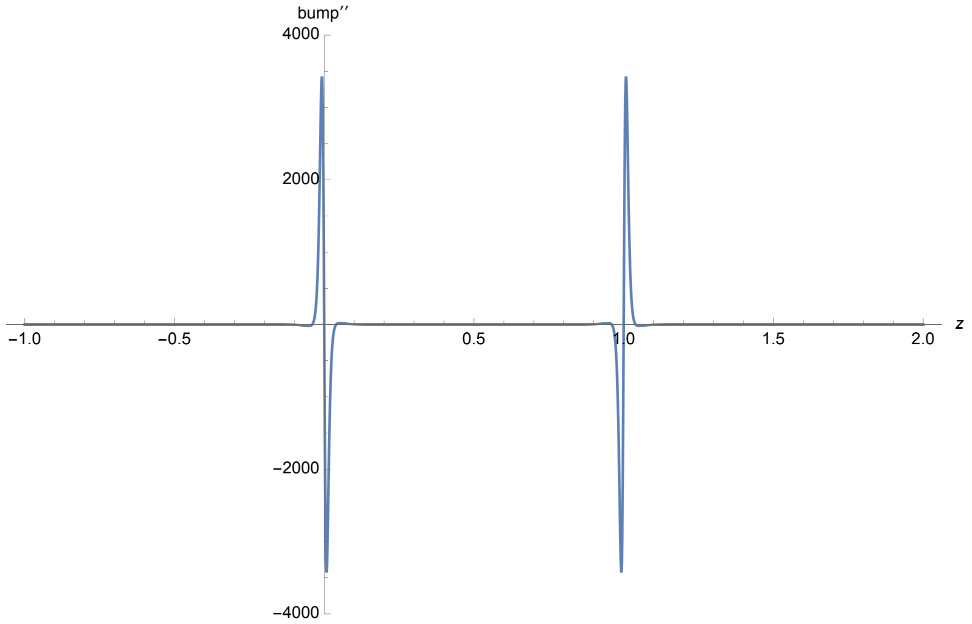


Figure 16.3.16: The function bump$''(z, a = 0.02, L = 1)$ associated with (3.38).

16.3.8 Terminating Dipole End Fields

As already described, cylindrical harmonic expansions are of limited use for dipoles because the design orbit in a dipole is bent. Correspondingly, it is generally not useful to describe the termination of dipole end fields in terms of cylindrical harmonic expansions. For the special case of wigglers/undulators, where the use of cylindrical harmonic expansions may be appropriate for describing the termination of end fields, see Subsection 4.3. For a treatment of dipole end-field termination in the general case without the use of cylindrical harmonic expansions, see Section 22.8.

16.3.9 Limited Utility of Hard-Edge Models for Dipole Fringe Fields

Exercises

16.3.1. Verify (3.11) and (3.12). Verify that (3.12) also follows from (15.8.21) and (15.8.33).

16.3.2. Verify (3.14) through (3.17).

16.3.3. Verify (3.20) through (3.22).

16.3.4. Verify (3.24) by shoing that

$$\partial_\zeta \mathrm{sgn}(\zeta, a) = 2\delta(\zeta, a). \tag{16.3.48}$$

Verify (3.25).

16.3.5. Material re $\cos(\phi)$ current distribution.

16.3.6. Verify $\cdots$.

16.3.7. Verify the relation (3.33).

16.4 Air-Core Wiggler/Undulator Models

16.4.1 Simple Air-Core Wiggler/Undulator Model

The fields of individual monopole doublets or lines of monopole doublets or idealized air-core dipoles may be used to create model fields for wigglers/undulators. We will consider the simplest case where individual monopole doublet fields are employed.

A possible simple three-pole model of a wiggler/undulator may be taken to be a string of three equally spaced monopole doublets having relative strengths $+1/2, -1, +1/2$. That is we may define a three-pole wiggler/undulator profile function $\text{wig}(3, z, a, L)$ by the rule

$$\text{wig}(3, z, a, L) = (1/2)\delta(z + L, a) - \delta(z, a) + (1/2)\delta(z - L, a) \tag{16.4.1}$$

where $2L$ is the wiggler/undulator period. Here $\delta(z, a)$ is the approximating delta function given by (3.13). The sum of the pole strengths is zero so that the wiggler/undulator produces no net bending, and the end poles are given half strengths so that the wiggler/undulator produces no net translation in x. For this profile there is the asymptotic fall off

$$\text{wig}(3, z, a, L) = (3a^2L^2)/|z|^5 + O(1/|z|^6) \text{ as } |z| \to \infty. \tag{16.4.2}$$

Figure 4.1 displays the profile function $\text{wig}(3, z, a, L)$ for the case $a = .1$ and $L = .5$. Evidently in this case, as expected from (4.2), the fringe field falls off quite rapidly. For example, at a distance of one wiggler/undulator period from the end, there is the result

$$\text{wig}(3, 1.5, .1, .5)/\text{wig}(3, 0, .1, .5) = -2.6 \times 10^{-4}. \tag{16.4.3}$$

At this point $C_{1,s}^{[0]}$ has fallen from its peak value by almost four orders of magnitude.

Another simple model is a string of five equally spaced monopole doublets having relative strengths $+1/2, -1, +1, -1, +1/2$. In this model we define a five-pole wiggler/undulator profile function $\text{wig}(5, z, a, L)$ by the rule

$$\text{wig}(5, z, a, L) = (1/2)\delta(z + 2L, a) - \delta(z + L, a) + \delta(z, a) - \delta(z - L, a) + (1/2)\delta(z - 2L, a). \tag{16.4.4}$$

For this profile there is the asymptotic fall off

$$\text{wig}(5, z, a, L) = (6a^2L^2)/|z|^5 + O(1/|z|^6) \text{ as } |z| \to \infty. \tag{16.4.5}$$

Note that the fall off for both the three-pole and five-pole wiggler/undulator goes as $1/|z|^5$, which is two orders higher in $1/|z|$ than that for a single monopole doublet. Compare (3.19), (4.2), and (4.5). This higher fall-off rate arises from cancellations that occur between the doublets because the sum of the pole strengths is zero. That is, we have enforced the relation

$$\int_{-\infty}^{\infty} dz \, \text{wig}(n, z, a, L) = 0. \tag{16.4.6}$$

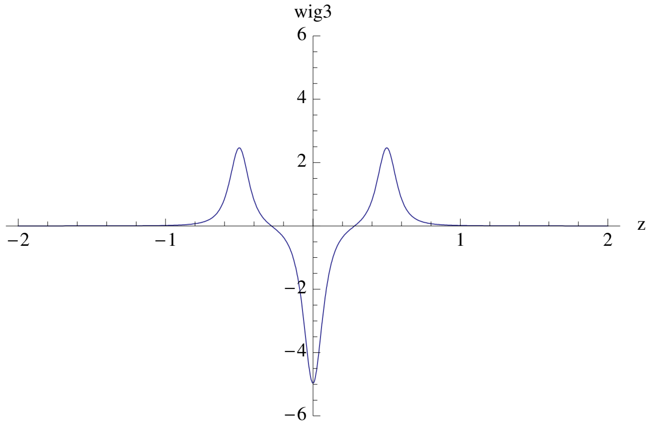


Figure 16.4.1: The three-pole wiggler/undulator profile function (4.1) when $a = 0.1$ and $L = 0.5$.

Figure 4.2 displays the five-pole profile function for the case $a = .1$ and $L = .5$. Evidently the fringe field again falls off quite rapidly. For example, at a distance of one wiggler/undulator period from the end, there is the result

$$\mathrm{wig}(5, 2, .1, .5)/\mathrm{wig}(5, 0, .1, .5) = 2.8 \times 10^{-4}. \tag{16.4.7}$$

At this point $C^{[0]}_{1,s}$ has again fallen from its peak value by almost four orders of magnitude.

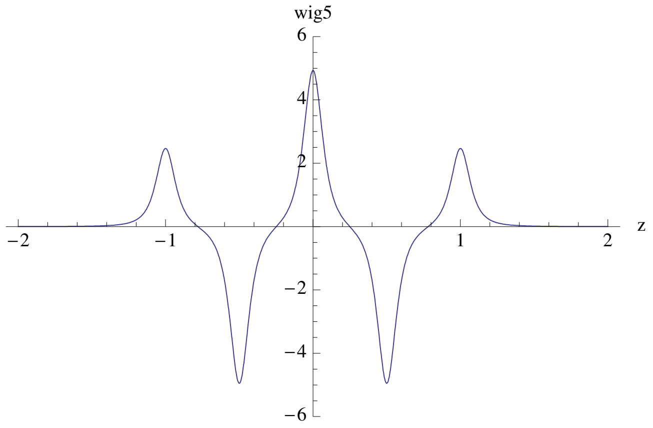


Figure 16.4.2: The five-pole wiggler/undulator profile function (4.4) when $a = 0.1$ and $L = 0.5$.

16.4.2 Iron-Free Rare Earth Cobalt (REC) Wiggler/Undulator

16.4.3 Terminating Wiggler/Undulator End Fields

Preliminaries

There is an application for which expansions employing $m = 1$ cylindrical harmonics may be useful, namely the case of wigglers/undulators when the excursion of the design orbit from the axis may be treated as small. That is, it is assumed that the design orbit enters and exits the wiggler/undulator on axis and nearly along the axis, and the excursions of the design orbit from the axis while within the wiggler/undulator may be treated as small.

At this point it is necessary to make a change of notation. In the discussion of most of the preceding sections and subsections, save for that of Sections 15.2, 15.7 and 15.9, we have used the coordinates x, y, z as global coordinates. For this subsection, as described in Section 15,2, we will use

$$\boldsymbol{R} = (X, Y, Z) \tag{16.4.8}$$

as global coordinates and

$$\boldsymbol{r} = (x, y, z) \tag{16.4.9}$$

as local coordinates. In analogy with the notation of Subsection 1.2, we assume the leading wiggler/undulator fringe field begins at $Z = Z^{\rm en}$ and the trailing wiggler/undulator fringe field ends at $Z = Z^{\rm ex}$. In the interval $[Z^{\rm en}, Z^{\rm ex}]$ the design orbit and map integrations will be carried out using the vector potential in the Coulomb gauge of Section 15.5. The entering transition at $Z^{\rm en}$ from the leading no-field region to the leading fringe-field region, and the exiting transition at $Z^{\rm ex}$ from the trailing fringe-field region to the trailing no-field region, will be made using the minimum vector potential, namely the Poincaré-Coulomb gauge vector potential.

Entering a Leading Fringe-Field Region

If we wish to make the transition from the leading no-field region to the leading fringe-field region using the minimum vector potential (the vector potential in the Poincaré-Coulomb gauge), and also wish to carry out the design orbit and map integrations using the vector potential in the Coulomb gauge of Section 15.5.2, then we need to find at $Z = Z^{\rm en}$ the gauge transformation that relates the Poincaré-Coulomb gauge and the Coulomb gauge of Section 15.5.2. The general problem of changing gauges has already been discussed in Subsection 1.3, the general relation between the Poincaré-Coulomb gauge vector potential and the Coulomb gauge vector potential of Section 15.5 has been described in Section 15.7, and the specific $m = 1$ and $\alpha = s$ relation has been treated in Section 15.9.2.

We seek relations in the vicinity of the point $(0, 0, Z^{\rm en})$. Let us recapitulate some of what we have learned. In the vicinity of this point, and for the $\alpha = s$ component of the magnetic field, the Poincaré-Coulomb gauge vector potential and the Coulomb gauge vector potential of Section 15.5 are related by the gauge transformation

$$^{P}\boldsymbol{A}^{1,s}(x, y, z; Z^{\rm en}) = \hat{\boldsymbol{A}}^{1,s}(x, y, Z^{\rm en} + z) + \nabla\chi_{1,s}. \tag{16.4.10}$$

Recall (15.7.5). Also, we found that the gauge term $\chi_{1,s}$ is related to the $m = 1$ and $\alpha = s$ on-axis gradient by the expansion

$$\chi_{1,s}(x, y, z; Z^{\rm en}) = \{[(1/2)xz]C^{[0]}_{1,s}(Z^{\rm en})\} + \{[(1/3)xz^2 - (1/12)x\rho^2]C^{[1]}_{1,s}(Z^{\rm en})\} + \cdots\ . \tag{16.4.11}$$

See (15.9.21).

Next, see (1.75), we recall the relation

$$\boldsymbol{A}^a - \boldsymbol{A}^b = \nabla\chi, \tag{16.4.12}$$

and compare it with the relation (4.10) rewritten in the form

$$\hat{\boldsymbol{A}}^{1,s}(x,y,Z^{\rm en}+z)\ -\ {}^{P}\!\boldsymbol{A}^{1,s}(x,y,z;Z^{\rm en}) = -\nabla\chi_{1,s}. \tag{16.4.13}$$

We conclude that if we wish to identify $\hat{\boldsymbol{A}}^{1,s}$ with $\boldsymbol{A}^{a}$, and identify ${}^{P}\!\boldsymbol{A}^{1,s}$ with $\boldsymbol{A}^{b}$, then we should require the relation

$$\chi = -\chi_{1,s}. \tag{16.4.14}$$

Finally, define the function $\chi^{\rm en}$ by the rule

$$\chi^{\rm en}(x,y;Z^{\rm en}) = \chi(x,y,0;Z^{\rm en}). \tag{16.4.15}$$

With this definition we see from (4.11) and (4.14) that there is the result

$$\chi^{\rm en}(x,y;Z^{\rm en}) = \{[(1/12)x\rho^2]C^{[1]}_{1,s}(Z^{\rm en})\} - \cdots . \tag{16.4.16}$$

We are now ready to invoke the results (1.77) through (1.79). So doing, we find that the canonical coordinates $(x,y,t;p_x^{\rm can},p_y^{\rm can},p_t^{\rm can})$ *after* and *before* $Z^{\rm en}$ are connected by the symplectic map $\mathcal{T}^{\rm en}$,

$$\begin{aligned}
x^a(Z) &= \mathcal{T}^{\rm en}x^b(Z) \text{ with } Z = Z^{\rm en},\\
y^a(Z) &= \mathcal{T}^{\rm en}y^b(Z) \text{ with } Z = Z^{\rm en},\\
t^a(Z) &= \mathcal{T}^{\rm en}t^b(Z) \text{ with } Z = Z^{\rm en};
\end{aligned} \tag{16.4.17}$$

$$\begin{aligned}
p_x^{\rm cana}(Z) &= \mathcal{T}^{\rm en}p_x^{\rm canb}(Z) \text{ with } Z = Z^{\rm en},\\
p_y^{\rm cana}(Z) &= \mathcal{T}^{\rm en}p_y^{\rm canb}(Z) \text{ with } Z = Z^{\rm en},\\
p_t^{\rm cana}(Z) &= \mathcal{T}^{\rm en}p_t^{\rm canb}(Z) \text{ with } Z = Z^{\rm en},
\end{aligned} \tag{16.4.18}$$

where

$$\mathcal{T}^{\rm en} = \exp(q : \chi^{\rm en} :). \tag{16.4.19}$$

Exiting a Trailing Fringe-Field Region

The general considerations for the transition associated with exiting a trailing fringe-field region are similar to those employed earlier for entering a leading fringe-field region. If we wish to make the transition from the trailing fringe-field region to the trailing no-field region using the minimum vector potential (the vector potential in the Poincaré-Coulomb gauge), and also wish to carry out the design orbit and map integrations using the vector potential in the Coulomb gauge of Section 15.5.2, then we need to find at $Z = Z^{\rm ex}$ the gauge transformation that relates the Poincaré-Coulomb gauge and the Coulomb gauge of Section 15.5.2.

We now seek relations in the vicinity of the point $(0,0,Z^{\rm ex})$. Let us again recapitulate some of what we have learned. In the vicinity of this point, and for the $\alpha = s$ component of the magnetic field, the Poincaré-Coulomb gauge vector potential and the Coulomb gauge vector potential of Section 15.5 are related by the gauge transformation

$${}^{P}\!\boldsymbol{A}^{1,s}(x,y,z;Z^{\rm ex}) = \hat{\boldsymbol{A}}^{1,s}(x,y,Z^{\rm ex}+z) + \nabla\chi_{1,s}. \tag{16.4.20}$$

Also, the gauge term $\chi_{1,s}$ is related to the $m = 1$ and $\alpha = s$ on-axis gradient by the expansion

$$\chi_{1,s}(x, y, z; Z^{\rm ex}) = \{[(1/2)xz]C^{[0]}_{1,s}(Z^{\rm ex})\} + \{[(1/3)xz^2 - (1/12)x\rho^2]C^{[1]}_{1,s}(Z^{\rm ex})\} + \cdots . \quad (16.4.21)$$

Next we again recall the relation

$$\boldsymbol{A}^a - \boldsymbol{A}^b = \nabla\chi, \quad (16.4.22)$$

and compare it with the relation (4.20) rewritten in the form

$${}^P\!\boldsymbol{A}^{1,s}(x, y, z; Z^{\rm ex}) - \hat{\boldsymbol{A}}^{1,s}(x, y, Z^{\rm ex} + z) = \nabla\chi_{1,s}. \quad (16.4.23)$$

We conclude that if we wish to identify ${}^P\!\boldsymbol{A}^{1,s}$ with $\boldsymbol{A}^a$, and identify $\hat{\boldsymbol{A}}^{1,s}$ with $\boldsymbol{A}^b$, then we should now require the relation

$$\chi = \chi_{1,s}. \quad (16.4.24)$$

Finally, define the function $\chi^{\rm ex}$ by the rule

$$\chi^{\rm ex}(x, y; Z^{\rm ex}) = \chi(x, y, 0; Z^{\rm ex}). \quad (16.4.25)$$

With this definition we see from (4.21) and (4.24) that there is the result

$$\chi^{\rm ex}(x, y; Z^{\rm ex}) = -\ \{[(1/12)x\rho^2]C^{[1]}_{1,s}(Z^{\rm ex})\} + \cdots . \quad (16.4.26)$$

We are again ready to invoke the results (1.77) through (1.79). So doing, we find that the canonical coordinates $(x, y, t; p^{\rm can}_x, p^{\rm can}_y, p^{\rm can}_t)$ after and before $Z^{\rm ex}$ are connected by the symplectic map $\mathcal{T}^{\rm ex}$,

$$\begin{aligned} x^a(Z) &= \mathcal{T}^{\rm ex}x^b(Z) \text{ with } Z = Z^{\rm ex},\\ y^a(Z) &= \mathcal{T}^{\rm ex}y^b(Z) \text{ with } Z = Z^{\rm ex},\\ t^a(Z) &= \mathcal{T}^{\rm ex}t^b(Z) \text{ with } Z = Z^{\rm ex}; \end{aligned} \quad (16.4.27)$$

$$\begin{aligned} p^{\rm cana}_x(Z) &= \mathcal{T}^{\rm ex}p^{\rm canb}_x(Z) \text{ with } Z = Z^{\rm ex},\\ p^{\rm cana}_y(Z) &= \mathcal{T}^{\rm ex}p^{\rm canb}_y(Z) \text{ with } Z = Z^{\rm ex},\\ p^{\rm cana}_t(Z) &= \mathcal{T}^{\rm ex}p^{\rm canb}_t(Z) \text{ with } Z = Z^{\rm ex}, \end{aligned} \quad (16.4.28)$$

where

$$\mathcal{T}^{\rm ex} = \exp(q : \chi^{\rm ex} :). \quad (16.4.29)$$

Behavior of $C^{[1]}_{1,s}(Z)$

According to (4.16) and (4.26) both $\chi^{\rm en}$ and $\chi^{\rm ex}$ are proportional to $C^{[1]}_{1,s}(Z)$. Let us explore the behavior of $C^{[1]}_{1,s}(Z)$ for the simplest wiggler/undulator models described in Subsection 4.1. In accord with (3.13), (4.1), and (4.4) we may write for these models

$$C^{[0]}_{1,s}(Z) = -(4g/a){\rm wig}(3, Z, a, L) \quad (16.4.30)$$

and

$$C^{[0]}_{1,s}(Z) = -(4g/a)\mathrm{wig}(5, Z, a, L) \tag{16.4.31}$$

for the three-pole and five-pole cases, respectively. It follows that for the three-pole model there is the relation

$$C^{[1]}_{1,s}(Z) = -(4g/a)(\partial/\partial Z)\mathrm{wig}(3, Z, a, L) = -(4g/a)\mathrm{wig}'(3, Z, a, L), \tag{16.4.32}$$

and there is an analogous relation for the five-pole model. Figure 4.3 displays the profile function $\mathrm{wig}'(3, z, a, L)$ for the case $a = 0.1$ and $L = 0.5$. Evidently $\mathrm{wig}'(3, z, a, L)$ falls of quite rapidly for large $|Z|$. From (4.2) we expect the asymptotic fall off behavior to go as $1/|Z|^6$ for large $|Z|$. For example, at a distance of one wiggler/undulator period from the end, there is the result

$$\mathrm{wig}'(3, 1.5, .1, .5)/\mathrm{wig}(3, 0, .1, .5) = *. \tag{16.4.33}$$

And, at a distance of two wiggler/undulator periods from the end, there is the result

$$\mathrm{wig}'(3, 1.5, .1, .5)/\mathrm{wig}(3, 0, .1, .5) = *. \tag{16.4.34}$$

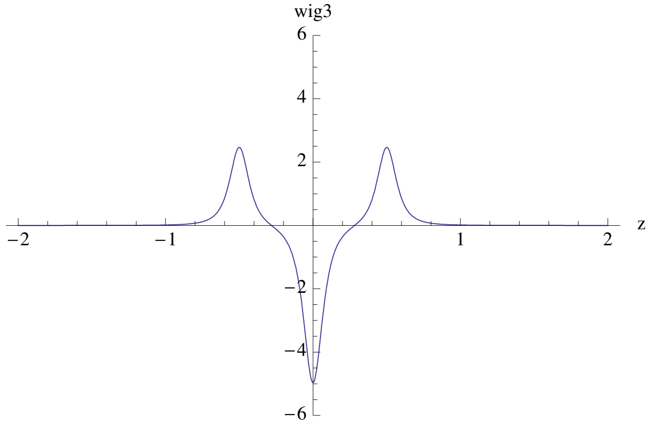


Figure 16.4.3: (Place Holder) The three-pole wiggler/undulator profile function $\mathrm{wig}'(3, z, a, L)$ when $a = 0.1$ and $L = 0.5$.

Net Total Map

Let $\mathcal{M}_{\mathrm{en}\to\mathrm{ex}}$ denote the map obtained by integrating for a wiggler/undulator the design orbit and map equations from $Z = Z^{\mathrm{en}}$ to $Z = Z^{\mathrm{ex}}$ using the Coulomb gauge vector potential of Section 15.5. Then the full net map $\mathcal{M}$ for the wiggler/undulator, including end-field termination effects, is given by the product

$$\mathcal{M} = \mathcal{T}^{\mathrm{en}}\mathcal{M}_{\mathrm{en}\to\mathrm{ex}}\mathcal{T}^{\mathrm{ex}}. \tag{16.4.35}$$

Discontinuities in Mechanical Momenta Associated with Termination Approximation

As described in Subsection 1.2, there are discontinuities in the mechanical momenta associated with the use of a symplectic termination procedure. Recall (1.30), (1.31), (1.41), and (1.42). Here, for our simple wiggler/undulator models, we will study the consequences of terminating end fields using the minimum (Poincaré-Coulomb gauge) vector potential.

The $m = 1$ and $\alpha = s$ Poincaré-Coulomb gauge vector potential about the expansion point $(0, 0, Z_0)$ is given in terms of on-axis gradients by (15.9.25) through (15.9.27). And, for our simple wiggler/undulator models, the on-axis gradients are given by relations of the form (4.30) through (4.32). Combining these relations with (1.30) and (1.31) gives, upon entry and for the 3-pole case, the discontinuity results

$$\begin{aligned}\Delta p_x^{\rm mech} &= q[{}^PA_x^{1,s}(x,y,0;Z^{\rm en})]\\ &= q\{[-(1/3)y^2]C_{1,s}^{[1]}(Z^{\rm en})+\cdots\}\\ &= q\{[-(1/3)y^2](-4g/a){\rm wig}'(3,Z^{\rm en},a,L)+\cdots\},\end{aligned} \tag{16.4.36}$$

$$\begin{aligned}\Delta p_y^{\rm mech} &= q[{}^PA_y^{1,s}(x,y,0;Z^{\rm en})]\\ &= q\{[(1/3)xy]C_{1,s}^{[1]}(Z^{\rm en})+\cdots\}\\ &= q\{[(1/3)xy](-4g/a){\rm wig}'(3,Z^{\rm en},a,L)+\cdots\}.\end{aligned} \tag{16.4.37}$$

Similarly, upon exit, we find from (15.9.25) through (15.9.27), (1.40), and (1.41) the discontinuity results

$$\begin{aligned}\Delta p_x^{\rm mech} &= q[{}^PA_x^{1,s}(x,y,0;Z^{\rm ex})]\\ &= q\{[-(1/3)y^2]C_{1,s}^{[1]}(Z^{\rm ex})+\cdots\}\\ &= q\{[-(1/3)y^2](-4g/a){\rm wig}'(3,Z^{\rm en},a,L)+\cdots\},\end{aligned} \tag{16.4.38}$$

$$\begin{aligned}\Delta p_y^{\rm mech} &= q[{}^PA_y^{1,s}(x,y,0;Z^{\rm ex})]\\ &= q\{[(1/3)xy]C_{1,s}^{[1]}(Z^{\rm ex})+\cdots\}\\ &= q\{[(1/3)xy](-4g/a){\rm wig}'(3,Z^{\rm ex},a,L)+\cdots\}.\end{aligned} \tag{16.4.39}$$

Recall the relation (15.9.35) which we rewrite in the form

$$C_{1,s}^{[0]}(Z) = B_y(0,0,Z). \tag{16.4.40}$$

We see from (4.36) through (4.39) and (4.40) that in all cases the discontinuities are proportional to $B_y'(0,0,Z)$ and its derivatives at $Z = Z^{\rm en}$ or $Z = Z^{\rm ex}$. Recall (4.32) and see Figure 4.3 for an example of how these functions behave (fall off) in the case of the simplest 3-pole wiggler/undulator model. Moreover, the discontinuities also vanish as the spatial deviations from the z axis (the design orbit) become small.

Exercises

16.4.1. Verify (4.11) and (4.12). Verify that (4.12) also follows from (15.8.21) and (15.8.33).

16.4.2. Verify (4.14) through (4.17).

16.4.3. Verify (4.20) through (4.22).

16.4.4. Verify (4.24) and (4.25).

16.4.5. Verify $\cdots$.

16.4.6. Verify the relation (4.33).

16.5 Iron-Free Quadrupoles

16.5.1 Preliminaries

A quadrupole is a straight beam-line element whose field is described by a cylindrical harmonic expansion that contains primarily an $m = 2$ term. We recall from Section 15.2.3 that in this case the magnetic scalar potential ψ has the expansion

$$\psi(x, y, z) = \psi_{2,s}(x, y, z) = 2xy[C_{2,s}^{[0]}(z) - (1/24)(x^2 + y^2)C_{2,s}^{[2]}(z) + \cdots]. \tag{16.5.1}$$

(Here we have retained only the *normal* term, but a skew term is also possible.) See (15.2.65). Correspondingly, the associated magnetic field has the expansion

$$B_x = \partial_x \psi_{2,s} = 2yC_{2,s}^{[0]}(z) - (1/12)(3x^2y + y^3)C_{2,s}^{[2]}(z) + \cdots , \tag{16.5.2}$$

$$B_y = \partial_y \psi_{2,s} = 2xC_{2,s}^{[0]}(z) - (1/12)(x^3 + 3xy^2)C_{2,s}^{[2]}(z) + \cdots , \tag{16.5.3}$$

$$B_z = \partial_z \psi_{2,s} = 2xy[C_{2,s}^{[1]}(z) - (1/24)(x^2 + y^2)C_{2,s}^{[3]}(z) + \cdots]. \tag{16.5.4}$$

From (6.2) through (6.4) we see that $\boldsymbol{B}$ is completely specified in terms of a single "master" function $C_{2,s}^{[0]}(z)$ and its derivatives. Moreover, according to (6.2) and (6.3), the on-axis field is characterized by a quadrupole strength Q given by the relation

$$Q(0, 0, z) = 2C_{2,s}^{[0]}(z). \tag{16.5.5}$$

See Exercise 1.5.9. We will next examine what can be said about the master function $C_{2,s}^{[0]}(z)$ in various cases.

16.5.2 Single Monopole Quartet

The simplest way to produce a quadrupole field, as the name again suggests, is to properly locate and assign strengths to four monopoles. Suppose two monopoles with strengths g are placed at the diametrically opposed (x,y,z) locations $(a\sqrt{2},a\sqrt{2},0)$ and $(-a\sqrt{2},-a\sqrt{2},0)$. Suppose two more monopoles of strength $-g$ are placed at the diametrically opposed locations $(a\sqrt{2},-a\sqrt{2},0)$ and $(-a\sqrt{2},a\sqrt{2},0)$. Note that all these locations lie on a cylinder of radius a and are spaced 90° apart. Such an array of monopoles will be called a monopole *quartet*. (For a discussion of the case of a monopole *half quartet*, which is what we will call two diametrically opposed monopoles of the same strength, see Exercise *.)

We will now see that a monopole quartet produces an interior field whose cylindrical multipole expansion begins with an $m=2$ term. From the symmetry of this array we expect that, in addition to the leading $m=2$ term, there may also be $m=6,10,\cdots$ terms. Recall Subsection 15.2.5. As before, let $\chi(\boldsymbol{r};\boldsymbol{r}')$ be the potential for a monopole having strength g and located at the point $\boldsymbol{r}'$. Recall (4.9). Then the potential $\psi^{\rm quart}$ for a monopole quartet centered on the origin is given by the relation

$$\begin{aligned}\psi^{\rm quart}(x,y,z) \;\;=\;\; & g\chi(x,y,z;a\sqrt{2},a\sqrt{2},0)+g\chi(x,y,z;-a\sqrt{2},-a\sqrt{2},0)\\ & -g\chi(x,y,z;a\sqrt{2},-a\sqrt{2},0)-g\chi(x,y,z;-a\sqrt{2},a\sqrt{2},0). \qquad (16.5.6)\end{aligned}$$

Next expand $\psi^{\rm quart}(x,y,z)$ as a power series in x and y. So doing yields the result

$$\psi^{\rm quart}(x,y,z)=-6ga^2xy/(z^2+a^2)^{5/2}+\text{higher order terms}. \qquad (16.5.7)$$

Comparison of (6.1) with (6.10) reveals that for a monopole quartet centered on the origin its $m=2$ on-axis gradient is given by the relation

$$C_{2,s}^{[0]}(z)=-3ga^2/(z^2+a^2)^{5/2}=-4(g/a^2)\delta(z,a) \qquad (16.5.8)$$

where $\delta(z,a)$ is an approximating delta function defined by the relation

$$\delta(z,a)=(3/4)a^4/(z^2+a^2)^{5/2}. \qquad (16.5.9)$$

Calculation shows that this approximating delta function has the properties

$$\delta(-z,a)=\delta(z,a), \qquad (16.5.10)$$

$$\delta(0,a)=3/(4a) \qquad (16.5.11)$$

$$\delta(z,a)=(3a^4/4)/|z|^5+O(1/|z|^7) \text{ as } |z|\to\infty, \qquad (16.5.12)$$

$$\int_{-\infty}^{\infty} dz\;\delta(z,a)=1. \qquad (16.5.13)$$

Figures 5.1 and 5.2 illustrate the behavior of this approximating delta function for two different values of a. Evidently this approximating delta function (5.9) becomes the true delta function in the limit $a\to 0$,

$$\lim_{a\to 0}\delta(z,a)=\delta(z), \qquad (16.5.14)$$

and a controls the fall-off rate. From (5.11) we see that the $m=2$ on-axis gradient for a monopole quartet has the fall-off behavior

$$C_{2,s}^{[0]}(z)=-3ga^2/|z|^5+O(1/|z|^7) \text{ as } |z|\to\infty. \qquad (16.5.15)$$

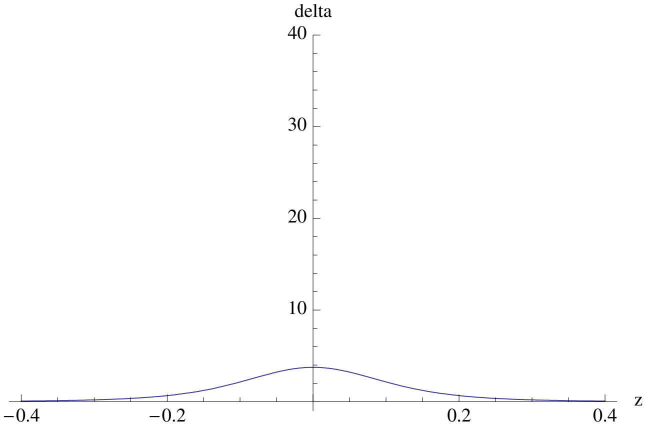


Figure 16.5.1: The approximating delta function (5.9) when $a = .2$.

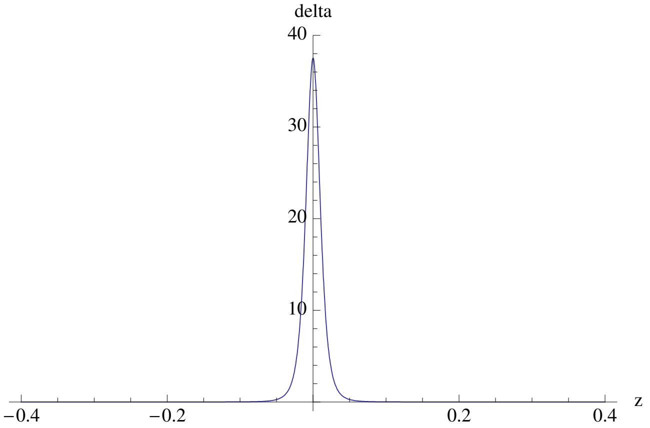


Figure 16.5.2: The approximating delta function (5.9) when $a = .02$.

16.5.3 Line of Monopole Quartets

Next consider the case of a *line* of monopole quartets extending from $z = 0$ to $z = L$. It can be treated analogously to the treatment of a line of monopole doublets provided in Subsection 4.3. A line of monopole quartets will produce a profile function described by the relation

$$\text{bump}(z, a, L) = \int_0^L dz' \, \delta(z - z', a) \tag{16.5.16}$$

where now $\delta(z, a)$ is the approximating delta function given by (5.9). In terms of this soft-edge bump function the $m = 1$ on-axis field gradient $C_{2,s}^{[0]}(z)$ for a line of monopole quartets can be written in the form

$$C_{2,s}^{[0]}(z) = (Q/2)\text{bump}(z, a, L) \tag{16.5.17}$$

where Q is the quadrupole strength in the infinite length limit. Recall Exercise 1.5.9 and (15.2.66) through (15.2.68). It is given in terms of the quartet parameters by the relation

$$Q = -8G/a^2. \tag{16.5.18}$$

Here again G is the monopole strength per unit length.

To evaluate (5.16) again make make the change of variables $\zeta = z - z'$ so that (5.16) becomes

$$\text{bump}(z, a, L) = -\int_z^{z-L} d\zeta \; \delta(\zeta, a) = \int_{z-L}^{z} d\zeta \; \delta(\zeta, a). \tag{16.5.19}$$

The approximating delta function (5.9) has as an indefinite integral the result

$$\int d\zeta \; \delta(\zeta, a) = [(1/2)\zeta^3 + (3/4)a^2\zeta]/(\zeta^2 + a^2)^{3/2} \tag{16.5.20}$$

so that we may now define an associated approximating signum function by the rule

$$\text{sgn}(z, a) = z[z^2 + (3/2)a^2]/(z^2 + a^2)^{3/2}. \tag{16.5.21}$$

We conclude from (5.19) and (5.20) that, in terms of the approximating signum function (5.21), there is the relation.

$$\text{bump}(z, a, L) = [\text{sgn}(z, a) - \text{sgn}(z - L, a)]/2. \tag{16.5.22}$$

It follows from (5.21) and (5.22) that, like the soft-edge bump functions already discussed, the soft-edge bump function for a line of monopole quartets satisfies the relations (1.15) through (1.18). Recall Exercise 1.2. We see from (1.18) and (5.17) that for line of monopole quartets there is the relation

$$\int_{-\infty}^{\infty} C_{2,s}^{[0]}(z)dz = (Q/2)L. \tag{16.5.23}$$

Figures 5.3 and 5.4 illustrate the behavior of the approximating signum function (5.21) for two different values of a. Evidently it becomes the true signum function in the limit

$a \to 0$. Figures 5.5 and 5.6 illustrate the behavior of the corresponding soft-edge bump function (5.22). It becomes the true bump function in the limit $a \to 0$. These sigum and bump functions have the asymptotic behaviors

$$\mathrm{sgn}(z,a) = 1 - (3/8)(a/z)^4 + O(1/z^6) \text{ as } z \to \infty, \tag{16.5.24}$$

$$\mathrm{bump}(z,a,L) = (3/4)La^4/|z|^5 + O(1/|z|^6) \text{ as } |z| \to \infty. \tag{16.5.25}$$

It follows that

$$C_{2,s}^{[0]}(z) = (3/8)QLa^4/|z|^5 + O(1/|z|^6) \text{ as } |z| \to \infty. \tag{16.5.26}$$

The fall off for the $m = 2$ on-axis gradient arising from a line of monopole quartets goes as $1/|z|^5$, just as it does for a single monopole quartet.

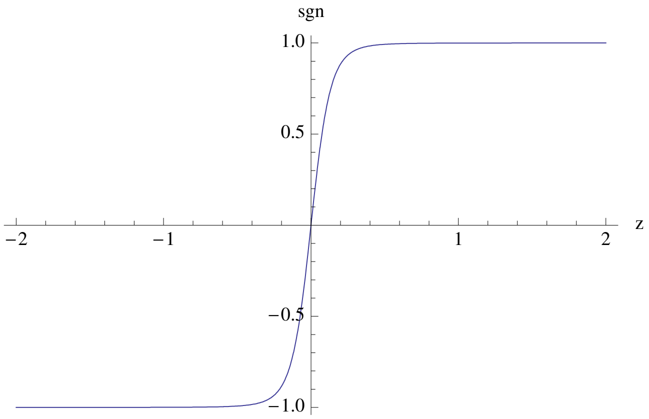


Figure 16.5.3: The approximating signum function (5.21) when $a = .2$.

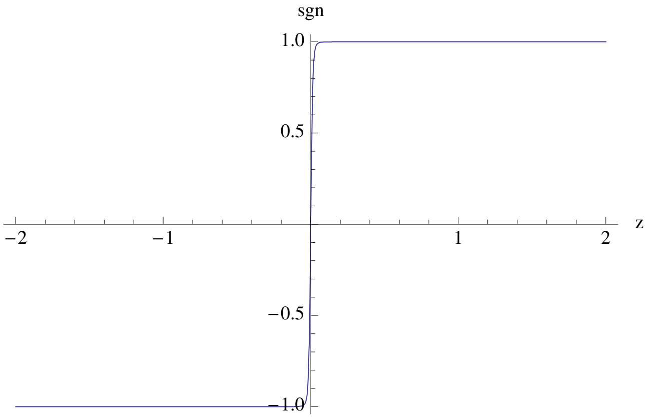


Figure 16.5.4: The approximating signum function (5.21) when $a = .02$.

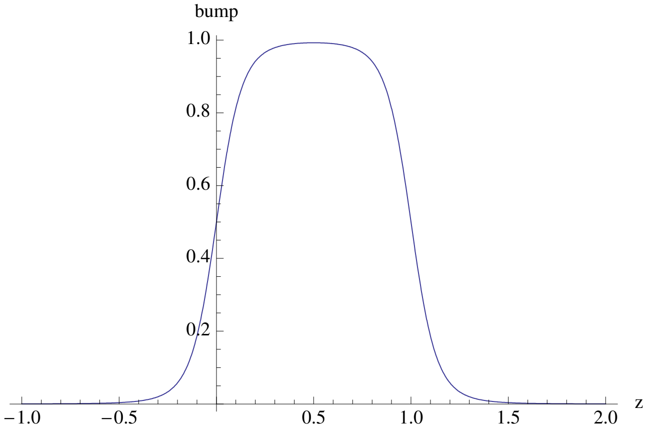


Figure 16.5.5: The soft-edge bump function (5.22) when $a = .2$.

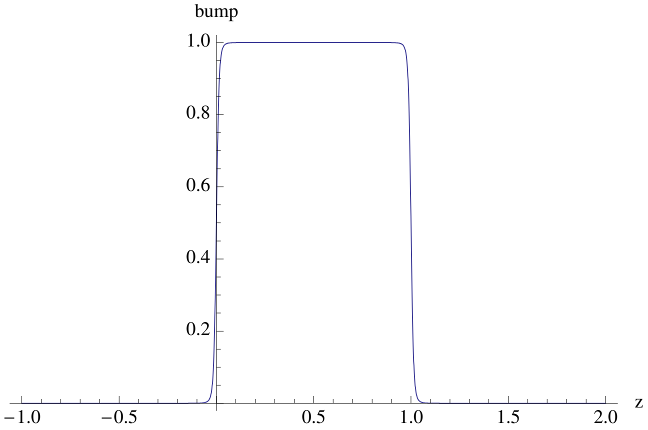


Figure 16.5.6: The soft-edge bump function function (5.22) when $a = .02$.

16.5.4 Idealized Air-Core Quadrupole

According to Subsection 15.2.5 we expect that a single monopole quartet may have nonzero on-axis gradients for the values $m = 6, 10, \cdots$ as well as the value $m = 2$. The same is true for a line of monopole quartets. An idealized air-core quadrupole consists of a thin-shell winding having a circular cross section of radius $\rho = a$ and length L. The current in the winding ideally has a $\cos(2\phi)$ dependence, which results in a "pure" quadrupole field within the bore. That is, *only* the $m = 2$ on-axis gradient is nonzero for the field produced by such a winding. However, it is not the case that the field is that of a *perfect* quadrupole. According to (15.2.65) through (15.2.68) there are additional components in the fringe-field regions at the ends of the quadrupole where $C_{2,s}^{[0]}(z)$ is changing.

If the winding begins at $z = 0$ and extends to $z = L$, then the on-axis field gradient for such a quadrupole can again be described in terms of a soft-edge bump function which we will again call $\text{bump}(z, a, L)$. That is, the on-axis field gradient $C_{2,s}^{[0]}(z)$ can again be written in the form

$$C_{2,s}^{[0]}(z) = (Q/2)\text{bump}(z, a, L) \tag{16.5.27}$$

where Q is again the quadrupole strength in the infinite length limit.

Like the previous examples of soft-edge bump functions, the soft-edge bump function for an idealized air-core quadrupole can be written in terms of an associated approximating signum function in the form

$$\text{bump}(z, a, L) = [\text{sgn}(z, a) - \text{sgn}(z - L, a)]/2. \tag{16.5.28}$$

It can be shown that for an idealized air-core quadrupole the approximating signum function $\text{sgn}(z, a)$ is given by the relation

$$\begin{aligned} \text{sgn}(z, a) &= [z^5 + (5/2)z^3a^2 + (9/4)za^4]/(z^2 + a^2)^{5/2} \\ &= z[z^4 + (5/2)z^2a^2 + (9/4)a^4]/(z^2 + a^2)^{5/2}. \end{aligned} \tag{16.5.29}$$

Figures 5.7 and 5.8 illustrate the behavior of this approximating signum function for two different values of a. Evidently the approximating signum function (5.29) becomes the true signum function in the limit $a \to 0$.

It follows from (5.28) and (5.29) that, like the previous soft-edge bump functions, the soft-edge bump function for an idealized air-core quadrupole satisfies the relations (1.15) through (1.18). Recall Exercise 1.2. Also, the relation (5.23) continues to hold.

Figures 5.9 and 5.10 illustrate the properties (1.15) through (1.17) for a fixed value of L and two different values of the radius a. Evidently the ideal air-core quadrupole soft-edge bump function becomes a hard-edge bump function in the limit $a \to 0$. The radius a plays the role of a characteristic length that controls the rate of fall off. The fringe-field region is large if a is large, and vanishes as a goes to zero. From (5.29) and (5.28) we find the asymptotic behaviors

$$\text{sgn}(z, a) = 1 + (3/8)(a/z)^4 + O(1/z^6) \text{ as } z \to \infty, \tag{16.5.30}$$

$$\text{bump}(z, a, L) = -(3/4)La^4/|z|^5 + O(1/|z|^6) \text{ as } |z| \to \infty. \tag{16.5.31}$$

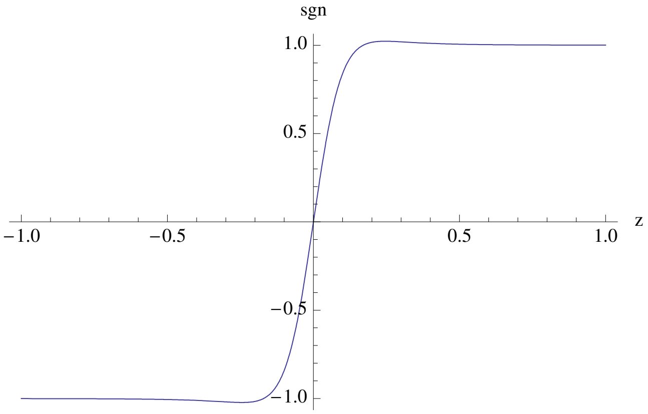


Figure 16.5.7: The approximating signum function (5.29) when $a = .2$.

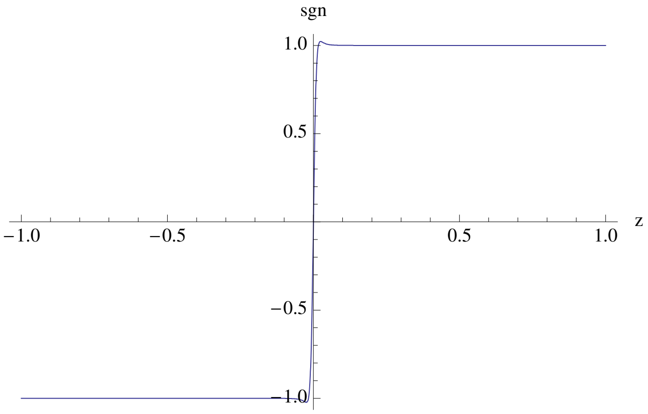


Figure 16.5.8: The approximating signum function (5.29) when $a = .02$.

Consequently $C^{[0]}_{2,s}(z)$ falls off for large distances as

$$C^{[0]}_{2,s}(z) = -(3/4)(Q/2)La^4/|z|^5 + O(1/|z|^6) \text{ as } |z| \to \infty. \tag{16.5.32}$$

We see that the fall off goes as $1/|z|^5$, which is the same rate as that for a monopole quartet and a line of monopole quartets. Note, however, that the sign on the right side of (5.31) is negative just as it is for the right side of (3.31). Therefore, the bump function for an idealized air-core quadrupole also cannot be modeled by an Enge function profile.

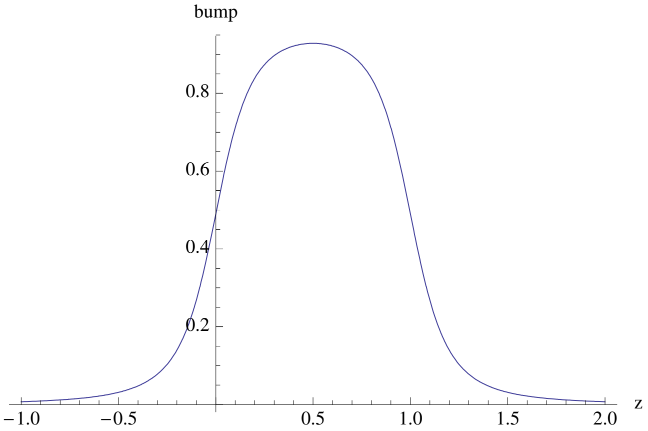


Figure 16.5.9: The soft-edge bump function given by (5.28) and (5.29) when $a = .2$ and $L = 1$.

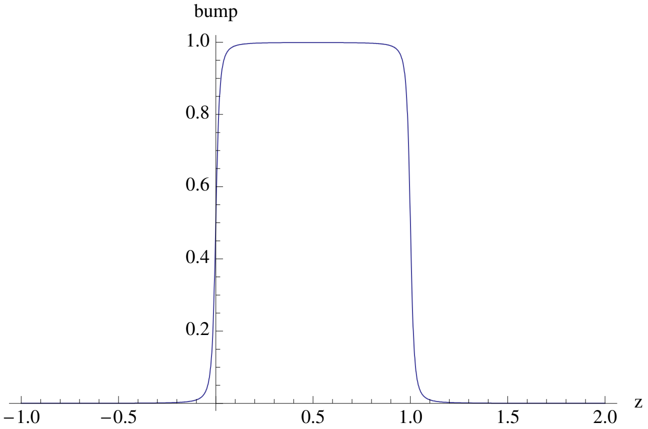


Figure 16.5.10: The soft-edge bump function given by (5.28) and (5.29) when $a = .02$ and $L = 1$.

Our discussion of quadrupole fringe fields so far has treated the iron-free case, and we have found a $1/|z|^5$ fall off in all cases. When iron is present, and the coils are buried in iron, the fall off can in principle be much faster including the possibility of essentially exponential fall off.

16.5.5 Rare Earth Cobalt (REC) Quadrupoles

A rare earth cobalt (REC) quadrupole typically has a circular annular cross section with outer radius r_2 and inner/bore radius r_1. The space between r_1 and r_2 is filled with REC material magnetized and arranged so as to produce a pure quadrupole magnetic field within the bore. That is, ideally *only* the $m = 2$ on-axis gradient is nonzero for the field produced by such an arrangement of REC material.

It can be shown that the on-axis gradient can again be described in terms of a soft-edge bump function which we will call bump(z, r_1, r_2, L) where L is the quadrupole length. That is, the on-axis field gradient $C_{2,s}^{[0]}(z)$ can be written in the form

$$C_{2,s}^{[0]}(z) = (Q/2)\text{bump}(z, r_1, r_2, L) \tag{16.5.33}$$

where Q is the strength of the REC quadrupole in the infinite length limit. Moreover, as before, the soft-edge bump function for a REC quadrupole can be written in terms of an associated approximating signum function in the form

$$\text{bump}(z, r_1, r_2, L) = [\text{sgn}(z, r_1, r_2) - \text{sgn}(z - L, r_1, r_2)]/2. \tag{16.5.34}$$

Finally, it can be shown that for the REC quadrupole the approximating signum function sgn(z, r_1, r_2) is given by the relation

$$\text{sgn}(z, r_1, r_2) = z[(r_1 + r_2)/(r_1 r_2)][(v_1 v_2)/(v_1 + v_2)][1 + (1/8)v_1 v_2(4 + v_1^2 + v_1 v_2 + v_2^2)] \tag{16.5.35}$$

where v_1 and v_2 are defined by the relations

$$v_1 = 1/\sqrt{1 + (z/r_1)^2}, \tag{16.5.36}$$

$$v_2 = 1/\sqrt{1 + (z/r_2)^2}. \tag{16.5.37}$$

It can be easily checked that the approximating signum function sgn(z, r_1, r_2) becomes the true signum function in the limit $r_1 \to 0$. See Exercise 5.4. For example, Figures 5.11 and 5.12 illustrate the behavior of this approximating signum function for two different values of r_1 and fixed values of r_2 and L.

It follows from (5.34) that the soft-edge bump function for a REC quadrupole also satisfies relations analogous to (1.15) through (1.18) and (5.23) remains true. Figures 5.13 and 5.14 illustrate the properties (1.15) through (1.17) for fixed values of r_2 and L and two different values of the inner radius r_1. Evidently the REC soft-edge bump function becomes a hard-edge bump function in the limit $r_1 \to 0$. The inner radius r_1 (as well as r_2) plays the role of a characteristic length that controls the rate of fall off. The fringe-field region is large if r_1 is large, and vanishes as r_1 goes to zero. From (5.34) through (5.37) we find the asymptotic behaviors

$$\begin{aligned}\text{sgn}(z, r_1, r_2) &= 1 - (1/16)r_1 r_2[(r_1^5 - r_2^5)/(r_1 - r_2)](1/z)^6 + O(1/z^8) \\ &= 1 - (1/16)r_1 r_2(r_1^4 + r_1^3 r_2 + r_1^2 r_2^2 + r_1 r_2^3 + r_2^4)(1/z^6) + O(1/z^8) \\ &\quad \text{as } z \to \infty,\end{aligned} \tag{16.5.38}$$

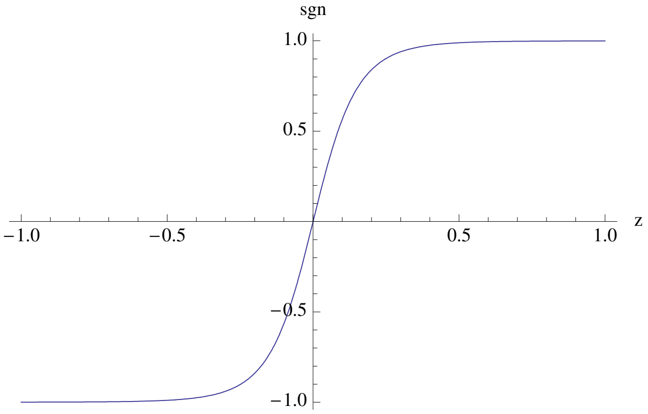


Figure 16.5.11: The approximating signum function (5.35) when $r_1 = .2$ and $r_2 = .5$.

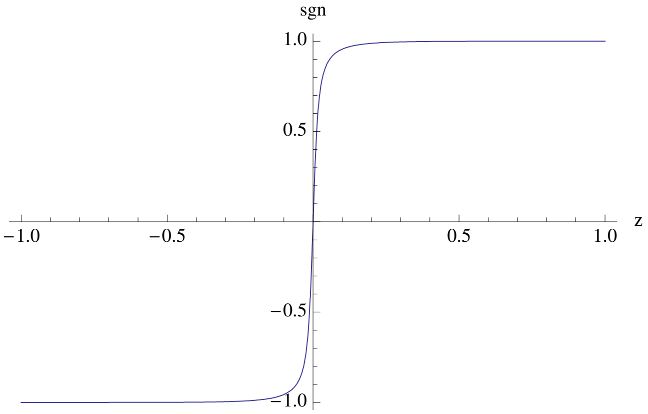


Figure 16.5.12: The approximating signum function (5.35) when $r_1 = .02$ and $r_2 = .5$.

$$\begin{aligned}\text{bump}(z,r_1,r_2,L) &= (3/16)Lr_1r_2[(r_1^5-r_2^5)/(r_1-r_2)](1/|z|^7)+O(1/|z|^8)\\ &= (3/16)Lr_1r_2(r_1^4+r_1^3r_2+r_1^2r_2^2+r_1r_2^3+r_2^4)(1/|z|^7)+O(1/|z|^8)\\ &\quad \text{as } |z|\to\infty. \qquad (16.5.39)\end{aligned}$$

Consequently $C_{2,s}^{[0]}(z)$ falls off for large distances as

$$\begin{aligned}C_{2,s}^{[0]}(z) &= (3/16)(Q/2)Lr_1r_2[(r_1^5-r_2^5)/(r_1-r_2)](1/|z|^7)+O(1/|z|^8)\\ &= (3/16)(Q/2)Lr_1r_2(r_1^4+r_1^3r_2+r_1^2r_2^2+r_1r_2^3+r_2^4)(1/|z|^7)+O(1/|z|^8).\end{aligned} \qquad (16.5.40)$$

We see that the fall off goes as $1/|z|^7$, which is pleasantly rapid. Remarkably, this rate of fall off for a REC quadrupole is two orders higher in $1/|z|$ than that for an idealized air-core quadrupole. Compare (5.32) and (5.40).

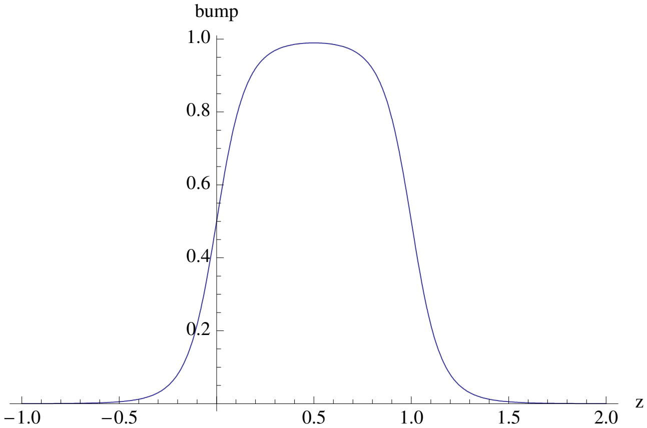


Figure 16.5.13: The soft-edge bump function (5.34) when $r_1 = .2$, $r_2 = .5$, and $L = 1$.

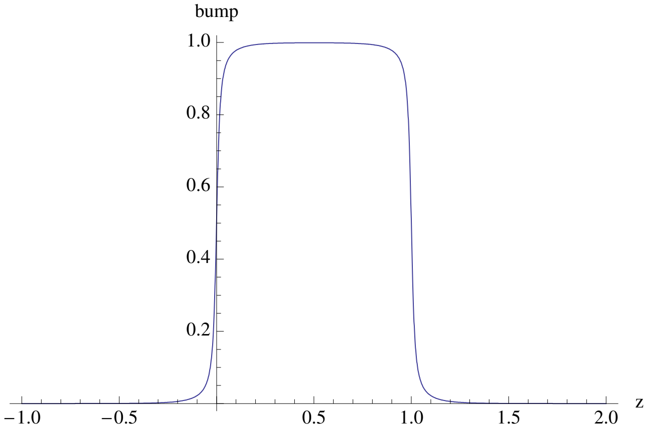


Figure 16.5.14: The soft-edge bump function (5.34) when $r_1 = .02$, $r_2 = .5$, and $L = 1$.

16.5.6 Overlapping Fringe Fields

16.5.7 Terminating Quadrupole End Fields

Preliminaries

For this subsection, as we did in Subsection 4.3, we will use the global coordinates $\boldsymbol{R}$ and local coordinates $\boldsymbol{r}$ as given by (4.8) and (4.9). Following previous notation, we assume the leading quadruple fringe field begins at $Z = Z^{\rm en}$ and the trailing quadrupole fringe field ends at $Z = Z^{\rm ex}$. In the interval $[Z^{\rm en}, Z^{\rm ex}]$ the design orbit and map integrations will be carried out using the vector potential in the Coulomb gauge of Section 15.5. The entering transition at $Z^{\rm en}$ from the leading no-field region to the leading fringe-field region, and the exiting transition at $Z^{\rm ex}$ from the trailing fringe-field region to the trailing no-field region, will be made using the minimum vector potential, namely the Poincaré-Coulomb gauge vector potential.

If we wish to make the transition from the leading no-field region to the leading fringe-field region using the minimum vector potential (the vector potential in the Poincaré-Coulomb gauge), and also wish to carry out the design orbit and map integrations using the vector potential in the Coulomb gauge of Section 15.5.2, then we need to find at $Z = Z^{\rm en}$ the gauge transformation that relates the Poincaré-Coulomb gauge and the Coulomb gauge of Section 15.5.2. Similarly, If we wish to make the transition from the trailing fringe-field region to the trailing no-field region using the minimum vector potential (the vector potential in the Poincaré-Coulomb gauge), and also wish to carry out the design orbit and map integrations using the vector potential in the Coulomb gauge of Section 15.5.2, then we need to find at $Z = Z^{\rm ex}$ the gauge transformation that relates the Poincaré-Coulomb gauge and the Coulomb gauge of Section 15.5.2. For simplicity, we will continue to assume that we are dealing with the case of a *normal* quadrupole; namely $\alpha = s$.

For both transitions there is a relation of the form

$$ {}^{P}\!\boldsymbol{A}^{2,s}(x, y, z; Z^\beta) = \hat{\boldsymbol{A}}^{2,s}(x, y, Z^\beta + z) + \nabla\chi_{2,s} \tag{16.5.41}$$

where $\beta = $ en or $\beta = $ ex. See (15.7.5). There is also the relation

$$\begin{aligned} \chi_{2,s}(x, y, z; Z^\beta) &= \cos(2\phi)\sum_{k=0}^{\infty}(-1)^k \frac{2!}{2^{2k}k!(k+2)!} D^{[2k]}_{2,s}(z; Z^\beta)\rho^{2k+2} \\ &= \rho^2\cos(2\phi)[D^{[0]}_{2,s}(z; Z^\beta) - (1/6)\rho^2 D^{[2]}_{2,s}(z; Z^\beta) + (*)\rho^4 D^{[4]}_{2,s}(z; Z^\beta) + \cdots] \\ &= (x^2 - y^2)[D^{[0]}_{2,s}(z; Z^\beta) - (1/6)\rho^2 D^{[2]}_{2,s}(z; Z^\beta) + (*)\rho^4 D^{[4]}_{2,s}(z; Z^\beta) + \cdots]. \end{aligned} \tag{16.5.42}$$

See (15.7.20). Moreover we have found, see (15.7.60), the relation

$$\begin{aligned} D^{[0]}_{2,s}(z; Z^\beta) &= \{[1/(2+1)]C^{[0]}_{2,s}(Z^\beta)\}z + \{[1/(2+2)]C^{[1]}_{2.s}(Z^\beta)\}z^2 \\ &+ \{[1/(2+3)](1/2!)C^{[2]}_{2,s}(Z^\beta)\}z^3 + \{[1/(2+4)](1/3!)C^{[3]}_{2,s}(Z^\beta)\}z^4 + \cdots \\ &= (1/3)C^{[0]}_{2,s}(Z^\beta)z + (1/4)C^{[1]}_{2,s}(Z^\beta)z^2 + (1/10)C^{[2]}_{2,s}(Z^\beta)z^3 + (1/36)C^{[3]}_{2,s}(Z^\beta)z^4 + \cdots , \end{aligned} \tag{16.5.43}$$

from which it follows that

$$D_{2,s}^{[0]}(0;Z^\beta)=0, \tag{16.5.44}$$

$$D_{2,s}^{[2]}(0;Z^\beta)=(1/2)C_{2,s}^{[1]}(Z^\beta), \tag{16.5.45}$$

$$D_{2,s}^{[4]}(0;Z^\beta)=(2/3)C_{2,s}^{[3]}(Z^\beta). \tag{16.5.46}$$

Inserting these results in (5.42) gives the expansion

$$\begin{aligned}\chi_{2,s}(x,y,0;Z^\beta) &= (x^2-y^2)[(*)\rho^2C_{2,s}^{[1]}(Z^\beta)+(*)\rho^4C_{2,s}^{[3]}(Z^\beta)+\cdots]\\ &= (x^4-y^4)[(*)C_{2,s}^{[1]}(Z^\beta)+(*)\rho^2C_{2,s}^{[3]}(Z^\beta)+\cdots].\end{aligned} \tag{16.5.47}$$

Finally there is the relation (4.12), which we repeat below:

$$\boldsymbol{A}^a-\boldsymbol{A}^b=\nabla\chi, \tag{16.5.48}$$

Entering a Leading Fringe-Field Region

Compare (5.48) with the relation (5.41) evaluated for the case $\beta=\text{en}$ and rewritten in the form

$$\hat{\boldsymbol{A}}^{2,s}(x,y,Z^{\text{en}}+z)\ -\ {}^P\!\boldsymbol{A}^{2,s}(x,y,z;Z^{\text{en}})=-\nabla\chi_{2,s}. \tag{16.5.49}$$

We conclude that if we wish to identify $\hat{\boldsymbol{A}}^{2,s}$ with $\boldsymbol{A}^a$, and identify ${}^P\!\boldsymbol{A}^{2,s}$ with $\boldsymbol{A}^b$, then we should require the relation

$$\chi=-\chi_{2,s}. \tag{16.5.50}$$

Next define the function χ^{en} by the rule

$$\chi^{\text{en}}(x,y;Z^{\text{en}})=\chi(x,y,0;;Z^{\text{en}}). \tag{16.5.51}$$

With this definition we see from (5.47), (5.50), and (5.51) that there is the result

$$\chi^{\text{en}}(x,y;Z^{\text{en}})=-(x^4-y^4)[(*)C_{2,s}^{[1]}(Z^{\text{en}})+(*)\rho^2C_{2,s}^{[3]}(Z^{\text{en}})+\cdots]. \tag{16.5.52}$$

We are now ready to invoke the results (1.77) through (1.79). So doing, we find that the canonical coordinates $(x,y,t;p_x^{\text{can}},p_y^{\text{can}},p_t^{\text{can}})$ *after* and *before* Z^{en} are connected by the symplectic map $\mathcal{T}^{\text{en}}$,

$$\begin{aligned}x^a(Z)&=\mathcal{T}^{\text{en}}x^b(Z)\text{ with }Z=Z^{\text{en}},\\ y^a(Z)&=\mathcal{T}^{\text{en}}y^b(Z)\text{ with }Z=Z^{\text{en}},\\ t^a(Z)&=\mathcal{T}^{\text{en}}t^b(Z)\text{ with }Z=Z^{\text{en}};\end{aligned} \tag{16.5.53}$$

$$\begin{aligned}p_x^{\text{cana}}(Z)&=\mathcal{T}^{\text{en}}p_x^{\text{canb}}(Z)\text{ with }Z=Z^{\text{en}},\\ p_y^{\text{cana}}(Z)&=\mathcal{T}^{\text{en}}p_y^{\text{canb}}(Z)\text{ with }Z=Z^{\text{en}},\\ p_t^{\text{cana}}(Z)&=\mathcal{T}^{\text{en}}p_t^{\text{canb}}(Z)\text{ with }Z=Z^{\text{en}},\end{aligned} \tag{16.5.54}$$

where

$$\mathcal{T}^{\text{en}}=\exp(q:\chi^{\text{en}}:). \tag{16.5.55}$$

Exiting a Trailing Fringe-Field Region

Compare (5.48) with the relation (5.41) evaluated for the case $\beta = \text{ex}$ and rewritten in the form

$$^{P}\boldsymbol{A}^{2,s}(x,y,z) - \hat{\boldsymbol{A}}^{2,s}(x,y,Z^{\text{ex}}+z) = \nabla\chi_{2,s}. \tag{16.5.56}$$

We conclude that if we wish to identify $^{P}\boldsymbol{A}^{2,s}$ with $\boldsymbol{A}^{a}$, and identify $\hat{\boldsymbol{A}}^{2,s}$ with $\boldsymbol{A}^{b}$, then we should now require the relation

$$\chi = \chi_{2,s}. \tag{16.5.57}$$

Next define the function χ^{ex} by the rule

$$\chi^{\text{ex}}(x,y) = \chi(x,y,0). \tag{16.5.58}$$

With this definition we see from (5.47), (5.57), and (5.58) that there is the result

$$\chi^{\text{ex}}(x,y) = (x^4 - y^4)[(*)C_{2,s}^{[1]}(Z^{\text{ex}}) + (*)\rho^2 C_{2,s}^{[3]}(Z^{\text{ex}}) + \cdots]. \tag{16.5.59}$$

We are again ready to invoke the results (1.77) through (1.79). So doing, we find that the canonical coordinates $(x,y,t;p_x^{\text{can}},p_y^{\text{can}},p_t^{\text{can}})$ after and before Z^{ex} are connected by the symplectic map $\mathcal{T}^{\text{ex}}$,

$$\begin{aligned}
x^a(Z) &= \mathcal{T}^{\text{ex}} x^b(Z) \text{ with } Z = Z^{\text{ex}},\\
y^a(Z) &= \mathcal{T}^{\text{ex}} y^b(Z) \text{ with } Z = Z^{\text{ex}},\\
t^a(Z) &= \mathcal{T}^{\text{ex}} t^b(Z) \text{ with } Z = Z^{\text{ex}};
\end{aligned} \tag{16.5.60}$$

$$\begin{aligned}
p_x^{\text{cana}}(Z) &= \mathcal{T}^{\text{ex}} p_x^{\text{canb}}(Z) \text{ with } Z = Z^{\text{ex}},\\
p_y^{\text{cana}}(Z) &= \mathcal{T}^{\text{ex}} p_y^{\text{canb}}(Z) \text{ with } Z = Z^{\text{ex}},\\
p_t^{\text{cana}}(Z) &= \mathcal{T}^{\text{ex}} p_t^{\text{canb}}(Z) \text{ with } Z = Z^{\text{ex}},
\end{aligned} \tag{16.5.61}$$

where

$$\mathcal{T}^{\text{ex}} = \exp(q : \chi^{\text{ex}} :). \tag{16.5.62}$$

Behavior of $C_{2,s}^{[1]}(Z)$

According to (5.52) and (5.59) both χ^{en} and χ^{ex} involve $C_{2,s}^{[1]}(Z)$ and its derivatives. Let us explore the behavior of $C_{2,s}^{[1]}(Z)$ for the cases of idealized air-core quadrupoles and REC quadrupoles, which are described in Subsections 5.4 and 5.6, respectively. In both cases $C_{2,s}^{[0]}(Z)$ is proportional to an associated bump function. See (5.27) and (5.33). Therefore, in both cases we are interested in the Z dependence of bump′, the derivative of the bump function. Figures 5.15 and 5.16 display the derivative of the soft-edge bump functions shown in Figures 5.9 and 5.10 for two idealized air-core quadrupoles; and Figures 5.17 and 5.18 display the derivative of the soft-edge bump functions shown in Figures 5.13 and 5.14 for two REC quadrupoles.

Evidently bump′, the derivative of the bump function, falls of quite rapidly beyond the quadrupole body. For example we expect, according to (5.31), that in the case of an idealized air-core quadrupole the function bump′ will fall off like $1/|z|^6$ as $z \to -\infty$. And, according to (5.39), we expect a fall off like $1/|z|^8$ for the case of a REC quadrupole.

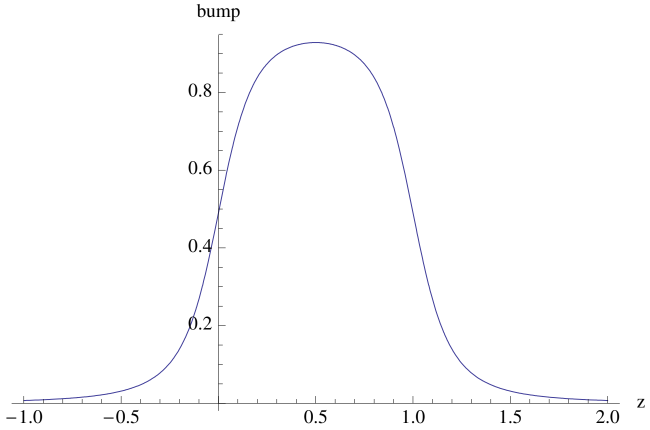


Figure 16.5.15: (Place Holder) Derivative of the soft-edge bump function given by (5.28) and (5.29) when $a = .2$ and $L = 1$, and shown in Figure 5.9.

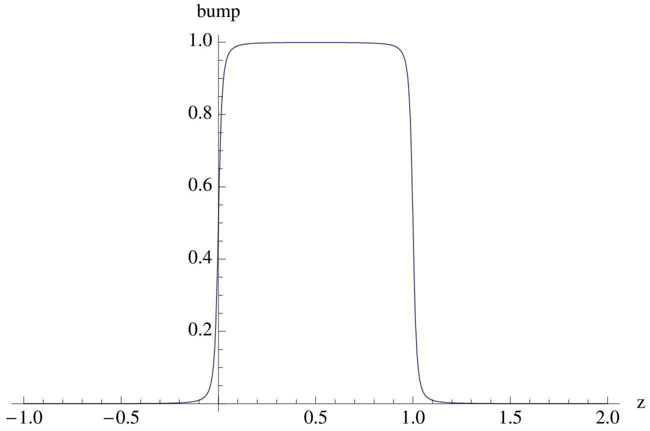


Figure 16.5.16: (Place Holder) Derivative of the soft-edge bump function given by (5.28) and (5.29) when $a = .02$ and $L = 1$, and shown in Figure 5.10.

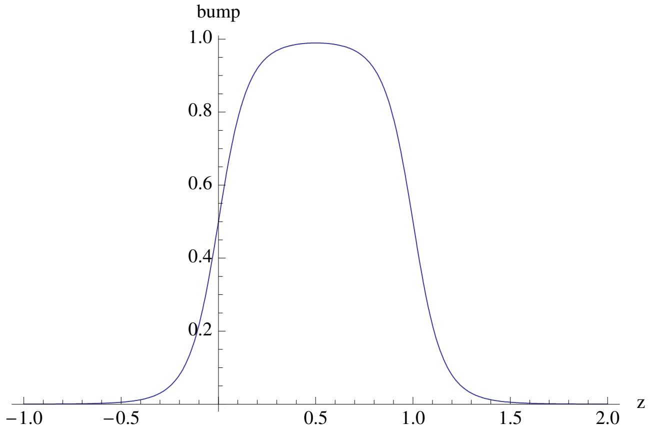


Figure 16.5.17: (Place Holder) Derivative of the soft-edge bump function (5.34) when $r_1 = .2$, $r_2 = .5$, and $L = 1$, and shown in Figure 5.13.

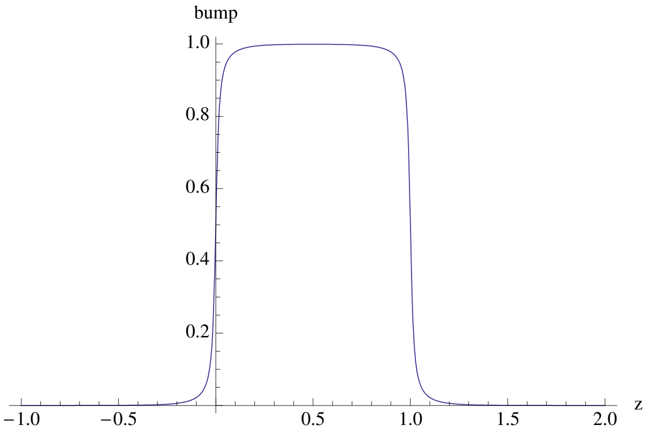


Figure 16.5.18: (Place Holder) Derivative of the soft-edge bump function (5.34) when $r_1 = .02$, $r_2 = .5$, and $L = 1$, and shown in Figure 5.14.

Net Total Map

Let $\mathcal{M}_{\text{en}\rightarrow\text{ex}}$ denote the map obtained by integrating for a quadrupole the design orbit and map equations from $Z = Z^{\text{en}}$ to $Z = Z^{\text{ex}}$ using the Coulomb gauge vector potential of Section 15.5. Then the full net map $\mathcal{M}$ for the quadrupole, including end-field termination effects, is given by the product

$$\mathcal{M} = \mathcal{T}^{\text{en}}\mathcal{M}_{\text{en}\rightarrow\text{ex}}\mathcal{T}^{\text{ex}}. \tag{16.5.63}$$

Discontinuities in Mechanical Momenta Associated with Termination Approximation

As described in Subsection 1.2, there are discontinuities in the mechanical momenta associated with the use of a symplectic termination procedure. Recall (1.30), (1.31), (1.41), and (1.42). Here we will study for quadrupoles the consequences of terminating end fields using the minimum (Poincaré-Coulomb gauge) vector potential. To do so we will need the $m = 2$ and $\alpha = s$ Poincaré-Coulomb gauge vector potential about the expansion point $(0, 0, Z_0)$.

For fun let us compute ${}^{P}\boldsymbol{A}^{2,s}(x, y, z)$ from scratch starting with $\psi_{2,s}(x, y, Z_0 + z)$. From (5.1) we have the result

$$\psi_{2,s}(x, y, Z_0 + z) = 2xy[C^{[0]}_{2,s}(Z_0 + z) - (1/24)(x^2 + y^2)C^{[2]}_{2,s}(Z_0 + z) + \cdots]. \tag{16.5.64}$$

Let $\boldsymbol{B}(\boldsymbol{r}; Z_0)$ be the associated magnet field and employ the notation

$$\boldsymbol{B}(\boldsymbol{r}; Z_0) = \boldsymbol{B}(x, y, z; Z_0) = \boldsymbol{B}(x, y, Z_0 + z). \tag{16.5.65}$$

We then have the relations

$$B_x(x, y, z; Z_0) = \partial_x\psi_{2,s} = 2yC^{[0]}_{2,s}(Z_0 + z) - (1/12)(3x^2y + y^3)C^{[2]}_{2,s}(Z_0 + z) + \cdots , \tag{16.5.66}$$

$$B_y(x, y, z; Z_0) = \partial_y\psi_{2,s} = 2xC^{[0]}_{2,s}(Z_0 + z) - (1/12)(x^3 + 3xy^2)C^{[2]}_{2,s}(Z_0 + z) + \cdots , \tag{16.5.67}$$

$$B_z(x,y,z;Z_0) = \partial_z \psi_{2,s} = 2xy[C_{2,s}^{[1]}(Z_0+z) - (1/24)(x^2+y^2)C_{2,s}^{[3]}(Z_0+z) + \cdots]. \quad (16.5.68)$$

Next expand $\boldsymbol{B}(x,y,Z_0+z)$ in homogeneous polynomials by writing,

$$\boldsymbol{B}(\boldsymbol{r};Z_0) = \boldsymbol{B}^1(\boldsymbol{r};Z_0) + \boldsymbol{B}^2(\boldsymbol{r};Z_0) + \boldsymbol{B}^3(\boldsymbol{r};Z_0) + \cdots . \quad (16.5.69)$$

From (5.65) through (5.67) we see that there are the relations

$$\boldsymbol{B}^1(\boldsymbol{r};Z_0) = 2C_{2,s}^{[0]}(Z_0)(y\boldsymbol{e}_x + x\boldsymbol{e}_y), \quad (16.5.70)$$

$$\boldsymbol{B}^2(\boldsymbol{r};Z_0) = 2C_{2,s}^{[1]}(Z_0)(yz\boldsymbol{e}_x + xz\boldsymbol{e}_y + xy\boldsymbol{e}_z). \quad (16.5.71)$$

Now we may use (15.2.111) to find the results

$$\begin{aligned}\boldsymbol{A}^2(\boldsymbol{r};Z_0) &= -(1/3)[\boldsymbol{r}\times\boldsymbol{B}^1(\boldsymbol{r};Z_0)] \\ &= (-2/3)C_{2,s}^{[0]}(Z_0)[-zx\boldsymbol{e}_x + zy\boldsymbol{e}_y + (x^2-y^2)\boldsymbol{e}_z], \quad (16.5.72)\end{aligned}$$

$$\begin{aligned}\boldsymbol{A}^3(\boldsymbol{r};Z_0) &= -(1/4)[\boldsymbol{r}\times\boldsymbol{B}^2(\boldsymbol{r};Z_0)] \\ &= (-1/2)C_{2,s}^{[1]}(Z_0)[(xy^2-xz^2)\boldsymbol{e}_x + (yz^2-yx^2)\boldsymbol{e}_y + (zx^2-zy^2)\boldsymbol{e}_z]. \quad (16.5.73)\end{aligned}$$

[Note that, in view of the relation(5.5), (5.72) agrees with (15.2.165), as it should.] Finally, we write

$$^{P}\boldsymbol{A}^{2,s}(\boldsymbol{r};Z_0) = \boldsymbol{A}^2(\boldsymbol{r};Z_0) + \boldsymbol{A}^3(\boldsymbol{r};Z_0) + \cdots . \quad (16.5.74)$$

Let us find what this knowledge of $^{P}\boldsymbol{A}^{2,s}$ entails for discontinuities in mechanical momenta. Observe that, according to (5.72) and (5.73), there are the results

$$\boldsymbol{A}^2(x,y,0;Z_0) = 0, \quad (16.5.75)$$

$$\boldsymbol{A}^3(x,y,0;Z_0) = (-1/2)C_{2,s}^{[1]}(Z_0)[(xy^2)\boldsymbol{e}_x + (-yx^2)\boldsymbol{e}_y], \quad (16.5.76)$$

so that

$$^{P}\boldsymbol{A}^{2,s}(x,y,0;Z_0) = (-1/2)C_{2,s}^{[1]}(Z_0)[(xy^2)\boldsymbol{e}_x + (-yx^2)\boldsymbol{e}_y] + \cdots . \quad (16.5.77)$$

We conclude from * that upon entry there are the discontinuity results

$$\begin{aligned}\Delta p_x^{\rm mech} &= q[^{P}A_x^{2,s}(x,y,0;Z^{\rm en})] \\ &= q\{[-(1/2)xy^2]C_{2,s}^{[1]}(Z^{\rm en}) + \cdots\}, \quad (16.5.78)\end{aligned}$$

$$\begin{aligned}\Delta p_y^{\rm mech} &= q[^{P}A_y^{2,s}(x,y,0;Z^{\rm en})] \\ &= q\{[(1/2)x^2y]C_{2,s}^{[1]}(Z^{\rm en}) + \cdots\}. \quad (16.5.79)\end{aligned}$$

Similarly, upon exit, we find from * the discontinuity results

$$\begin{aligned}\Delta p_x^{\rm mech} &= q[^{P}A_x^{2,s}(x,y,0;Z^{\rm ex})] \\ &= q\{[-(1/2)xy^2]C_{2,s}^{[1]}(Z^{\rm ex}) + \cdots\}, \quad (16.5.80)\end{aligned}$$

$$\begin{aligned}\Delta p_y^{\rm mech} &= q[{}^P A_y^{2,s}(x,y,0;Z^{\rm ex})] \\ &= q\{[(1/2)x^2y]C_{2,s}^{[1]}(Z^{\rm ex})+\cdots\}.\end{aligned} \tag{16.5.81}$$

Recall the relations (5.5), (5.27), and (5.33). We see that in all cases the discontinuities are proportional to $Q'(0,0,Z)$, or, equivalently bump$'$, and its derivatives at $Z = Z^{\rm en}$ or $Z = Z^{\rm ex}$. We have already seen examples, in Figures 15 through 18, of how these functions behave (fall off) in the cases of idealized air-core and REC quadrupoles. Moreover, the discontinuities also vanish as the spatial deviations from the z axis (the design orbit) become small.

Exercises

16.5.1. Verify the relations (7.32) through (7.39).

16.5.2. Evidently the second-order portion of $\psi^e(\boldsymbol{r}^d;x_0,z_0)$ as given in (7.61) is composed of the monomials $\xi\eta$ and $\eta\zeta$. Show that these are the only monomials allowed at this order based on symmetry considerations. Verify that each monomial is an harmonic polynomial. Indeed, making the usual correspondence between ξ,η,ζ and x,y,z show, following the harmonic polynomial labeling scheme (U.2.9), that there are the relations

$$\xi\eta = [1/(4i)][\sqrt{32\pi/15}][H_2^2(\boldsymbol{r}) - H_2^{-2}(\boldsymbol{r})], \tag{16.5.82}$$

$$\eta\zeta = [-1/(2i)][\sqrt{8\pi/15}][H_2^1(\boldsymbol{r}) + H_2^{-1}(\boldsymbol{r})]. \tag{16.5.83}$$

Would these relations have been simpler had the polar axis, used to set up spherical polar coordinates, been taken to be the y axis instead of the z axis?

16.5.3. Verify (6.13) through (6.16.

16.5.4. Verify (6.19) through (6.24).

16.5.5. Verify that sgn(z,r) as given by (6.24) becomes the true signum function in the limit $r \to 0$. Verify the relations (6.27) through (6.29).

16.5.6. The purpose of this exercise is to verify that the approximating signum function sgn(z,r_1,r_2) becomes the true signum function in the limit that r_1 goes to zero,

$$\lim_{r_1\to 0} \mathrm{sgn}(z,r_1,r_2) = \mathrm{sgn}(z). \tag{16.5.84}$$

Along the way we will also verify some other expected properties of sgn(z,r_1,r_2).

16.5.7. Verify (6.41) through (6.43).

16.6 Sextupoles and Beyond

A sextupole is a beam-line element whose field is described by a cylindrical harmonic expansion that contains primarily an $m = 3$ term. The simplest way to produce a sextupole field, as the name again suggests, is to properly locate and assign strengths to six monopoles. In the case of a sextupole the sextet of monopoles can be taken to be three doublets rotated successively by 60 degrees. We already know from the work of Section 15.8.2 that a monopole doublet produces an $m = 3$ term. See (15.8.33). And (15.8.34) evaluated with $m = 3$ describes how this term falls off for large $|z|$. We conclude that the on-axis gradient for a monopole sextet falls off as $1/|z|^7$ for large $|z|$. In analogy with the case of dipoles and quadrupoles, we expect that the on-axis gradients for a line of monopole sextets and an idealized air-core sextupole will also fall off as $1/|z|^7$ for large $|z|$.

Moreover, the general pattern is now clear. We expect that the on-axis gradient for an idealized air-core $2m$-pole magnet will fall off for large $|z|$ as $1/|z|^{2m+1}$.

16.7 Lithium Lenses

Ackowledgement

We have benefitted greatly from many conversations with Peter Walstrom, the reading of some of his internal Los Alamos Technical Notes, and the use of his Bessel function routines. We are also grateful to Thomas Mottershead, Filippo Neri, and Peter Walstrom for their many contributions to MARYLIE 5.0.

Bibliography

General References

[1] M. Venturini, "Lie Methods, Exact Map computation, and the Problem of Dispersion in space Charge Dominated Beams", University of Maryland College Park Physics Department Ph.D. thesis (1998).

[2] M. Venturini and A. Dragt, "Accurate Computation of Transfer Maps from Magnetic Field Data", *Nuclear Instruments and Methods* **A427**, p. 387 (1999).

[3] É. Forest, *Beam Dynamics: A New Attitude and Framework*, Harwood Academic Publishers (1998).

[4] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover (1972). Also available on the Web by Googling "abramowitz and stegun 1972".

Solenoids

[5] A. El-Kareh and J. El-Kareh, *Electron Beams, Lenses, and Optics*, Vols. 1 and 2, Academic Press (1970).

[6] A. Dragt, "Numerical third-order transfer map for solenoid", *Nuclear Instruments and Methods in Physics Research* **A298**, p. 441 (1990).

[7] P. W. Hawkes and E. Kasper, *Principles of Electron Optics*, Vols. 1 through 3, Academic Press (1996).

[8] W. Smythe, *Static and Dynamic Electricity*, McGraw-Hill (1939).

Dipoles

[9] P. L. Walstrom, "Dipole-magnet field models based on a conformal map", *Physical Review Special Topics-Accelerators and Beams* **15**, 102401 (2012).

[10] L. N. Brouwer, "Canted-Cosine-Theta Superconducting Accelerator Magnets for High Energy Physics and Ion Beam Cancer Therapy", Ph.D. Thesis, University of California, Berkeley (2015). https://escholarship.org/uc/item/8jp4g75g

[11] M. Bassetti and C. Biscari, "Analytical Formulae for Magnetic Multipoles", *Particle Accelerators* **52**, pp. 221-250 (1996). http://cds.cern.ch/record/1120230/files/p221.pdf

[12] M. Bassetti and C. Biscari, “Cylinder Model of Multipoles”, *Handbook of Accelerator Physics and Engineering*, First Edition, Section 2.2.2, A. Chao and M. Tigner Edit., World Scientific (1999). Unfortunately, subsequent editions of this book no longer contain this material.

Wigggglers

Air-Core and Lambertson Windings

[13] R.P. Avery, B.R. Lambertson, C.D. Pike, PAC 1971 Proceedings, p. 885.

[14] T.G. Godlove, S. Bernal, M. Reiser, Printed Circuit Quadrupole Design, PAC 1995 Proceedings, p. 2117

[15] M. Venturini, Transfer Map for Printed-Circuit Magnetic Quadrupoles, Technical Note, Dept. of Physics, Univ. of Maryland (1995).

Rare Earth Cobalt Magnets

[16] K. Halbach, “Physical and Optical Properties of Rare Earth Cobalt Magnets”, *Nuclear Instruments and Methods* **187**, pp. 109-117 (1981).

Computation of Charged-Particle Beam Transport

[17] A. Dragt et al., *MaryLie 3.0 Users’ Manual* (2003). See www.physics.umd.edu/dsat/.

Chapter 17

Surface Methods for General Straight Beam-Line Elements

17.1 Introduction

Section 15.1 described the need for Taylor expansions of the vector potential $\boldsymbol{A}$ in order to determine the transfer map $\mathcal{M}$. As illustrated in Chapter 16, there are cases in which these Taylor expansions can be found analytically. However, for most cases, all that we can hope to have are magnetic field values determined numerically at points on some regular 3-dimensional grid with the aid of some electromagnetic code.[1] This places us in what might appear to be a hopeless position: it is well known that it is generally difficult to extract reliable information about derivatives from numerical data on a grid. And we want to know about high derivatives! Hildebrand, author of *Introduction to Numerical Analysis*, writes

> *Once an interpolating polynomial $y(x)$ has been determined so that it satisfactorily approximates a given function $f(x)$ over a certain interval I, it may be hoped that the results of differentiating $y(x)$ $\cdots$ will also satisfactorily approximate the corresponding derivative $\cdots$ of $f(x)$. However $\cdots$ we may anticipate the fact that, even though the deviation between $y(x)$ and $f(x)$ will be small throughout the interval, still the slopes of the two curves representing them may differ quite appreciably. Further, it is seen that roundoff errors (or errors of observation) of alternating sign in consecutive ordinates could affect the calculation of the derivative quite strongly if those ordinates were fairly closely spaced $\cdots$. In particular, numerical differentiation should be avoided whenever possible, particularly when the data are empirical and subject to appreciable errors of observation.*

Remarkably, we will find that this problem can be overcome to some satisfactory aberration order with the use of *surface* data.[2] We will fit field data onto some surface, and then use

[1] Alternatively, see Section 17.2, we may have numerically-determined values of the magnetic scalar potential.

[2] The determination of the solution of Laplace's equation in terms of surface data is called the *Dirichlet* (1805-1859) problem. Dirichlet was the thesis advisor of, among others, Kronecker and Lipschitz. It is also interesting to note that he married Rebecka Mendelssohn, one of the sisters of Felix Mendelssohn.

this surface data to compute interior fields. Specifically, and in summary, we will find that surface methods have the following virtues:

- Only functions with known (orthonormal) completeness properties and known (optimal) convergence properties are employed.
- The Maxwell equations are exactly satisfied.
- The results are manifestly analytic in all variables.
- The error is globally controlled. Fields that satisfy the Laplace equation take their extrema on boundaries. Both the exact and computed fields satisfy the Laplace equation. Therefore their difference, the error field, also satisfies the Laplace equation, and must take its extrema on the boundary. But this is precisely where a controlled fit is made. Thus, the error on the boundary is controlled, and the interior error must be even smaller.
- Because fields take their extrema on boundaries, interior values inferred from surface data are relatively insensitive to errors/noise in the surface data. Put another way, the inverse Laplacian (Laplace Green function), which relates interior data to surface data, is *smoothing*. It is this smoothing that we seek to exploit. We will find that the sensitivity to noise in the data decreases rapidly (as some high inverse power of distance) with increasing distance from the surface, and this property improves the accuracy of the high-order interior derivatives needed to compute high-order transfer maps.

In this chapter, devoted to the case of straight beam-line elements, we will develop methods for computing high-order transfer maps based on data provided on a 3-dimensional grid. See Figure 1.1. These methods make it possible to compute realistic transfer maps for real (straight) beam-line elements including all fringe-field and higher-order multipole effects. In Chapter 15 we learned how to characterize magnetic fields in terms of cylindrical harmonics described by on-axis gradients, and also how to determine vector potentials in terms of on-axis gradients. In this chapter we will see how on-axis gradients can in turn be computed from numerical data provided on a 3-dimensional grid. Chapters 18 through 21 will elaborate on these methods and apply them to a variety of straight beam-line elements. In Chapters 22 through 25 we will consider realistic transfer maps for curved beam-line elements.

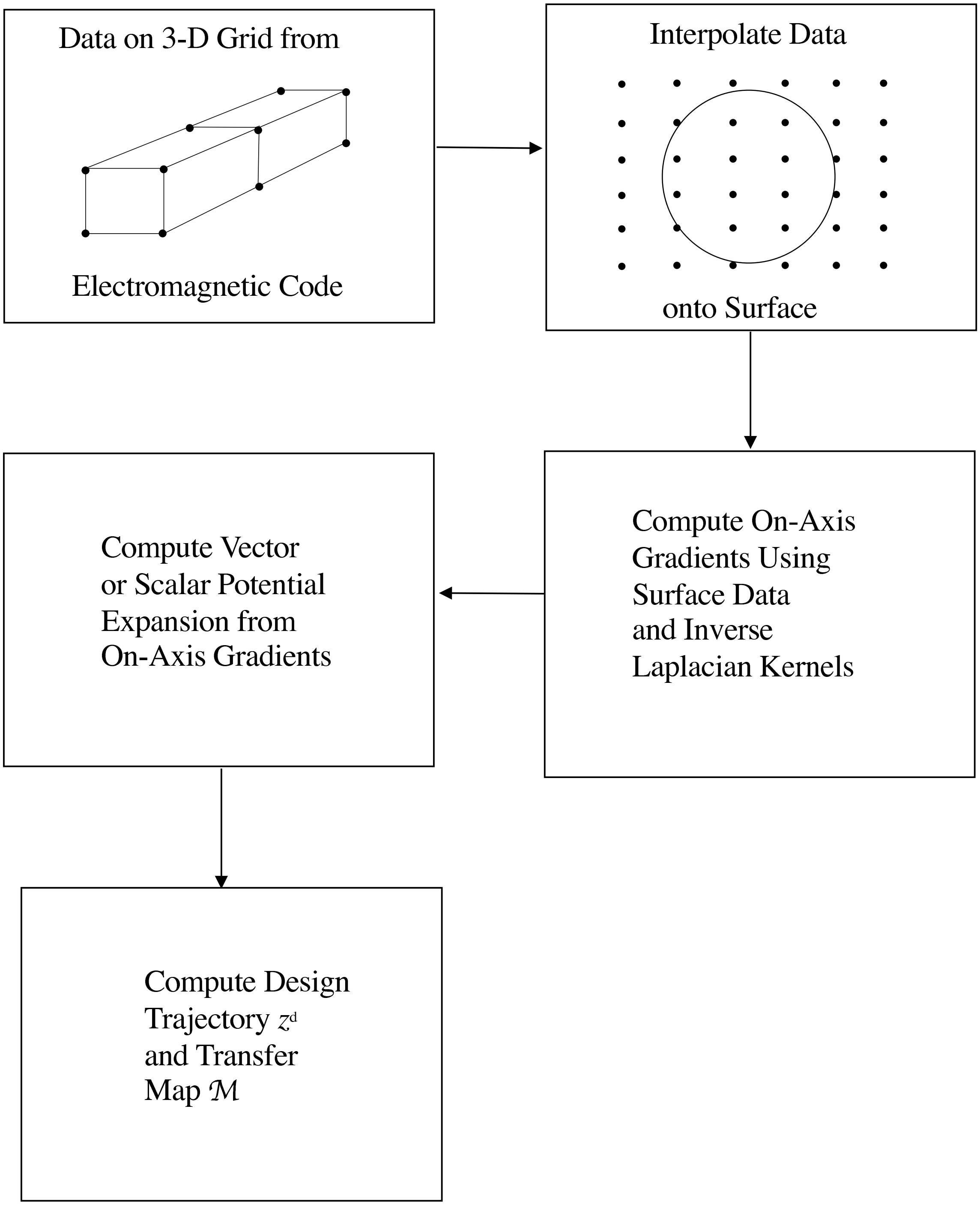

Data on 3-D Grid from
Electromagnetic Code
Interpolate Data
onto Surface
Compute On-Axis Gradients Using Surface Data and Inverse Laplacian Kernels
Compute Vector or Scalar Potential Expansion from On-Axis Gradients
Compute Design Trajectory z^d and Transfer Map $\mathcal{M}$

Figure 17.1.1: Calculation of realistic design trajectory z^d and its associated realistic transfer map $\mathcal{M}$ based on data provided on a 3-dimensional grid for a real beam-line element. Only a few points on the 3-dimensional grid are shown. In this illustration, data from the 3-dimensional grid is interpolated onto the surface of a cylinder with circular cross section, and this surface data is then processed to compute the design trajectory and the transfer map. The use of other surfaces is also possible, and may offer various advantages.

At this point one might wonder about a seemingly simpler approach: Suppose some fine grid of possible initial conditions is laid out in phase space. Next suppose the final conditions associated with these initial conditions are computed numerically, say by integrating Newton's equations with a Lorentz force, using volume magnetic field data interpolated off a three-dimensional grid. Based on the collection of initial and final conditions, make a polynomial expansion of the final conditions in terms of the initial conditions. Announce that these truncated Taylor series, generally six in number assuming a six-dimensional phase space, constitute a (Taylor) transfer map. What could be wrong with that?

There are three reasons why such an approach is problematic:

1. After some reflection, we see that this procedure essentially amounts to high-order numerical differentiation, and therefore Hildebrand's warning still holds. The problem of error associated with high-order numerical differentiation remains.

2. If Newton's equations are integrated, the symplectic symmetry inherent in a Hamiltonian formulation cannot be exploited.

3. Again assuming Newton's equations are integrated, and even in the absence of the error associated with high-order numerical differentiation, the result will not be symplectic if there is a residual magnetic field at the beginning or end of the integration region. See Exercise 6.4.11. Therefore the result may not be suitable for long-term tracking. Perhaps this possible lack of symplecticity could in principle be handled by factorizing the resulting Taylor map into symplectic and nonsymplectic parts, and then using only the symplectic part for any subsequent calculations. See Section 29.1. As a bonus, examination of the size of the nonsymplectic part might give some indication of the error involved in the calculation.

There are also some other approaches that have sometimes been attempted to obtain transfer maps based on 3-d field data on a grid. They are described in Section 17.6. They too involve high-order numerical differentiation, and therefore are unlikely to succeed beyond modest order, at best.

Finally, we mention that there are two other possible ways of determining on-axis gradients that warrant exploration. The first is to infer on-axis gradients from experimental spinning coil data. The second, applicable in the case of air-core magnets, is to compute on-axis gradients based on data describing coil winding geometry and currents flowing in the windings. See Appendix K.

Exercises

17.1.1. This exercise explores some aspects of the Laplace/Poisson equation. We will consider solutions $\psi(x, y, z)$ about some point which, without loss of generality, may be taken to be the origin $\boldsymbol{r} = 0$.

Suppose that ψ is analytic in the Cartesian components of $\boldsymbol{r}$, has at the origin the value

$$\psi(0) = \psi_0, \tag{17.1.1}$$

and is harmonic is some volume V surrounding the origin,

$$\nabla^2\psi(\boldsymbol{r}) = 0. \tag{17.1.2}$$

Introduce spherical coordinates in the usual way and let $Q_{\ell,m,c}(\theta,\phi)$ and $Q_{\ell,m,s}(\theta,\phi)$ denote the functions

$$Q_{\ell,m,c}(\theta,\phi) = \Re[Y_{\ell,m}(\theta,\phi)] = O_{\ell,m}(\theta)\cos(m\phi), \tag{17.1.3}$$

$$Q_{\ell,m,s}(\theta,\phi) = \Im[Y_{\ell,m}(\theta,\phi)] = O_{\ell,m}(\theta)\sin(m\phi). \tag{17.1.4}$$

Here we impose the requirement $m \geq 0$ and make the definitions

$$Q_{\ell,0,s}(\theta,\phi) = 0, \tag{17.1.5}$$

$$O_{\ell,m}(\theta) = (-1)^m\sqrt{[(2\ell+1)(\ell-m)!]/[(4\pi)(\ell+m)!]}P_\ell^m(\cos\theta). \tag{17.1.6}$$

Expand $\psi(x,y,z)$ in a Taylor series about the origin and group terms of like degree so as to yield an expansion in homogeneous polynomials. Show that rewriting this expansion in spherical coordinates gives the result

$$\psi = \psi_0 + \sum_{\ell=1}^{\infty}\sum_{m=0}^{\ell}\sum_{\alpha=c,s} d_{\ell,m,\alpha}[r^\ell Q_{\ell,m,\alpha}(\theta,\phi)]. \tag{17.1.7}$$

Here the quantities $d_{\ell,m,\alpha}$ are arbitrary coefficients and we have enforced the conditions (1.1) and (1.2).

Next, integrate ψ over the surface of a sphere of radius R centered on the origin. Show, recalling the orthogonality properties of the $Y_{\ell,m}$, that using the expansion (1.7) yields the result

$$\int_S \psi dS = 4\pi R^2\psi_0. \tag{17.1.8}$$

Consequently, there is the relation

$$[1/(4\pi R^2]\int_S \psi dS = \psi_0. \tag{17.1.9}$$

The average of ψ over the surface of a sphere equals it value at the center of the sphere. It follows that if $\psi > \psi_0$ at some point on the surface of the sphere, then it must be the case that $\psi < \psi_0$ at some other point on the surface of the sphere, and vice versa, in order for (1.9) to hold. Finally, an analogous result must be true for any expansion point within V. Consequently, ψ has no local minima or maxima, and must take its extrema on the boundary of V.

Suppose we replace the harmonic requirement (1.2) by the condition

$$\nabla^2\psi|_{\boldsymbol{r}=0} = \rho_0. \tag{17.1.10}$$

What happens now? In this case expand $\psi(x,y,z)$ in a Taylor series about the origin through terms of degree 2 and group terms of like degree so as to again yield an expansion in homogeneous polynomials. Show that rewriting this expansion in spherical coordinates gives, through terms of degree 2, the result

$$\psi = \psi_0 + \sum_{m=0}^{1} \sum_{\alpha=c,s} d_{1,m,\alpha}[rQ_{1,m,\alpha}(\theta,\phi)] + (\rho_0/6)r^2 + \sum_{m=0}^{2} \sum_{\alpha=c,s} d_{2,m,\alpha}[r^2 Q_{2,m,\alpha}(\theta,\phi)] + \cdots . \tag{17.1.11}$$

Here the quantities $d_{\ell,m,\alpha}$ are again arbitrary coefficients and we have enforced the conditions (1.1) and (1.10).

Again integrate ψ over the surface of a sphere of radius R centered on the origin. Show that using the expansion (1.11) yields the result

$$\int_S \psi dS = 4\pi R^2 \psi_0 + 4\pi(\rho_0/6)R^4 + O(R^6). \tag{17.1.12}$$

Consequently, there is the relation

$$[1/(4\pi R^2)] \int_S \psi dS = \psi_0 + (\rho_0/6)R^2 + O(R^4). \tag{17.1.13}$$

The average of ψ over the surface of a small sphere equals it value at the center of the sphere, plus a correction of order R^2 that involves ρ_0, plus corrections of order R^4. In lowest order, the difference between the spherical average of ψ and its central value ψ_0 involves ρ_0. For this reason, the quantity ρ_0 is called the *concentration* of ψ at $\boldsymbol{r} = 0$.

17.2 Use of Potential Data on Surface of Circular Cylinder

We will begin our discussion with the use of the surface of a cylinder with circular cross section, and the use of scalar potential data on this surface. This is conceptually the simplest case, and will give us opportunity to develop various needed concepts. Moreover, some electromagnetic codes calculate directly the scalar potential on some regular three-dimensional grid, and this data can be interpolated onto the surface of a cylinder. Therefore, this method can also be of practical use.

Consider a circular cylinder of radius R, centered on the z-axis, fitting within the bore of the beam-line element in question, and extending beyond the fringe-field regions at the ends of the beam-line element. The beam-line element could be any straight element such as a solenoid, quadrupole, sextupole, octupole, etc., or it could be wiggler with no net bending. See Figure 2.1, which illustrates the case of a wiggler. Write

$$\psi(x,y,z) = \psi(\rho,\phi,z), \tag{17.2.1}$$

and suppose $\psi(R,\phi,z)$ is known. Here we have used the coordinates (15.2.12) through (15.2.16). In general, determination of $\psi(R,\phi,z)$ will require interpolation onto a circle of data on a square (or rectangular) grid in x and y for each z value on the grid. See the second frame of Figure 1.1 which depicts a square or rectangular grid in the x,y plane for a fixed z

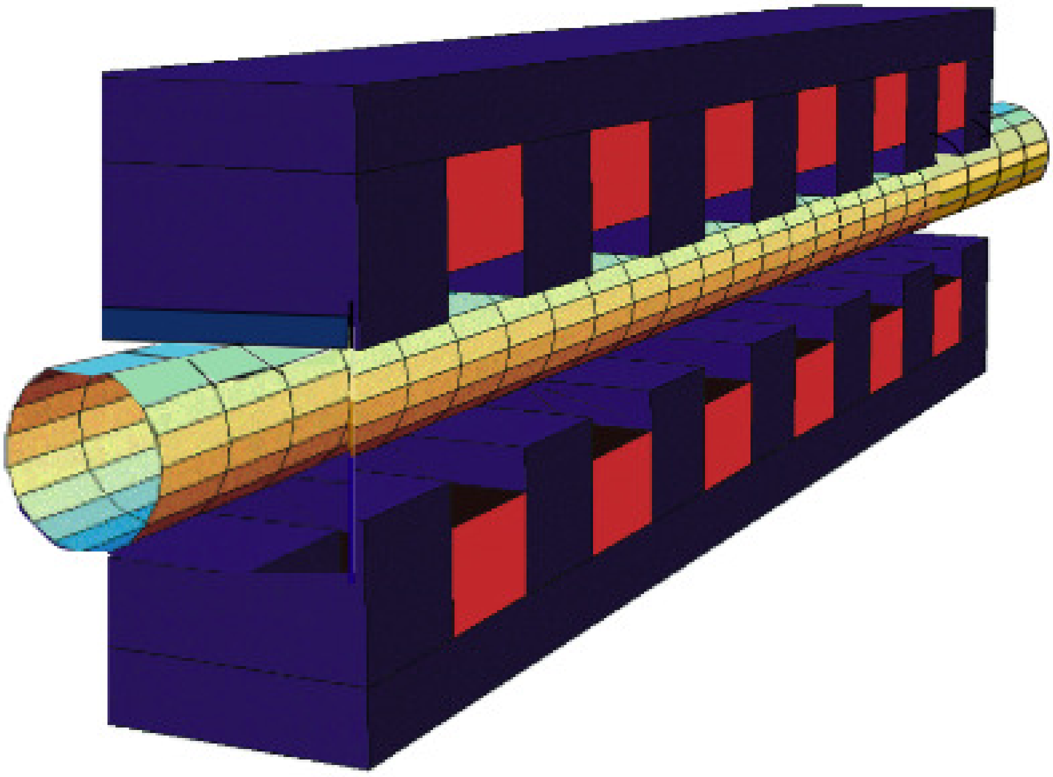


Figure 17.2.1: A circular cylinder of radius R, centered on the z-axis, fitting within the bore of a beam-line element, in this case a wiggler, and extending beyond the fringe-field regions at the ends of the beam-line element.

value on the 3-dimensional grid. Values at data points near the circle are to be interpolated onto the circle.

From this given function $\psi(R,\phi,z)$, obtained by interpolation, form/define the function $\tilde{\tilde{\psi}}(R,m',k')$ by the rule

$$\tilde{\tilde{\psi}}(R,m',k') = [1/(2\pi)]^2 \int_{-\infty}^{\infty} dz \exp(-ik'z) \int_0^{2\pi} d\phi \exp(-im'\phi)\psi(R,\phi,z). \tag{17.2.2}$$

Here we pause a moment to describe our nomenclature and notation: We will refer to the operation of Fourier transforming over the *line* $[-\infty,\infty]$ as performing a *linear* Fourier transform, and the result of this transform will be labeled by a continuous variable usually called k. We will refer to the operation of Fourier transforming over the angular domain $[0,2\pi]$ as performing an *angular* Fourier transform.[3] Moreover, the result of performing an angular Fourier transform will be called a Fourier coefficient, and these coefficients will be labeled by integers such as m and n. Finally, we have used the symbol $\tilde{}$ to denote a linear or angular Fourier transform, and the symbol $\tilde{\tilde{}}$ to denote that both have been performed.

To continue, we know from (15.3.7) that $\psi(R,\phi,z)$ has the representation

$$\psi(R,\phi,z) = \sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} dk G_m(k) \exp(ikz) \exp(im\phi) I_m(kR). \tag{17.2.3}$$

Employing this representation in (2.2) and performing the indicated integrations give the result

$$\tilde{\tilde{\psi}}(R,m',k') = G_{m'}(k') I_{m'}(k'R), \tag{17.2.4}$$

from which we conclude that

$$G_m(k) = \tilde{\tilde{\psi}}(R,m,k)/I_m(kR). \tag{17.2.5}$$

This relation for $G_m(k)$ can now be employed in (15.3.15) to give the result

$$C_m^{[n]}(z) = i^n (1/2)^{|m|} (1/|m|!) \int_{-\infty}^{\infty} dk [k^{n+|m|}/I_m(kR)] \tilde{\tilde{\psi}}(R,m,k) \exp(ikz). \tag{17.2.6}$$

We have found an expression for the generalized on-axis gradients in terms of potential data on the surface of the cylinder. Equation (2.6) may be viewed as the convolution of Fourier surface data $\tilde{\tilde{\psi}}(R,m,k)$ with the *inverse Laplacian* kernel $[k^{n+|m|}/I_m(kR)]$. Moreover, this kernel has a very desirable property. The Bessel functions $I_m(kR)$ have the asymptotic behavior

$$|I_m(kR)| \sim \exp(|k|R)/\sqrt{2\pi|k|R} \text{ as } |k| \to \infty. \tag{17.2.7}$$

[3] Joseph Fourier (1768-1830) was a student of Lagrange. Fourier was the first to make extensive use of the trigonometric series that bear his name, and to make the claim that they could be used to represent arbitrary functions. This claim his elders and contemporaries found hard to believe. In reviewing one of his fundamental papers on the theory of heat that employed these series the referees Lagrange, Laplace, Legendre, and others complained that $\cdots$ *the manner in which the author arrives at these equations is not exempt of difficulties and that his analysis to integrate them still leaves something to be desired on the score of generality and even rigor.* As a result, the paper was not published.

Since $I_m(kR)$ appears in the denominator of (2.6), we see that the integrand is exponentially damped for large $|k|$. Now suppose there is uncorrelated point-to-point noise in the surface data. Such noise will result in anomalously large $|k|$ contributions to the $\tilde{\tilde{\psi}}(R, m, k)$. But, because of the exponential damping arising from $I_m(kR)$ in the denominator, the effect of this noise is effectively filtered out. Moreover, this filtering action is improved by making R as large as possible. This filtering, or *smoothing*, feature will be discussed in more detail in Chapter 18.

17.3 Use of Field Data on Surface of Circular Cylinder

All three-dimensional electromagnetic codes calculate all three components of the field on some three-dimensional grid. Also, such data is in principle available from actual field measurements. In this section we will describe how to compute the on-axis gradients from field data.

Again we will employ a cylinder of radius R centered on the z axis. Suppose the magnetic field $\boldsymbol{B}(x, y, z)$ is interpolated onto the surface of the cylinder using values at the grid points near the surface. Next, from the values on the surface, compute $B_\rho(x, y, z) = B_\rho(R, \phi, z)$, the component of $\boldsymbol{B}(x, y, z)$ *normal* to the surface. We will now see how to compute the generalized gradients from a knowledge of $B_\rho(x, y, z) = B_\rho(R, \phi, z)$.

From this known function form the functions $\tilde{B}_\rho(R, m', z)$ and $\tilde{\tilde{B}}_\rho(R, m', k')$ by the rules

$$\tilde{B}_\rho(R, m', z) = [1/(2\pi)] \int_0^{2\pi} d\phi \exp(-im'\phi) B_\rho(R, \phi, z), \tag{17.3.1}$$

$$\tilde{\tilde{B}}_\rho(R, m', k') = [1/(2\pi)] \int_{-\infty}^{\infty} dz \exp(-ik'z) \tilde{B}_\rho(R, m', z). \tag{17.3.2}$$

Note that we may also directly write that

$$\tilde{\tilde{B}}_\rho(R, m', k') = [1/(2\pi)]^2 \int_{-\infty}^{\infty} dz \exp(-ik'z) \int_0^{2\pi} d\phi \exp(-im'\phi) B_\rho(R, \phi, z), \tag{17.3.3}$$

and the indicated integrations may be performed in either order. We also know that

$$B_\rho(R, \phi, z) = [\partial_\rho \psi(\rho, \phi, z)]|_{\rho=R}, \tag{17.3.4}$$

from which it follows, using the representation (15.3.7), that

$$B_\rho(R, \phi, z) = \sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} dk G_m(k) \exp(ikz) \exp(im\phi) k I'_m(kR). \tag{17.3.5}$$

Now substitute (3.5) into the right side of (3.3) and perform the indicated integrations to get the result

$$\tilde{\tilde{B}}_\rho(R, m', k') = G_{m'}(k') k' I'_{m'}(k'R), \tag{17.3.6}$$

from which it follows that

$$G_m(k) = \tilde{\tilde{B}}_\rho(R, m, k)/[kI'_m(kR)]. \tag{17.3.7}$$

This relation for $G_m(k)$ can be employed in (15.3.15) to give the result

$$C_m^{[n]}(z) = i^n(1/2)^{|m|}(1/|m|!)\int_{-\infty}^{\infty} dk[k^{n+|m|-1}/I'_m(kR)]\tilde{\tilde{B}}_\rho(R, m, k)\exp(ikz). \tag{17.3.8}$$

We have found an expression for the generalized on-axis gradients in terms of field data (normal component) on the surface of the cylinder. Moreover, this expression again has the smoothing property since the denominator functions $I'_m(kR)$ also have the asymptotic behavior (2.7) and therefore also provide exponential damping,

$$|I'_m(kR)| \sim \exp(|k|R)/\sqrt{2\pi|k|R} \text{ as } |k| \to \infty. \tag{17.3.9}$$

For future use it is also convenient to have explicit formulas for the $C_{m,\alpha}^{[n]}(z)$. Motivated by (15.3.28) and (15.3.31), define quantities $\tilde{\tilde{B}}_\rho^\alpha(R, m', k')$ and $\tilde{B}_\rho^\alpha(R, m', z)$ with $m' \geq 1$ by the rules

$$\tilde{\tilde{B}}_\rho^s(R, m', k') = i[\tilde{\tilde{B}}_\rho(R, m', k') - \tilde{\tilde{B}}_\rho(R, -m', k')], \tag{17.3.10}$$

$$\tilde{\tilde{B}}_\rho^c(R, m', k') = [\tilde{\tilde{B}}_\rho(R, m', k') + \tilde{\tilde{B}}_\rho(R, -m', k')], \tag{17.3.11}$$

$$\tilde{B}_\rho^s(R, m', z) = i[\tilde{B}_\rho(R, m', z) - \tilde{B}_\rho(R, -m', z)], \tag{17.3.12}$$

$$\tilde{B}_\rho^c(R, m', z) = [\tilde{B}_\rho(R, m', z) + \tilde{B}_\rho(R, -m', z)]. \tag{17.3.13}$$

Then we have the results

$$\tilde{\tilde{B}}_\rho^\alpha(R, m', k') = [1/(2\pi)]\int_{-\infty}^{\infty} dz\exp(-ik'z)\tilde{B}_\rho^\alpha(R, m', z) \tag{17.3.14}$$

with

$$\tilde{B}_\rho^s(R, m', z) = (1/\pi)\int_0^{2\pi} d\phi\sin(m'\phi)B_\rho(R, \phi, z), \tag{17.3.15}$$

$$\tilde{B}_\rho^c(R, m', z) = (1/\pi)\int_0^{2\pi} d\phi\cos(m'\phi)B_\rho(R, \phi, z). \tag{17.3.16}$$

And, in accord with (15.3.35) and (15.3.36), for $m' = 0$ make the definitions

$$\tilde{\tilde{B}}_\rho^s(R, m' = 0, k') = 0, \tag{17.3.17}$$

$$\tilde{\tilde{B}}_\rho^c(R, m' = 0, k') = \tilde{\tilde{B}}_\rho(R, m' = 0, k'), \tag{17.3.18}$$

$$\tilde{B}_\rho^s(R, m' = 0, z) = 0, \tag{17.3.19}$$

$$\tilde{B}_\rho^c(R, m' = 0, z) = \tilde{B}_\rho(R, m' = 0, z). \tag{17.3.20}$$

Then we have the further results

$$\tilde{\tilde{B}}_\rho^c(R, m' = 0, k') = [1/(2\pi)]\int_{-\infty}^{\infty} dz\exp(-ik'z)\tilde{B}_\rho^c(R, m' = 0, z), \tag{17.3.21}$$

$$\tilde{B}^c_\rho(R, m' = 0, z) = \tilde{B}_\rho(R, m' = 0, z) = [1/(2\pi)] \int_0^{2\pi} d\phi B_\rho(R, \phi, z). \tag{17.3.22}$$

Note that the quantities $\tilde{B}^\alpha_\rho(R, m', z)$ are real. Correspondingly, we see form (3.14) that the real part of $\tilde{\tilde{B}}^\alpha_\rho(R, m', k')$ is even in k and the imaginary part is odd in k.

With these definitions in hand, we are ready to state the final results:

$$C^{[n]}_{m,\alpha}(z) = i^n(1/2)^m(1/m!) \int_{-\infty}^{\infty} dk[k^{n+m-1}/I'_m(kR)]\tilde{\tilde{B}}^\alpha_\rho(R, m, k)\exp(ikz) \tag{17.3.23}$$

for $m > 0$, and

$$C^{[n]}_{m=0,s}(z) = 0, \tag{17.3.24}$$

$$C^{[n]}_{m=0,c}(z) = C^{[n]}_0(z) = i^n \int_{-\infty}^{\infty} dk[k^{n-1}/I'_0(kR)]\tilde{\tilde{B}}^c_\rho(R, m = 0, k)\exp(ikz). \tag{17.3.25}$$

We close this section with the remark that if one wishes to extract the $C^{[n]}_0(z)$ (monopole) on-axis gradients from field data, it may be preferable to use the longitudinal component $B_z(R, \phi, z)$ on the surface of the cylinder rather than the normal component $B_\rho(R, \phi, z)$.[4] See Section 19.2.

17.4 Use of Field Data on Surface of Elliptical Cylinder

17.4.1 Background

In the previous two sections we employed a cylinder with circular cross section, and observed mathematically that it is desirable for error insensitivity to use a cylinder with a large radius R. Physically, this is because we want the data points to be as far from the axis as possible since the effect of inhomogeneities (noise) in the data decays with distance from the inhomogeneity. Evidently the use of a large circular cylinder is optimal for beam-line elements with a circular bore. However, for dipoles or wigglers with small gaps and wide pole faces, use of a cylinder with elliptical cross section should give improved error insensitivity. See Figure 4.1. In this section we will set up the machinery required for the use of elliptical cylinders, and apply it to the calculation of on-axis gradients based on field data.

We will see that the use of elliptic cylinders requires a knowledge of Mathieu functions. Since these functions may well be relatively unfamiliar to the reader, considerable effort will be devoted to describing their properties.

For brevity, we will omit treatment of the related case where potential data is used on the surface of the elliptic cylinder. The reader should be able to solve this simpler problem based on the work of the current section and what was done in Section 14.2.

[4] Note that in any case we only need the $C^{[n]}_0(z)$ with $n \geq 1$ because they are what is required to compute the vector potential. See (15.5.32) through (15.5.34). Thus, (3.6) is well defined for all values of m and n of physical interest.

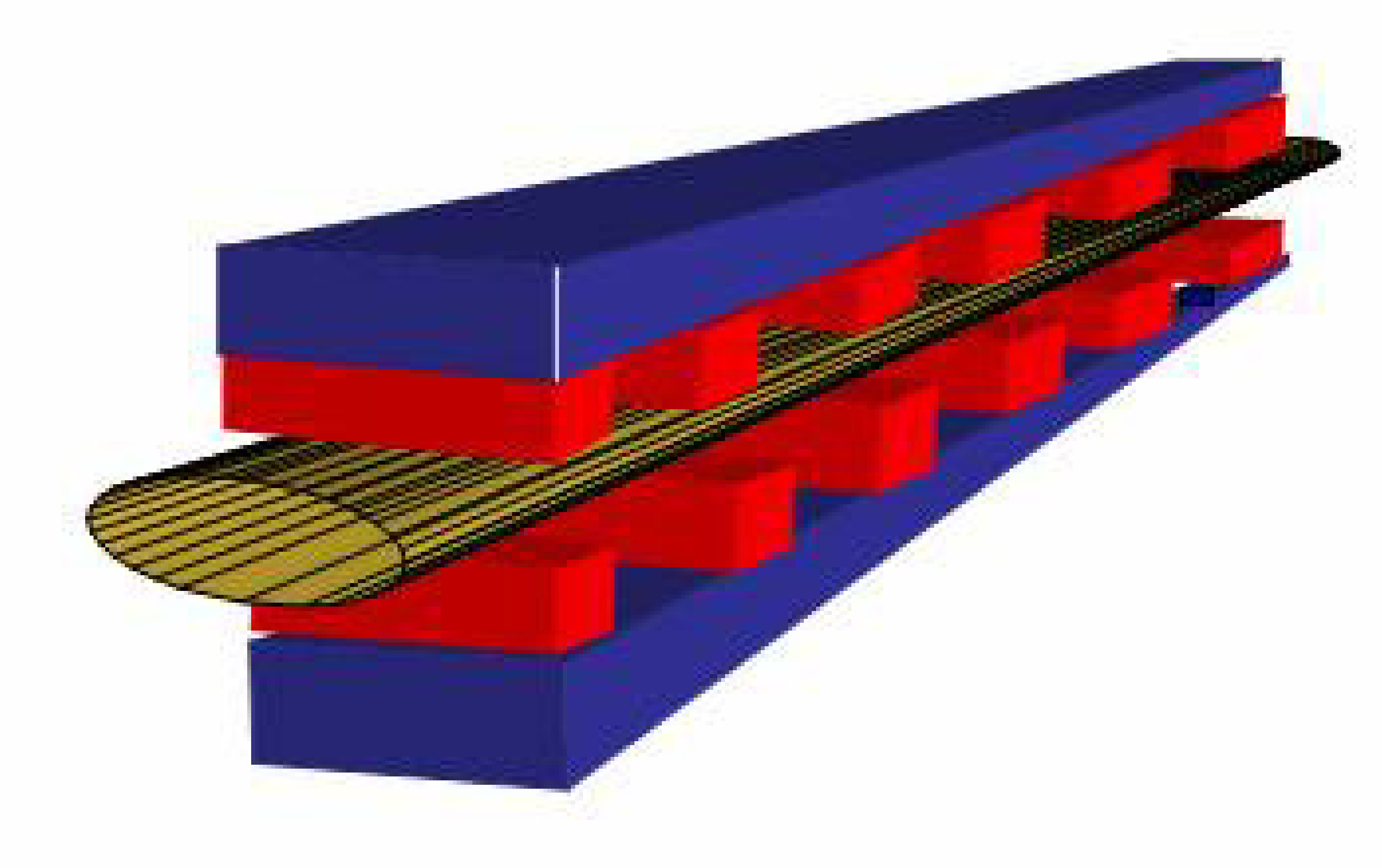


Figure 17.4.1: An elliptical cylinder, centered on the z-axis, fitting within the bore of a wiggler, and extending beyond the fringe-field regions at the ends of the wiggler.

17.4.2 Elliptic Coordinates

Elliptic coordinates in the x, y plane are described by the relations

$$x = f\cosh(u)\cos(v), \tag{17.4.1}$$

$$y = f\sinh(u)\sin(v). \tag{17.4.2}$$

Contours of constant u, with $u \in [0, \infty]$, are nested ellipses with common foci located at $(x; y) = (\pm f; 0)$. Contours of constant v, with $v \in [0, 2\pi]$, are hyperbolae. Together these contours form an orthogonal coordinate system. See Figure 4.2. Data is to be interpolated onto the ellipse whose cross section is that of the elliptical cylinder of Figure 4.1. See Figure 4.3.

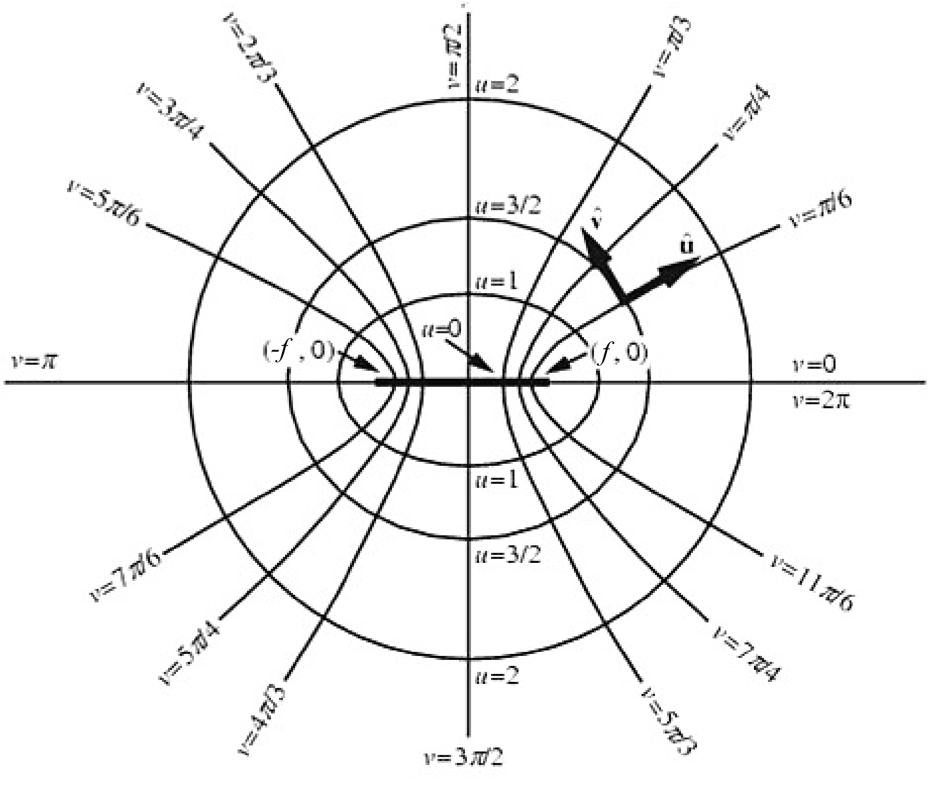


Figure 17.4.2: Elliptical coordinates showing contours of constant u and constant v.

For our work we will need the unit vector $\hat{\boldsymbol{e}}_u$, the unit vector (outwardly) normal to the surface of the elliptical cylinder. Write

$$\begin{aligned}\boldsymbol{r} &= x\hat{\boldsymbol{e}}_x + y\hat{\boldsymbol{e}}_y + z\hat{\boldsymbol{e}}_z \\ &= f\cosh(u)\cos(v)\hat{\boldsymbol{e}}_x + f\sinh(u)\sin(v)\hat{\boldsymbol{e}}_y + z\hat{\boldsymbol{e}}_z.\end{aligned} \tag{17.4.3}$$

Then, by definition, we have the result

$$\begin{aligned}\hat{\boldsymbol{e}}_u &= (\partial\boldsymbol{r}/\partial u)/||(\partial\boldsymbol{r}/\partial u)|| \\ &= [\sinh(u)\cos(v)\hat{\boldsymbol{e}}_x + \cosh(u)\sin(v)\hat{\boldsymbol{e}}_y]/[\cosh^2(u) - \cos^2(v)]^{1/2}.\end{aligned} \tag{17.4.4}$$

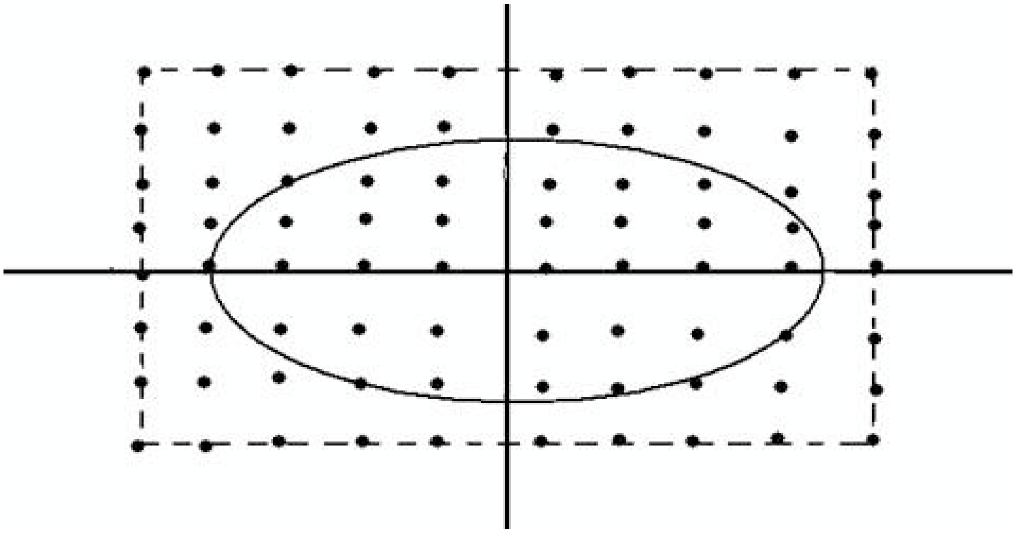


Figure 17.4.3: A square or rectangular grid in the x,y plane for a fixed z value on the 3-dimensional grid. Values at data points near the ellipse are to be interpolated onto the ellipse.

It is also convenient to employ the complex variables

$$\zeta = x + iy, \tag{17.4.5}$$

and

$$w = u + iv. \tag{17.4.6}$$

In these variables, the relations (4.1) and (4.2) can be written in the more compact form

$$\zeta = f\cosh(w). \tag{17.4.7}$$

[For a discussion of the analytic properties of $\zeta(w)$ and its inverse $w(\zeta)$, see Exercise 4.2.] Form differentials of both sides of (4.7). Doing so gives the result

$$dx + idy = f\sinh(w)(du + idv) \tag{17.4.8}$$

and the complex conjugate result

$$dx - idy = f\sinh(\bar{w})(du - idv). \tag{17.4.9}$$

Now form the product of (4.8) and (4.9) to get the transverse line-element relation

$$\begin{aligned} ds_\perp^2 &= dx^2 + dy^2 = f^2\sinh(u+iv)\sinh(u-iv)(du^2+dv^2) \\ &= f^2[\cosh^2(u) - \cos^2(v)](du^2+dv^2). \end{aligned} \tag{17.4.10}$$

From this relation we infer the results

$$B_u = \hat{\boldsymbol{e}}_u \cdot \boldsymbol{B} = (\nabla\psi)_u = (1/f)[\cosh^2(u) - \cos^2(v)]^{-1/2}(\partial\psi/\partial u), \tag{17.4.11}$$

$$\nabla^2\psi = (1/f^2)[\cosh^2(u) - \cos^2(v)]^{-1}[(\partial_u)^2 + (\partial_v)^2]\psi + (\partial_z)^2\psi. \tag{17.4.12}$$

17.4.3 Mathieu Equations

Let us seek to construct harmonic functions of the form

$$\psi \sim P(u)Q(v)\exp(ikz) \tag{17.4.13}$$

where the functions P and Q are yet to be determined. Employing the Ansatz (4.13) in Laplace's equation and use of (4.12) yields the requirement

$$[(\partial_u)^2 + (\partial_v)^2][P(u)Q(v)] = k^2 f^2[\cosh^2(u) - \cos^2(v)]P(u)Q(v). \tag{17.4.14}$$

We also observe that there is the trigonometric identity

$$\cosh^2(u) - \cos^2(v) = (1/2)[\cosh(2u) - \cos(2v)] \tag{17.4.15}$$

so that the requirement (4.14) can be rewritten in the form

$$[(\partial_u)^2 + (\partial_v)^2][P(u)Q(v)] = (k^2 f^2/4)[2\cosh(2u) - 2\cos(2v)]P(u)Q(v). \tag{17.4.16}$$

Upon dividing both sides by PQ, (4.16) becomes

$$(1/P)(\partial_u)^2 P + (1/Q)(\partial_v)^2 Q = (k^2 f^2/4)[2\cosh(2u) - 2\cos(2v)], \tag{17.4.17}$$

from which it follows that

$$(1/P)(\partial_u)^2 P - (k^2 f^2/4)[2\cosh(2u)] = -(1/Q)(\partial_v)^2 Q - (k^2 f^2/4)[2\cos(2v)]. \tag{17.4.18}$$

Therefore, there is a common *separation* constant a such that

$$(1/P)(\partial_u)^2 P - (k^2 f^2/4)[2\cosh(2u)] = a \tag{17.4.19}$$

and

$$-(1/Q)(\partial_v)^2 Q - (k^2 f^2/4)[2\cos(2v)] = a. \tag{17.4.20}$$

Correspondingly, P and Q must satisfy the ordinary and linear differential equations

$$d^2P/du^2 - [a - 2q\cosh(2u)]P = 0, \tag{17.4.21}$$

$$d^2Q/dv^2 + [a - 2q\cos(2v)]Q = 0, \tag{17.4.22}$$

where

$$q = -k^2 f^2/4. \tag{17.4.23}$$

Equation (4.22) for Q is called the *Mathieu* equation, and Equation (4.21) for P is called the *modified Mathieu* equation.[5]

[5] We remark that many, and probably the majority, of the special functions ordinarily encountered in Mathematical Physics are particular cases of the hypergeometric function. The Mathieu functions do not fall in this category. In some sense, they are *more transcendental* than the hypergeometric function.

17.4.4 Periodic Mathieu Functions and Separation Constants

For our purposes, we will need solutions $Q(v)$ of (4.22) that are *periodic* with period 2π. See Figure 4.2. Such solutions exist only for certain specific values of the separation constant a. These values are called $a_n(q)$ for $n = 0, 1, 2, 3, \cdots$ and $b_n(q)$ for $n = 1, 2, 3, \cdots$.[6] The functions $a_n(q)$ and $b_n(q)$ are all *real* for real values of q. As the notation indicates, their values depend on q (and on n). For small q they have expansions of the form

$$a_0(q) = -(1/2)q^2 + (7/128)q^4 + \cdots\,, \tag{17.4.24}$$

$$a_1(q) = 1 + q - (1/8)q^2 - (1/64)q^3 - (1/1536)q^4 + \cdots\,, \tag{17.4.25}$$

$$a_2(q) = 4 + (5/12)q^2 - (763/13824)q^4 + \cdots\,, \text{ etc.;} \tag{17.4.26}$$

$$b_1(q) = 1 - q - (1/8)q^2 + (1/64)q^3 - (1/1536)q^4 + \cdots\,, \tag{17.4.27}$$

$$b_2(q) = 4 - (1/12)q^2 + (5/13824)q^4 + \cdots\,, \text{ etc.} \tag{17.4.28}$$

In each case the leading (the q independent) term is n^2.

Note that, according to (4.23), for our purposes we are interested in negative, and possibly quite negative, values of q.[7] Figures 4.4 and 4.5 display the first few $a_n(q)$ and $b_n(q)$ for negative values of q. Observe that, as $q \to -\infty$, the quantities $a_{2m}(q)$ and $a_{2m+1}(q)$, for $m = 0, 1, 2, 3, \cdots$, tend to agree. Similarly, for large negative q, the quantities $b_{2m+1}(q)$ and $b_{2m+2}(q)$, for $m = 0, 1, 2, 3, \cdots$, tend to agree. Indeed, it can be shown that there is the asymptotic behavior

$$\begin{aligned}
&a_{2m}(q) \sim a_{2m+1}(q)\\
&\sim 2q + (8m+2)(-q)^{1/2} - (1/4)(8m^2 + 4m + 1)\\
&-(1/32)(4m^2 + 2m + 1)(4m + 1)(-q)^{-1/2} + O(1/q)\\
&\text{ as } q \to -\infty \text{ for } m = 0, 1, 2, 3, \cdots\,,
\end{aligned} \tag{17.4.29}$$

$$\begin{aligned}
&b_{2m+1}(q) \sim b_{2m+2}(q)\\
&\sim 2q + (8m+6)(-q)^{1/2} - (1/4)(8m^2 + 12m + 5)\\
&-(1/32)(4m^2 + 6m + 3)(4m + 3)(-q)^{-1/2} + O(1/q)\\
&\text{ as } q \to -\infty \text{ for } m = 0, 1, 2, 3, \cdots\,.
\end{aligned} \tag{17.4.30}$$

We also remark, in passing, that there are the relations

$$a_n(-q) = a_n(q) \text{ for } n \text{ even,} \tag{17.4.31}$$

$$b_n(-q) = b_n(q) \text{ for } n \text{ even,} \tag{17.4.32}$$

$$a_n(-q) = b_n(q) \text{ for } n \text{ odd,} \tag{17.4.33}$$

$$b_n(-q) = a_n(q) \text{ for } n \text{ odd.} \tag{17.4.34}$$

[6] The reader might find confusing the use of the symbols a, a_n, and b_n to denote separation constants. We agree, but it is standard in the Mathieu-equation literature.

[7] Unfortunately for our purposes, the Mathieu function literature treats primarily the $q > 0$ case because this is the case that arises in the solution of the wave equation. See Exercise 4.1.

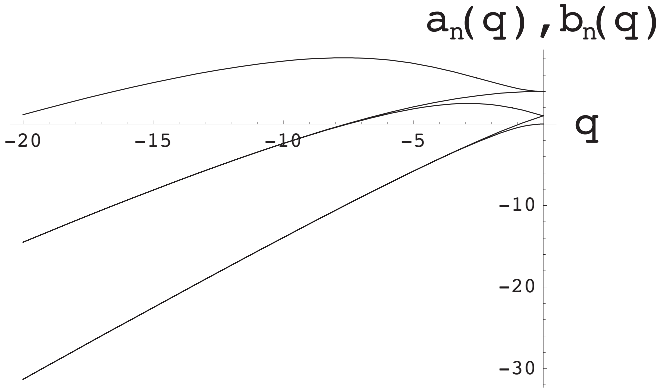


Figure 17.4.4: The functions $a_0(q)$ through $a_2(q)$ and $b_1(q)$ and $b_2(q)$ for negative values of q.

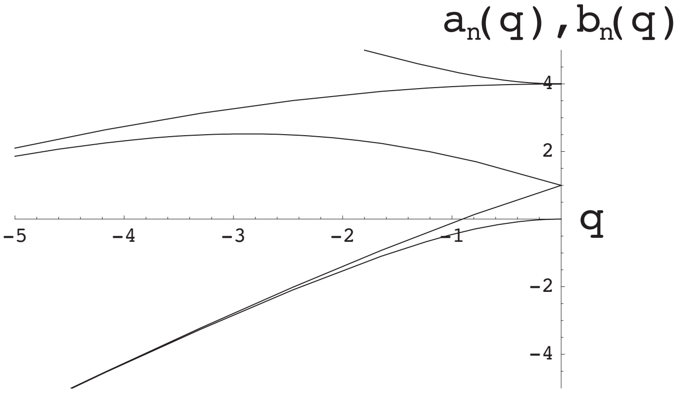


Figure 17.4.5: An enlargement of a portion of Figure 4.4. For q fixed and slightly negative, the curves, in order of increasing value, are $a_0(q)$, $a_1(q)$, $b_1(q)$, $b_2(q)$, and $a_2(q)$. See (4.24) through (4.28). Note that the pair $a_0(q)$ and $a_1(q)$ tends to merge for large negative q, as does the pair $b_1(q)$ and $b_2(q)$. Similarly, although not shown in this figure, the pair $a_2(q)$ and $a_3(q)$ tends to merge as does the pair $b_3(q)$ and $b_4(q)$, etc. See (4.29) and (4.30).

The solutions associated with the separation constants $a = a_n(q)$ are called $\mathrm{ce}_0(v,q)$, $\mathrm{ce}_1(v,q)$, $\mathrm{ce}_2(v,q)$, $\mathrm{ce}_3(v,q)$ $\cdots$. They are *even* functions of v and, in the small q limit, are proportional to the functions 1, $\cos(v)$, $\cos(2v)$, $\cos(3v)$, $\cdots$. The solutions associated with the separation constants $a = b_n(q)$ are called $\mathrm{se}_1(v,q)$, $\mathrm{se}_2(v,q)$, $\mathrm{se}_3(v,q)$, $\cdots$. They are *odd* functions of v and, in the small q limit, are proportional to the functions $\sin(v)$, $\sin(2v)$, $\sin(3v)$, $\cdots$.[8] Indeed, the $\mathrm{ce}_n(v,q)$ and $\mathrm{se}_n(v,q)$ are normalized so that in the limit $q \to 0$ there are the relations

$$\mathrm{ce}_0(v,0) = 1/\sqrt{2}, \tag{17.4.35}$$

$$\mathrm{ce}_n(v,0) = \cos(nv) \text{ for } n \geq 1, \tag{17.4.36}$$

$$\mathrm{se}_n(v,0) = \sin(nv) \text{ for } n \geq 1. \tag{17.4.37}$$

Moreover, like their trigonometric counterparts, the functions $\mathrm{ce}_n(v,q)$ and $\mathrm{se}_n(v,q)$ form a complete set over the interval $[0, 2\pi]$. In fact, they form a complete orthogonal set and are normalized so that

$$\int_0^{2\pi} dv\ \mathrm{ce}_m(v,q)\ \mathrm{ce}_n(v,q) = \pi\delta_{mn}, \tag{17.4.38}$$

$$\int_0^{2\pi} dv\ \mathrm{se}_m(v,q)\ \mathrm{se}_n(v,q) = \pi\delta_{mn}, \tag{17.4.39}$$

$$\int_0^{2\pi} dv\ \mathrm{ce}_m(v,q)\ \mathrm{se}_n(v,q) = 0. \tag{17.4.40}$$

Apart from $\mathrm{ce}_0(v,q)$, this normalization is like that of their trigonometric counterparts. See (4.35) through (4.37). Finally, again like their trigonometric counterparts, it can be shown that the functions $\mathrm{ce}_n(v,q)$ and $\mathrm{se}_n(v,q)$ have n zeroes in the half-open interval $v \in [0, \pi)$.

As noted earlier, we are primarily interested in the case $q \leq 0$. However we note for the record that, in concert with the relations (4.31) through (4.34), there are the relations

$$ce_{2n}(v,-q) = (-1)^n ce_{2n}(\pi/2 - v, q), \tag{17.4.41}$$

$$ce_{2n+1}(v,-q) = (-1)^n se_{2n+1}(\pi/2 - v, q), \tag{17.4.42}$$

$$se_{2n+1}(v,-q) = (-1)^n ce_{2n+1}(\pi/2 - v, q), \tag{17.4.43}$$

$$se_{2n+2}(v,-q) = (-1)^n se_{2n+2}(\pi/2 - v, q). \tag{17.4.44}$$

We will shortly present figures that display the first few $\mathrm{ce}_n(v,q)$ and $\mathrm{se}_n(v,q)$ as functions of v. Before doing so it is useful to look more closely at the terms appearing in the Mathieu equations. Inspired by both the analogy to Schrödinger's equation and the harmonic oscillator, rewrite (4.22) in the form

$$d^2Q/dv^2 - \lambda(v,q)Q = 0 \tag{17.4.45}$$

where

$$\lambda(v,q) = -[a - 2q\cos(2v)]. \tag{17.4.46}$$

[8] Note that, unlike their trigonometric counterparts $\cos(nv)$ and $\sin(nv)$, the functions $\mathrm{ce}_n(v,q)$ and $\mathrm{se}_n(v,q)$ do not satisfy the *same* differential equation. This is because $a_n(q) \neq b_n(q)$.

From a Schrödinger perspective, we may view $Q(v)$ as the wave function and $\lambda(v, q)$ as the 'potential'. In the harmonic oscillator analogy, we may view Q as the oscillator coordinate and $-\lambda(v, q)$ as the instantaneous square of the time (v) dependent frequency. With this background in mind, Figure 4.6 shows $\lambda(v, q = -2)$ for various n values with $a = a_n(q)$. These are the potentials appropriate to the $\text{ce}_n(v, q)$. Similarly, Figure 4.7 shows $\lambda(v, q = -2)$ for various n values with $a = b_n(q)$. These are the potentials appropriate to the $\text{se}_n(v, q)$. According to (4.45), understood in the harmonic oscillator analogy, when $\lambda < 0$ we expect oscillatory behavior; and when $\lambda > 0$ we expect exponentially growing or decaying behavior. From the Schrödinger perspective, the region where $\lambda < 0$ is an allowed region, and the region where $\lambda > 0$ is a forbidden or tunneling region. Inspection of Figures 4.6 and 4.7 shows that (when $q = -2$) part of the v axis is forbidden for small n values, but that all of it is allowed once n becomes sufficiently large.

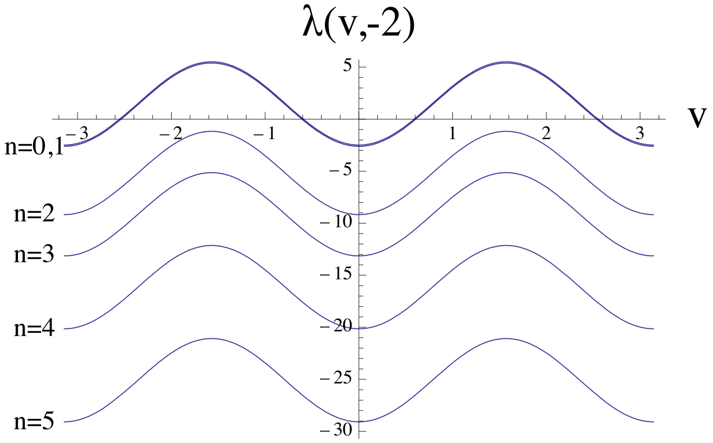


Figure 17.4.6: The effective potentials $\lambda(v, q)$ for the $\text{ce}_n(v, q)$ in the case $q = -2$. They are displayed as a function of v, over the interval $[-\pi, \pi]$, for various n values with $a = a_n(q)$. The top two curves, which very nearly coincide so as to almost look identical on the scale of the figure, are for the cases $n = 0$ and $n = 1$. According to Figure 4.5, the curve for $n = 0$ lies just slightly above that for $n = 1$. The bottom curve is that for $n = 5$. The curves in between are for $n = 2, 3, 4$ in that order.

Figures 4.8 through 4.10 display the first few $\text{ce}_n(v, q)$ as a function of v for $q = -2$. Figures 4.11 and 4.12 do the same for $\text{se}_1(v, q)$ and $\text{se}_2(v, q)$. It can be shown, as a consequence of Poincaré's theorem (see Section 1.3), that the $\text{ce}_n(v, q)$ and $\text{se}_n(v, q)$ are *entire* functions (analytic everywhere in the complex plane except at infinity) of v. Also, since the differential equation (4.22) has real coefficients for q real, the solutions $\text{ce}_n(v, q)$ and $\text{se}_n(v, q)$ are taken to be real for real q and real v.[9]

Observe from Figure 4.10 that $\text{ce}_2(v, q)$ is freely oscillating. This is to be expected from Figure 4.6 because we see that for $n \geq 2$ all of the v axis allowed. By contrast, Figure

[9] Since (4.22) is a second-order differential equation, there will also be second solutions that are linearly independent of the $\text{ce}_n(v, q)$ when $a = a_n(q)$, and second solutions that are linearly independent of the $\text{se}_n(v, q)$ when $a = b_n(q)$. Since the differential equation is invariant under parity (it is even in v), these solutions could, for example, be taken to have the opposite parity of the $\text{ce}_n(v, q)$ and the $\text{se}_n(v, q)$. They will not have period 2π.

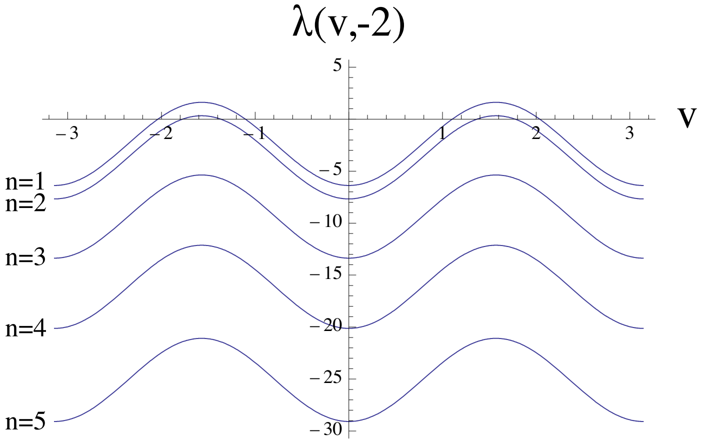


Figure 17.4.7: The effective potentials $\lambda(v,q)$ for the $\mathrm{se}_n(v,q)$ in the case $q=-2$. They are displayed as a function of v, over the interval $[-\pi,\pi]$, for various n values with $a=b_n(q)$. The top curve is that for $n=1$, and the bottom that for $n=5$. The curves in between are for $n=2,3,4$ in that order.

4.8 shows that $\mathrm{ce}_0(v,q)$ does not change sign. This is because (as v increases) this solution enters a forbidden region for $v \approx .6$ and at this point the function begins to decay. Again see Figure 4.6. Moreover, in the forbidden region, there is also a small exponentially growing part, with *positive* coefficient, that eventually dominates the solution by the time $v=\pi/2$ so that the solution begins to grow beyond this point. Finally, for $v >\approx 2.5$ the solution again enters an allowed region and begins to oscillate so that it has zero slope by the time $v=\pi$.

The case of $\mathrm{ce}_1(v,q)$ is more delicate. As has already been noted in the caption to Figure 4.6, when $q=-2$ the potentials λ for $a=a_0$ and for $a=a_1$ are almost the same. Yet, inspection of Figures 4.8 and 4.9 shows that $\mathrm{ce}_0(v,q)$ and $\mathrm{ce}_1(v,q)$ are very different! Because $\lambda|_{a_1} < \lambda|_{a_0}$, the forbidden region for $\mathrm{ce}_1(v,q)$ is somewhat smaller than for $\mathrm{ce}_0(v,q)$. Therefore $\mathrm{ce}_1(v,q)$ 'oscillates' a bit more before entering the forbidden region, and does so in such a way that the exponentially growing part in the forbidden region now has a negative sign. This exponentially growing part, although initially small in magnitude, eventually dominates at $v=\pi/2$ so that $\mathrm{ce}_1(v,q)$ crosses through zero and continues on to become negative. Eventually v again reaches an allowed region and $\mathrm{ce}_1(v,q)$ begins to oscillate so that it has zero slope by the time $v=\pi$.

What about the behavior of the $\mathrm{se}_n(v,q)$? Figure 4.7 shows their effective potentials for the case $q=-2$. Evidently these potentials are all completely negative when $n \geq 3$ and therefore the $\mathrm{se}_n(v,q)$ will be freely oscillatory when $n \geq 3$. Moreover, for $n=1$ the forbidden region is small, and for $n=2$ it is smaller yet. Therefore we expect the effects of the forbidden regions will be small. For example, the dips in $\mathrm{se}_1(v,q)$ at $v=\pm\pi/2$, see Figure 4.11, arise from the solution momentarily tunneling in forbidden regions. And examination of Figure 4.12 shows that, for $\mathrm{se}_2(v,q)$, passage through the forbidden regions has little noticeable effect.

It is also instructive to examine the behavior of the $\mathrm{ce}_n(v,q)$ and $\mathrm{se}_n(v,q)$ when q has a much more negative value. Figures 4.13 and 4.14 show their effective potentials for the case

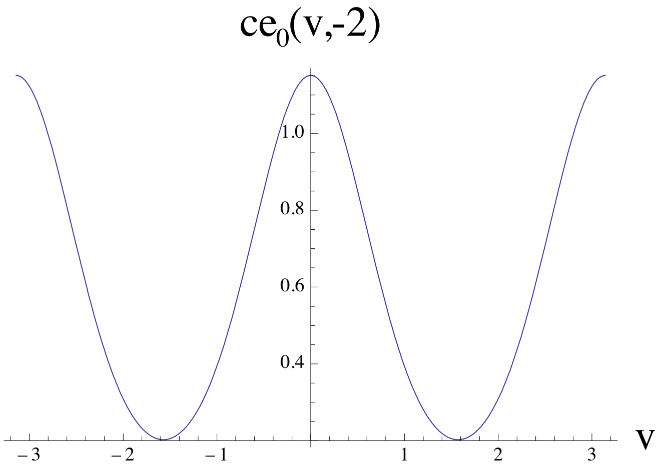


Figure 17.4.8: The function $\mathrm{ce}_0(v,q)$ as a function of v, over the interval $[-\pi,\pi]$, for $q=-2$. High magnification of this figure would reveal that the graph of $\mathrm{ce}_0(v,q)$ never touches or crosses, but always lies above, the v axis so that $\mathrm{ce}_0(v,q)$ has no zeroes.

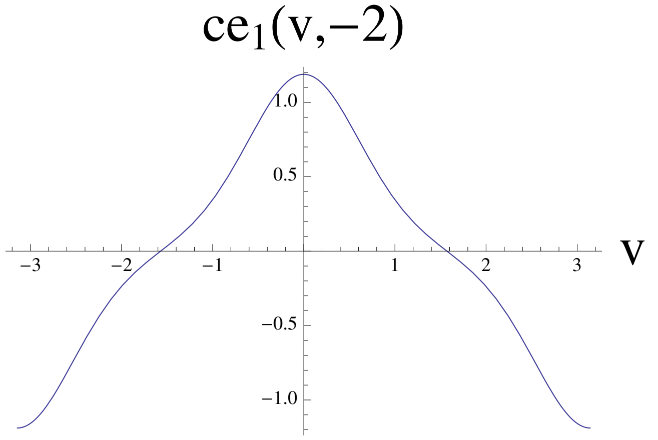


Figure 17.4.9: The function $\mathrm{ce}_1(v,q)$ as a function of v, over the interval $[-\pi,\pi]$, for $q=-2$.

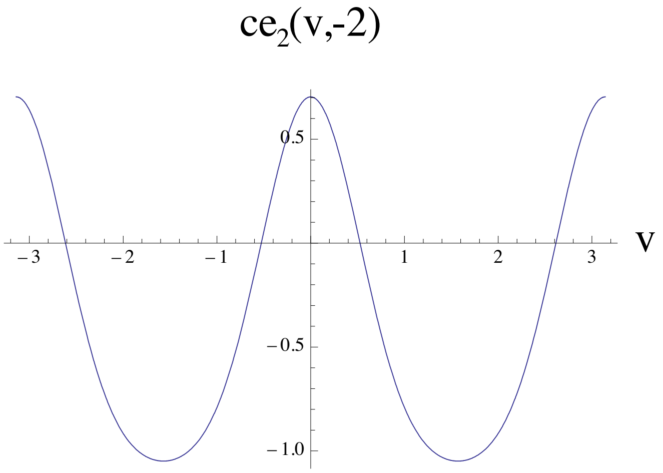


Figure 17.4.10: The function $\mathrm{ce}_2(v,q)$ as a function of v, over the interval $[-\pi,\pi]$, for $q=-2$.

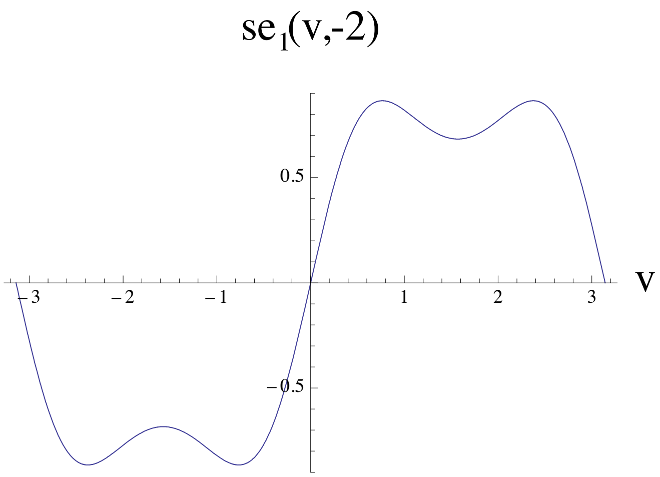


Figure 17.4.11: The function $\mathrm{se}_1(v,q)$ as a function of v, over the interval $[-\pi,\pi]$, for $q=-2$. The small dips at $v=\pm\pi/2$ arise from passage through forbidden regions.

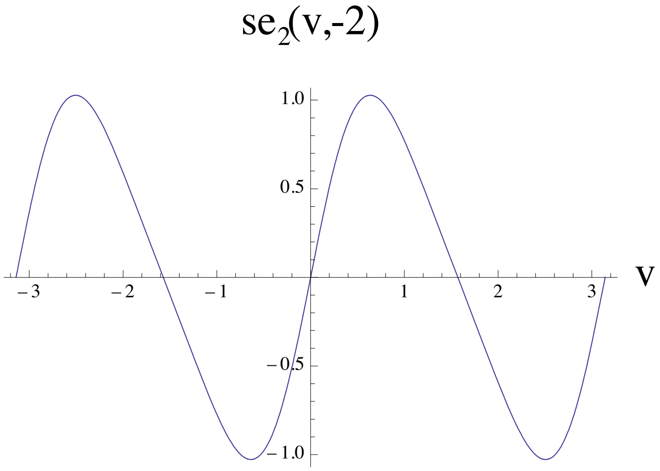


Figure 17.4.12: The function $\mathrm{se}_2(v,q)$ as a function of v, over the interval $[-\pi,\pi]$, for $q=-2$.

$q = -300$. Now we see that, for modest values of n, the potentials are *positive* for most values of v save for small intervals where they are negative. Therefore large portions of the v axis are forbidden regions. Consequently, for modest n values, the functions $\mathrm{ce}_n(v,q)$ and $\mathrm{se}_n(v,q)$ contain exponentially decaying terms for most values of v, and are oscillatory only over small intervals. On the other hand, as n is increased, the forbidden regions become smaller and the allowed regions become larger until for sufficiently large n the entire v axis becomes an allowed region. Therefore, for sufficiently large n, the functions $\mathrm{ce}_n(v,q)$ and $\mathrm{se}_n(v,q)$ are fully oscillatory.

As an illustration of this expected behavior, Figures 4.15 through 4.17 display the $\mathrm{ce}_n(v,q)$ for $q = -300$ and $n = 0, 1, 2$. We see that these functions begin bravely in the small allowed region about $v = 0$, rapidly decay to very nearly zero values in the forbidden regions centered about $v = \pm\pi/2$, and then rapidly revive in the allowed region centered about the (equivalent, due to periodicity) points $v = \pm\pi$. Compare these figures with their $q = -2$ counterparts, Figures 4.8 through 4.10. By contrast, Figure 4.18 shows $\mathrm{ce}_{22}(v,q)$ for $q = -300$. It can be shown that in this case the effective potential $\lambda(v, q = -300)$ is negative for all v. Therefore, in accord with Figure 4.18, $\mathrm{ce}_{22}(v,q)$ is fully oscillatory.

Similarly, Figures 4.19 and 4.20 display the $\mathrm{se}_n(v,q)$ for $q = -300$ and $n = 1, 2$. Again we see these functions are very nearly zero in the forbidden regions. For example, the dips in Figure 4.11 have become, in Figure 4.19, *canyons* with very steep walls and very flat floors. By contrast, Figure 4.21 shows $\mathrm{se}_{23}(v,q)$ for $q = -300$. It can be shown that in this case the effective potential $\lambda(v, q = -300)$ is negative for all v. Therefore, in accord with Figure 4.21, $\mathrm{se}_{23}(v,q)$ is fully oscillatory.

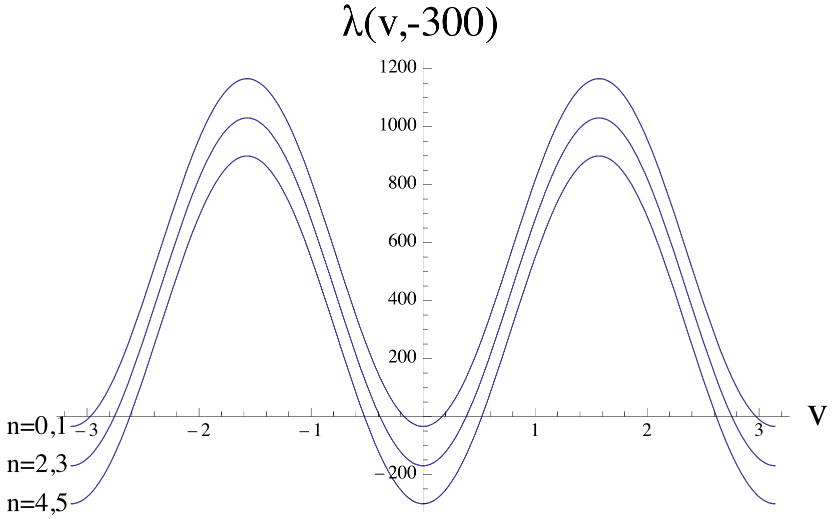


Figure 17.4.13: The effective potentials $\lambda(v,q)$ for the $\mathrm{ce}_n(v,q)$ in the case $q = -300$. They are displayed as a function of v, over the interval $[-\pi, \pi]$, for the n values $n = 0, 1, 2, 3, 4, 5$ with $a = a_n(q)$. The top two curves, which very nearly coincide so as to almost look identical on the scale of the figure, are for the cases $n = 0$ and $n = 1$. The next two curves, which also nearly coincide, are for $n = 2$ and $n = 3$. Finally, the bottom two curves also nearly coincide and are for the cases $n = 4$ and $n = 5$. As in Figure 4.6, the higher the n value, the lower the curve.

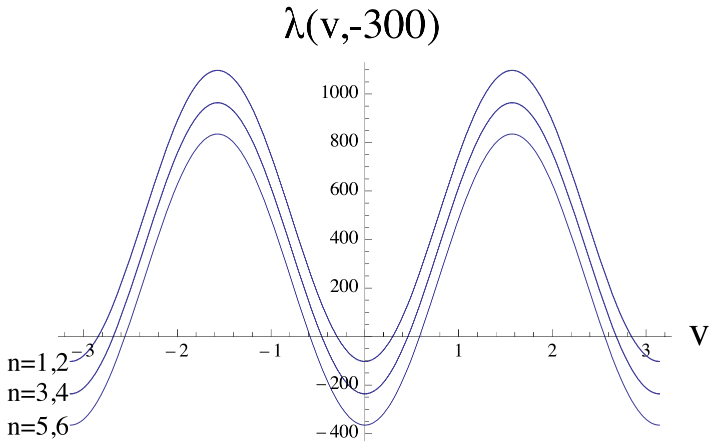


Figure 17.4.14: The effective potentials $\lambda(v,q)$ for the $\mathrm{se}_n(v,q)$ in the case $q=-300$. They are displayed as a function of v, over the interval $[-\pi,\pi]$, for the n values $n=1,2,3,4,5,6$ with $a=b_n(q)$. The top two curves, which very nearly coincide so as to almost look identical on the scale of the figure, are for the cases $n=1$ and $n=2$. The next two curves, which also nearly coincide, are for $n=3$ and $n=4$. Finally, the bottom two curves also nearly coincide and are for the cases $n=5$ and $n=6$. As in Figure 4.7, the higher the n value, the lower the curve.

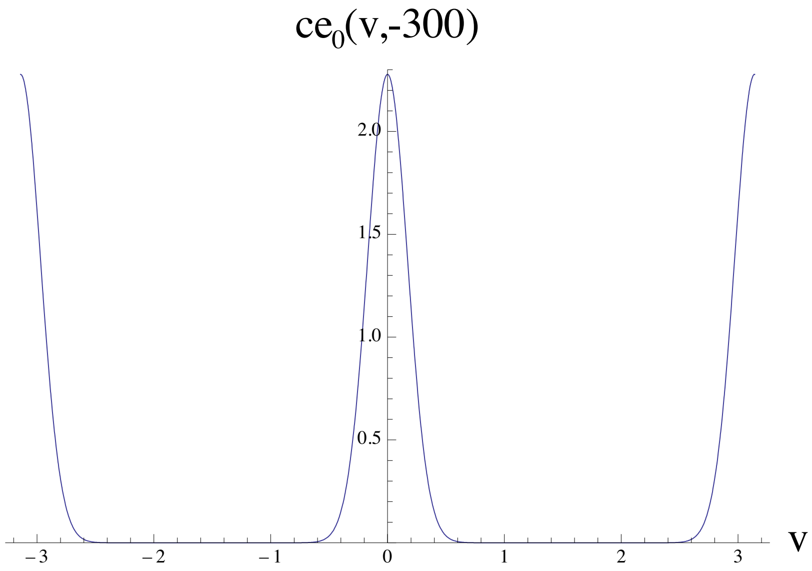


Figure 17.4.15: The function $\mathrm{ce}_0(v,q)$ as a function of v, over the interval $[-\pi,\pi]$, for $q=-300$. Most of the v axis is forbidden.

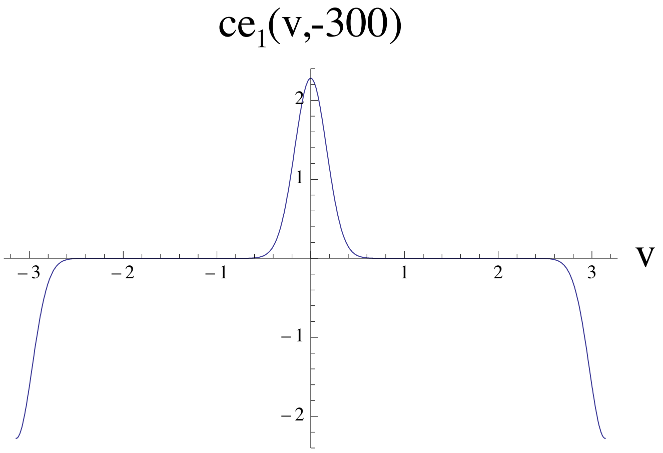


Figure 17.4.16: The function $\mathrm{ce}_1(v, q)$ as a function of v, over the interval $[-\pi, \pi]$, for $q = -300$. Most of the v axis is forbidden.

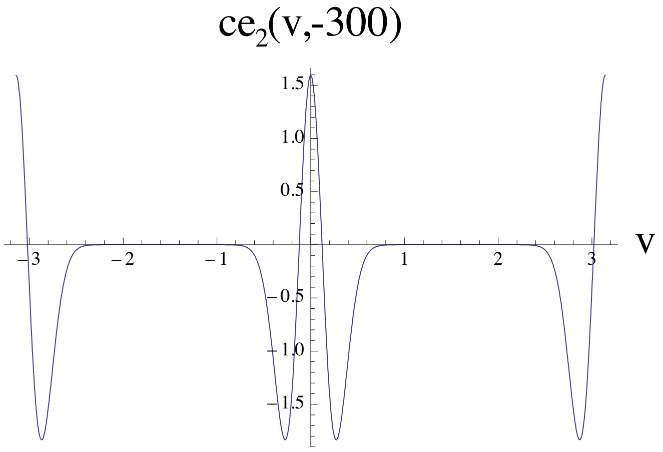


Figure 17.4.17: The function $\mathrm{ce}_2(v, q)$ as a function of v, over the interval $[-\pi, \pi]$, for $q = -300$. Most of the v axis is forbidden.

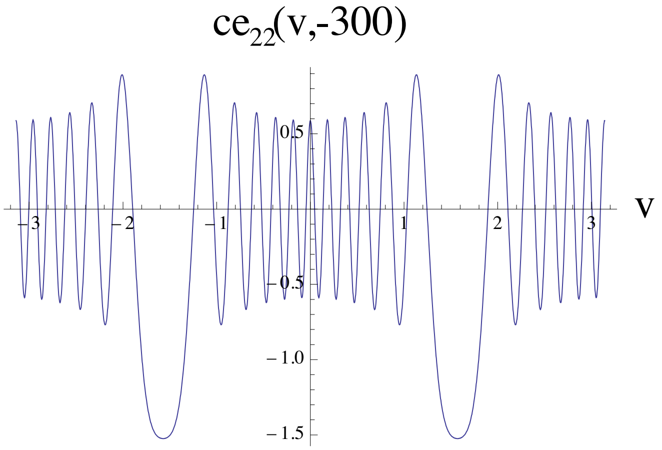


Figure 17.4.18: The function $\text{ce}_{22}(v, q)$ as a function of v, over the interval $[-\pi, \pi]$, for $q = -300$. For these q and n values all of the v axis is allowed, and the function is fully oscillatory.

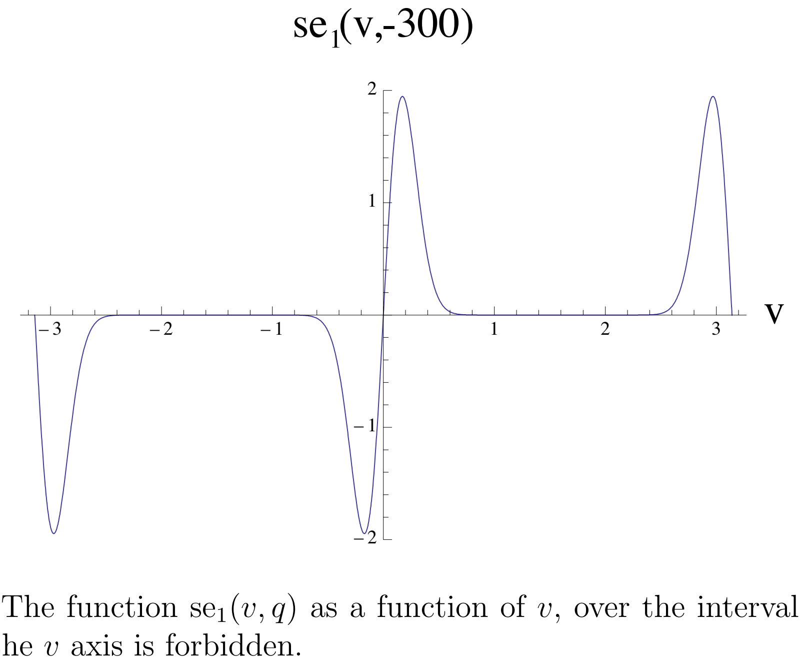


Figure 17.4.19: The function $\text{se}_1(v, q)$ as a function of v, over the interval $[-\pi, \pi]$, for $q = -300$. Most of the v axis is forbidden.

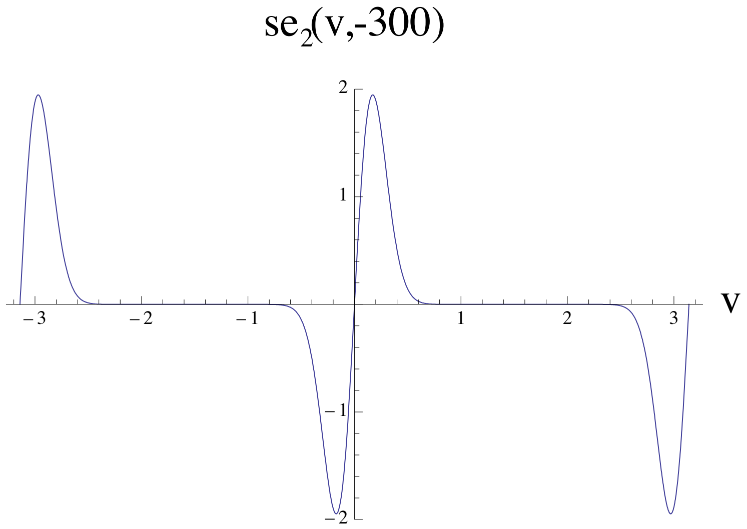


Figure 17.4.20: The function $\mathrm{se}_2(v, q)$ as a function of v, over the interval $[-\pi, \pi]$, for $q = -300$. Most of the v axis is forbidden.

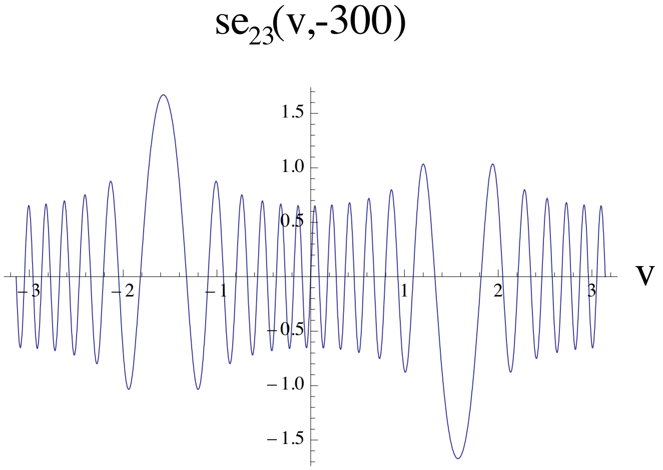


Figure 17.4.21: The function $\mathrm{se}_{23}(v, q)$ as a function of v, over the interval $[-\pi, \pi]$, for $q = -300$. For these q and n values all of the v axis is allowed, and the function is fully oscillatory.

Look again at the potentials $\lambda(v,q)$ shown in Figures 4.6, 4.7, 4.13, and 4.14. We see, as is also evident from (4.46), that they have maxima at $v = \pm\pi/2$, and at these points they have the maximum values

$$\lambda_{\max}(q) = \lambda(\pm\pi/2, q) = -a - 2q. \tag{17.4.47}$$

Figure 4.22 shows $\lambda_{\max}(q)$ for various n values in the case $a = a_n(q)$, and Figure 4.23 does the same in the case $a = b_n(q)$. From (4.29) and (4.30) we have the asymptotic formulas

$$\begin{aligned}&\lambda_{\max}(q) \sim -4q - (8m+2)(-q)^{1/2} + \cdots\\ &\text{as } q \to -\infty \text{ for } a = a_{2m}(q) \sim a_{2m+1}(q) \text{ and } m = 0,1,2,3,\cdots,\end{aligned} \tag{17.4.48}$$

$$\begin{aligned}&\lambda_{\max}(q) \sim -4q - (8m+6)(-q)^{1/2} + \cdots\\ &\text{as } q \to -\infty \text{ for } a = b_{2m+1}(q) \sim b_{2m+2}(q) \text{ and } m = 0,1,2,3,\cdots.\end{aligned} \tag{17.4.49}$$

We know that all of the interval $v \in [-\pi, \pi]$ is allowed when $\lambda_{\max}(q) < 0$, and part of it becomes forbidden when $\lambda_{\max}(q) > 0$. Thus, for each n value and each alternative $a = a_n(q)$ or $a = b_n(q)$, there is a *critical* value $q_{\rm cr}$ such that $\lambda_{\max}(q_{\rm cr}) = 0$. From (4.47) we see that these critical values, in the two alternatives, are given (implicitly) by the relations

$$q_{\rm cr}(n) = -(1/2)a_n[q_{\rm cr}(n)], \tag{17.4.50}$$

$$q_{\rm cr}(n) = -(1/2)b_n[q_{\rm cr}(n)]. \tag{17.4.51}$$

These critical values, which can be read off from the 'x' intercepts of the curves in Figures 4.22 and 4.23, are listed in Table 4.1. For a given value of n, all of the v axis is allowed and $\mathrm{ce}_n(v,q)$ is fully oscillatory if $q > q_{\rm cr}(n)$, and otherwise part of the v axis is forbidden. Here $q_{\rm cr}(n)$ is to be calculated using (4.50). An analogous statement holds for $\mathrm{se}_n(v,q)$ where now $q_{\rm cr}(n)$ is to be calculated using (4.51).

At this point we are prepared to comment on the symmetry properties of the $\mathrm{ce}_n(v,q)$ and $\mathrm{se}_n(v,q)$. We begin with the $\mathrm{ce}_n(v,q)$. We know they are even and periodic with period 2π. They therefore have Fourier series expansions consisting only of cosine terms. Also, consistent with the behavior of $\mathrm{ce}_0(v,q)$ and $\mathrm{ce}_2(v,q)$ displayed in Figures 4.8, 4.10, 4.15, and 4.17, it can be shown that the $\mathrm{ce}_n(v,q)$ for *even n* are *symmetric* about the point $v = \pi/2$. Specifically, the $\mathrm{ce}_n(v,q)$ for even n have Fourier expansions of the form

$$\mathrm{ce}_n(v,q) = *1 + *\cos(2v) + *\cos(4v) + \cdots \text{ for even } n \tag{17.4.52}$$

where the *'s denote q and n dependent coefficients. That is, there is the relation

$$\mathrm{ce}_n(\pi/2 + \Delta, q) = \mathrm{ce}_n(\pi/2 - \Delta, q) \text{ for even } n. \tag{17.4.53}$$

It follows that the $\mathrm{ce}_n(v,q)$ for even n have vanishing first derivative at $v = \pi/2$,

$$\mathrm{ce}_n'(\pi/2, q) = 0 \text{ for even } n. \tag{17.4.54}$$

By contrast, as illustrated in Figures 4.9 and 4.16, the $\mathrm{ce}_n(v,q)$ for *odd n* are *antisymmetric* about the point $v = \pi/2$ and have Fourier expansions of the form

$$\mathrm{ce}_n(v,q) = *\cos(v) + *\cos(3v) + *\cos(5v) + \cdots \text{ for odd } n. \tag{17.4.55}$$

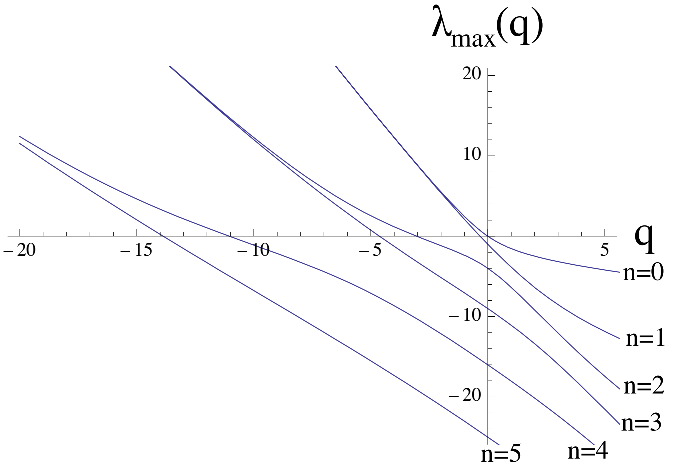


Figure 17.4.22: The function $\lambda_{\max}(q)$ for the n values 0 through 5 in the case $a = a_n(q)$. When $\lambda_{\max}(q) < 0$, all of the v axis is allowed, and the function $\mathrm{ce}_n(v, q)$ is fully oscillatory. When $\lambda_{\max}(q) > 0$, part of the v axis is forbidden. The higher the n value, the lower the curve. Note that the 'y' intercepts have the values $-n^2$ in accord with (4.24) through (4.26) and (4.47). The 'x' intercepts are the values $q_{\mathrm{cr}}(n)$. Note also that the values of $\lambda_{\max}(q)$ for $n = 0$ and $n = 1$ tend to merge for large negative q, as do the values for $n = 2$ and $n = 3$, etc. See Figure 4.5 and (4.48).

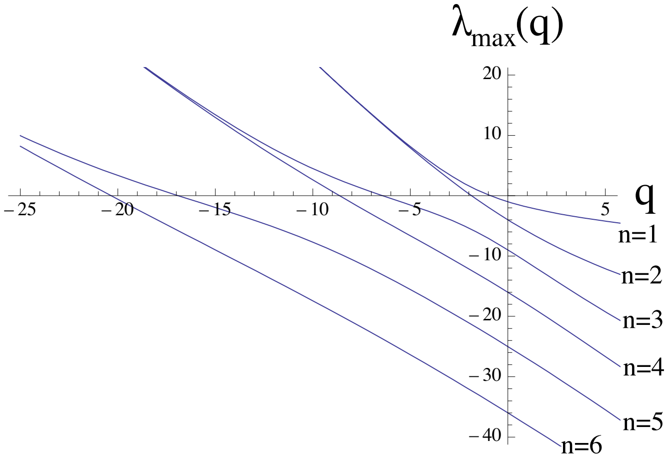


Figure 17.4.23: The function $\lambda_{\max}(q)$ for the n values 1 through 6 in the case $a = b_n(q)$. When $\lambda_{\max}(q) < 0$, all of the v axis is allowed, and the function $\mathrm{se}_n(v, q)$ is fully oscillatory. When $\lambda_{\max}(q) > 0$, part of the v axis is forbidden. The higher the n value, the lower the curve. Note that the 'y' intercepts have the values $-n^2$ in accord with (4.27), (4.28), and (4.47). The 'x' intercepts are the values $q_{\mathrm{cr}}(n)$. Note also that the values of $\lambda_{\max}(q)$ for $n = 1$ and $n = 2$ tend to merge for large negative q, as do the values for $n = 3$ and $n = 4$, etc. See Figure 4.5 and (4.49).

Table 17.4.1: The quantity $q_{\rm cr}(n)$ for various values of n.

n	$q_{\rm cr}(n)$ when $a = a_n(q)$	$q_{\rm cr}(n)$ when $a = b_n(q)$
0	0	*
1	-0.329005727826915	-0.889819993831662
2	-3.039073671630782	-1.8582116914842934
3	-4.626950799904568	-6.425863307211811
4	-11.047992936386709	-8.6316091625993501
5	-13.871128836603399	-16.904741557017985
6	-23.995780230075020	-20.345062417526364
7	-28.0531793998642485	-32.320930596434941
8	-41.880084880521011	-36.995345508020719
9	-47.171475670427398	-52.673172894843788
10	-64.7001463432813892	-58.581512590132760

That is, there is the relation

$$\mathrm{ce}_n(\pi/2 + \Delta, q) = -\mathrm{ce}_n(\pi/2 - \Delta, q) \text{ for odd } n. \tag{17.4.56}$$

It follows that the $\mathrm{ce}_n(v, q)$ for odd n vanish at $v = \pi/2$,

$$\mathrm{ce}_n(\pi/2, q) = 0 \text{ for odd } n. \tag{17.4.57}$$

Next consider the symmetry properties of the $\mathrm{se}_n(v, q)$. We know they are odd and periodic with period 2π. They therefore have Fourier series expansions consisting only of sine terms. Also, consistent with the behavior of $\mathrm{se}_2(v, q)$ displayed in Figures 4.12 and 4.20, it can be shown that the $\mathrm{se}_n(v, q)$ for even n are antisymmetric about the point $v = \pi/2$. Specifically, the $\mathrm{se}_n(v, q)$ for even n have Fourier expansions of the form

$$\mathrm{se}_n(v, q) = * \sin(2v) + * \sin(4v) + * \sin(6v) + \cdots \text{ for even } n. \tag{17.4.58}$$

That is, there is the relation

$$\mathrm{se}_n(\pi/2 + \Delta, q) = -\mathrm{se}_n(\pi/2 - \Delta, q) \text{ for even } n, \tag{17.4.59}$$

from which it follows that

$$\mathrm{se}_n(\pi/2, q) = 0 \text{ for even } n. \tag{17.4.60}$$

By contrast, as ilustrated in Figures 4.11 and 4.19 for $\mathrm{se}_1(v, q)$, the $\mathrm{se}_n(v, q)$ for odd n are symmetric about the point $v = \pi/2$ and have Fourier expansions of the form

$$\mathrm{se}_n(v, q) = * \sin(v) + * \sin(3v) + * \sin(5v) + \cdots \text{ for odd } n. \tag{17.4.61}$$

That is, there is the relation

$$\mathrm{se}_n(\pi/2 + \Delta, q) = \mathrm{se}_n(\pi/2 - \Delta, q) \text{ for odd } n, \tag{17.4.62}$$

from which it follows that

$$\mathrm{se}_n'(\pi/2, q) = 0 \text{ for odd } n. \tag{17.4.63}$$

Finally, it follows from (4.52) and (4.58) that

$$\mathrm{ce}_n(v + \pi, q) = \mathrm{ce}_n(v, q) \text{ for even } n \tag{17.4.64}$$

and

$$\mathrm{se}_n(v + \pi, q) = \mathrm{se}_n(v, q) \text{ for even } n. \tag{17.4.65}$$

Thus the $\mathrm{se}_n(v, q)$ and $\mathrm{se}_n(v, q)$ for even n have period π as well as period 2π. By contrast, we see from (4.55) and (4.61) that there are the relations

$$\mathrm{ce}_n(v + \pi, q) = -\mathrm{ce}_n(v, q) \text{ for odd } n \tag{17.4.66}$$

and

$$\mathrm{se}_n(v + \pi, q) = -\mathrm{se}_n(v, q) \text{ for odd } n. \tag{17.4.67}$$

The relations (4.64) through (4.67) can be written more succinctly in the form

$$\mathrm{ce}_n(v + \pi, q) = (-1)^n \mathrm{ce}_n(v, q), \tag{17.4.68}$$

$$\mathrm{se}_n(v + \pi, q) = (-1)^n \mathrm{se}_n(v, q). \tag{17.4.69}$$

From a computational perspective, an important consequence of these symmetry properties of the $\mathrm{ce}_n(v, q)$ and $\mathrm{se}_n(v, q)$ is that they only need to be computed over the interval $v \in [0, \pi/2]$. Their values elsewhere are then determined by their symmetry properties. Moreover, if q and n are such that a value of v is deep within a strongly forbidden region, then we may set the associated value of $\mathrm{ce}_n(v, q)$ or $\mathrm{se}_n(v, q)$ to zero for these values of v. Recall that the forbidden regions are centered about the values $v = \pm\pi/2$. Thus, if there are such strongly forbidden regions, we only need to compute $\mathrm{ce}_n(v, q)$ or $\mathrm{se}_n(v, q)$ over the smaller interval $v \in [0, v_{\text{deep}}]$ where v_{deep} is the smallest v value deep within the strongly forbidden region.

17.4.5 Modified Mathieu Functions

Now that the possible values of the separation constant a have been determined by the periodicity requirement, these values of a can be employed in (4.21) to determine the functions $P(u)$. The so-called solutions of the *first kind* for (4.21), when $a = a_n(q)$, are denoted as $\mathrm{Ce}_n(u, q)$; and the solutions of the first kind, when $a = b_n(q)$, are denoted as $\mathrm{Se}_n(u, q)$. The functions $\mathrm{Ce}_n(u, q)$ are even functions of u and the functions $\mathrm{Se}_n(u, q)$ are odd functions of u. They can be conveniently arranged to satisfy the relations

$$\mathrm{Ce}_n(u, q) = \mathrm{ce}_n(iu, q), \tag{17.4.70}$$

$$\mathrm{Se}_n(u, q) = -i\mathrm{se}_n(iu, q). \tag{17.4.71}$$

Evidently, they are also entire functions, and they are also real for q and u real. Since the $\mathrm{ce}_n(v, q)$ and $\mathrm{se}_n(v, q)$ are analogous to cosines and sines, see (4.52), (4.55), (4.58), and

(4.61), the relations (4.70) and (4.71) indicate that the $\mathrm{Ce}_n(u,q)$ and $\mathrm{Se}_n(u,q)$ are analogous to hyperbolic cosines and hyperbolic sines.

Suppose we write (4.21) in the form

$$d^2P/du^2 - \Lambda(u,q)P = 0 \tag{17.4.72}$$

where

$$\Lambda(u,q) = a - 2q\cosh(2u). \tag{17.4.73}$$

Figure 4.24 shows these $\Lambda(u,q)$ for various n values with $a = a_n(q)$ and $q = -2$. These are the potentials appropriate to the $\mathrm{Ce}_n(u,q)$. The potentials for the $\mathrm{Se}_n(u,q)$ computed with $a = b_n(q)$ are similar. Inspection of these $\mathrm{Ce}_n(u,q)$ potentials shows that they are all *positive* for *all* u. The same can be shown to be true for the $\mathrm{Se}_n(v,q)$ potentials.

In fact, more can be said. From (4.73) it is evident that, when $q < 0$ (which is what we have assumed), Λ has a minimum at $u = 0$ and, at this point has the value

$$\Lambda(0,q) = a - 2q. \tag{17.4.74}$$

Therefore, if we can show that

$$\Lambda(0,q) = a - 2q > 0 \tag{17.4.75}$$

for all $q < 0$ and all n with $a = a_n(q)$ or $a = b_n(q)$, then we will have shown that all Λ are positive for all u. Figure 4.25 displays $\Lambda(0,q)$ for various n values when $a = a_n(q)$, and Figure 4.26 does the same for the case $a = b_n(q)$. Note also that from (4.29) and (4.30) there is the asymptotic behavior

$$\begin{aligned}&\Lambda(0,q) \sim (8m+2)(-q)^{1/2} + \cdots\\ &\text{as } q \to -\infty \text{ for } a = a_{2m}(q) \sim a_{2m+1}(q) \text{ and } m = 0,1,2,3,\cdots,\end{aligned} \tag{17.4.76}$$

$$\begin{aligned}&\Lambda(0,q) \sim (8m+6)(-q)^{1/2} + \cdots\\ &\text{as } q \to -\infty \text{ for } a = b_{2m+1}(q) \sim b_{2m+2}(q) \text{ and } m = 0,1,2,3,\cdots.\end{aligned} \tag{17.4.77}$$

Evidently (4.75) is always satisfied when $q < 0$. We conclude that for $q \le 0$ the $\mathrm{Ce}_n(u,q)$ and $\mathrm{Se}_n(u,q)$ are non-oscillatory and must all be exponentially growing.[10]

As examples, Figures 4.27 and 4.28 display the first few $\mathrm{Ce}_n(u,q)$ and $\mathrm{Se}_n(u,q)$ as a function of u for $q = -2$.[11] We see that, as predicted, they are non-oscillatory and their magnitudes do indeed become large for large values of $|u|$.

[10] Also, as Figures 4.25 and 4.26 illustrate, oscillatory behavior for the $\mathrm{Ce}_n(u,q)$ and $\mathrm{Se}_n(u,q)$ is possible if $q > 0$. See Exercise 4.1.

[11] Since the modified Mathieu equation (4.21) is also of second order, it will also have second solutions that are linearly independent of the $\mathrm{Ce}_n(u,q)$ and $\mathrm{Se}_n(u,q)$. Because it too is invariant under parity, these solutions could be constructed to have parities opposite to those of the $\mathrm{Ce}_n(u,q)$ and $\mathrm{Se}_n(u,q)$.

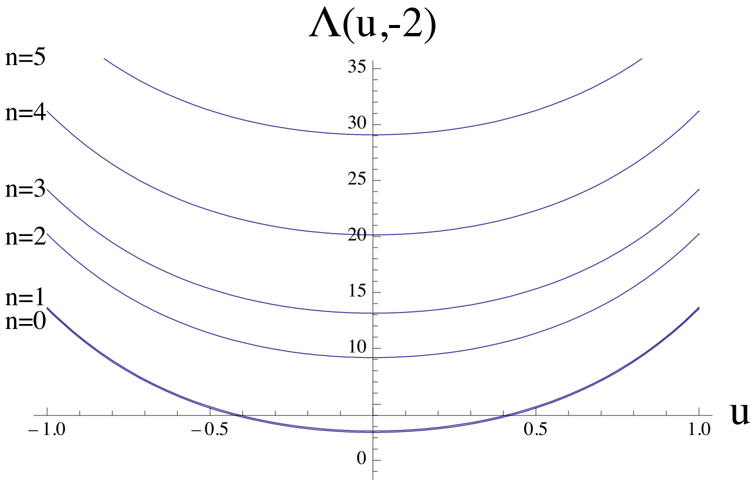


Figure 17.4.24: The effective potentials $\Lambda(u, q)$ for the $\mathrm{Ce}_n(v, q)$ in the case $q = -2$. They are displayed as a function of u for the n values $n = 0, 1, 2, 3, 4, 5$ with $a = a_n(q)$. As in Figure 4.6, the curves for $n = 0$ and $n = 1$ nearly coincide. Now, because of the difference in sign between (4.46) and (4.73), the higher the n value the higher the curve.

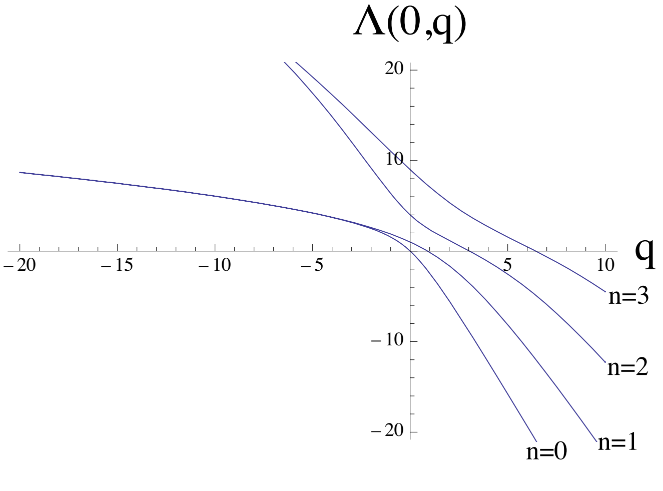


Figure 17.4.25: The quantities $\Lambda(0, q)$ with $a = a_n(q)$ and $n = 0, 1, 2, 3$. The higher the n value, the higher the curve. Note that the 'y' intercepts have the values n^2 in agreement with (4.24) through (4.26). Also, values of $\Lambda(0, q)$ for $n = 0$ and $n = 1$ tend to merge for large negative q, as do the values for $n = 2$ and $n = 3$, etc. See Figure 4.5 and (4.76).

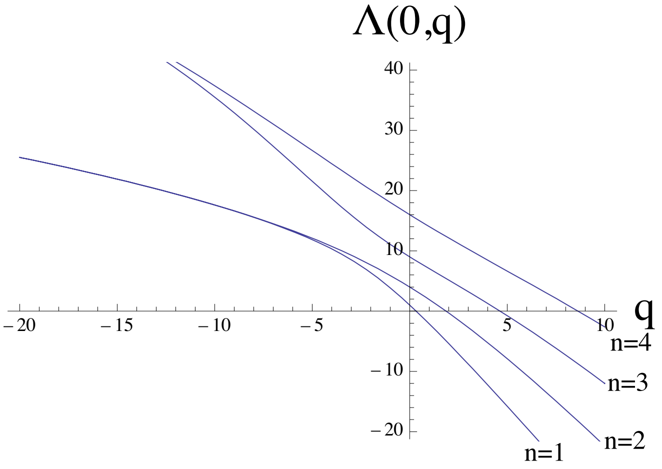


Figure 17.4.26: The quantities $\Lambda(0, q)$ with $a = b_n(q)$ and $n = 1, 2, 3, 4$. The higher the n value, the higher the curve. Note that the 'y' intercepts have the values n^2 in agreement with (4.27) and (4.28). Also, values of $\Lambda(0, q)$ for $n = 1$ and $n = 2$ tend to merge for large negative q, as do the values for $n = 3$ and $n = 4$, etc. See Figure 4.5 and (4.77).

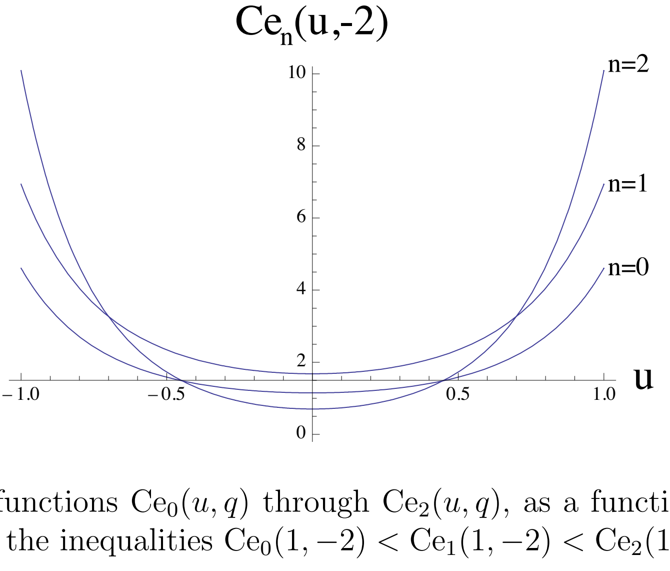


Figure 17.4.27: The functions $\mathrm{Ce}_0(u, q)$ through $\mathrm{Ce}_2(u, q)$, as a function of u, for $q = -2$. At $u = 1$ they satisfy the inequalities $\mathrm{Ce}_0(1, -2) < \mathrm{Ce}_1(1, -2) < \mathrm{Ce}_2(1, -2)$.

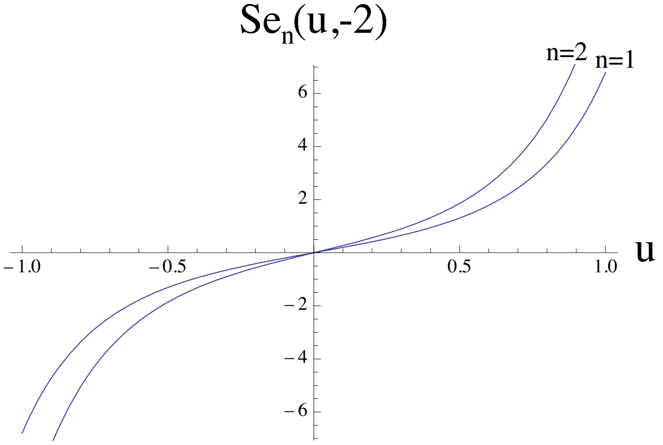


Figure 17.4.28: The functions $\mathrm{Se}_1(u,q)$ and $\mathrm{Se}_2(u,q)$, as a function of u, for $q=-2$. At $u=1$ they satisfy the inequality $\mathrm{Se}_1(1,-2)<\mathrm{Se}_2(1,-2)$.

17.4.6 Analyticity in x and y

So far we have described the functions $\mathrm{ce}_n(v,q)$ and $\mathrm{se}_n(v,q)$, and the functions $\mathrm{Ce}_n(u,q)$ and $\mathrm{Se}_n(u,q)$, and have seen that they are entire functions of the variables u and v. But this does not mean that they are entire functions of the variables x and y because the relation (4.7) that connects u,v to x,y has singularities. See Exercise 4.2. Remarkably, however, the products $[\mathrm{Ce}_n(u,q)\times\mathrm{ce}_n(v,q)]$ and $[\mathrm{Se}_n(u,q)\times\mathrm{se}_n(v,q)]$, which according to (4.13) is what we hope to use to construct solutions of the Laplace equation, are entire functions of the variables x and y. This situation is analogous to the case of cylindrical coordinates where the functions $\exp(im\phi)$ and $I_m(k\rho)$ are entire functions of ϕ and ρ, respectively, but are not entire functions of x and y. However, the products $[\exp(im\phi)\times I_m(k\rho)]$ are entire functions of x and y. In the case of the Mathieu functions, for example, there is an integral representation of the form

$$\mathrm{Ce}_{2n}(u,q)\,\mathrm{ce}_{2n}(v,q)=p_{2n}(q)\int_0^{\pi/2}d\tau\;\mathrm{ce}_{2n}(\tau,q)\cosh[kx\cos(\tau)]\cosh[ky\sin(\tau)] \qquad (17.4.78)$$

where $p_{2n}(q)$ is some q-dependent coefficient. There are similar representations for the other relevant products. The right side of (4.78) is manifestly an entire function of x and y.

17.4.7 Elliptic Cylinder Harmonic Expansion and On-Axis Gradients

The stage is now set to describe the expansion of any harmonic function ψ in terms of Mathieu functions. The general harmonic function that is analytic in x and y near the origin can be written in the form

$$\begin{aligned}\psi(x,y,z)=\psi(u,v,z) &= \sum_{n=0}^{\infty}\int_{-\infty}^{\infty} dk\; c_n(k)\exp(ikz)\mathrm{Ce}_n(u,q)\;\mathrm{ce}_n(v,q)\\ &+ \sum_{n=1}^{\infty}\int_{-\infty}^{\infty} dk\; s_n(k)\exp(ikz)\mathrm{Se}_n(u,q)\;\mathrm{se}_n(v,q)\end{aligned}\tag{17.4.79}$$

where the functions $c_n(k)$ and $s_n(k)$ are arbitrary. We will call (4.79) an *elliptic cylinder harmonic* expansion.

To exploit this expansion, suppose the magnetic field $\boldsymbol{B}(x,y,z)$ is interpolated onto the surface $u=U$ of an elliptic cylinder using values at the grid points near the surface. See Figure 4.3. Let us employ the notation $\boldsymbol{B}(x,y,z)=\boldsymbol{B}(u,v,z)$ so that the magnetic field on the surface can be written as $\boldsymbol{B}(U,v,z)$. Next, from the values on the surface, compute $B_u(U,v,z)$, the component of $\boldsymbol{B}(x,y,z)$ *normal* to the surface. Our aim will be to determine the generalized on-axis gradients from a knowledge of $B_u(U,v,z)$.

Let us begin by solving (4.11) for $(\partial\psi/\partial u)$. We find, using (4.4), the result,

$$\begin{aligned}(\partial\psi/\partial u) &= f[\cosh^2(u)-\cos^2(v)]^{1/2}B_u\\ &= f(\sinh u\cos v)B_x(u,v,z)+f(\cosh u\sin v)B_y(u,v,z).\end{aligned}\tag{17.4.80}$$

We see that the right side of (4.80) is a well-behaved function $F(u,v,z)$ whose values are known for $u=U$,

$$F(U,v,z)=f(\sinh U\cos v)B_x(U,v,z)+f(\cosh U\sin v)B_y(U,v,z).\tag{17.4.81}$$

Moreover, using the representation (4.79) in (4.80) and (4.81), we may also write

$$\begin{aligned}F(U,v,z) &= \sum_{n=0}^{\infty}\int_{-\infty}^{\infty} dk\; c_n(k)\exp(ikz)\mathrm{Ce}_n'(U,q)\;\mathrm{ce}_n(v,q)\\ &+ \sum_{n=1}^{\infty}\int_{-\infty}^{\infty} dk\; s_n(k)\exp(ikz)\mathrm{Se}_n'(U,q)\;\mathrm{se}_n(v,q).\end{aligned}\tag{17.4.82}$$

Next multiply both sides of (4.82) by $\exp(-ik'z)$ and integrate over z. So doing gives the result

$$\begin{aligned}&(1/2\pi)\int_{-\infty}^{\infty} dz\exp(-ik'z)F(U,v,z)=\\ &\sum_{n=0}^{\infty}c_n(k')\mathrm{Ce}_n'(U,q')\;\mathrm{ce}_n(v,q')+\sum_{n=1}^{\infty}s_n(k')\mathrm{Se}_n'(U,q')\;\mathrm{se}_n(v,q').\end{aligned}\tag{17.4.83}$$

Now, employ the orthogonality properties (4.38) through (4.40) to obtain the relations

$$c_r(k')\mathrm{Ce}_r'(U,q')=[1/(2\pi^2)]\int_0^{2\pi}dv\int_{-\infty}^{\infty}dz\exp(-ik'z)\mathrm{ce}_r(v,q')F(U,v,z),\tag{17.4.84}$$

$$s_r(k')\mathrm{Se}'_r(U,q') = [1/(2\pi^2)]\int_0^{2\pi} dv \int_{-\infty}^{\infty} dz \exp(-ik'z)\mathrm{se}_r(v,q')F(U,v,z). \tag{17.4.85}$$

In view of (4.84) and (4.85), define the function $\tilde{F}(v,k)$ by the rule

$$\tilde{F}(v,k) = [1/(2\pi)]\int_{-\infty}^{\infty} dz \exp(-ikz)F(U,v,z), \tag{17.4.86}$$

and define functions $\tilde{\tilde{F}}{}^c_r(k)$ and $\tilde{\tilde{F}}{}^s_r(k)$ by the rules

$$\begin{aligned}\tilde{\tilde{F}}{}^c_r(k) &= (1/\pi)\int_0^{2\pi} dv\ \mathrm{ce}_r(v,q)\tilde{F}(v,k)\\ &= [1/(2\pi^2)]\int_0^{2\pi} dv \int_{-\infty}^{\infty} dz \exp(-ikz)\mathrm{ce}_r(v,q)F(U,v,z),\end{aligned} \tag{17.4.87}$$

$$\begin{aligned}\tilde{\tilde{F}}{}^s_r(k) &= (1/\pi)\int_0^{2\pi} dv\ \mathrm{se}_r(v,q)\tilde{F}(v,k)\\ &= [1/(2\pi^2)]\int_0^{2\pi} dv \int_{-\infty}^{\infty} dz \exp(-ikz)\mathrm{se}_r(v,q)F(U,v,z).\end{aligned} \tag{17.4.88}$$

[Here we have extended the use of the ˜ notation to include angular *Mathieu* transforms, such as those in (4.87) and (4.88), where $\cos(r\phi)$ and $\sin(r\phi)$ are replaced by $\mathrm{ce}_r(v,q)$ and $\mathrm{se}_r(v,q)$.] We will call the functions $\tilde{\tilde{F}}{}^\alpha_r(k)$ *Mathieu coefficient* functions in analogy to the Fourier coefficients that arise in Fourier analysis. Note that, because F and the Mathieu functions are real, the real parts of the functions $\tilde{\tilde{F}}{}^\alpha_r(k)$ are even in k, and the imaginary parts are odd in k.

With these definitions, the relations (4.84) and (4.85) can be rewritten in the form

$$c_r(k) = \tilde{\tilde{F}}{}^c_r(k)/\mathrm{Ce}'_r(U,q), \tag{17.4.89}$$

$$s_r(k) = \tilde{\tilde{F}}{}^s_r(k)/\mathrm{Se}'_r(U,q). \tag{17.4.90}$$

Finally, employ (4.89) and (4.90) in (4.79). So doing gives the result

$$\begin{aligned}\psi(x,y,z) &= \sum_{r=0}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz)[\tilde{\tilde{F}}{}^c_r(k)/\mathrm{Ce}'_r(U,q)]\mathrm{Ce}_r(u,q)\ \mathrm{ce}_r(v,q)\\ &+ \sum_{r=1}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz)[\tilde{\tilde{F}}{}^s_r(k)/\mathrm{Se}'_r(U,q)]\mathrm{Se}_r(u,q)\ \mathrm{se}_r(v,q).\end{aligned} \tag{17.4.91}$$

We have obtained an elliptical cylinder harmonic expansion for ψ in terms of surface field data.

Of course, what we really want are the on-axis gradients. They can be found by employing two remarkable *connections* (identities) between elliptic and circular cylinder functions of the form

$$\mathrm{Ce}_r(u,q)\ \mathrm{ce}_r(v,q) = \sum_{m=0}^{\infty} \alpha_m^r(k) I_m(k\rho)\cos(m\phi), \tag{17.4.92}$$

$$\mathrm{Se}_r(u,q)\ \mathrm{se}_r(v,q) = \sum_{m=1}^{\infty} \beta_m^r(k) I_m(k\rho)\sin(m\phi). \tag{17.4.93}$$

For further reference, we will call the quantities $\alpha_m^r(k)$ and $\beta_m^r(k)$ *Mathieu-Bessel connection coefficients*. Let us employ these identities in (4.91) to find the results

$$\begin{aligned} &\sum_{r=0}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz)[\tilde{\tilde{F}}_r^c(k)/\mathrm{Ce}_r'(U,q)]\mathrm{Ce}_r(u,q)\ \mathrm{ce}_r(v,q) \\ = &\sum_{m=0}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz) I_m(k\rho)\cos(m\phi)\sum_{r=0}^{\infty}\alpha_m^r(k)[\tilde{\tilde{F}}_r^c(k)/\mathrm{Ce}_r'(U,q)], \end{aligned} \tag{17.4.94}$$

and

$$\begin{aligned} &\sum_{r=1}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz)[\tilde{\tilde{F}}_r^s(k)/\mathrm{Se}_r'(U,q)]\mathrm{Se}_r(u,q)\ \mathrm{se}_r(v,q) \\ = &\sum_{m=1}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz) I_m(k\rho)\sin(m\phi)\sum_{r=1}^{\infty}\beta_m^r(k)[\tilde{\tilde{F}}_r^s(k)/\mathrm{Se}_r'(U,q)]. \end{aligned} \tag{17.4.95}$$

Using these results, (4.91) can be rewritten in the form

$$\begin{aligned} \psi(x,y,z) \ = \ &\sum_{m=0}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz) I_m(k\rho)\cos(m\phi)\sum_{r=0}^{\infty}\alpha_m^r(k)[\tilde{\tilde{F}}_r^c(k)/\mathrm{Ce}_r'(U,q)] \\ + &\sum_{m=1}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz) I_m(k\rho)\sin(m\phi)\sum_{r=1}^{\infty}\beta_m^r(k)[\tilde{\tilde{F}}_r^s(k)/\mathrm{Se}_r'(U,q)]. \end{aligned} \tag{17.4.96}$$

Upon comparing (4.96) with (14.2.54), we conclude that there are the relations

$$G_{m,c}(k) = \sum_{r=0}^{\infty}\alpha_m^r(k)[\tilde{\tilde{F}}_r^c(k)/\mathrm{Ce}_r'(U,q)], \tag{17.4.97}$$

and

$$G_{m,s}(k) = \sum_{r=1}^{\infty}\beta_m^r(k)[\tilde{\tilde{F}}_r^s(k)/\mathrm{Se}_r'(U,q)]. \tag{17.4.98}$$

We remark that it can be shown that the real parts of the $G_{m,\alpha}$ are even in k, and the imaginary parts are odd in k. See Exercise 4.2. Finally, in view of (14.2.55) and (14.2.56), we have the desired results

$$\begin{aligned} C_{m,c}^{[n]}(z) &= i^n(1/2)^m(1/m!)\int_{-\infty}^{\infty} dk\exp(ikz)k^{n+m}\sum_{r=0}^{\infty}\alpha_m^r(k)[\tilde{\tilde{F}}_r^c(k)/\mathrm{Ce}_r'(U,q)] \\ &= i^n(1/2)^m(1/m!)\int_{-\infty}^{\infty} dk\exp(ikz)k^{n+m}G_{m,c}(k), \end{aligned} \tag{17.4.99}$$

$$\begin{aligned} C_{m,s}^{[n]}(z) &= i^n(1/2)^m(1/m!)\int_{-\infty}^{\infty} dk\exp(ikz)k^{n+m}\sum_{r=1}^{\infty}\beta_m^r(k)[\tilde{\tilde{F}}_r^s(k)/\mathrm{Se}_r'(U,q)] \\ &= i^n(1/2)^m(1/m!)\int_{-\infty}^{\infty} dk\exp(ikz)k^{n+m}G_{m,s}(k). \end{aligned} \tag{17.4.100}$$

We have found expressions for the generalized gradients in terms of field data (normal component) on the surface of an elliptic cylinder. These results hold for the cases $m \geq 1$. When $m = 0$ there are the results

$$\begin{aligned} C_{m=0,c}^{[n]}(z) &= C_0^{[n]}(z) = i^n\int_{-\infty}^{\infty} dk\exp(ikz)k^n\sum_{r=0}^{\infty}\alpha_0^r(k)[\tilde{\tilde{F}}_r^c(k)/\mathrm{Ce}_r'(U,q)] \\ &= i^n\int_{-\infty}^{\infty} dk\exp(ikz)k^nG_{0,c}(k), \end{aligned} \tag{17.4.101}$$

$$C_{m=0,s}^{[n]}(z) = 0. \tag{17.4.102}$$

Just as in the $m = 0$ case for the circular cylinder, so too here it may be better to derive and employ formulas based on the tangential component B_z rather than the normal component. See Section 19.2.

Exercises

17.4.1. Verify that (4.1) and (4.2) can be written in the form (4.7).

17.4.2. The purpose of this exercise is to study the analytic properties of elliptic coordinates. Our discussion is based on the relation (4.7). According to (4.7), the function $\zeta(w)$ is an entire function of w. What can be said about its inverse $w(\zeta)$? Verify that (4.7) has the inverse

$$w = \cosh^{-1}(\zeta/f) = \log[(\zeta/f) + \sqrt{(\zeta/f)^2 - 1}]. \tag{17.4.103}$$

Evidently $w(\zeta)$ has branch points at $\zeta = \pm f$. Verify that $w(\zeta)$ is analytic in the cut ζ plane with a cut consisting of a straight line extending from $\zeta = -f$ to $\zeta = f$ as illustrated in Figure 4.2.

17.4.3. Exercise on wave equation.

17.5 Use of Field Data on Surface of Rectangular Cylinder

17.5.1 Finding the Magnetic Scalar Potential $\psi(x, y, z)$

Consider the domain $x \in [-W/2, W/2]$, $y \in [-H/2, H/2]$, $z \in [-\infty, \infty]$. This domain is the interior and surface of a cylinder of infinite extent in the $\pm z$ direction, centered on the z axis, and having rectangular cross section with width W and height H. In this section we will describe the use of field data on the surface of this cylinder. This cylinder has 4 surfaces (sides) which we will call t and b for *top* and *bottom*, and ℓ and r for *left* and *right*.

Suppose that we are given the normal component of the magnetic field on the top, bottom, left, and right surfaces. That is, we are given the field data

$$B_y^t(x, z) = B_y(x, y, z)|_{y=H/2} \text{ with } x \in [-W/2, W/2] \text{ and } z \in [-\infty, \infty], \tag{17.5.1}$$

$$B_y^b(x, z) = B_y(x, y, z)|_{y=-H/2} \text{ with } x \in [-W/2, W/2] \text{ and } z \in [-\infty, \infty], \tag{17.5.2}$$

$$B_x^\ell(y, z) = B_x(x, y, z)|_{x=-W/2} \text{ with } y \in [-H/2, H/2] \text{ and } z \in [-\infty, \infty], \tag{17.5.3}$$

$$B_x^r(y, z) = B_x(x, y, z)|_{x=W/2} \text{ with } y \in [-H/2, H/2] \text{ and } z \in [-\infty, \infty]. \tag{17.5.4}$$

Our goal is to find the scalar magnetic potential $\psi(x, y, z)$ in terms of this field data.

For purposes of Fourier analysis, we will *extend* the normal field surface data beyond the ranges indicated above as follows:

- Extend $B_y^t(x, z)$ and $B_y^b(x, z)$ from the interval $x \in [-W/2, W/2]$ to the extended interval $x \in [-W/2, 3W/2]$ by requiring that the extension be *even* in x about the value $x = W/2$. Note that the extended interval $x \in [-W/2, 3W/2]$ can be written in the form
$$[-W/2, 3W/2] = [W/2 - W, W/2 + W]. \tag{17.5.5}$$

- Extend $B_x^\ell(y, z)$ and $B_x^b(y, z)$ from the interval $y \in [-H/2, H/2]$ to the extended interval $y \in [-H/2, 3H/2]$ by requiring that the extension be *even* in y about the value $y = H/2$. Note that the extended interval $[-H/2, 3H/2]$ can be written in the form
$$[-H/2, 3H/2] = [H/2 - H, H/2 + H]. \tag{17.5.6}$$

So doing will produce functions that are continuous over the extended intervals

$$x \in [W/2 - W, W/2 + W] \text{ and } y \in [H/2 - H, H/2 + H]. \tag{17.5.7}$$

Moreover, by construction, the extended functions will take the *same* value at both ends of each extended interval. Therefore they can be *further* extended beyond their extended intervals, both to the left and the right, in a *continuous* way by requiring that their further extensions be periodic with periods $2W$ and $2H$, respectively. (Moreover, these extensions will also be even about the values $x = -W/2$ and $y = -H/2$, respectively.) In summary, we have produced extensions that are even about $\pm W/2$ or $\pm H/2$, are periodic with periods

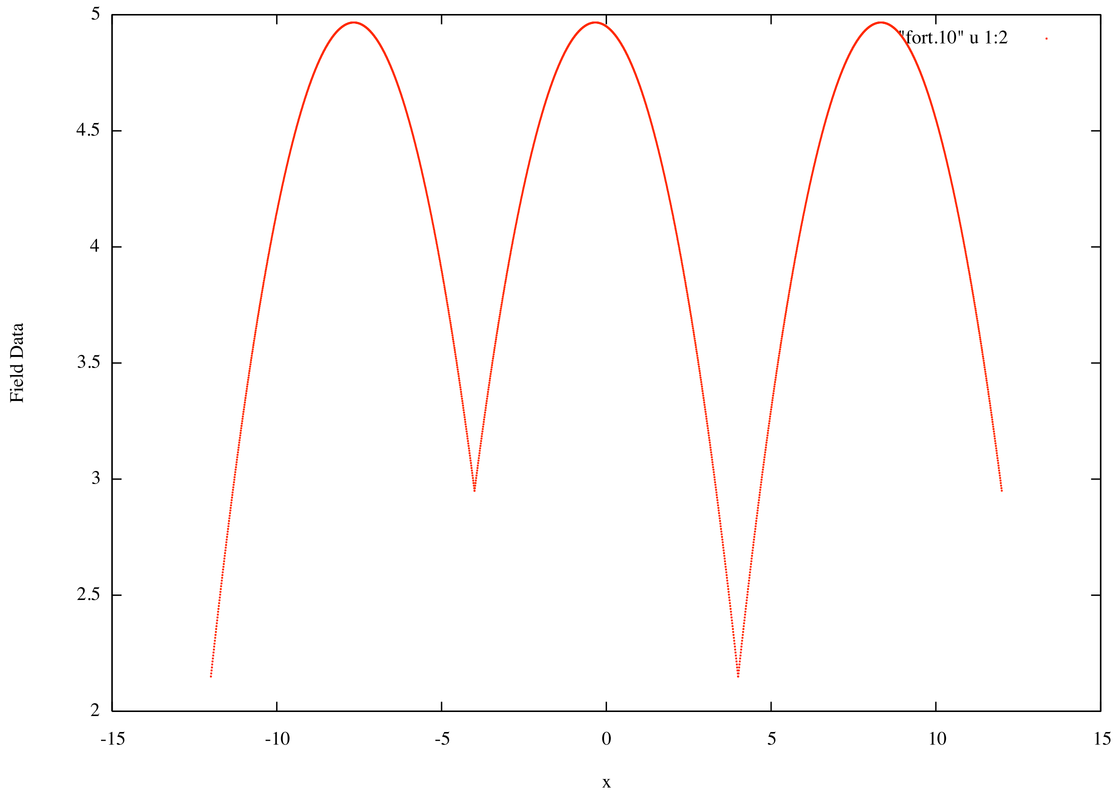


Figure 17.5.1: Hypothetical $B_y^t(x,z)$ data in the interval $x \in [-W/2, W/2]$ and its extension to the full x axis to facilitate Fourier analysis. In this example, $W = 8$ so that $[-W/2, W/2] = [-4, 4]$. The extension has period $2W = 16$, is even about the points $x = \pm W/2 = \pm 4$ and their periodic counterparts, and is continuous. Generally, the first derivative is discontinuous at the points $x = \pm W/2 = \pm 4$ and their periodic counterparts.

$2W$ or $2H$, and are continuous. See, for example, Figure 5.1. They are therefore ideally suited to Fourier analysis over the extended intervals (5.7).

Now consider the functions

$$\cos[(x + W/2)(n\pi/W)] \text{ with } n = 0, 1, 2, \cdots \tag{17.5.8}$$

and the functions

$$\sin[(x + W/2)(n\pi/W)] \text{ with } n = 1, 2, \cdots . \tag{17.5.9}$$

They have period $2W$ and form a complete orthogonal set over the interval (5.5). Therefore $B_y^t(x, z)$ and $B_y^b(x, z)$, when extended as described above, can be expanded in terms of them. Note also that the cosine functions (5.8) are even about $x = W/2$ and the sine functions (5.9) are odd. Since (by construction) the extended $B_y^t(x, z)$ and $B_y^b(x, z)$ are even about $x = W/2$, it follows that only the cosine terms will appear in the Fourier expansion. Thus, we have the Fourier representations

$$B_y^t(x, z) = \sum_{n=0}^{\infty} \tilde{B}_y^t(n, z) \cos[(x + W/2)(n\pi/W)], \tag{17.5.10}$$

$$B_y^b(x, z) = \sum_{n=0}^{\infty} \tilde{B}_y^b(n, z) \cos[(x + W/2)(n\pi/W)], \tag{17.5.11}$$

where

$$\tilde{B}_y^t(0, z) = [1/(2W)] \int_{-W/2}^{3W/2} dx\ B_y^t(x, z) = (1/W) \int_{-W/2}^{W/2} dx\ B_y^t(x, z), \tag{17.5.12}$$

$$\begin{aligned} \tilde{B}_y^t(n, z) &= (1/W) \int_{-W/2}^{3W/2} dx\ B_y^t(x, z) \cos[(x + W/2)(n\pi/W)] \\ &= (2/W) \int_{-W/2}^{W/2} dx\ B_y^t(x, z) \cos[(x + W/2)(n\pi/W)] \text{ for } n > 0, \end{aligned} \tag{17.5.13}$$

$$\tilde{B}_y^b(0, z) = [1/(2W)] \int_{-W/2}^{3W/2} dx\ B_y^b(x, z) = (1/W) \int_{-W/2}^{W/2} dx\ B_y^b(x, z), \tag{17.5.14}$$

$$\begin{aligned} \tilde{B}_y^b(n, z) &= (1/W) \int_{-W/2}^{3W/2} dx\ B_y^b(x, z) \cos[(x + W/2)(n\pi/W)] \\ &= (2/W) \int_{-W/2}^{W/2} dx\ B_y^b(x, z) \cos[(x + W/2)(n\pi/W)] \text{ for } n > 0. \end{aligned} \tag{17.5.15}$$

Here we have used the fact that $B_y^t(x, z)$ and $B_y^b(x, z)$ and the cosine functions are even about $x = W/2$.

We have already seen that the fully extended $B_y^t(x, z)$ and $B_y^b(x, z)$ are continuous and periodic with period $2W$. However, they will generally not have continuous first derivatives across the joins at $x = \pm W/2$, $x = \pm 3W/2$, etc. It follows from standard Fourier analysis theory that the coefficients $\tilde{B}^t_y(n, z)$ and $\tilde{B}^b_y(n, z)$ must, in general, fall off like $1/n^2$,

$$\tilde{B}^t_y(n, z) \sim (1/n^2) \text{ as } n \to \infty, \text{ etc.} \tag{17.5.16}$$

Therefore the series (5.10) and (5.11) are point-wise absolutely, but not wonderfully, convergent. The slow falloff (5.16) is the price to be paid for working with a bounding surface that has sharp corners. By contrast, it can be shown that the analogous coefficients in the cases of circular and elliptic cylinders fall off much more rapidly, namely as $(1/\Lambda)^{|n|}$ for some $\Lambda > 1$. See Exercises 16.2.3 and 16.2.4.

In a similar way, we have the Fourier representations

$$B^\ell_x(y, z) = \sum_{n=0}^{\infty} \tilde{B}^\ell_y(n, z) \cos[(y + H/2)(n\pi/H)], \tag{17.5.17}$$

$$B^r_x(y, z) = \sum_{n=0}^{\infty} \tilde{B}^r_y(n, z) \cos[(y + H/2)(n\pi/H)], \tag{17.5.18}$$

where

$$\tilde{B}^\ell_x(0, z) = [1/(2H)] \int_{-H/2}^{3H/2} dy\; B^\ell_x(y, z) = (1/H) \int_{-H/2}^{H/2} dx\; B^\ell_x(y, z), \tag{17.5.19}$$

$$\begin{aligned} \tilde{B}^\ell_x(n, z) &= (1/H) \int_{-H/2}^{3H/2} dy\; B^\ell_x(y, z) \cos[(y + H/2)(n\pi/H)] \\ &= (2/H) \int_{-H/2}^{H/2} dy\; B^\ell_x(y, z) \cos[(y + H/2)(n\pi/H)] \text{ for } n > 0, \end{aligned} \tag{17.5.20}$$

$$\tilde{B}^r_x(0, z) = [1/(2H)] \int_{-H/2}^{3H/2} dy\; B^r_x(y, z) = (1/H) \int_{-H/2}^{H/2} dx\; B^r_x(y, z), \tag{17.5.21}$$

$$\begin{aligned} \tilde{B}^r_x(n, z) &= (1/H) \int_{-H/2}^{3H/2} dy\; B^r_x(y, z) \cos[(y + H/2)(n\pi/H)] \\ &= (2/H) \int_{-H/2}^{H/2} dy\; B^r_x(y, z) \cos[(y + H/2)(n\pi/H)] \text{ for } n > 0. \end{aligned} \tag{17.5.22}$$

To proceed further, we perform Fourier transforms in z. Thus, we make the definitions

$$\tilde{\tilde{B}}^t_y(n, k) = [1/(2\pi)] \int_{-\infty}^{\infty} dz\; \exp(-ikz) \tilde{B}^t_y(n, z), \tag{17.5.23}$$

$$\tilde{\tilde{B}}^b_y(n, k) = [1/(2\pi)] \int_{-\infty}^{\infty} dz\; \exp(-ikz) \tilde{B}^b_y(n, z), \tag{17.5.24}$$

$$\tilde{\tilde{B}}_x^{\ell}(n,k) = [1/(2\pi)]\int_{-\infty}^{\infty} dz\ \exp(-ikz)\tilde{B}_x^{\ell}(n,z), \tag{17.5.25}$$

$$\tilde{\tilde{B}}_x^{r}(n,k) = [1/(2\pi)]\int_{-\infty}^{\infty} dz\ \exp(-ikz)\tilde{B}_x^{r}(n,z), \tag{17.5.26}$$

Note that the various $\tilde{B}$ terms on the right sides of (5.23) through (5.26) are real. It follows that the real parts of the various $\tilde{\tilde{B}}$ terms on the left sides of (5.23) through (5.26) are even in k, and the imaginary parts are odd in k.

With these definitions in hand, we are ready to determine the scalar potential $\psi(x,y,z)$ in terms of surface field values. First we note, as is easily checked, that functions of the form

$$\exp(ikz)\cos[(x+W/2)(n\pi/W)]\cosh[\sigma_n(y \pm H/2)] \tag{17.5.27}$$

and

$$\exp(ikz)\cos[(y+H/2)(n\pi/H)]\cosh[\tau_n(x \pm W/2)], \tag{17.5.28}$$

where

$$\sigma_n = [k^2 + (n\pi/W)^2]^{1/2} \tag{17.5.29}$$

and

$$\tau_n = [k^2 + (n\pi/H)^2]^{1/2}, \tag{17.5.30}$$

satisfy Laplace's equation. Next define functions $\psi^t(x,y,z)$, $\psi^b(x,y,z)$, $\psi^{\ell}(x,y,z)$, and $\psi^r(x,y,z)$ by the relations

$$\psi^t(x,y,z) = \int_{-\infty}^{\infty} dk \exp(ikz)\sum_{n=0}^{\infty}\frac{\tilde{\tilde{B}}_y^{t}(n,k)\cos[(x+W/2)(n\pi/W)]\cosh[\sigma_n(y+H/2)]}{\sigma_n\sinh(H\sigma_n)}, \tag{17.5.31}$$

$$\psi^b(x,y,z) = \int_{-\infty}^{\infty} dk \exp(ikz)\sum_{n=0}^{\infty}\frac{\tilde{\tilde{B}}_y^{b}(n,k)\cos[(x+W/2)(n\pi/W)]\cosh[\sigma_n(y-H/2)]}{\sigma_n\sinh(H\sigma_n)}, \tag{17.5.32}$$

$$\psi^{\ell}(x,y,z) = \int_{-\infty}^{\infty} dk \exp(ikz)\sum_{n=0}^{\infty}\frac{\tilde{\tilde{B}}_y^{\ell}(n,k)\cos[(y+H/2)(n\pi/H)]\cosh[\tau_n(x+W/2)]}{\tau_n\sinh(W\tau_n)}, \tag{17.5.33}$$

$$\psi^{r}(x,y,z) = \int_{-\infty}^{\infty} dk \exp(ikz)\sum_{n=0}^{\infty}\frac{\tilde{\tilde{B}}_y^{r}(n,k)\cos[(y+H/2)(n\pi/H)]\cosh[\tau_n(x-W/2)]}{\tau_n\sinh(W\tau_n)}. \tag{17.5.34}$$

Evidently by construction they all satisfy Laplace's equation.

Now watch closely: From the definition (5.31) we have the result

$$\partial\psi^t(x,y,z)/\partial x = -\int_{-\infty}^{\infty} dk \exp(ikz)\sum_{n=0}^{\infty}\frac{\tilde{\tilde{B}}_y^{t}(n,k)(n\pi/W)\sin[(x+W/2)(n\pi/W)]\cosh[\sigma_n(y+H/2)]}{\sigma_n\sinh(H\sigma_n)}. \tag{17.5.35}$$

[Note that we may interchange summation and differentiation: the series (5.31) converges absolutely and uniformly due to the $\sigma_n \sinh(H\sigma_n)$ denominator.] It follows that

$$[\partial\psi^t(x,y,z)/\partial x]|_{x=\pm W/2} = 0. \tag{17.5.36}$$

Similarly, we find that

$$[\partial\psi^b(x,y,z)/\partial x]|_{x=\pm W/2} = 0. \tag{17.5.37}$$

Also from (5.31) we find that

$$\partial\psi^t(x,y,z)/\partial y =$$
$$\int_{-\infty}^{\infty} dk\exp(ikz)\sum_{n=0}^{\infty}\frac{\tilde{\tilde{B}}{}_y^t(n,k)\cos[(x+W/2)(n\pi/W)]\sinh[\sigma_n(y+H/2)]}{\sinh(H\sigma_n)}. \tag{17.5.38}$$

It follows that

$$[\partial\psi^t(x,y,z)/\partial y]|_{y=-H/2} = 0 \tag{17.5.39}$$

and

$$\begin{aligned}
[\partial\psi^t(x,y,z)/\partial y]|_{y=H/2} &= \int_{-\infty}^{\infty} dk\exp(ikz)\sum_{n=0}^{\infty}\tilde{\tilde{B}}{}_y^t(n,k)\cos[(x+W/2)(n\pi/W)] \\
&= \sum_{n=0}^{\infty}\tilde{B}_y^t(n,z)\cos[(x+W/2)(n\pi/W)] \\
&= B_y^t(x,z).
\end{aligned} \tag{17.5.40}$$

Similarly, we find that

$$[\partial\psi^b(x,y,z)/\partial y]|_{y=H/2} = 0 \tag{17.5.41}$$

and

$$\begin{aligned}
[\partial\psi^b(x,y,z)/\partial y]|_{y=-H/2} &= \int_{-\infty}^{\infty} dk\exp(ikz)\sum_{n=0}^{\infty}\tilde{\tilde{B}}{}_y^b(n,k)\cos[(x+W/2)(n\pi/W)] \\
&= \sum_{n=0}^{\infty}\tilde{B}_y^b(n,z)\cos[(x+W/2)(n\pi/W)] \\
&= B_y^b(x,z).
\end{aligned} \tag{17.5.42}$$

Finally, analogous results hold for $\psi^\ell(x,y,z)$ and $\psi^r(x,y,z)$. They satisfy the relations

$$[\partial\psi^\ell(x,y,z)/\partial y]|_{y=\pm H/2} = 0, \tag{17.5.43}$$

$$[\partial\psi^r(x,y,z)/\partial y]|_{y=\pm H/2} = 0, \tag{17.5.44}$$

$$[\partial\psi^\ell(x,y,z)/\partial x]|_{x=W/2} = 0, \tag{17.5.45}$$

$$[\partial\psi^\ell(x,y,z)/\partial x]|_{x=-W/2} = B_x^\ell(y,z), \tag{17.5.46}$$

$$[\partial\psi^r(x,y,z)/\partial x]|_{x=-W/2} = 0, \tag{17.5.47}$$

$$[\partial\psi^r(x,y,z)/\partial x]|_{x=W/2} = B_x^r(y,z). \tag{17.5.48}$$

At last we are ready to construct $\psi(x,y,z)$. We make the definition

$$\psi(x,y,z) = \psi^t(x,y,z) + \psi^b(x,y,z) + \psi^\ell(x,y,z) + \psi^r(x,y,z). \tag{17.5.49}$$

Evidently this ψ satisfies Laplace's equation. Also we find that

$$\begin{aligned}&[\partial\psi(x,y,z)/\partial y]|_{y=H/2} = \\ &[\partial\psi^t(x,y,z)/\partial y]|_{y=H/2} + [\partial\psi^b(x,y,z)/\partial y]|_{y=H/2} \\ &+[\partial\psi^\ell(x,y,z)/\partial y]|_{y=H/2} + [\partial\psi^r(x,y,z)/\partial y]|_{y=H/2} \\ &= B_y^t(x,z).\end{aligned} \tag{17.5.50}$$

Here we have used (5.35) through (5.48). Similarly we find that

$$[\partial\psi(x,y,z)/\partial y]|_{y=-H/2} = B_y^b(x,z), \tag{17.5.51}$$

$$[\partial\psi(x,y,z)/\partial x]|_{x=-W/2} = B_x^\ell(y,z), \tag{17.5.52}$$

$$[\partial\psi(x,y,z)/\partial x]|_{x=W/2} = B_x^r(y,z). \tag{17.5.53}$$

Thus ψ satisfies the required boundary conditions on all four surfaces. Finally, observe that the quantities appearing on the right sides of (5.31) through (5.34) are *entire* functions of x, y, and z. Now suppose that x, y have values corresponding to a point *inside* the cylinder. Then, thanks to the denominators appearing on the right sides of (5.31) through (5.34), the integrals and sums are rapidly convergent. It follows that $\psi(x,y,z)$ is *analytic* in x, y, z for all points within the cylinder.

17.5.2 Finding the On-Axis Gradients

The remaining task is to determine the on-axis gradients by finding the cylindrical harmonic expansion for ψ as given by (5.31) through (5.34) and (5.49). In analogy to the case of the elliptical cylinder, this will be done with the aid of what may be called *Fourier-Bessel connection coefficients*. Namely, there are the formulas

$$\begin{aligned}&\cos[(x+W/2)(j\pi/W)]\cosh[\sigma_j(y+H/2)] = \\ &\sum_{m=0}^{\infty} d_{mj}^{tc}(k) I_m(k\rho)\cos(m\phi) + \sum_{m=1}^{\infty} d_{mj}^{ts}(k) I_m(k\rho)\sin(m\phi),\end{aligned} \tag{17.5.54}$$

$$\begin{aligned}&\cos[(x+W/2)(j\pi/W)]\cosh[\sigma_j(y-H/2)] = \\ &\sum_{m=0}^{\infty} d_{mj}^{bc}(k) I_m(k\rho)\cos(m\phi) + \sum_{m=1}^{\infty} d_{mj}^{bs}(k) I_m(k\rho)\sin(m\phi),\end{aligned} \tag{17.5.55}$$

$$\begin{aligned}&\cos[(y+H/2)(j\pi/H)]\cosh[\tau_j(x+W/2)] = \\ &\sum_{m=0}^{\infty} d_{mj}^{\ell c}(k) I_m(k\rho)\cos(m\phi) + \sum_{m=1}^{\infty} d_{mj}^{\ell s}(k) I_m(k\rho)\sin(m\phi),\end{aligned} \tag{17.5.56}$$

$$\cos[(y+H/2)(j\pi/H)]\cosh[\tau_j(x-W/2)] =$$
$$\sum_{m=0}^{\infty} d_{mj}^{rc}(k) I_m(k\rho)\cos(m\phi) + \sum_{m=1}^{\infty} d_{mj}^{rs}(k) I_m(k\rho)\sin(m\phi). \tag{17.5.57}$$

We will derive them shortly. Before doing so, we will use them to find the on-axis gradients. Suppose we employ (5.54) in (5.31). Doing so gives the result

$$\psi^t(x,y,z) = \sum_{m=0}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz) I_m(k\rho)\cos(m\phi)\sum_{j=0}^{\infty}\frac{d_{mj}^{tc}(k)\tilde{\tilde{B}}_y^t(j,k)}{\sigma_j\sinh(H\sigma_j)}$$
$$+\sum_{m=1}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz) I_m(k\rho)\sin(m\phi)\sum_{j=0}^{\infty}\frac{d_{mj}^{ts}(k)\tilde{\tilde{B}}_y^t(j,k)}{\sigma_j\sinh(H\sigma_j)}. \tag{17.5.58}$$

Similarly, we find the relations

$$\psi^b(x,y,z) = \sum_{m=0}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz) I_m(k\rho)\cos(m\phi)\sum_{j=0}^{\infty}\frac{d_{mj}^{bc}(k)\tilde{\tilde{B}}_y^b(j,k)}{\sigma_j\sinh(H\sigma_j)}$$
$$+\sum_{m=1}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz) I_m(k\rho)\sin(m\phi)\sum_{j=0}^{\infty}\frac{d_{mj}^{bs}(k)\tilde{\tilde{B}}_y^b(j,k)}{\sigma_j\sinh(H\sigma_j)}, \tag{17.5.59}$$

$$\psi^\ell(x,y,z) = \sum_{m=0}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz) I_m(k\rho)\cos(m\phi)\sum_{j=0}^{\infty}\frac{d_{mj}^{\ell c}(k)\tilde{\tilde{B}}_y^\ell(j,k)}{\tau_j\sinh(W\tau_j)}$$
$$+\sum_{m=1}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz) I_m(k\rho)\sin(m\phi)\sum_{j=0}^{\infty}\frac{d_{mj}^{\ell s}(k)\tilde{\tilde{B}}_y^\ell(j,k)}{\tau_j\sinh(W\tau_j)}, \tag{17.5.60}$$

$$\psi^r(x,y,z) = \sum_{m=0}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz) I_m(k\rho)\cos(m\phi)\sum_{j=0}^{\infty}\frac{d_{mj}^{rc}(k)\tilde{\tilde{B}}_y^r(j,k)}{\tau_j\sinh(W\tau_j)}$$
$$+\sum_{m=1}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz) I_m(k\rho)\sin(m\phi)\sum_{j=0}^{\infty}\frac{c_{mj}^{rs}(k)\tilde{\tilde{B}}_y^r(j,k)}{\tau_j\sinh(W\tau_j)}. \tag{17.5.61}$$

It follows that ψ can be written in the form

$$\psi(x,y,z) = \sum_{m=0}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz) I_m(k\rho)\cos(m\phi) G_{m,c}(k)$$
$$+\sum_{m=1}^{\infty}\int_{-\infty}^{\infty} dk\ \exp(ikz) I_m(k\rho)\sin(m\phi) G_{m,s}(k) \tag{17.5.62}$$

where

$$G_{m,c}(k) = \sum_{j=0}^{\infty}[\frac{d^{tc}_{mj}(k)\tilde{\tilde{B}}^{t}_y(j,k) + d^{bc}_{mj}(k)\tilde{\tilde{B}}^{b}_y(j,k)}{\sigma_j \sinh(H\sigma_j)} + \frac{d^{\ell c}_{mj}(k)\tilde{\tilde{B}}^{\ell}_y(j,k) + d^{rc}_{mj}(k)\tilde{\tilde{B}}^{r}_y(j,k)}{\tau_j \sinh(W\tau_j)}], \tag{17.5.63}$$

$$G_{m,s}(k) = \sum_{j=0}^{\infty}[\frac{d^{ts}_{mj}(k)\tilde{\tilde{B}}^{t}_y(j,k) + d^{bs}_{mj}(k)\tilde{\tilde{B}}^{b}_y(j,k)}{\sigma_j \sinh(H\sigma_j)} + \frac{d^{\ell s}_{mj}(k)\tilde{\tilde{B}}^{\ell}_y(j,k) + d^{rs}_{mj}(k)\tilde{\tilde{B}}^{r}_y(j,k)}{\tau_j \sinh(W\tau_j)}]. \tag{17.5.64}$$

We remark that it can be shown that the real parts of the $G_{m,\alpha}$ are even in k, and the imaginary parts are odd in k. See Exercise 5.2. Finally, in view of (14.2.55) and (14.2.56), we have the desired results

$$C^{[n]}_{m,c}(z) = i^n(1/2)^m(1/m!)\int_{-\infty}^{\infty} dk \exp(ikz)k^{n+m}G_{m,c}(k), \tag{17.5.65}$$

$$C^{[n]}_{m,s}(z) = i^n(1/2)^m(1/m!)\int_{-\infty}^{\infty} dk \exp(ikz)k^{n+m}G_{m,s}(k). \tag{17.5.66}$$

We have found expressions for the generalized gradients in terms of field data (normal component) on the surface of a rectangular cylinder. These results hold for the cases $m \geq 1$. When $m = 0$ there are the results

$$C^{[n]}_{m=0,c}(z) \quad = \quad C^{[n]}_0(z) = i^n \int_{-\infty}^{\infty} dk \exp(ikz)k^n G_{0,c}(k), \tag{17.5.67}$$

$$C^{[n]}_{m=0,s}(z) = 0. \tag{17.5.68}$$

Just as in the $m = 0$ case for the circular and elliptical cylinder, in the rectangular case it may also be better to derive and employ formulas based on the tangential component B_z rather than the normal component. See Section 18.2.

17.5.3 Fourier-Bessel Connection Coefficients

The purpose of this subsection is to derive the Fourier-Bessel connection coefficients postulated in (5.54) through (5.57). We will do so in pieces. First, recall the Bessel generating-function relation

$$\begin{aligned}\exp[z\cos(\theta)] \quad &= \quad I_0(z) + 2\sum_{m=1}^{\infty}\cos(m\theta)I_m(z) \\ &= \quad \sum_{m=-\infty}^{\infty}\exp(im\theta)I_m(z).\end{aligned} \tag{17.5.69}$$

Here we have again employed (14.2.12). Second, we have been taught from the cradle that circular and hyperbolic functions, which appear on the left sides of (5.54) through (5.57), are made of exponential functions. We will see that the combinations of these exponential functions that occur in (5.54) through (5.57) can in turn be written in a form that enables the use of (5.69).

Begin with (5.54), whose left side can be written in the expanded form

$$\begin{aligned}&\cos[(x+W/2)(j\pi/W)]\cosh[\sigma_j(y+H/2)] = (1/4)\times\\ &\{\exp[i(x+W/2)(j\pi/W)]+\exp[-i(x+W/2)(j\pi/W)]\}\times\\ &\{\exp[\sigma_j(y+H/2)]+\exp[-\sigma_j(y+H/2)]\}.\end{aligned} \tag{17.5.70}$$

Multiplying out the factors that occur on the right side of (5.70) produces a sum of four terms:

$$\begin{aligned}&\cos[(x+W/2)(j\pi/W)]\cosh[\sigma_j(y+H/2)] = (1/4)\times\\ &[(i)^j\exp(\sigma_jH/2)\exp(ij\pi x/W)\exp(\sigma_jy)+\\ &(i)^j\exp(-\sigma_jH/2)\exp(ij\pi x/W)\exp(-\sigma_jy)+\\ &(-i)^j\exp(\sigma_jH/2)\exp(-ij\pi x/W)\exp(\sigma_jy)+\\ &(-i)^j\exp(-\sigma_jH/2)\exp(-ij\pi x/W)\exp(-\sigma_jy)].\end{aligned} \tag{17.5.71}$$

Here we have used the the result

$$\exp[(\pm iW/2)(j\pi/W)] = \exp(\pm ij\pi/2) = (\pm i)^j. \tag{17.5.72}$$

We see that we have to deal with products of the form $\exp(\pm ij\pi x/W)\exp(\pm\sigma_jy)$ where the signs are to be taken independently. We will treat each of these four possibilities separately.

For the ++ possibility we write

$$\exp(ij\pi x/W)\exp(\sigma_jy) = \exp(ij\pi x/W+\sigma_jy) = \exp[(ij\pi\rho/W)\cos\phi+(\rho\sigma_j)\sin\phi]. \tag{17.5.73}$$

Here we have made the substitutions (14.2.3) and (14.2.4). Next recall the identity

$$\cos(\phi+\psi) = \cos\psi\cos\phi - \sin\psi\sin\phi. \tag{17.5.74}$$

Let us write the argument appearing on the right side of (5.73) in the form

$$(ij\pi\rho/W)\cos\phi + (\rho\sigma_j)\sin\phi = \lambda\cos(\phi+\psi) = \lambda\cos\psi\cos\phi - \lambda\sin\psi\sin\phi \tag{17.5.75}$$

where λ, $\sin\psi$, and $\cos\psi$ are yet to be determined. Equating like terms in ϕ yields the relations

$$\lambda\cos\psi = ij\pi\rho/W, \tag{17.5.76}$$

$$\lambda\sin\psi = -\rho\sigma_j. \tag{17.5.77}$$

Now square both sides of (5.76) and (5.77) and add the results to obtain the relation

$$\lambda^2 = \rho^2[-(j\pi/W)^2+\sigma_j^2] = (k\rho)^2. \tag{17.5.78}$$

Here we have used (5.29). It follows that

$$\lambda = k\rho, \tag{17.5.79}$$

and we may also write

$$\cos\psi = ij\pi/(kW), \tag{17.5.80}$$

$$\sin\psi = -\sigma_j/k. \tag{17.5.81}$$

For future use, we invoke Euler to write

$$\exp(i\psi) = \cos\psi + i\sin\psi = ij\pi/(kW) - i\sigma_j/k = (i/k)(j\pi/W - \sigma_j), \tag{17.5.82}$$

$$\exp(-i\psi) = \cos\psi - i\sin\psi = ij\pi/(kW) + i\sigma_j/k = (i/k)(j\pi/W + \sigma_j). \tag{17.5.83}$$

Finally employing (5.79), first in (5.75) and then in (5.73), yields the results

$$(ij\pi\rho/W)\cos\phi + (\rho\sigma_j)\sin\phi = k\rho\cos(\phi + \psi) \tag{17.5.84}$$

and

$$\exp(ij\pi x/W)\exp(\sigma_j y) = \exp[k\rho\cos(\phi + \psi)]. \tag{17.5.85}$$

We next deal with the other sign possibilities. Consider the substitutions $\phi \rightarrow -\phi$, $\phi \rightarrow \phi + \pi$, and $\phi \rightarrow \pi - \phi$. It is readily verified from (14.2.3) and (14.2.4) that these substitutions correspond to the following substitutions in x and y:

$$\phi \rightarrow -\phi \Longleftrightarrow x \rightarrow x, y \rightarrow -y; \tag{17.5.86}$$

$$\phi \rightarrow \phi + \pi \Longleftrightarrow x \rightarrow -x, y \rightarrow -y; \tag{17.5.87}$$

$$\phi \rightarrow \pi - \phi \Longleftrightarrow x \rightarrow -x, y \rightarrow y. \tag{17.5.88}$$

Therefore, from (5.85), we find the results

$$\exp(ij\pi x/W)\exp(-\sigma_j y) = \exp[k\rho\cos(-\phi + \psi)] = \exp[k\rho\cos(\phi - \psi)], \tag{17.5.89}$$

$$\exp(-ij\pi x/W)\exp(\sigma_j y) = \exp[k\rho\cos(\pi - \phi + \psi)] = \exp[k\rho\cos(\phi - \psi - \pi)], \tag{17.5.90}$$

$$\exp(-ij\pi x/W)\exp(-\sigma_j y) = \exp[k\rho\cos(\phi + \psi + \pi)]. \tag{17.5.91}$$

Let us now see how the results for the products $\exp(\pm ij\pi x/W)\exp(\pm\sigma_j y)$, as given by (5.85) and (5.89) through (5.91), can be used in conjunction with (5.69). For the ++ case use of (5.69) and (5.85) gives the result

$$\begin{aligned} &\exp(ij\pi x/W)\exp(\sigma_j y) = \exp[k\rho\cos(\phi + \psi)] = \\ &\sum_{m=-\infty}^{\infty} \exp[im(\phi + \psi)] I_m(k\rho) = \\ &\sum_{m=-\infty}^{\infty} \exp(im\psi)\exp(im\phi) I_m(k\rho) = \\ &\sum_{m=-\infty}^{\infty} (i/k)^m (j\pi/W - \sigma_j)^m \exp(im\phi) I_m(k\rho). \end{aligned} \tag{17.5.92}$$

Here we have also used (5.82). Similarly, for the remaining cases, we find the results

$$\begin{aligned}&\exp(ij\pi x/W)\exp(-\sigma_j y) = \exp[k\rho\cos(\phi-\psi)] = \\ ,\quad &\sum_{m=-\infty}^{\infty}\exp(-im\psi)\exp(im\phi)I_m(k\rho) = \\ &\sum_{m=-\infty}^{\infty}(i/k)^m(j\pi/W+\sigma_j)^m\exp(im\phi)I_m(k\rho),\end{aligned} \tag{17.5.93}$$

$$\begin{aligned}&\exp(-ij\pi x/W)\exp(\sigma_j y) = \exp[k\rho\cos(\phi-\psi-\pi)] = \\ &\sum_{m=-\infty}^{\infty}\exp(-im\pi)\exp(-im\psi)\exp(im\phi)I_m(k\rho) = \\ &\sum_{m=-\infty}^{\infty}(-1)^m(i/k)^m(j\pi/W+\sigma_j)^m\exp(im\phi)I_m(k\rho),\end{aligned} \tag{17.5.94}$$

$$\begin{aligned}&\exp(-ij\pi x/W)\exp(-\sigma_j y) = \exp[k\rho\cos(\phi+\psi+\pi)] = \\ &\sum_{m=-\infty}^{\infty}\exp(im\pi)\exp(im\psi)\exp(im\phi)I_m(k\rho) = \\ &\sum_{m=-\infty}^{\infty}(-1)^m(i/k)^m(j\pi/W-\sigma_j)^m\exp(im\phi)I_m(k\rho).\end{aligned} \tag{17.5.95}$$

Bessel expansions have now been obtained for all the various pieces that result from expanding in exponentials the terms on the left side of (5.54). We now combine them to find a Bessel expansion for the left side of (5.54). From (5.71) and (5.92) through (5.95), we find the result

$$\begin{aligned}&\cos[(x+W/2)(j\pi/W)]\cosh[\sigma_j(y+H/2)] = \\ &\sum_{m=-\infty}^{\infty}d^t_{mj}(k)I_m(k\rho)\exp(im\phi)\end{aligned} \tag{17.5.96}$$

where

$$\begin{aligned}d^t_{mj}(k) &= (1/4)(i)^{j+m}\{[(j\pi/W-\sigma_j)/k]^m + (-1)^{j+m}[(j\pi/W+\sigma_j)/k]^m\}\exp(\sigma_j H/2) \\ &+ (1/4)(i)^{j+m}\{[(j\pi/W+\sigma_j)/k]^m + (-1)^{j+m}[(j\pi/W-\sigma_j)/k]^m\}\exp(-\sigma_j H/2).\end{aligned} \tag{17.5.97}$$

Upon comparing (5.54) and (5.96) we see that for $m \geq 1$ there are the relations

$$d^{tc}_{mj} = d^t_{mj} + d^t_{-mj}, \tag{17.5.98}$$

$$d^{ts}_{mj} = id^t_{mj} - id^t_{-mj}. \tag{17.5.99}$$

And for $m = 0$ there is the result

$$d^{tc}_{0j} = d^{t}_{0j}, \tag{17.5.100}$$

$$d^{ts}_{0j} = 0. \tag{17.5.101}$$

Let us first deal with the simplest case, that for $m = 0$. Use of (5.97) then gives the result

$$\begin{aligned} d^{t}_{0j}(k) &= (1/4)(i)^{j}\{[1 + (-1)^{j}\}\exp(\sigma_j H/2) \\ &+ (1/4)(i)^{j}\{[1 + (-1)^{j}\}\exp(-\sigma_j H/2). \end{aligned} \tag{17.5.102}$$

Therefore, for $m = 0$, we conclude that

$$d^{tc}_{0j} = 0 \text{ for } j \text{ odd}, \tag{17.5.103}$$

$$d^{tc}_{0j} = (-1)^{j/2}\cosh(\sigma_j H/2) \text{ for } j \text{ even}. \tag{17.5.104}$$

Now tackle the more complicated case $m \geq 1$. Note that there is the relation

$$[(j\pi/W - \sigma_j)/k][(j\pi/W + \sigma_j)/k] = [(j\pi/W)^2 - (\sigma_j^2)]/k^2 = -k^2/k^2 = -1. \tag{17.5.105}$$

Here we have again used (5.29). Consequently, we find that

$$\begin{aligned} d^{t}_{-mj}(k) &= (1/4)(i)^{j-m}\{[(j\pi/W - \sigma_j)/k]^{-m} + (-1)^{j-m}[(j\pi/W + \sigma_j)/k]^{-m}\}\exp(\sigma_j H/2) \\ &+ (1/4)(i)^{j-m}\{[(j\pi/W + \sigma_j)/k]^{-m} + (-1)^{j-m}[(j\pi/W - \sigma_j)/k]^{-m}\}\exp(-\sigma_j H/2) \\ &= (1/4)(i)^{j-m}(-1)^{m}\{[(j\pi/W + \sigma_j)/k]^{m} + (-1)^{j-m}[(j\pi/W - \sigma_j)/k]^{m}\}\exp(\sigma_j H/2) \\ &+ (1/4)(i)^{j-m}(-1)^{m}\{[(j\pi/W - \sigma_j)/k]^{m} + (-1)^{j-m}[(j\pi/W + \sigma_j)/k]^{m}\}\exp(-\sigma_j H/2). \end{aligned} \tag{17.5.106}$$

We also note that

$$(i)^{j-m}(-1)^{m} = (i)^{j-m}(i)^{2m} = (i)^{j+m} \tag{17.5.107}$$

and

$$(-1)^{j-m} = (-1)^{2m}(-1)^{j-m} = (-1)^{j+m}. \tag{17.5.108}$$

Therefore, we may also write

$$\begin{aligned} d^{t}_{-mj}(k) &= (1/4)(i)^{j+m}\{[(j\pi/W + \sigma_j)/k]^{m} + (-1)^{j+m}[(j\pi/W - \sigma_j)/k]^{m}\}\exp(\sigma_j H/2) \\ &+ (1/4)(i)^{j+m}\{[(j\pi/W - \sigma_j)/k]^{m} + (-1)^{j+m}[(j\pi/W + \sigma_j)/k]^{m}\}\exp(-\sigma_j H/2). \end{aligned} \tag{17.5.109}$$

Let us now compute $d^{tc}_{mj}(k)$. It follows from (5.97), (5.98), and (5.109) that for $m \geq 1$ there is the result

$$\begin{aligned} d^{tc}_{mj}(k) &= (1/4)(i)^{j+m}[1 + (-1)^{j+m}] \times \\ &\quad \{[(j\pi/W - \sigma_j)/k]^{m} + [(j\pi/W + \sigma_j)/k]^{m}\}\exp(\sigma_j H/2) \\ &+ (1/4)(i)^{j+m}[1 + (-1)^{j+m}] \times \\ &\quad \{[(j\pi/W + \sigma_j)/k]^{m} + [(j\pi/W - \sigma_j)/k]^{m}\}\exp(-\sigma_j H/2) \\ &= (1/2)(i)^{j+m}[1 + (-1)^{j+m}] \times \\ &\quad \{[(j\pi/W - \sigma_j)/k]^{m} + [(j\pi/W + \sigma_j)/k]^{m}\}\cosh(\sigma_j H/2). \end{aligned} \tag{17.5.110}$$

Therefore, when $(j+m)$ is odd,

$$d^{tc}_{mj} = 0. \tag{17.5.111}$$

And, when $(j+m)$ is even,

$$d^{tc}_{mj} = (-1)^{(j+m)/2}\{[(j\pi/W - \sigma_j)/k]^m + [(j\pi/W + \sigma_j)/k]^m\}\cosh(\sigma_j H/2). \tag{17.5.112}$$

Finally, let us compute d^{ts}_{mj}. It follows from (5.97), (5.99), and (5.109) that for $m \geq 1$ there is the result

$$\begin{aligned} d^{ts}_{mj}(k) &= (1/4)i(i)^{j+m}[1-(-1)^{j+m}] \times \\ &\quad \{[(j\pi/W - \sigma_j)/k]^m - [(j\pi/W + \sigma_j)/k]^m\}\exp(\sigma_j H/2) \\ &+ (1/4)i(i)^{j+m}[1-(-1)^{j+m}] \times \\ &\quad \{[(j\pi/W + \sigma_j)/k]^m - [(j\pi/W - \sigma_j)/k]^m\}\exp(-\sigma_j H/2) \\ &= (1/2)i(i)^{j+m}[1-(-1)^{j+m}] \times \\ &\quad \{[(j\pi/W - \sigma_j)/k]^m - [(j\pi/W + \sigma_j)/k]^m\}\sinh(\sigma_j H/2). \end{aligned} \tag{17.5.113}$$

Therefore, when $(j+m)$ is even,

$$d^{ts}_{mj} = 0. \tag{17.5.114}$$

And, when $(j+m)$ is odd,

$$d^{ts}_{mj} = (-1)^{(j+m+1)/2}\{[(j\pi/W - \sigma_j)/k]^m - [(j\pi/W + \sigma_j)/k]^m\}\sinh(\sigma_j H/2). \tag{17.5.115}$$

We have found the Fourier-Bessel coefficients $d^{t\alpha}_{mj}(k)$. Next observe that the left sides of (5.54) and (5.55) are interchanged under the substitution $H \leftrightarrow -H$. Therefore, for $m \geq 1$, there are also the relations

$$d^{bc}_{mj} = d^{tc}_{mj}, \tag{17.5.116}$$

$$d^{bs}_{mj} = -d^{ts}_{mj}. \tag{17.5.117}$$

And for $m = 0$ there are the results

$$d^{bc}_{0j} = d^{tc}_{0j}, \tag{17.5.118}$$

$$d^{bs}_{0j} = 0. \tag{17.5.119}$$

Analogous calculations can be made to find Bessel expansions for the right sides of (5.56) and (5.57). Instead, for variety, we will take a different approach that utilizes results already obtained. Consider the relation (5.54). Using (14.2.3) and (14.2.4) we may rewrite it in the form

$$\begin{gathered} \cos\{[\rho\cos(\phi) + W/2][j\pi/W]\}\cosh\{\sigma_j[\rho\sin(\phi) + H/2]\} = \\ \sum_{m=0}^{\infty} d^{tc}_{mj}(k) I_m(k\rho)\cos(m\phi) + \sum_{m=1}^{\infty} d^{ts}_{mj}(k) I_m(k\rho)\sin(m\phi). \end{gathered} \tag{17.5.120}$$

Now make the substitution $\phi \to \phi + \pi/2$. So doing gives the result

$$\begin{gathered} \cos\{[\rho\cos(\phi+\pi/2) + W/2][j\pi/W]\}\cosh\{\sigma_j[\rho\sin(\phi+\pi/2) + H/2]\} = \\ \sum_{m=0}^{\infty} d^{tc}_{mj}(k) I_m(k\rho)\cos[m(\phi+\pi/2)] + \sum_{m=1}^{\infty} d^{ts}_{mj}(k) I_m(k\rho)\sin[m(\phi+\pi/2)]. \end{gathered} \tag{17.5.121}$$

Next employ the identities

$$\rho\cos(\phi+\pi/2) = -\rho\sin(\phi) = -y, \tag{17.5.122}$$

$$\rho\sin(\phi+\pi/2) = \rho\cos(\phi) = x, \tag{17.5.123}$$

$$\cos[m(\phi+\pi/2)] = \cos(m\phi)\cos(m\pi/2) - \sin(m\phi)\sin(m\pi/2), \tag{17.5.124}$$

$$\sin[m(\phi+\pi/2)] = \sin(m\phi)\cos(m\pi/2) + \cos(m\phi)\sin(m\pi/2). \tag{17.5.125}$$

We see that (5.120) can be rewritten in the form

$$\begin{aligned}&\cos\{[-y+W/2][j\pi/W]\}\cosh\{\sigma_j[x+H/2]\} = \\ &\sum_{m=0}^{\infty} D^c_{mj}(k)I_m(k\rho)\cos(m\phi) + \sum_{m=1}^{\infty} D^s_{mj}(k)I_m(k\rho)\sin(m\phi)\end{aligned} \tag{17.5.126}$$

where

$$D^c_{mj}(k) = d^{tc}_{mj}(k)\cos(m\pi/2) + d^{ts}_{mj}(k)\sin(m\pi/2), \tag{17.5.127}$$

$$D^s_{mj}(k) = -d^{tc}_{mj}(k)\sin(m\pi/2) + d^{ts}_{mj}(k)\cos(m\pi/2). \tag{17.5.128}$$

Now employ the already known results (5.101), (5.103), (5.104), (5.111), (5.112), (5.114), and (5.115). First, for $m=0$, we find the relations

$$D^c_{0j}(k) = d^{tc}_{0j} = (-1)^{j/2}\cosh(\sigma_j H/2) \text{ for } j \text{ even}, \tag{17.5.129}$$

$$D^c_{0j}(k) = d^{tc}_{0j} = 0 \text{ for } j \text{ odd}, \tag{17.5.130}$$

$$D^s_{0j}(k) = d^{ts}_{0j} = 0. \tag{17.5.131}$$

For the remaining $D^\alpha_{mj}(k)$ we must distinguish the cases $(j+m)$ odd and even. When $(j+m)$ is odd we find the results

$$\begin{aligned}&D^c_{mj}(k) = d^{ts}_{mj}(k)\sin(m\pi/2) \\ &= (-1)^{(j+m+1)/2}\sin(m\pi/2)\{[(j\pi/W-\sigma_j)/k]^m - [(j\pi/W+\sigma_j)/k]^m\}\sinh(\sigma_j H/2),\end{aligned} \tag{17.5.132}$$

$$\begin{aligned}&D^s_{mj}(k) = d^{ts}_{mj}(k)\cos(m\pi/2) \\ &= (-1)^{(j+m+1)/2}\cos(m\pi/2)\{[(j\pi/W-\sigma_j)/k]^m - [(j\pi/W+\sigma_j)/k]^m\}\sinh(\sigma_j H/2).\end{aligned} \tag{17.5.133}$$

And, when $(j+m)$ is even, we find the results

$$\begin{aligned}&D^c_{mj}(k) = d^{tc}_{mj}(k)\cos(m\pi/2) \\ &= (-1)^{(j+m)/2}\cos(m\pi/2)\{[(j\pi/W-\sigma_j)/k]^m + [(j\pi/W+\sigma_j)/k]^m\}\cosh(\sigma_j H/2),\end{aligned} \tag{17.5.134}$$

$$\begin{aligned} D^{s}_{mj}(k) &= -d^{tc}_{mj}(k)\sin(m\pi/2) \\ &= (-1)^{(j+m)/2}\sin(m\pi/2)\{[(j\pi/W-\sigma_j)/k]^m+[(j\pi/W+\sigma_j)/k]^m\}\cosh(\sigma_j H/2). \end{aligned} \tag{17.5.135}$$

For the penultimate step, compare the left side of (5.126) with the left side of (5.56). We see that the first is transformed into the second under the substitutions $W \to -H$, $\sigma_j \to \tau_j$, and $H \to W$. It follows that there are the relations

$$d^{\ell c}_{0j} = (-1)^{j/2}\cosh(\tau_j W/2) \text{ for } j \text{ even}, \tag{17.5.136}$$

$$d^{\ell c}_{0j} = 0 \text{ for } j \text{ odd}, \tag{17.5.137}$$

$$d^{\ell s}_{0j} = 0. \tag{17.5.138}$$

For the remaining $d^{\ell\alpha}_{mj}$ we must again distinguish the cases $(j+m)$ odd and even. When $(j+m)$ is odd we find the results

$$d^{\ell c}_{mj}(k) = (-1)^{(j+m+1)/2}(-1)^m\sin(m\pi/2)\{[(j\pi/H+\tau_j)/k]^m-[(j\pi/H-\tau_j)/k]^m\}\sinh(\tau_j W/2), \tag{17.5.139}$$

$$d^{\ell s}_{mj}(k) = (-1)^{(j+m+1)/2}(-1)^m\cos(m\pi/2)\{[(j\pi/H+\tau_j)/k]^m-[(j\pi/H-\tau_j)/k]^m\}\sinh(\tau_j W/2). \tag{17.5.140}$$

And, when $(j+m)$ is even, we find the results

$$d^{\ell c}_{mj}(k) = (-1)^{(j+m)/2}(-1)^m\cos(m\pi/2)\{[(j\pi/H+\tau_j)/k]^m+[(j\pi/H-\tau_j)/k]^m\}\cosh(\tau_j W/2), \tag{17.5.141}$$

$$d^{\ell s}_{mj}(k) = -(-1)^{(j+m)/2}(-1)^m\sin(m\pi/2)\{[(j\pi/H+\tau_j)/k]^m+[(j\pi/H-\tau_j)/k]^m\}\cosh(\tau_j W/2). \tag{17.5.142}$$

Finally, observe that the left sides of (5.56) and (5.57) are interchanged under the substitution $W \leftrightarrow -W$. It follows that for $m=0$ there are the relations

$$d^{rc}_{0j} = d^{\ell c}_{0j}, \tag{17.5.143}$$

$$d^{rs}_{0j} = 0. \tag{17.5.144}$$

For the remaining $d^{r\alpha}_{mj}$, when $(j+m)$ is odd, we find the results

$$d^{rc}_{mj}(k) = -d^{\ell c}_{mj}(k), \tag{17.5.145}$$

$$d^{rs}_{mj}(k) = -d^{\ell s}_{mj}(k). \tag{17.5.146}$$

And, when $(j+m)$ is even, we find the results

$$d^{rc}_{mj}(k) = d^{\ell c}_{mj}(k), \tag{17.5.147}$$

$$d^{rs}_{mj}(k) = d^{\ell s}_{mj}(k). \tag{17.5.148}$$

Exercises

17.5.1. How could one have known that Fourier-Bessel expansions of the form (5.54) through (5.57) must exist? Consider, for example, (5.54). Multiply both sides by $\exp(ikz)$. Show that both sides then become harmonic functions. Moreover, the left side is analytic in the vicinity of the z axis. But we know from Section 14.2.1 that such functions must have an expansion of the form (14.2.11).

17.5.2. Verify the relations (5.116) through (5.119).

17.5.3. Check the consistency of the relations (5.92) through (5.95) by verifying that they are transformed among themselves by the substitutions $W \leftrightarrow -W$ and $\sigma_j \leftrightarrow -\sigma_j$.

17.5.4. Verify that the functions $\sigma_j \sinh(H\sigma_j)$ and $\tau_j \sinh(W\tau_j)$ and the Fourier-Bessel coefficients $d_{mj}^{\beta\alpha}$ are even functions of k. Use these facts to show that the real parts of the $G_{m,\alpha}$ are even in k, and the imaginary parts are odd in k.

17.5.5. Consider the functions $\psi_j(x, y, z)$ for $j = 1, 2, 3$ defined by the relations

$$\psi_1(x, y, z) = a\cos(k_x x)\sinh(k_y y)\cos(kz + \chi) \tag{17.5.149}$$

with

$$-k_x^2 + k_y^2 = k^2; \tag{17.5.150}$$

$$\psi_2(x, y, z) = a\cosh(k_x x)\sinh(k_y y)\cos(kz + \chi) \tag{17.5.151}$$

with

$$k_x^2 + k_y^2 = k^2; \tag{17.5.152}$$

$$\psi_3(x, y, z) = a\cosh(k_x x)\sin(k_y y)\cos(kz + \chi) \tag{17.5.153}$$

with

$$k_x^2 - k_y^2 = k^2. \tag{17.5.154}$$

Verify that each ψ_j satisfies Laplace's equation, is analytic everywhere, and in particular is analytic in x, y near the z axis. Verify that each ψ_j can be written in the form (14.2.11). Verify that each ψ_j produces a vertical ($\pm y$ direction) field in the midplane $y = 0$ that oscillates in z, and therefore is some approximation (at least near the z axis) to the field of an infinitely long wiggler.

17.5.6. Consider a ψ of the form

$$\psi(x, y, z) = (a + by)\exp(kx)\exp(ikz). \tag{17.5.155}$$

Verify that this ψ satisfies Laplace's equation, is analytic everywhere, and in particular is analytic in x, y near the z axis. Verify that this ψ can be written in the form (14.2.11). In particular, verify that this ψ can be written in the form

$$\begin{aligned}\psi &= a[\mathrm{I}_0(k\rho) + 2\sum_{m=1}^{\infty}\cos(m\phi)\mathrm{I}_m(k\rho)]\exp(ikz) \\ &+ [(2b/k)\sum_{m=1}^{\infty} m\sin(m\phi)\mathrm{I}_m(k\rho)]\exp(ikz).\end{aligned} \tag{17.5.156}$$

17.6 Attempted Use of Nearly On-Axis and Midplane Field Data

As promised at the end of Section 14, here we examine other attempted approaches. All will be seen to involve what, in essence, is high-order numerical differentiation. Therefore, they are unlikely to yield reliable results beyond modest order, at best.

17.6.1 Use of Nearly On-Axis Data

Let us begin with the cylindrical multipole expansion

$$\begin{aligned}\psi(\rho,\phi,z) &= \psi(x,y,z)=\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_0^{[2\ell]}(z)\rho^{2\ell}\\ &+ \sum_{m=1}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell+m}\\ &+ \sum_{m=1}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell+m}.\end{aligned}\tag{17.6.1}$$

Suppose (6.1) is multiplied by factors of $\cos(m\phi)$ or $\sin(m\phi)$ and the integrated over ϕ. Doing so gives the results

$$\tilde{\psi}(\rho,0,z)=\int_0^{2\pi}d\phi\;\psi(\rho,\phi,z)=\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2\pi}{2^{2\ell}\ell!\ell!}C_0^{[2\ell]}(z)\rho^{2\ell},\tag{17.6.2}$$

$$\tilde{\psi}_c(\rho,m,z)=\int_0^{2\pi}d\phi\;\cos(m\phi)\psi(\rho,\phi,z)=\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{\pi\;m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell+m},\tag{17.6.3}$$

$$\tilde{\psi}_s(\rho,m,z)=\int_0^{2\pi}d\phi\;\sin(m\phi)\psi(\rho,\phi,z)=\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{\pi\;m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell+m}.\tag{17.6.4}$$

If $\psi(\rho,\phi,z)$ is known, and it is in fact provided at grid points by some three-dimensional codes, then the integrals in (6.2) through (6.4) can be computed. Moreover, we have the relations

$$C_0^{[0]}(z)=[1/(2\pi)]\lim_{\rho\to 0}\tilde{\psi}(\rho,0,z),\tag{17.6.5}$$

$$C_{m.c}^{[0]}(z)=(1/\pi)\lim_{\rho\to 0}(1/\rho^m)\tilde{\psi}_c(\rho,m,z),\tag{17.6.6}$$

$$C_{m,s}^{[0]}(z)=(1/\pi)\lim_{\rho\to 0}(1/\rho^m)\tilde{\psi}_s(\rho,m,z).\tag{17.6.7}$$

It is also in principle possible to compute the on-axis gradients from field data. Let $B_\rho(\rho,\phi,z)$ be the ρ component of $\boldsymbol{B}$. It is defined by the relation

$$B_\rho(\rho,\phi,z)=B_\rho(x,y,z)=\boldsymbol{e}_\rho\cdot\boldsymbol{B}=(x/\rho)B_x+(y/\rho)B_y=\cos(\phi)B_x+\sin(\phi)B_y.\tag{17.6.8}$$

In terms of $B_\rho(\rho,\phi,z)$, define the quantities

$$\tilde{B}_\rho(\rho,0,z) = \int_0^{2\pi} d\phi \; B_\rho(\rho,\phi,z), \tag{17.6.9}$$

$$\tilde{B}_{\rho c}(\rho,m,z) = \int_0^{2\pi} d\phi \; \cos(m\phi) B_\rho(\rho,\phi,z), \tag{17.6.10}$$

$$\tilde{B}_{\rho s}(\rho,m,z) = \int_0^{2\pi} d\phi \; \sin(m\phi) B_\rho(\rho,\phi,z). \tag{17.6.11}$$

We also know that

$$B_\rho(\rho,\phi,z) = (\partial/\partial\rho)\psi(\rho,\phi,z). \tag{17.6.12}$$

It follows that there are the relations

$$\tilde{B}_\rho(\rho,0,z) = \sum_{\ell=0}^{\infty}(-1)^\ell \frac{4\pi\ell}{2^{2\ell}\ell!\ell!} C_0^{[2\ell]}(z)\rho^{2\ell-1}, \tag{17.6.13}$$

$$\tilde{B}_{\rho c}(\rho,m,z) = \sum_{\ell=0}^{\infty}(-1)^\ell \frac{\pi \; (2\ell+m)m!}{2^{2\ell}\ell!(\ell+m)!} C_{m,c}^{[2\ell]}(z)\rho^{2\ell+m-1}, \tag{17.6.14}$$

$$\tilde{B}_{\rho s}(\rho,m,z) = \sum_{\ell=0}^{\infty}(-1)^\ell \frac{\pi \; (2\ell+m)m!}{2^{2\ell}\ell!(\ell+m)!} C_{m,s}^{[2\ell]}(z)\rho^{2\ell+m-1}. \tag{17.6.15}$$

Consequently, there are the relations

$$C_0^{[2]}(z) = (1/\pi)\lim_{\rho\to 0}(1/\rho^m)\tilde{B}_\rho(\rho,0,z), \tag{17.6.16}$$

$$C_{m,c}^{[0]}(z) = (1/\pi)\lim_{\rho\to 0}(1/\rho^?)\tilde{B}_{\rho c}(\rho,m,z), \tag{17.6.17}$$

$$C_{m,s}^{[0]}(z) = (1/\pi)\lim_{\rho\to 0}(1/\rho^m)\tilde{B}_{\rho s}(\rho,m,z). \tag{17.6.18}$$

Also, from *, we have the relation.

$$C_0^{[1]}(z) = \lim_{x,y\to 0} B_z(x,y,z) = B_z(0,0,z). \tag{17.6.19}$$

Let us now examine the feasibility of carrying out the indicated calculations based on numerical data provided on a 3-d grid. We see that, apart from (6.5) and (6.19), which can be reliably estimated, particularly if the grid is chosen so that there are grid points on the z axis, it is necessary to perform a limiting process in which both a numerator and denominator approach zero. Such a limiting process is akin to numerical differentiation. Also, to compute the $C_{m,\alpha}^{[n]}(z)$ for larger values of n, it is necessary to repeatedly differentiate the above relations with respect to z. When based on grid data, this again involves multiple numerical differentiation.

17.6.2 Use of Midplane Field Data

In place of nearly on-axis data, one might consider the use of Midplane Field Data. Use of the relations (H.1.17) through (H.1.19) gives for the nearly midplane field the expansions

$$\begin{aligned}&B_x(x,y,z)=\partial_x\psi=C_{1,c}^{[0]}(z)+x[2C_{2,c}^{[0]}(z)-(1/2)C_0^{[2]}(z)]+2yC_{2,s}^{[0]}(z)\\&+3x^2[C_{3,c}^{[0]}(z)-(1/8)C_{1,c}^{[2]}(z)]-y^2[3C_{3,c}^{[0]}(z)+(1/8)C_{1,c}^{[2]}(z)]\\&+2xy[3C_{3,s}^{[0]}(z)-(1/8)C_{1,s}^{[2]}(z)]\cdots,\end{aligned}\tag{17.6.20}$$

$$\begin{aligned}&B_y(x,y,z)=\partial_y\psi=+C_{1,s}^{[0]}(z)-y[2C_{2,c}^{[0]}(z)+(1/2)C_0^{[2]}(z)]+2xC_{2,s}^{[0]}(z)\\&-3y^2[C_{3,s}^{[0]}(z)+(1/8)C_{1,s}^{[2]}(z)]+x^2[3C_{3,s}^{[0]}(z)-(1/8)C_{1,s}^{[2]}(z)]\\&-2xy[3C_{3,c}^{[0]}(z)+(1/8)C_{1,c}^{[2](z)}]+\cdots,\end{aligned}\tag{17.6.21}$$

$$\begin{aligned}&B_z(x,y,z)=\partial_z\psi=C_0^{[1]}(z)+xC_{1,c}^{[1]}(z)+yC_{1,s}^{[1]}(z)\\&+(x^2-y^2)C_{2,c}^{[1]}(z)+2xyC_{2,s}^{[1]}(z)-(1/4)(x^2+y^2)C_0^{[3]}(z)+\cdots.\end{aligned}\tag{17.6.22}$$

Evaluating these expansions in the midplane gives the results

$$\begin{aligned}&B_x(x,y=0,z)=C_{1,c}^{[0]}(z)+x[2C_{2,c}^{[0]}(z)-(1/2)C_0^{[2]}(z)]\\&+3x^2[C_{3,c}^{[0]}(z)-(1/8)C_{1,c}^{[2]}(z)]\cdots,\end{aligned}\tag{17.6.23}$$

$$\begin{aligned}&B_y(x,y=0,z)=C_{1,s}^{[0]}(z)+2xC_{2,s}^{[0]}(z)\\&+x^2[3C_{3,s}^{[0]}(z)-(1/8)C_{1,s}^{[2]}(z)]+\cdots,\end{aligned}\tag{17.6.24}$$

$$\begin{aligned}&B_z(x,y=0,z)=C_0^{[1]}(z)+xC_{1,c}^{[1]}(z)\\&+x^2[C_{2,c}^{[1]}(z)-(1/4)C_0^{[3]}(z)]+\cdots.\end{aligned}\tag{17.6.25}$$

The relations (6.4) through (6.6) express the midplane fields in terms of on-axis gradients. These relations can be inverted to determine the on-axis gradients in terms of the midplane fields. By repeatedly differentiating them with respect to x and z and then setting $x=0$, one finds the results

$$C_0^{[1]}(z)=B_z(x=0,y=0,z),\tag{17.6.26}$$

$$C_{1,c}^{[0]}(z)=B_x(x=0,y=0,z),\tag{17.6.27}$$

$$C_{1,s}^{[0]}(z)=B_y(x=0,y=0,z),\tag{17.6.28}$$

$$C_{2,c}^{[0]}(z)=(1/2)(\partial B_x/\partial x)\Big|_{(0,0,z)}+(1/4)(\partial B_z/\partial z)\Big|_{(0,0,z)},\tag{17.6.29}$$

$$C_{2,s}^{[0]}(z) = (1/2)(\partial B_y/\partial x)\Big|_{(0,0,z)}, \tag{17.6.30}$$

$$C_{3,c}^{[0]}(z) = (1/6)(\partial^2 B_x/\partial x^2)\Big|_{(0,0,z)} + (1/8)(\partial^2 B_x/\partial z^2)\Big|_{(0,0,z)}, \tag{17.6.31}$$

$$C_{3,s}^{[0]}(z) = (1/6)(\partial^2 B_y/\partial x^2)\Big|_{(0,0,z)} + (1/24)(\partial^2 B_y/\partial z^2)\Big|_{(0,0,z)}, \text{etc.} \tag{17.6.32}$$

See Exercise 6.1. Finally, by repeatedly differentiating these relations with respect to z, one can obtain the $C_{m,\alpha}^{[n]}(z)$ for $n > 0$. In general, the computation of the $C_{m,\alpha}^{[n]}(z)$ requires $m+n-1$ differentiations. Again, when based on grid data, this involves multiple numerical differentiation, and therefore is expected to be unreliable.

Exercises

17.6.1. The aim of this exercise is to verify the relations (6.26) through (6.32). Begin by setting $x = 0$ in the relations (6.4) through (6.6). Show that so doing yields the results

$$B_x(x=0, y=0, z) = C_{1,c}^{[0]}(z), \tag{17.6.33}$$

$$B_y(x=0, y=0, z) = C_{1,s}^{[0]}(z), \tag{17.6.34}$$

$$B_z(x=0, y=0, z) = C_0^{[1]}(z). \tag{17.6.35}$$

Next, differentiate (6.4) through (6.6) with respect to x and then set $x = 0$. Show that so doing gives the results

$$(\partial B_x/\partial x)|_{0,0,z} = -(1/2)C_0^{[2]}(z) + 2C_{2,c}^{[0]}(z), \tag{17.6.36}$$

$$(\partial B_y/\partial x)|_{0,0,z} = 2C_{2,s}^{[0]}(z), \tag{17.6.37}$$

$$(\partial B_z/\partial x)|_{0,0,z} = C_{1,c}^{[1]}(z). \tag{17.6.38}$$

Show that solving (6.7) through (6.12) for the on-axis gradients gives, so far, the results

$$C_{1,c}^{[0]}(z) = B_x(x=0, y=0, z), \tag{17.6.39}$$

$$C_{1,s}^{[0]}(z) = B_y(x=0, y=0, z), \tag{17.6.40}$$

$$C_0^{[1]}(z) = B_z(x=0, y=0, z), \tag{17.6.41}$$

$$C_{2,c}^{[0]}(z) = (1/2)(\partial B_x/\partial x)|_{0,0,z} + (1/4)(\partial B_z/\partial z)|_{0,0,z}, \tag{17.6.42}$$

$$C_{2,s}^{[0]}(z) = (1/2)(\partial B_y/\partial x)|_{0,0,z}, \tag{17.6.43}$$

$$C_{1,c}^{[1]}(z) = (\partial B_z/\partial x)|_{0,0,z}. \tag{17.6.44}$$

Verify that (6.18) is redundant because from (6.13) we also have the relation

$$C_{1,c}^{[1]}(z) = (\partial B_x/\partial z)|_{0,0,z}. \tag{17.6.45}$$

Alternatively, (6.19) serves as a consistency check on (6.18).[12] Next differentiate the mid-plane fields twice with respect to x and then set $x = 0$. Show that so doing yields the relations

$$(\partial^2 B_x/\partial x^2)|_{0,0,z} = 6[C_{3,c}^{[0]}(z) - (1/8)C_{1,c}^{[2]}(z)], \tag{17.6.46}$$

$$(\partial^2 B_y/\partial x^2)|_{0,0,z} = 2[3C_{3,s}^{[0]}(z) - (1/8)C_{1,s}^{[2]}(z)], \tag{17.6.47}$$

$$(\partial^2 B_z/\partial x^2)|_{0,0,z} = 2[C_{2,c}^{[1]}(z) - (1/4)C_{0}^{[3]}(z)]. \tag{17.6.48}$$

Show that these relations, with the aid of the previous relations, can be solved for the next set of on-axis gradients to give the results

$$(\partial^2 B_x/\partial x^2)|_{0,0,z} = 6[C_{3,c}^{[0]}(z) - (1/8)C_{1,c}^{[2]}(z)], \tag{17.6.49}$$

$$(\partial^2 B_y/\partial x^2)|_{0,0,z} = 2[3C_{3,s}^{[0]}(z) - (1/8)C_{1,s}^{[2]}(z)], \tag{17.6.50}$$

$$(\partial^2 B_z/\partial x^2)|_{0,0,z} = 2[C_{2,c}^{[1]}(z) - (1/4)C_{0}^{[3]}(z)], \tag{17.6.51}$$

$$[C_{3,c}^{[0]}(z) - (1/8)C_{1,c}^{[2]}(z)] = (1/6)(\partial^2 B_x/\partial x^2)|_{0,0,z}, \tag{17.6.52}$$

$$C_{3,c}^{[0]}(z) = (1/6)(\partial^2 B_x/\partial x^2)|_{0,0,z} + (1/8)C_{1,c}^{[2]}(z), \tag{17.6.53}$$

$$C_{3,c}^{[0]}(z) = (1/6)(\partial^2 B_x/\partial x^2)|_{0,0,z} + (1/8)(\partial^2 B_x/\partial z^2)\Big|_{(0,0,z)}, \tag{17.6.54}$$

$$(\partial^2 B_y/\partial x^2)|_{0,0,z} = 2[3C_{3,s}^{[0]}(z) - (1/8)C_{1,s}^{[2]}(z)]. \tag{17.6.55}$$

17.7 Terminating End Fields

In principle, the fringe field of an individual beam-line element at either end of the element has infinite extent. But in practice in many instances we may wish to regard a beam line as a collection of separated/isolated elements. To do this it is necessary to make an approximation in which leading and trailing end fields are "terminated" in some way. The crucial problem is how to relate canonical coordinates in the absence of a magnetic field with canonical coordinates in the presence of a magnetic field.

17.7.1 Preliminary Observations

We begin with some preliminary observations. In Cartesian coordinates the Hamiltonian describing charged-particle motion with z as the independent variable is given by the relation

$$K = -[(p_t^{\rm can})^2/c^2 - m^2c^2 - (p_x^{\rm can} - qA_x)^2 - (p_y^{\rm can} - qA_y)^2]^{1/2} - qA_z. \tag{17.7.1}$$

Here we have assumed that the electric scalar potential ψ vanishes and $\boldsymbol{A}$ is static so that there is no electric field. Also, we have used the notation $p_x^{\rm can}$, $p_y^{\rm can}$, and $p_t^{\rm can}$ to indicate that it is the components of the *canonical* momenta that are involved in a Hamiltonian description of motion. See (1.6.16).

[12] This check arises from the requirement $\nabla \times B = 0$.

According to Hamilton's equations of motion, the change of a coordinate, say $x(z)$, with z is given by

$$\begin{aligned} dx/dz &= \partial K/\partial p_x^{\rm can} \\ &= (p_x^{\rm can} - qA_x)/[(p_t^{\rm can})^2/c^2 - m^2c^2 - (p_x^{\rm can} - qA_x)^2 - (p_y^{\rm can} - qA_y)^2]^{1/2} \\ &= (p_x^{\rm can} - qA_x)/[-(K + qA_z)]. \end{aligned} \tag{17.7.2}$$

Let us verify that this result agrees with what we already know. Recall that

$$K = -p_z^{\rm can}. \tag{17.7.3}$$

See (1.6.6). It follows that (7.2) can be rewritten in the form

$$dx/dz = (p_x^{\rm can} - qA_x)/(p_z^{\rm can} - qA_z). \tag{17.7.4}$$

According to (1.5.27) through (1.5.30) there is the relation

$$\boldsymbol{p}^{\rm can} - q\boldsymbol{A} = \boldsymbol{p}^{\rm mech} \tag{17.7.5}$$

where $\boldsymbol{p}^{\rm mech}$ is the *mechanical* momentum given by

$$\boldsymbol{p}^{\rm mech} = \gamma m\boldsymbol{v}. \tag{17.7.6}$$

Consequently, (7.4) can be rewritten in the form

$$dx/dz = p_x^{\rm mech}/p_z^{\rm mech} = \gamma m v_x/(\gamma m v_z) = v_x/v_z = \frac{dx/dt}{dz/dt}. \tag{17.7.7}$$

Evidently, the far left and far right sides of (7.7) agree. It is also easy to see that results analogous to those just found also hold for $y(z)$.

To complete the story we need to examine also the equation of motion for $t(z)$. In this case application of the standard Hamiltonian rules gives the result

$$\begin{aligned} dt/dz &= \partial K/\partial p_t^{\rm can} \\ &= (-p_t^{\rm can}/c^2)/[(p_t^{\rm can})^2/c^2 - m^2c^2 - (p_x^{\rm can} - qA_x)^2 - (p_y^{\rm can} - qA_y)^2]^{1/2} \\ &= (-p_t^{\rm can}/c^2)/[-(K + qA_z)]. \end{aligned} \tag{17.7.8}$$

Now use of (7.3), (7.5), and (1.6.17) yields the relation

$$dt/dz = (-p_t^{\rm can}/c^2)/p_z^{\rm mech} = \gamma m/(\gamma m v_z) = \frac{1}{dz/dt} \tag{17.7.9}$$

so that the far left and far right sides of (7.9) also agree.

17.7.2 Changing Gauge

It may be useful to change gauges at various points during the course of integrating a trajectory and computing an associated transfer map. Suppose the gauge is to be *changed* at the point $z = z^c$. Let x^b, y^b, and t^b denote coordinate functions *before* the change, and let x^a, y^a, and t^a denote coordinate functions *after* the change. Also, let $\boldsymbol{A}^b(x^b, y^b; z)$ and $\boldsymbol{A}^a(x^a, y^a; z)$ be the vector potentials before $(z < z^c)$ and after $(z > z^c)$ the change point z^c. Finally, let $p_x^{\rm canb}$, $p_y^{\rm canb}$, $p_t^{\rm canb}$ be the canonical momentum functions before the change, and let $p_x^{\rm cana}$, $p_y^{\rm cana}$, $p_t^{\rm cana}$ be the canonical momentum functions after the change. In terms of these quantities, the before and after Hamiltonians K^b and K^a are given by the relations

$$K^b = -[(p_t^{\rm canb})^2/c^2 - m^2c^2 - (p_x^{\rm canb} - qA_x^b)^2 - (p_y^{\rm canb} - qA_y^b)^2]^{1/2} - qA_z^b \text{ for } z < z^c, \quad (17.7.10)$$

$$K^a = -[(p_t^{\rm cana})^2/c^2 - m^2c^2 - (p_x^{\rm cana} - qA_x^a)^2 - (p_y^{\rm cana} - qA_y^a)^2]^{1/2} - qA_z^a \text{ for } z > z^c. \quad (17.7.11)$$

What should be the matching relations between the phase-space quantities before and after? Since the choice of gauge should have no physical effect, there is the immediate requirement that the coordinate functions be continuous:

$$\begin{aligned} x^a(z) &= x^b(z) \text{ when } z = z^c, \\ y^a(z) &= y^b(z) \text{ when } z = z^c, \\ t^a(z) &= t^b(z) \text{ when } z = z^c. \end{aligned} \quad (17.7.12)$$

For the same reason, we require that the velocities, and hence the mechanical momenta, be continuous. From (7.5) and (7.6) we see that this requirement is equivalent to the relations

$$\boldsymbol{p}^{\rm cana} - q\boldsymbol{A^a} = \boldsymbol{p}^{\rm canb} - q\boldsymbol{A^b} \text{ when } z = z^c. \quad (17.7.13)$$

In terms of components, (7.13) yields the matching relations

$$\begin{aligned} p_x^{\rm cana} &= p_x^{\rm canb} + q(A_x^a - A_x^b) \text{ when } z = z^c, \\ p_y^{\rm cana} &= p_y^{\rm canb} + q(A_y^a - A_y^b) \text{ when } z = z^c. \end{aligned} \quad (17.7.14)$$

Finally, the total energy cannot change under a gauge transformation and therefore, since we have assumed that the scalar potential ψ vanishes, there is the matching relation

$$p_t^{\rm cana} = p_t^{\rm canb} \text{ when } z = z^c. \quad (17.7.15)$$

We note that this relation also follows from (1.6.17), (7.5), (7.6), and (7.13).

We assume there is some common overlap region where both $\boldsymbol{A}^b$ and $\boldsymbol{A}^a$ are defined. Since they both give rise to the same magnetic field, there is the relation

$$\nabla \times (\boldsymbol{A}^a - \boldsymbol{A}^b) = 0. \quad (17.7.16)$$

It follows that there is a function χ such that

$$\boldsymbol{A}^a - \boldsymbol{A}^b = \nabla\chi. \quad (17.7.17)$$

Consequently, the relations (7.14) can be rewritten in the form

$$\begin{aligned} p_x^{\rm cana} &= p_x^{\rm canb} + q(\partial/\partial x)\chi \text{ when } z = z^c, \\ p_y^{\rm cana} &= p_y^{\rm canb} + q(\partial/\partial y)\chi \text{ when } z = z^c. \end{aligned} \tag{17.7.18}$$

There is one last step. Let $\mathcal{T}^c$ be the symplectic *transformation* map defined by the relation

$$\mathcal{T}^c = \exp(q : \chi :). \tag{17.7.19}$$

With aid of this map it is easily verified that the relations (7.12), (7.14), and (7.15) can be rewritten in the form

$$\begin{aligned} x^a(z) &= \exp(q : \chi :)x^b(z) \text{ with } z = z^c, \\ y^a(z) &= \exp(q : \chi :)y^b(z) \text{ with } z = z^c, \\ t^a(z) &= \exp(q : \chi :)t^b(z) \text{ with } z = z^c; \end{aligned} \tag{17.7.20}$$

$$\begin{aligned} p_x^{\rm cana}(z) &= \exp(q : \chi :)p_x^{\rm canb}(z) \text{ with } z = z^c, \\ p_y^{\rm cana}(z) &= \exp(q : \chi :)p_y^{\rm canb}(z) \text{ with } z = z^c, \\ p_t^{\rm cana}(z) &= \exp(q : \chi :)p_t^{\rm canb}(z) \text{ with } z = z^c. \end{aligned} \tag{17.7.21}$$

We have determined that a change in gauge amounts to making a symplectic transformation. Review Exercises 6.2.8 and 6.5.3.

17.7.3 Finding the Minimal Vector Potential

The goal of this subsection is, given $\boldsymbol{B}(\boldsymbol{r})$ in some region, to find an associated vector potential $\boldsymbol{A}^s$ that is as *small*/minimal as possible in the sense that $\boldsymbol{A}^s$ is small if $\boldsymbol{B}(\boldsymbol{r})$ is small. The reason for this goal will become apparent in following subsections.

Our plan is as follows: Make Taylor expansions, with initially unknown coefficients, for the Cartesian components of $\boldsymbol{A}^s$, organize these expansions into homogeneous polynomials, and then further organize them as spherical polynomial vector fields. Then use this representation to compute and organize $\nabla \times \boldsymbol{A}^s$ in terms of spherical polynomial vector fields. At the same time parameterize $\boldsymbol{B}(\boldsymbol{r})$ in terms of a scalar potential ψ expanded in harmonic polynomials. Finally, compare the two expansions for $\boldsymbol{B}(\boldsymbol{r})$ given by $\boldsymbol{B} = \nabla\psi$ and $\boldsymbol{B} = \nabla \times \boldsymbol{A}^s$, equate coefficients of like terms, and thereby determine the coefficients in the Taylor expansion for the components of $\boldsymbol{A}^s$ in terms of the coefficients in the expansion for ψ. For the notation and machinery required for the execution of this plan, see Appendix U.

We begin with an expansion for $\boldsymbol{B}(\boldsymbol{r})$ based on the use of a scalar potential. Without loss of generality, we may take the region of interest to be centered at the origin. We also assume that $\boldsymbol{B}(\boldsymbol{r})$ has a Taylor expansion in the components of $\boldsymbol{r}$ and is divergence and curl free. In this case there is a harmonic magnetic scalar potential $\psi(\boldsymbol{r})$ such that

$$\boldsymbol{B} = \nabla\psi. \tag{17.7.22}$$

Recall the beginning of Section 15.2. Employing the notation of Appendix U, we may assume without loss of generality that ψ has a spherical polynomial expansion of the form

$$\psi(\boldsymbol{r}) = \sum_{n=1}^{n_{\max}} \sum_{m} d_{nm} S_{nn}^{m}(\boldsymbol{r}) \tag{17.7.23}$$

where the quantities d_{nm} are arbitrary coefficients. Here we assume an expansion through terms of degree $n_{\max}$ and omit $n = 0$ terms since constant terms make no contribution to $\boldsymbol{B}$ as given by (7.22).

For the associated vector potential $\boldsymbol{A}^s$ we make the spherical polynomial vector field expansion

$$\boldsymbol{A}^s(\boldsymbol{r}) = \sum_{n=1}^{n_{\max}} \sum_{\ell} \sum_{J} \sum_{M} c_{n\ell JM} \boldsymbol{S}_{n\ell J}^{M}(\boldsymbol{r}). \tag{17.7.24}$$

Again see Appendix U. Given the coefficients d_{nm}, our task is to use the equality

$$\nabla \times \boldsymbol{A}^s(\boldsymbol{r}) = \nabla \times \sum_{n=1}^{n_{\max}} \sum_{\ell} \sum_{J} \sum_{M} c_{n\ell JM} \boldsymbol{S}_{n\ell J}^{M}(\boldsymbol{r}) = \nabla \sum_{n=1}^{n_{\max}} \sum_{m} d_{nm} S_{nn}^{m}(\boldsymbol{r}) = \nabla\psi(\boldsymbol{r}) \tag{17.7.25}$$

to find the coefficients $c_{n\ell JM}$.

Let us begin by evaluating the right side of (7.25). We find the results

$$\boldsymbol{B}(r) = \nabla\psi(\boldsymbol{r}) = \nabla \sum_{n=1}^{n_{\max}} \sum_{m} d_{nm} S_{nn}^{m}(\boldsymbol{r}) = \sum_{n=1}^{n_{\max}} \sum_{m} d_{nm} \sqrt{n(2n+1)} \boldsymbol{S}_{n-1,n-1,n}^{m}(\boldsymbol{r}). \tag{17.7.26}$$

Here we have used (U.5.3).

Next work on evaluating the left side of (7.25). This is a more complicated task. In accord with the range rules (U.3.7) and (U.3.8) we decompose the expansion into the sum of four pieces with each containing a particular kind of term:

a) All terms for which $\ell = 0$ and hence $J = 1$. Also, therefore, $n = 2k$ with $k > 0$. The associated spherical polynomial vectors are of the form $\boldsymbol{S}_{2k,0,1}^{M}(\boldsymbol{r})$.

b) All terms for which $\ell > 0$ and $J = \ell + 1$. The associated spherical polynomial vectors are of the form $\boldsymbol{S}_{n,\ell,\ell+1}^{M}(\boldsymbol{r})$.

c) All terms for which $\ell > 0$ and $J = \ell$. The associated spherical polynomial vectors are of the form $\boldsymbol{S}_{n,\ell,\ell}^{M}(\boldsymbol{r})$.

d) All terms for which $\ell > 0$ and $J = \ell - 1$. The associated spherical polynomial vectors are of the form $\boldsymbol{S}_{n,\ell,\ell-1}^{M}(\boldsymbol{r})$.

Thus, we write

$$\boldsymbol{A}^s = \boldsymbol{A}^{sa} + \boldsymbol{A}^{sb} + \boldsymbol{A}^{sc} + \boldsymbol{A}^{sd} \tag{17.7.27}$$

where

$$\boldsymbol{A}^{sa}(\boldsymbol{r}) = \sum_{k=1}^{k_{\max}} \sum_{M} c_{2k,0,1,M} \boldsymbol{S}_{2k,0,1}^{M}(\boldsymbol{r}), \tag{17.7.28}$$

$$\boldsymbol{A}^{sb}(\boldsymbol{r}) = \sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} c_{n,\ell,\ell+1,M} \boldsymbol{S}^M_{n,\ell,\ell+1}(\boldsymbol{r}), \tag{17.7.29}$$

$$\boldsymbol{A}^{sc}(\boldsymbol{r}) = \sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} c_{n,\ell,\ell,M} \boldsymbol{S}^M_{n,\ell,\ell}(\boldsymbol{r}), \tag{17.7.30}$$

$$\boldsymbol{A}^{sd}(\boldsymbol{r}) = \sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} c_{n,\ell,\ell-1,M} \boldsymbol{S}^M_{n,\ell,\ell-1}(\boldsymbol{r}). \tag{17.7.31}$$

We are now ready to proceed. For the $\boldsymbol{A}^{sa}$ term we find, using (U.5.20), the result

$$\nabla \times \boldsymbol{A}^{sa}(\boldsymbol{r}) = \nabla \times \sum_{k=1}^{k_{\max}} \sum_{M} c_{2k,0,1,M} \boldsymbol{S}^M_{2k,0,1}(\boldsymbol{r}) = \sum_{k=1}^{k_{\max}} \sum_{M} c_{2k,0,1,M} [i(\sqrt{2/3})(2k)] \boldsymbol{S}^M_{2k-1,1,1}(\boldsymbol{r}). \tag{17.7.32}$$

For the $\boldsymbol{A}^{sb}$ term we find, using (U.5.17), the result

$$\nabla \times \boldsymbol{A}^{sb}(\boldsymbol{r}) = \nabla \times \sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} c_{n,\ell,\ell+1,M} \boldsymbol{S}^M_{n,\ell,\ell+1}(\boldsymbol{r}) =$$
$$\sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} c_{n,\ell,\ell+1,M} [i\sqrt{(\ell+2)/(2\ell+3)}(n-\ell)] \boldsymbol{S}^M_{n-1,\ell+1,\ell+1}(\boldsymbol{r}). \tag{17.7.33}$$

For the $\boldsymbol{A}^{sc}$ term we find, using (U.5.18), the result

$$\nabla \times \boldsymbol{A}^{sc}(\boldsymbol{r}) = \nabla \times \sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} c_{n,\ell,\ell,M} \boldsymbol{S}^M_{n,\ell,\ell}(\boldsymbol{r}) =$$
$$\sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} c_{n,\ell,\ell,M} [i\sqrt{(\ell+1)/(2\ell+1)}(n+\ell+1)] \boldsymbol{S}^M_{n-1,\ell-1,\ell}(\boldsymbol{r})$$
$$+ \sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} c_{n,\ell,\ell,M} [i\sqrt{\ell/(2\ell+1)}(n-\ell)] \boldsymbol{S}^M_{n-1,\ell+1,\ell}(\boldsymbol{r}). \tag{17.7.34}$$

Finally, for the $\boldsymbol{A}^{sd}$ term we find, using (U.5.19), the result

$$\nabla \times \boldsymbol{A}^{sd}(\boldsymbol{r}) = \nabla \times \sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} c_{n,\ell,\ell-1,M} \boldsymbol{S}^M_{n,\ell,\ell-1}(\boldsymbol{r}) =$$
$$\sum_{n=1}^{n_{\max}} \sum_{\ell>0} \sum_{M} c_{n,\ell,\ell-1,M} [i\sqrt{(\ell-1/(2\ell-1)}(n+\ell+1)] \boldsymbol{S}^M_{n-1,\ell-1,\ell-1}(\boldsymbol{r}). \tag{17.7.35}$$

We are now prepared to equate coefficients of like terms. Let us begin with the first few corresponding to small values of n. The first of these, corresponding to $n = 0$, is $\boldsymbol{S}^M_{0,0,1}$. From (7.26) we see that

$$\text{coefficient of } \boldsymbol{S}^M_{0,0,1} \text{ in } \nabla\psi = \sqrt{3}\, d_{1M}. \tag{17.7.36}$$

We next examine the terms in $\nabla \times \boldsymbol{A}^s$: From (7.32) we see that there are no terms of the desired kind, namely terms involving $\boldsymbol{S}^M_{0,0,1}$, in $\nabla \times \boldsymbol{A}^{sa}$. From (7.33) we see that there are no terms of the desired kind in $\nabla \times \boldsymbol{A}^{sb}$. From (7.34) we see that there are terms of the desired kind in $\nabla \times \boldsymbol{A}^{sc}$, and find the relation

$$\text{coefficient of } \boldsymbol{S}^M_{0,0,1} \text{ in } \nabla \times \boldsymbol{A}^{sc} = i\sqrt{6}\, c_{1,1,1,M}. \tag{17.7.37}$$

Finally, from (7.35) we see that there are no terms of the desired kind in $\nabla \times \boldsymbol{A}^{sd}$.

Upon comparing (7.36) and (7.37) we conclude that there must be the relation

$$i\sqrt{6}\, c_{1,1,1,M} = \sqrt{3}\, d_{1M}, \tag{17.7.38}$$

and therefore

$$c_{1,1,1,M} = -i\sqrt{1/2}\, d_{1M}. \tag{17.7.39}$$

Note that this relation is consistent with (U.6.39). Moreover, we conclude that the six remaining $n = 1$ coefficients in $\boldsymbol{A}^s$, namely $c_{1,1,0,0}$ and the $c_{1,1,2,M}$, can be anything since there are the relations (U.6.38) and (U.6.40). For simplicity, we set these coefficients to zero. Then, so far, we have the result

$$\boldsymbol{A}^s(\boldsymbol{r}) = \sum_M (-i)\sqrt{1/2}\, d_{1M}\, \boldsymbol{S}^M_{111}(\boldsymbol{r}) + \text{ terms of degree } > 1. \tag{17.7.40}$$

In terms of Cartesian components, (7.40) yields the relation

$$\boldsymbol{A}^s(\boldsymbol{r}) = -(1/2)\boldsymbol{r} \times \boldsymbol{B}(0) + \text{ terms of degree } > 1. \tag{17.7.41}$$

Here we have used (7.22), (7.23), and (U.6.25) evaluated for $n = 1$. We observe that this choice for the leading term in $\boldsymbol{A}^s$ is in the symmetric/Poincaré/Coulomb gauge. See Exercise 28.2.7.

Let us push on to the case $n = 1$; in which case there are the spherical polynomial vector fields $\boldsymbol{S}^0_{110}$, $\boldsymbol{S}^M_{111}$ with $-1 \le M \le 1$, and $\boldsymbol{S}^M_{112}$ with $-2 \le M \le 2$. First see where/how they occur in $\nabla\psi$. Examination of (7.26) shows that the only such term in $\nabla\psi$ is $\boldsymbol{S}^M_{112}$, and we have the relation

$$\text{coefficient of } \boldsymbol{S}^M_{1,1,2} \text{ in } \nabla\psi = \sqrt{10}\, d_{2M}. \tag{17.7.42}$$

We next examine the terms in $\nabla \times \boldsymbol{A}^s$: From (7.32) we see that there are no terms of the desired kind, namely terms involving $\boldsymbol{S}^M_{1,1,2}$, in $\nabla \times \boldsymbol{A}^{sa}$. From (7.33) we see that there are no terms of the desired kind in $\nabla \times \boldsymbol{A}^{sb}$. From (7.34) we see that there are terms of the desired kind in $\nabla \times \boldsymbol{A}^{sc}$, and find the relation

$$\text{coefficient of } \boldsymbol{S}^M_{1,1,2} \text{ in } \nabla \times \boldsymbol{A}^{sc} = i\sqrt{15}\, c_{2,2,2,M}. \tag{17.7.43}$$

Finally, from (7.35) we see that there are no terms of the desired kind in $\nabla \times \boldsymbol{A}^{sd}$.

Upon comparing (7.42) and (7.43) we conclude that there must be the relation

$$i\sqrt{15}\, c_{2,2,2,M} = \sqrt{10}\, d_{2M}, \tag{17.7.44}$$

and therefore

$$c_{2,2,2,M} = -i\sqrt{2/3}\, d_{2M}. \tag{17.7.45}$$

What can be said about the thirteen remaining $n = 2$ coefficients in $\boldsymbol{A}^s$, namely the c_{201M}, $c_{2,2,3,M}$, and $c_{2,2,1,M}$? It can be shown that $\nabla \times \boldsymbol{S}_{223}^M(\boldsymbol{r}) = 0$, and therefore the terms with coefficients $c_{2,2,3,M}$ make no contribution to $\boldsymbol{B}(\boldsymbol{r})$. See Exercise (U.6.21). For simplicity, we set these coefficients to zero. It can be shown that terms with the coefficients c_{201M} and $c_{2,2,1,M}$ produce terms in $\boldsymbol{B}(\boldsymbol{r})$ having nonzero curl. Again see Exercise (U.6.21). We also set these coefficients to zero to ensure that $\boldsymbol{B}(\boldsymbol{r})$ is curl free. Then, so far, we have the result

$$\boldsymbol{A}^s(\boldsymbol{r}) = \sum_M (-i)\sqrt{1/2}\; d_{1M}\; \boldsymbol{S}_{111}^M(\boldsymbol{r}) + \sum_M (-i)\sqrt{2/3}\; d_{2M}\; \boldsymbol{S}_{222}^M(\boldsymbol{r}) + \text{ terms of degree } > 2. \tag{17.7.46}$$

The pattern should now be clear. There are the general relations

$$\nabla S_{nn}^M(\boldsymbol{r}) = \sqrt{n(2n+1)}\boldsymbol{S}_{n-1,n-1,n}^M(\boldsymbol{r}) \tag{17.7.47}$$

and

$$\nabla \times \boldsymbol{S}_{n,n,n}^M(\boldsymbol{r}) = i\sqrt{(n+1)(2n+1)}\boldsymbol{S}_{n-1,n-1,n}^M(\boldsymbol{r}). \tag{17.7.48}$$

Therefore there is the general relation

$$\boldsymbol{A}^s(\boldsymbol{r}) = \sum_{n=1}^{n_{\max}} \sum_{M=-n}^{n} (-i)\sqrt{n/(n+1)}\; d_{nM}\boldsymbol{S}_{nnn}^M(\boldsymbol{r}). \tag{17.7.49}$$

It can be verified that this particular choice of $\boldsymbol{A}^s(\boldsymbol{r})$ has the two properties

$$\nabla \cdot \boldsymbol{A}^s(\boldsymbol{r}) = 0 \tag{17.7.50}$$

and

$$\boldsymbol{r} \cdot \boldsymbol{A}^s(\boldsymbol{r}) = 0. \tag{17.7.51}$$

See (U.5.11) and (U.6.9). Thus this vector potential is in both a Coulomb and Poincaré gauge.

The relation (7.49) can be further manipulated using (U.6.25). Doing so gives the result

$$\boldsymbol{A}^s(\boldsymbol{r}) = -\sum_{n=1}^{n_{\max}} \sum_{M=-n}^{n} [1/(n+1)] d_{nM} [\boldsymbol{r} \times \nabla S_{nn}^M(\boldsymbol{r})]. \tag{17.7.52}$$

We observe that this result agrees with that found by Ansatz in Exercises (15.5.8) and (15.5.9). See (15.5.81) and (15.5.82).

Have we achieved our goal of finding a "minimal vector potential"? We have, in the following sense: Inspection of (7.26) shows that it provides an expansion of $\boldsymbol{B}(r)$ in terms of spherical polynomial vector fields $\boldsymbol{S}_{n-1,n-1,n}^m(\boldsymbol{r})$ with expansion coefficients proportional to the d_{nm}. Inspection of (7.49) shows that it provides an expansion of $\boldsymbol{A}^s(r)$ in terms of spherical polynomial vector fields $\boldsymbol{S}_{nnn}^M(\boldsymbol{r})$ with expansion coefficients again proportional to the d_{nM}. The vector potential $\boldsymbol{A}^s(\boldsymbol{r})$ has no constant part, and its non-constant parts are directly proportional to the coefficients d_{nm} that describe the constant and non-constant parts of $\boldsymbol{B}(r)$. Moreover, there is an order-by-order relation. Terms of order n in $\boldsymbol{A}^s(\boldsymbol{r})$ are proportional to terms of order $n - 1$ in $\boldsymbol{B}(r)$. Thus, $\boldsymbol{A}^s(\boldsymbol{r})$ is small if $\boldsymbol{B}(r)$ is small. In particular, if high-order terms in $\boldsymbol{B}(r)$ are negligible, they will also be negligible in $\boldsymbol{A}^s(\boldsymbol{r})$.

There is yet another sense in which the vector potential we have found is minimal. Suppose, for example, that we confine our attention to the case of a vector potential that is homogeneous of degree 1, which is the case we need to produce a constant magnetic field. When $n = 1$ we see from Table U.3.1 that $\ell = 1$ and $J = 0, 1, 2$. Therefore, such a vector potential, call it $\boldsymbol{A}^{[1]}$, can be written in the form

$$\boldsymbol{A}^{[1]}(\boldsymbol{r}) = \sum_J \sum_M c_{11JM} \boldsymbol{S}_{11J}^M(\boldsymbol{r}). \tag{17.7.53}$$

Recall (7.24). Let us compute the *norm* of $\boldsymbol{A}^{[1]}$ as defined by the relation

$$||\boldsymbol{A}^{[1]}(\boldsymbol{r})||^2 = \int d\Omega\, [\boldsymbol{A}^{[1]}(\boldsymbol{r})]^* \cdot \boldsymbol{A}^{[1]}(\boldsymbol{r}). \tag{17.7.54}$$

Since the $\boldsymbol{S}_{11J}^M(\boldsymbol{r})$ are mutually orthogonal under angular integration, we find from (7.53), (U.3.18), and (U.4.3) the result

$$||\boldsymbol{A}^{[1]}(\boldsymbol{r})||^2 = r^2 \sum_J \sum_M |c_{11JM}|^2. \tag{17.7.55}$$

We know the value of c_{111M} is fixed by(7.39), and we have chosen to set the remaining c_{11JM} to zero. We now see, since (7.55) is a sum of squares, that doing so *minimizes* $||\boldsymbol{A}^{[1]}(\boldsymbol{r})||$. Similar computations may be made for other values of n. The result is that the choice we have made for $\boldsymbol{A}^s$ minimizes $||\boldsymbol{A}^{s[n]}(\boldsymbol{r})||$ for each value of n.

17.7.4 The $m = 0$ Case: Solenoid Example

In this subsection we will explore the fringe fields for a solenoid. Our aim will be to compare the vector potential in the symmetric Coulomb gauge as given in Section 15.4 and the vector potential in the minimum gauge.[13] Reference to Section 20.1.2 shows that, for a simple air-core solenoid, $C_0^{[1]}(z)$ falls off as $1/|z|^3$ for large $|z|$. See (20.1.28) and Figures 20.1.3 and 20.1.4. Correspondingly, in this case, the $C_0^{[n+1]}(z)$ will fall off as $1/|z|^{n+3}$ for large $|z|$. We expect the simple air-core to be representative of the worst scenario in the sense that the fringe fields for other kinds of solenoids will fall of at this same rate or *faster*.

According to Subsection 15.2.3, the scalar potential for the $m = 0$ case is given by the relation

$$\psi_0(x, y, z) = C_0^{[0]}(z) - (1/4)(x^2 + y^2)C_0^{[2]}(z) + \cdots . \tag{17.7.56}$$

Let us expand ψ about the point $(0, 0, z_0)$. To do so, introduce local deviation variables ξ, η, and ζ by making the definitions

$$x = 0 + \xi, \tag{17.7.57}$$

$$y = 0 + \eta, \tag{17.7.58}$$

$$z = z_0 + \zeta. \tag{17.7.59}$$

[13] Note that according to Section 15.5 there are a variety of Coulomb gauges including vertical-free and horizontal-free Coulomb gauges. Here we treat the case where there is the greatest symmetry between the vertical and horizontal components of $\boldsymbol{A}$.

Also define a *deviation* vector $\boldsymbol{r}^d$ by writing

$$\boldsymbol{r}^d = \xi \boldsymbol{e}_x + \eta \boldsymbol{e}_y + \zeta \boldsymbol{e}_z. \tag{17.7.60}$$

We can then define a scalar potential ψ^e suitable for *expansion* by writing the relation

$$\psi^e(\boldsymbol{r}^d; z_0) = \psi(\xi, \eta, z_0 + \zeta). \tag{17.7.61}$$

Indeed, making use of (7.56) yields for ψ^e the expansion

$$\begin{aligned} \psi^e(\boldsymbol{r}^d; z_0) &= C_0^{[0]}(z_0+\zeta) - (1/4)(\xi^2+\eta^2)C_0^{[2]}(z_0+\zeta) + \text{ terms of order 3 and higher} \\ &= C_0^{[0]}(z_0) + C_0^{[1]}(z_0)\zeta \\ &\quad + C_0^{[2]}(z_0)(\zeta^2/2) - (1/4)(\xi^2+\eta^2)C_0^{[2]}(z_0) \\ &\quad + \text{ terms of order 3 and higher} \\ &= \psi^{e[0]} + \psi^{e[1]} + \psi^{e[2]} + \text{ terms of order 3 and higher.} \end{aligned} \tag{17.7.62}$$

Here the upper index in square brackets on a quantity denotes its degree.[14] And, from (7.22), the magnetic field associated with this expansion is given by the expansion

$$\boldsymbol{B} = \boldsymbol{B}^{[0]} + \boldsymbol{B}^{[1]} + \text{ terms of order 2 and higher} \tag{17.7.63}$$

where

$$\boldsymbol{B}^{[0]} = C_0^{[1]}(z_0)\boldsymbol{e}_z \tag{17.7.64}$$

and

$$\begin{aligned} \boldsymbol{B}^{[1]} &= -(1/2)C_0^{[2]}(z_0)(\xi\boldsymbol{e}_x + \eta\boldsymbol{e}_y) + C_0^{[2]}(z_0)\zeta\boldsymbol{e}_z \\ &= -(1/2)C_0^{[2]}(z_0)(\xi\boldsymbol{e}_x + \eta\boldsymbol{e}_y - 2\zeta\boldsymbol{e}_z) \\ &= -(1/2)C_0^{[2]}(z_0)(\xi\boldsymbol{e}_x + \eta\boldsymbol{e}_y + \zeta\boldsymbol{e}_z - 3\zeta\boldsymbol{e}_z) \\ &= -(1/2)C_0^{[2]}(z_0)(\boldsymbol{r}^d - 3\zeta\boldsymbol{e}_z). \end{aligned} \tag{17.7.65}$$

Let us find the associated minimum vector potential. According to (7.52), we expect that $\boldsymbol{A}^s$ will be of the form

$$\boldsymbol{A}^s(\boldsymbol{r}^d) = \boldsymbol{A}^{s[1]}(\boldsymbol{r}^d) + \boldsymbol{A}^{s[2]}(\boldsymbol{r}^d) \tag{17.7.66}$$

with

$$\boldsymbol{A}^{s[1]}(\boldsymbol{r}^d) = -(1/2)\boldsymbol{r}^d \times \boldsymbol{B}^{[0]}(\boldsymbol{r}^d) \tag{17.7.67}$$

and

$$\boldsymbol{A}^{s[2]}(\boldsymbol{r}) = -(1/3)\boldsymbol{r}^d \times \boldsymbol{B}^{[1]}(\boldsymbol{r}^d). \tag{17.7.68}$$

Working out the indicated cross products in (7.67) and (7.68) gives the results

$$\boldsymbol{A}^{s[1]}(\boldsymbol{r}^d) = -(1/2)C_0^{[1]}(z_0)(\eta\boldsymbol{e}_x - \xi\boldsymbol{e}_y), \tag{17.7.69}$$

[14] Note that according to Section U.7 the $\psi^{e[n]}$ will be homogenous harmonic polynomials of degree n.

$$\boldsymbol{A}^{s[2]}(\boldsymbol{r}^d) = -(1/2)C_0^{[2]}(z_0)(\zeta\eta\boldsymbol{e}_x - \zeta\xi\boldsymbol{e}_y). \tag{17.7.70}$$

Simple calculation verifies that there are indeed the relations

$$\nabla \times \boldsymbol{A}^{s[1]}(\boldsymbol{r}^d) = \boldsymbol{B}^{[0]}(\boldsymbol{r}^d), \tag{17.7.71}$$

$$\nabla \times \boldsymbol{A}^{s[2]}(\boldsymbol{r}^d) = \boldsymbol{B}^{[1]}(\boldsymbol{r}^d), \tag{17.7.72}$$

as desired.

How do the results given by (7.69) and (7.70) compare with those provided by the symmetric Coulomb gauge? According to Section 15.4, the vector potential in the symmetric Coulomb gauge for the $m = 0$ case is given by the expressions

$$\begin{aligned}\hat{A}_x^0(\boldsymbol{r}) &= -(y/2)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{1}{2^{2\ell}\ell!(\ell+1)!}C_0^{[2\ell+1]}(z)(x^2+y^2)^\ell \\ &= -(y/2)[C_0^{[1]}(z) - (1/8)C_0^{[3]}(z)(x^2+y^2) + \cdots],\end{aligned} \tag{17.7.73}$$

$$\begin{aligned}\hat{A}_y^0(\boldsymbol{r}) &= (x/2)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{1}{2^{2\ell}\ell!(\ell+1)!}C_0^{[2\ell+1]}(z)(x^2+y^2)^\ell \\ &= (x/2)[C_0^{[1]}(z) - (1/8)C_0^{[3]}(z)(x^2+y^2) + \cdots],\end{aligned} \tag{17.7.74}$$

$$\hat{A}_z^0(\boldsymbol{r}) = 0. \tag{17.7.75}$$

In terms of the expansion variables (7.57) through (7.59) these expressions become

$$\begin{aligned}\hat{A}_x^0(\xi,\eta,z_0+\zeta) &= -(\eta/2)[C_0^{[1]}(z_0+\zeta) - (1/8)C_0^{[3]}(z_0+\zeta)(\xi^2+\eta^2) + \cdots] \\ &= -(\eta/2)[C_0^{[1]}(z_0) + C_0^{[2]}(z_0)\zeta] + \cdots \\ &= -(1/2)C_0^{[1]}(z_0)\eta - (1/2)C_0^{[2]}(z_0)\zeta\eta + \text{terms of order 3 and higher},\end{aligned} \tag{17.7.76}$$

$$\begin{aligned}\hat{A}_y^0(\xi,\eta,z_0+\zeta) &= (\xi/2)[C_0^{[1]}(z_0+\zeta) - (1/8)C_0^{[3]}(z_0+\zeta)(\xi^2+\eta^2) + \cdots] \\ &= (\xi/2)[C_0^{[1]}(z_0) + C_0^{[2]}(z_0)\zeta] + \cdots \\ &= (1/2)C_0^{[1]}(z_0)\xi + (1/2)C_0^{[2]}(z_0)\zeta\xi + \text{terms of order 3 and higher},\end{aligned} \tag{17.7.77}$$

$$\hat{A}_z^0(\xi,\eta,z_0+\zeta) = 0. \tag{17.7.78}$$

Comparison of (7.69) and (7.70) with (7.76) through (7.78) shows that for the $m = 0$ case, at least for the orders computed, the minimum vector potential constructed from field expansions about on-axis points *agrees* with the symmetric Coulomb gauge vector potential constructed from on-axis field data. In retrospect, this result should not be surprising. We should expect agreement through *all* orders because, according to Subsection 15.6.2, the $m = 0$ symmetric Coulomb gauge vector potential constructed from on-axis field data is, in fact, in the Poincaré-Coulomb gauge.

17.7.5 The $m = 1$ Case: Magnetic Monopole Doublet and Wiggler Examples

In this subsection we will first study the behavior of the leading and trailing fringe fields for a magnetic monopole doublet. Subsequently we will examine the case of a wiggler.

Magnetic Monopole Doublet Example

The doublet will be located at the origin $(0, 0, 0)$ as in Subsection 15.8.1 and we will expand the scalar potential $\psi(x, y, z)$ given by (15.8.3) about the mid-plane point $(x_0, 0, z_0)$. Consequently, if $z_0 \ll 0$, we will obtain an expansion in the leading region, and if $z_0 \gg 0$, we will obtain an expansion in the trailing region. Moreover, if $x_0 = 0$, the expansion will be on axis; and setting $x_0 \neq 0$ allows for expansion about a point on the design orbit. See Subsection 21.5.1 and Figures 21.5.1 and 21.5.6.

As before, introduce local deviation variables ξ, η, and ζ and a deviation vector $\boldsymbol{r}^d$ by making the definitions

$$x = x_0 + \xi, \tag{17.7.79}$$

$$y = \eta, \tag{17.7.80}$$

$$z = z_0 + \zeta, \tag{17.7.81}$$

and (7.60). We can then define a scalar potential ψ^e by writing the relation

$$\psi^e(\boldsymbol{r}^d; x_0, z_0) = \psi(x_0 + \xi, \eta, z_0 + \zeta). \tag{17.7.82}$$

Indeed, making use of (15.8.3) and (7.82) yields the expansion

$$\begin{aligned}\psi^e(\boldsymbol{r}^d; x_0, z_0) \;\; = \;\; & [-2ga/(x_0^2 + z_0^2 + a^2)^{3/2}]\eta \\ & +[6ga/(x_0^2 + z_0^2 + a^2)^{5/2}][\eta(x_0\xi + z_0\zeta)] \\ & + \text{ terms of order 3 and higher.}\end{aligned} \tag{17.7.83}$$

Note that $\psi(x_0, 0, z_0)$ vanishes so that there is no constant term in the expansion (7.83). We observe that the first term in (7.83) falls off as $(1/x_0)^3$ or $(1/z_0)^3$ for large x_0 or z_0, and the second falls off as $(1/x_0)^4$ or $(1/z_0)^4$. In general, successive terms fall off with ever increasing powers of $(1/x_0)$ or $(1/z_0)$.

Let us compute the magnetic field $\boldsymbol{B}$ associated with the first two terms in (7.83). We find the result

$$\begin{aligned}\boldsymbol{B}(\boldsymbol{r}^d; x_0, z_0) \;\; = \;\; & -[2ga/(x_0^2 + z_0^2 + a^2)^{3/2}]\boldsymbol{e}_y \\ & +[6ga/(x_0^2 + z_0^2 + a^2)^{5/2}](x_0\xi + z_0\zeta)\boldsymbol{e}_y \\ & +[6ga/(x_0^2 + z_0^2 + a^2)^{5/2}][\eta(x_0\boldsymbol{e}_x + z_0\boldsymbol{e}_z)].\end{aligned} \tag{17.7.84}$$

Next let us find the minimum vector potential $\boldsymbol{A}^s$ associated with the first two terms in (7.83). Begin by decomposing $\boldsymbol{B}$ into homogeneous polynomials by rewriting (7.84) in the form (7.63) with

$$\boldsymbol{B}^{[0]}(\boldsymbol{r}^d) \;\; = \;\; -[2ga/(x_0^2 + z_0^2 + a^2)^{3/2}]\boldsymbol{e}_y \tag{17.7.85}$$

and

$$\boldsymbol{B}^{[1]}(\boldsymbol{r}^d) = [6ga/(x_0^2 + z_0^2 + a^2)^{5/2}][(x_0\xi + z_0\zeta)\boldsymbol{e}_y + \eta(x_0\boldsymbol{e}_x + z_0\boldsymbol{e}_z)]. \tag{17.7.86}$$

The minimum vector potential associated with this magnetic field will again be given by the relations (7.66) through (7.68). Working out the indicated cross products yields the results

$$\boldsymbol{A}^{s[1]}(\boldsymbol{r}^d) = [ga/(x_0^2 + z_0^2 + a^2)^{3/2}](-\zeta\boldsymbol{e}_x + \xi\boldsymbol{e}_z), \tag{17.7.87}$$

$$\begin{gathered}\boldsymbol{A}^{s[2]}(\boldsymbol{r}^d) = [-2ga/(x_0^2 + z_0^2 + a^2)^{5/2}] \times \\ [(z_0\eta^2 - z_0\zeta^2 - x_0\xi\zeta)\boldsymbol{e}_x + (x_0\eta\zeta - z_0\xi\eta)\boldsymbol{e}_y + (x_0\xi^2 + z_0\xi\zeta - x_0\eta^2)\boldsymbol{e}_z].\end{gathered} \tag{17.7.88}$$

Simple calculation verifies that there are indeed the relations (7.71) and (7.72) as desired.

At this point it is instructive to compare the minimum vector potential with other possible vector potentials. We note that the design orbit for a dipole field is *curved*, and therefore for the most part does not lie on axis. Consequently we must generally compare the minimum vector potential with other possible vector potentials at off-axis points. By construction, the minimum vector potential *vanishes* at every expansion point. In contrast, other vector potentials (for example those based on employing on-axis expansions of a scalar potential or the use of Dirac strings) generally not have this property. We conclude that problems involving curved design orbits are more complicated than those for straight beam-line elements, and their treatment requires special care. This treatment is deferred to Chapter 21.

On-axis Entry and Exit Wiggler Example

There is an application for which expansions of $m = 1$ cylindrical harmonics may be useful, namely the case of wigglers when the excursion of the design orbit from the axis may be treated as small. That is, it is assumed that the design orbit enters and exits the wiggler on axis and nearly along the axis, and the excursions of the design orbit from the axis while within the wiggler may be treated as small.

According to (15.2.61) the the scalar potential for a (normal) dipole is given by the relation

$$\psi_{1,s}(x, y, z) = y[C_{1,s}^{[0]}(z) - (1/8)(x^2 + y^2)C_{1,s}^{[2]}(z) + \cdots]. \tag{17.7.89}$$

Upon invoking the definitions (7.57) through (7.60), the expansion (7.89) yields the expansion

$$\begin{aligned}\psi^e(\boldsymbol{r}^d; z_0) &= \eta[C_{1,s}^{[0]}(z_0 + \zeta) - (1/8)(\xi^2 + \eta^2)C_{1,s}^{[2]}(z_0 + \zeta) + \cdots] \\ &= \eta C_{1,s}^{[0]}(z_0) + \eta\zeta C_{1,s}^{[1]}(z_0) + \text{terms of order 3 and higher}.\end{aligned} \tag{17.7.90}$$

The magnetic field associated with the scalar potential (7.90) has an expansion of the form (7.63) with

$$\boldsymbol{B}^{[0]} = C_{1,s}^{[0]}(z_0)\boldsymbol{e}_y \tag{17.7.91}$$

and

$$\boldsymbol{B}^{[1]} = \zeta C^{[1]}_{1,s}(z_0)\boldsymbol{e}_y + \eta C^{[1]}_{1,s}(z_0)\boldsymbol{e}_z. \tag{17.7.92}$$

The first two terms in minimum vector potential expansion associated with this magnetic field will again be given by the relations (7.66) through (7.68). Working out the indicated cross products now yields the results

$$\boldsymbol{A}^{s[1]}(\boldsymbol{r}^d) = (1/2)C^{[0]}_{1,s}(z_0)(\zeta\boldsymbol{e}_x - \xi\boldsymbol{e}_z), \tag{17.7.93}$$

$$\boldsymbol{A}^{s[2]}(\boldsymbol{r}^d) = (1/3)C^{[1]}_{1,s}(z_0)[(\zeta^2 - \eta^2)\boldsymbol{e}_x + \xi\eta\boldsymbol{e}_y - \xi\zeta\boldsymbol{e}_z]. \tag{17.7.94}$$

And again simple calculation verifies that there are indeed the relations (7.71) and (7.72) as desired.

How does the the vector potential for a normal dipole in the Coulomb gauge compare with the minimum vector potential just found? From (15.4.95) through (15.4.97) we find, in the Coulomb gauge, that $\hat{\boldsymbol{A}}^{1,s}$ has the expansion

$$\begin{aligned}\hat{A}^{1,s}_x(\xi,\eta,z_0+\zeta) &= (1/4)(\xi^2-\eta^2)C^{[1]}_{1,s}(z_0+\zeta)+\cdots\\ &= (1/4)(\xi^2-\eta^2)C^{[1]}_{1,s}(z_0) + \text{terms of order 3 and higher,}\end{aligned} \tag{17.7.95}$$

$$\begin{aligned}\hat{A}^{1,s}_y(\xi,\eta,z_0+\zeta) &= (1/2)\xi\eta C^{[1]}_{1,s}(z_0+\zeta)+\cdots\\ &= (1/2)\xi\eta C^{[1]}_{1,s}(z_0) + \text{terms of order 3 and higher,}\end{aligned} \tag{17.7.96}$$

$$\begin{aligned}\hat{A}^{1,s}_z(\xi,\eta,z_0+\zeta) &= -\xi C^{[0]}_{1,s}(z_0+\zeta)+\cdots\\ &= -\xi C^{[0]}_{1,s}(z_0) - \xi\zeta C^{[1]}_{1,s}(z_0) + \text{terms of order 3 and higher.}\end{aligned} \tag{17.7.97}$$

Comparison of (7.93) and (7.94) with (7.95) through (7.97) shows that for the $m=1$ case the minimum vector potential constructed from field expansions about on-axis points *differs* from the Coulomb gauge vector potential constructed from on-axis field data.

What can be said about the $m=1$ azimuthal-free gauge vector potential? From (15.3.31) through (15.3.33), we see that $\boldsymbol{A}^{1,s}$ has the expansion

$$A^{1,s}_x(\xi,\eta,z_0+\zeta) = \xi^2 C^{[1]}_{1,s}(z_0)+\cdots\ , \tag{17.7.98}$$

$$A^{1,s}_y(\xi,\eta,z_0+\zeta) = \xi\eta C^{[1]}_{1,s}(z_0)+\cdots\ , \tag{17.7.99}$$

$$\begin{aligned}A^{1,s}_z(\xi,\eta,z_0+\zeta) &= -\xi C^{[0]}_{1,s}(z_0+\zeta)+\cdots\\ &= -\xi C^{[0]}_{1,s}(z_0) - \xi\zeta C^{[1]}_{1,s}(z_0) + \text{terms of order 3 and higher.}\end{aligned} \tag{17.7.100}$$

Comparison of (7.93) and (7.94) with (7.98) through (7.100) shows that for the $m = 1$ case the minimum vector potential constructed from field expansions about on-axis points also *differs* the azimuthal-free gauge vector potential constructed from on-axis field data. In retrospect, this difference should be less surprising because the minimum vector potential satisfies the Coulomb gauge condition, and the azimuthal-free gauge vector potential does not.

17.7.6 The $m = 2$ Case

Text to be worked on:

As a second example of a vector potential in the azimuthal-free gauge, suppose all terms in (2.37) vanish save for the 'pure' quadrupole terms $C^{[n]}_{2,s}(z)$. Then, again using (3.28) through (3.30), we find through terms of degree four that $\boldsymbol{A}^{2,s}$ has the expansion

$$A^{2,s}_x = (1/2)(x^3 - xy^2)C^{[1]}_{2,s}(z) + \cdots , \tag{17.7.101}$$

$$A^{2,s}_y = -(1/2)(y^3 - yx^2)C^{[1]}_{2,s}(z) + \cdots , \tag{17.7.102}$$

$$A^{2,s}_z = -(x^2 - y^2)C^{[0]}_{2,s}(z) + (1/6)(x^4 - y^4)C^{[2]}_{2,s}(z) + \cdots . \tag{17.7.103}$$

Note that the results (3.34) through (3.36) agree with (1.5.59) if we make the identification $Q/2 = C^{[0]}_{2,s}$. However, we know that $C^{[0]}_{2,s}(z)$ must depend on z because the on-axis gradients must vanish far outside any magnet. Therefore the functions $C^{[1]}_{2,s}(z)$, $C^{[2]}_{2,s}(z)$, etc. must be nonzero, at least near the end and fringe-field regions of any quadrupole magnet. We conclude again that, as a consequence of Maxwell's equations, the vector potential must contain terms beyond degree two in the variables x, y. Correspondingly, the transfer map for any real quadrupole must contain nonlinear terms.

As a second example of the use of these relations, let us compute $\hat{\boldsymbol{A}}^{2,s}$ for the quadrupole case $m = 2$. As before, suppose all terms in (2.37) vanish save for the quadrupole terms $C^{[n]}_{2,s}(z)$. Then, again using (4.92) through (4.94), we find, through terms of degree four, that $\hat{\boldsymbol{A}}^{2,s}$ has the expansion

$$\hat{A}^{2,s}_x = (1/6)(x^3 - 3xy^2)C^{[1]}_{2,s}(z) + \cdots , \tag{17.7.104}$$

$$\hat{A}^{2,s}_y = -(1/6)(y^3 - 3x^2y)C^{[1]}_{2,s}(z) + \cdots , \tag{17.7.105}$$

$$\hat{A}^{2,s}_z = -(x^2 - y^2)C^{[0]}_{2,s}(z) + (1/12)(x^4 - y^4)C^{[2]}_{2,s}(z) + \cdots . \tag{17.7.106}$$

This expansion should be compared with the azimuthal-free gauge expansion given by (3.34) through (3.36). Direct calculation again verifies that (4.1) and (4.4) are satisfied by $\hat{\boldsymbol{A}}^{2,s}$ through the order of the terms that have been retained in the expansion

17.7.7 The $m = 3$ Case

Exercises

17.7.1. Verify the relations (7.32) through (7.39).

17.7.2. Evidently the second-order portion of $\psi^e(\boldsymbol{r}^d; x_0, z_0)$ as given in (7.61) is composed of the monomials $\xi\eta$ and $\eta\zeta$. Show that these are the only monomials alowed at this order based on symmetry considerations. Verify that each monomial is an harmonic polynomial. Indeed, making the usual correspondence between ξ, η, ζ and x, y, z show, following the harmonic polynomial labeling scheme (U.2.9), that there are the relations

$$\xi\eta = [1/(4i)][\sqrt{32\pi/15}][H_2^2(\boldsymbol{r}) - H_2^{-2}(\boldsymbol{r})], \tag{17.7.107}$$

$$\eta\zeta = [-1/(2i)][\sqrt{8\pi/15}][H_2^1(\boldsymbol{r}) + H_2^{-1}(\boldsymbol{r})]. \tag{17.7.108}$$

Would these relations have been simpler had the polar axis, used to set up spherical polar coordinates, been taken to be the y axis instead of the z axis?

17.7.8 More Text

To proceed further it is useful to introduce some notation. Let z^{en} denote the z value where a transition is to be made from a region where the magnetic field is *taken* to vanish to the beginning of the leading fringe-field region. That is, the charged particle in question *enters* the leading fringe-field region when $z = z^{\text{en}}$. We will also use the notation z^{ben} and z^{aen} to denote z values just *before* and just *after* entry. Similarly, let z^{ex} denote the z value where a transition is to be made from the end of a trailing fringe-field region to a region where the magnetic field is again taken to vanish. That is, the charged particle in question *exits* the trailing fringe-field region when $z = z^{\text{ex}}$.

Entering a Leading Fringe-Field Region

Suppose we begin with a consideration of the transition between a field-free region and a leading-fringe field region. Let K^{ben} be the Hamiltonian before entry into the fringe-field region, and let K^{aen} be the Hamiltonian after entry into the fringe-field region. Then, since the magnetic field and its associated vector potential are assumed to vanish before entry, we have the relation

$$K^{\text{ben}} = -[(p_t^{\text{canben}})^2/c^2 - m^2c^2 - (p_x^{\text{canben}})^2 - (p_y^{\text{canben}})^2]^{1/2}. \tag{17.7.109}$$

And, since the magnetic field (and therefore also the vector potential) is nonzero after entry, we have the relation

$$K^{\text{aen}} = -[(p_t^{\text{canaen}})^2/c^2 - m^2c^2 - (p_x^{\text{canaen}} - qA_x)^2 - (p_y^{\text{canean}} - qA_y)^2]^{1/2} - qA_z. \tag{17.7.110}$$

Here we have added the suffixes *ben* and *aen* to the phase-space coordinates to denote their values before and after entry. Our task is to relate these phase-space coordinates.

As a first step, we naturally require that the coordinates be continous at $z^{\rm en}$,

$$x^{\rm ben} = x^{\rm aen}, \tag{17.7.111}$$

$$y^{\rm ben} = y^{\rm aen}, \tag{17.7.112}$$

$$t^{\rm ben} = t^{\rm aen}, \tag{17.7.113}$$

when $z = z^{\rm en}$. The next step is specify what is to be done with the momenta.

One possibility is to require that the slopes/"velocities" dx/dz, dy/dz, and dt/dz be continuous at $z^{\rm en}$. Let us work out the consequences of such a requirement. Before entry we have the result

$$\begin{aligned} &dx/dz = \partial K^{\rm ben}/\partial p_x^{\rm canben} = \\ &(p_x^{\rm canben}/[(p_t^{\rm canben})^2/c^2 - m^2c^2 - (p_x^{\rm canben})^2 - (p_y^{\rm canben})^2]^{1/2}, \end{aligned} \tag{17.7.114}$$

and after entry there is the result

$$\begin{aligned} &dx/dz = \partial K^{\rm aen}/\partial p_x^{\rm canaen} = \\ &(p_x^{\rm canaen} - qA_x)/[(p_t^{\rm canaen})^2/c^2 - m^2c^2 - (p_x^{\rm canaen} - qA_x)^2 - (p_y^{\rm canaen} - qA_y)^2]^{1/2}. \end{aligned} \tag{17.7.115}$$

An analogous result holds for dy/dz. Finally, for dt/dz there is the before entry result

$$\begin{aligned} &dt/dz = \partial K^{\rm ben}/\partial p_t^{\rm canben} = \\ &(-p_t^{\rm canben}/c^2)/[(p_t^{\rm canben})^2/c^2 - m^2c^2 - (p_x^{\rm canben})^2 - (p_y^{\rm canben})^2]^{1/2}, \end{aligned} \tag{17.7.116}$$

and the after entry result

$$\begin{aligned} &dt/dz = \partial K^{\rm aen}/\partial p_t^{\rm canaen} = \\ &(-p_t^{\rm canaen}/c^2)/[(p_t^{\rm canaen})^2/c^2 - m^2c^2 - (p_x^{\rm canaen} - qA_x)^2 - (p_y^{\rm canaen} - qA_y)^2]^{1/2}. \end{aligned} \tag{17.7.117}$$

Now equate the far right sides of (7.16) and (7.17), the far right sides of there dy/dz counterparts, and the far right sides of (7.18) and (7.19). So doing yields the transition matching relations

$$p_x^{\rm canben} = p_x^{\rm canaen} - qA_x, \tag{17.7.118}$$

$$p_y^{\rm canben} = p_y^{\rm canaen} - qA_y, \tag{17.7.119}$$

$$p_t^{\rm canben} = p_t^{\rm canaen}. \tag{17.7.120}$$

In view of (7.3) and (7.5) these relations can also be written in the form

$$p_x^{\rm mechben} = p_x^{\rm mechaen}, \tag{17.7.121}$$

$$p_y^{\rm mechben} = p_y^{\rm mechaen}, \tag{17.7.122}$$

$$p_z^{\rm mechben} = p_z^{\rm mechaen}, \tag{17.7.123}$$

The relation (7.22) is satisfactory because magnetic forces do not change the energy. Recall (1.6.17). However, we also desire that the phase-space transformation given by (7.12) through (7.14) and (7.20) through (7.22) be symplectic. Calculation shows that it is not. Compute the Poisson bracket of the right sides of (7.20) and (7.21) to find the result

$$\begin{aligned}[p_x^{\rm canaen} - qA_x, p_y^{\rm canaen} - qA_y] &= [p_x^{\rm canaen}, -qA_y] + [-qA_x, p_y^{\rm canaen}] \\ &= q\{\partial A_y/\partial x^{\rm aen} - \partial A_x/\partial y^{\rm aen}\} = qB_z. \quad (17.7.124)\end{aligned}$$

While hopefully small, generally $B_z(x, y, z^{\rm en})$ differs from zero at the beginning of the leading fringe-field region. On the other hand, the Poisson bracket of the left sides of (7.20) and (7.21) must vanish since $p_x^{\rm canben}$ and $p_y^{\rm canben}$ are supposed to be canonical momenta. Therefore the the phase-space transformation given by (7.12) through (7.14) and (7.20) through (7.22) is generally not symplectic.

We expect that neglect of the magnetic field in the region $z < z^{\rm en}$ will lead to some error in trajectories. However, we do not want this error to violate the symplectic condition. The simplest way to maintain the symplectic condition is to retain the relations (7.12) through (7.14) and replace the relations (7.20) through (7.22) by the relations

$$p_x^{\rm canben} = p_x^{\rm canaen}, \quad (17.7.125)$$

$$p_y^{\rm canben} = p_y^{\rm canaen}, \quad (17.7.126)$$

$$p_t^{\rm canben} = p_t^{\rm canaen}. \quad (17.7.127)$$

In this case the transition matching relations (7.12) through (7.14) and (7.20) through (7.22) amount to the identity map $\mathcal{I}$, and the symplectic condition is trivially satisfied. Now, however, the error in trajectories manifests itself in that the slopes/"velocities" dx/dz, dy/dz, and dt/dz may be expected to be discontinuous at at $z^{\rm en}$. Inspection of (7.16) and (7.17), their dy/dz counterparts, and (7.18) and (7.19) shows that, in lowest approximation, these discontinuities are proportional to $A_x(x, y, z^{\rm en})$ and $A_y(x, y, z^{\rm en})$. Indeed, again in view of (7.3) and (7.5), the transition relations (7.26) through (7.28) can be written in the form

$$\Delta \boldsymbol{p}^{\rm mech} = \boldsymbol{p}^{\rm mechaen} - \boldsymbol{p}^{\rm mechben} = q\boldsymbol{A}(x, y, z^{\rm en}). \quad (17.7.128)$$

It is therefore desirable, where feasible, to work in a gauge where $\boldsymbol{A}(x, y, z^{\rm en})$ is as small as possible.

One way to view the symplectic matching relations (7.12) and (7.14) and (7.24) through (7.26) is to replace the Hamiltonian (7.1) by a modified Hamiltonian $K^{\rm mod}$ given by

$$K^{\rm mod} = -[(p_t^{\rm can})^2/c^2 - m^2c^2 - (p_x^{\rm can} - qA_x^{\rm mod})^2 - (p_y^{\rm can} - qA_y^{\rm mod})^2]^{1/2} - qA_z^{\rm mod} \quad (17.7.129)$$

where

$$\boldsymbol{A}^{\rm mod}(x, y, z) = \theta(z - z^{\rm en})\boldsymbol{A}(x, y, z). \quad (17.7.130)$$

That is, the vector potential is taken to vanish for $z < z^{\rm en}$ and turns on at $z = z^{\rm en}$. A little thought shows that integrating the equations of motion associated with this modified Hamiltonian automatically produces the matching relations (7.12) and (7.14) and (7.24) through (7.26).

What is the modified magnetic field $\boldsymbol{B}^{\rm mod}$ associated with this modified vector potential? Evaluation of $\nabla \times \boldsymbol{A}^{\rm mod}$ gives the relations

$$B_x^{\rm mod}(x,y,z) = \theta(z - z^{\rm en})B_x(x,y,z) - \delta(z - z^{\rm en})A_y(x,y,z), \tag{17.7.131}$$

$$B_y^{\rm mod}(x,y,z) = \theta(z - z^{\rm en})B_y(x,y,z) + \delta(z - z^{\rm en})A_x(x,y,z), \tag{17.7.132}$$

$$B_z^{\rm mod}(x,y,z) = \theta(z - z^{\rm en})B_z(x,y,z). \tag{17.7.133}$$

Calculation shows that $\boldsymbol{B}^{\rm mod}$ has divergence

$$\nabla \cdot \boldsymbol{B}^{\rm mod} = 0, \tag{17.7.134}$$

as desired. What current produces this modified magnetic field? The modified magnetic field satisfies the curl relation

$$\nabla \times \boldsymbol{B}^{\rm mod} = \boldsymbol{j}^{\rm mod} \tag{17.7.135}$$

where

$$\begin{aligned} j_x^{\rm mod} &= (\partial/\partial y)B_z^{\rm mod} - (\partial/\partial z)B_y^{\rm mod} \\ &= -\delta(z - z^{\rm en})[B_y(x,y,z) + (\partial/\partial z)A_x(x,y,z)] - \delta'(z - z^{\rm en})A_x(x,y,z) \\ &= -\delta(z - z^{\rm en})[2(\partial/\partial z)A_x(x,y,z) - (\partial/\partial x)A_z(x,y,z)] - \delta'(z - z^{\rm en})A_x(x,y,z), \end{aligned} \tag{17.7.136}$$

$$\begin{aligned} j_y^{\rm mod} &= (\partial/\partial z)B_x^{\rm mod} - (\partial/\partial x)B_z^{\rm mod} \\ &= -\delta(z - z^{\rm en})[-B_x(x,y,z) + (\partial/\partial z)A_y(x,y,z)] - \delta'(z - z^{\rm en})A_y(x,y,z) \\ &= -\delta(z - z^{\rm en})[2(\partial/\partial z)A_y(x,y,z) - (\partial/\partial y)A_z(x,y,z)] - \delta'(z - z^{\rm en})A_y(x,y,z), \end{aligned} \tag{17.7.137}$$

$$\begin{aligned} j_z^{\rm mod} &= (\partial/\partial x)B_y^{\rm mod} - (\partial/\partial y)B_x^{\rm mod} \\ &= \delta(z - z^{\rm en})[(\partial/\partial x)A_x(x,y,z) + (\partial/\partial y)A_y(x,y,z)]. \end{aligned} \tag{17.7.138}$$

Evidently terminating the vector potential at $z = z^{\rm en}$ is equivalent to introducing sheet (corresponding to the δ function) and double-sheet (corresponding to the δ' function) currents at $z = z^{\rm en}$. And the strengths of these currents are proportional to the values of $\boldsymbol{A}$ and its first derivatives at $z = z^{\rm en}$.

Exiting a Trailing Fringe-Field Region

Bibliography

Fitting Based on Use of Surface Data

[1] M. Venturini and A. Dragt, "Accurate Computation of Transfer Maps from Magnetic Field Data", *Nuclear Instruments and Methods* **A427**, p. 387 (1991).

[2] M. Venturini, "Lie Methods, Exact Map Computation, and the Problem of Dispersion in Space Charge Dominated Beams", University of Maryland Physics Department Ph.D. Thesis (1998).

[3] C. Mitchell, "Calculation of Realistic Charged-Particle Transfer Maps", University of Maryland Physics Department Ph.D. Thesis (2007).

[4] C.E. Mitchell and A.J. Dragt, "Accurate Transfer Maps for Realistic Beamline Elements: Part I, Straight Elements", 19 pages, *Phys. Rev. ST Accel. Beams* **13**, 064001 (2010).

Mathieu Functions

See also the references on Krein-Moser Theory and Periodic Linear Systems at the end of Chapter 3.

[5] N.W. McLachlan, *Theory and Application of Mathieu Functions*, Dover (1964).

[6] M. Strutt, *Lamésche-Mathieusche-und Verwandte Functionen in Physik und Technik*, Chelsea (1967).

[7] A. Erdélyi, Edit., *Higher Transcendental Functions*, Volume III, Chapter XVI, McGraw-Hill (1955).

[8] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Chapter 20, Dover (1972). Also available on the Web by Googling "abramowitz and stegun 1972".

[9] F. Olver, D. Lozier, R. Boisvert, and C. Clark, Editors, *NIST Handbook of Mathematical Functions*, Cambridge (2010). See also the Web site http://dlmf.nist.gov/.

[10] W. Magnus and S. Winkler, *Hill's Equation*, Dover (1979).

Other Fitting Methods

[11] L. Teng, "Expanded Form of Magnetic Field with Median Plane", Argonne National Laboratory Report ANL-LCT-28 (15 December 1962). https://www.osti.gov/scitech/servlets/purl/7364341.

[12] E. Akeley, "The Vector Potential of the Magnetic Field in the Mark V Accelerator", Midwestern Universities Research Association (MURA) Report MURA-70 ESA(MURA-3) (28 May 1955). http://lss.fnal.gov/archive/other/mura/MURA-070.pdf.

Chapter 18

Tools for Numerical Implementation

This chapter develops the tools that are necessary for the numerical implementation of the methods of Chapter 14. These tools include splines, bicubic interpolation, spline-based Fourier transforms, and routines for the calculation of Bessel and Mathieu functions.

18.1 Third-Order Splines

For our purposes splines are piecewise polynomial fits where various continuity conditions are imposed at the points the pieces join. There are two common possibilities: Either a fit is desired over some interval that may be viewed a portion of the real line; or a fit is desired over a full angular interval, in which case periodicity is to be imposed.

18.1.1 Fitting Over an Interval

Let $y = f(x)$ be a function of a single variable. Suppose its values y_j are specified at $N + 1$ equally spaced points x_0, x_1, $\cdots$, x_N over the interval $[x_0, x_N]$. (See Figure 2.1.1 for a similar setup employing the variable t.) Also suppose that on each subinterval $[x_j, x_{j+1}]$ we want to approximate f by a cubic polynomial with cubic polynomials on adjacent subintervals matched is such a way that f has continuous first and second derivatives at each *interior* point x_j. Such an approximation will be called a *cubic* or *third-order spline*. We will use these splines both for interpolation and for the calculation of direct and inverse Fourier transforms. See Sections 15.2 and 15.3.5.

Let us see what information is required to construct such a sequence of third-order polynomials (one for each interval). On the first subinterval, $[x_0, x_1]$, write

$$y = f_0(x) = a_0 + b_0(x - x_0) + c_0(x - x_0)^2 + d_0(x - x_0)^3 \tag{18.1.1}$$

where the coefficients a_0 through d_0 are to be determined. Then the condition

$$f_0(x_0) = y_0 \tag{18.1.2}$$

yields the relation

$$a_0 = y_0. \tag{18.1.3}$$

Next, for the moment, suppose we further require that

$$f_0'(x_0) = \beta_0 \tag{18.1.4}$$

and

$$f_0''(x_0) = \gamma_0. \tag{18.1.5}$$

These requirements yield the further relations

$$b_0 = \beta_0, \tag{18.1.6}$$

$$c_0 = \gamma_0/2. \tag{18.1.7}$$

Finally, the condition

$$f_0(x_1) = y_1 \tag{18.1.8}$$

yields the relation

$$y_1 = y_0 + \beta_0(x_1 - x_0) + (\gamma_0/2)(x_1 - x_0)^2 + d_0(x_1 - x_0)^3, \tag{18.1.9}$$

which can be solved to yield the value of d_0. The conditions (1.2) and (1.8), plus the requirements (1.4) and (1.5), have completely specified the first cubic polynomial (1.1).

Let us now move on to the second subinterval $[x_1, x_2]$. On this subinterval we assume that there is the cubic polynomial representation

$$y = f_1(x) = a_1 + b_1(x - x_1) + c_1(x - x_1)^2 + d_1(x - x_1)^3, \tag{18.1.10}$$

and we find from the condition

$$f_1(x_1) = y_1 \tag{18.1.11}$$

the relation

$$a_1 = y_1. \tag{18.1.12}$$

Also, since $f_0(x)$ has already been determined, the values $f_0'(x_1)$ and $f_0''(x_1)$ are already known. The relations (1.8) and (1.11) already guarantee continuity of f_0 and f_1 across the join at x_1. Next, as set forth in our initial statement of intent, let us require that

$$f_1'(x_1) = \beta_1 = f_0'(x_1) \tag{18.1.13}$$

and

$$f_1''(x_1) = \gamma_1 = f_0''(x_1). \tag{18.1.14}$$

That is, we also require continuity in the first and second derivatives. From (1.13) and (1.14) we conclude that

$$b_1 = \beta_1 = f_0'(x_1) \tag{18.1.15}$$

and

$$c_1 = \gamma_1/2 = f_0''(x_1)/2. \tag{18.1.16}$$

Finally, the condition

$$f_1(x_2) = y_2 \tag{18.1.17}$$

yields the relation

$$y_2 = y_1 + \beta_1(x_2 - x_1) + (\gamma_1/2)(x_2 - x_1)^2 + d_1(x_2 - x_1)^3, \tag{18.1.18}$$

which can be solved to yield the value of d_1. We see that the condition (1.17) plus the continuity requirements have completely specified the second cubic polynomial (1.10).

It is clear that this process can be continued for the subsequent subintervals $[x_2, x_3]$ $\cdots$ $[x_{N-1}, x_N]$ so that all the cubic polynomials are completely specified in terms of the y_0 $\cdots$ y_N and the two numbers β_0 and γ_0. Now we come to a subtle point. Since the cubic polynomials are all completely specified, the value $f'_{N-1}(x_N)$ is also specified in terms of the y_1 $\cdots$ y_N and the two numbers β_0 and γ_0. In fact, there will be a relation of the form

$$f'_{N-1}(x_N) = \delta + \epsilon\gamma_0 \tag{18.1.19}$$

where $\delta(\beta_0, y_0, \cdots, y_N)$ is some (linear) function of $\beta_0, y_0, \cdots, y_N$, and ϵ is some *nonzero* coefficient. Therefore, we may adjust γ_0 in such a way as to give $f'_{N-1}(x_N)$ any desired value. Put another way, we may replace a knowledge of γ_0 with a specification of $f'_{N-1}(x_N)$. Let us write

$$f'(x_0) = f'_0(x_0) = \beta_0 \tag{18.1.20}$$

and

$$f'(x_N) = f'_{N-1}(x_N). \tag{18.1.21}$$

With this notation in mind, we may view a cubic spline as being completely specified by the values y_0 $\cdots$ y_N and the two end-point derivatives $f'(x_0)$ and $f'(x_N)$.[1]

Of course, in general the end-point derivatives are unknown. Many users of cubic splines simply set end-point derivatives (either first or second) to zero on the grounds of convenience and the fact (to be demonstrated shortly) that their values actually have little effect on the spline approximation once one is a few grid points away from the ends.[2] For our purposes, we prefer to use the first few data points near the end points to estimate the end-point first derivatives. For example, upon deciding to use the first three points to estimate $f'(x_0)$ and the last three to estimate $f'(x_N)$, we use the approximations

$$\begin{aligned} f'(x_0) &= (1/h)[-(3/2)y_0 + 2y_1 - (1/2)y_2] + O(h^2), \\ f'(x_N) &= (1/h)[(3/2)y_N - 2y_{N-1} + (1/2)y_{N-2}] + O(h^2) \end{aligned} \tag{18.1.22}$$

where h is the spacing between successive grid points,

$$h = x_1 - x_0. \tag{18.1.23}$$

If we choose to employ the first and last four points, we use the approximations

$$\begin{aligned} f'(x_0) &= (1/h)[-(11/6)y_0 + 3y_1 - (3/2)y_2 + (1/3)y_3] + O(h^3), \\ f'(x_N) &= (1/h)[(11/6)y_N - 3y_{N-1} + (3/2)y_{N-2} - (1/3)y_{N-3}] + O(h^3). \end{aligned} \tag{18.1.24}$$

[1] Evidently, an alternate procedure is to specify the values y_0 $\cdots$ y_N and the two second-order end-point derivatives $f''(x_0)$ and $f''(x_N)$.

[2] If the second derivatives at each end point are set to zero, such a cubic spline is said to be *natural*.

See Exercise 1.1.

At this juncture we must remark that the algorithm we have been describing for computing cubic splines, while pedagogically instructive, is not numerically stable against roundoff errors. A stable spline routine is given in Appendix L.

As already alluded to, one of the advantages of spline fits is *localization* in that the fits over any subinterval in x depend primarily on the y_j values whose corresponding x_j lie within that subinterval. For example, consider the function $y = f(x)$ defined on the interval $x \in [0, 3]$ such that $f(1.5) = 1$ and $f = 0$ elsewhere. Figure 1.1 shows the function that is produced by a cubic spline fit when $h = .1$. In this case 31 points are used to make the fit with $x_0 = 0$, $x_{30} = 3$, and all y_j set to zero save for setting y_{15}=1. Also, $f'(0)$ and $f'(3)$ are set to zero. Evidently the spline fit falls rapidly to zero on either side of $x = 1.5$. In fact, it can be shown that the successive maxima *decay exponentially* as

$$y(x) \sim \exp[-\alpha(1/h)|x - 1.5|] \tag{18.1.25}$$

where

$$\alpha = \log(2 + \sqrt{3}) \simeq 1.317. \tag{18.1.26}$$

Similarly, Figure 1.2 shows the fit that is produced for the same setup when all y_j are set to zero, $f'(0)$ is set to 1, and $f'(3)$ is set to zero. Again the fit decays to zero exponentially with exponent $-\alpha$.

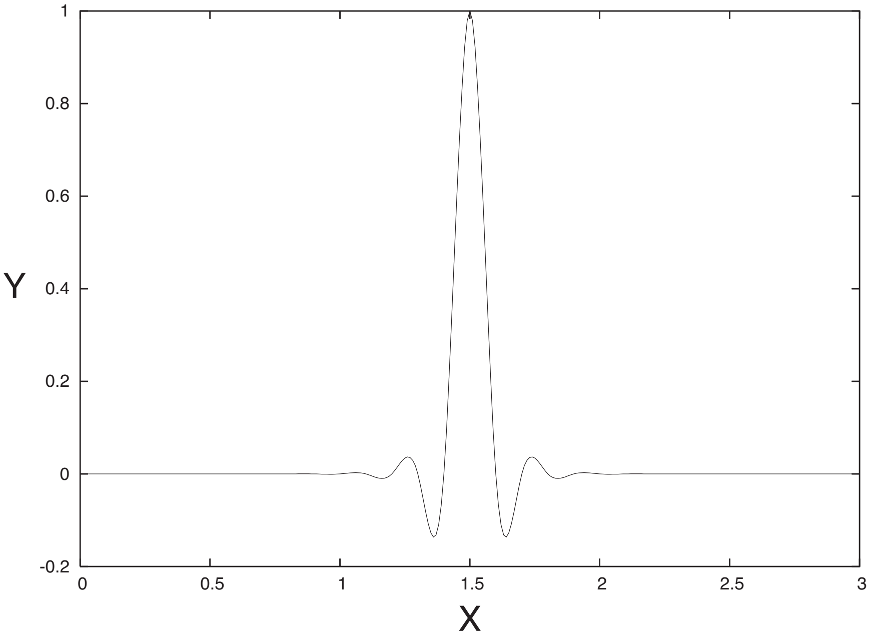


Figure 18.1.1: The 31-point spline fit associated with $y_{15} = 1$ and all other $y_j = 0$. Also, $f'(0)$ and $f'(3)$ are set to zero. Note that the fit falls rapidly to zero on either side of $x = 1.5$.

18.1.2 Periodic Splines

The splines defined so far are useful for fitting a general function over an interval. Suppose we instead want to fit a function which is known to be periodic. Such functions will typically

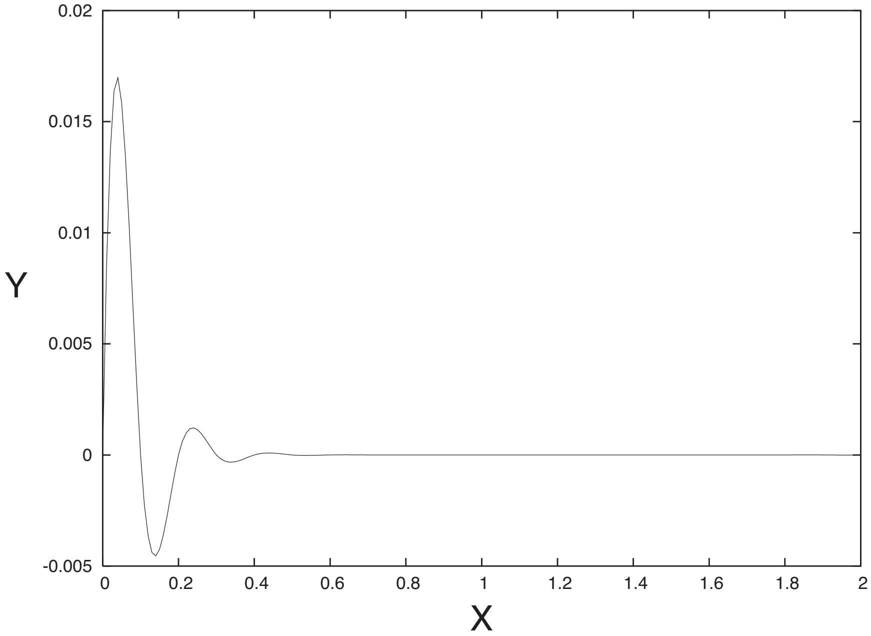


Figure 18.1.2: The spline fit associated with $f'(x_0) = 1$, all $y_j = 0$, and $f'(x_{30}) = 0$. Note that the fit falls rapidly to zero for x beyond $x_0 = 0$. Only the results over the interval $x \in [0, 2]$ are displayed.

depend on angular variables.

As before, we imagine that there are known function values y_j at the points x_0 to x_N, but that now $y_0 = y_N$. Begin the construction as before to find a unique set of cubic polynomials in terms of the y_j and β_0 and γ_0. With this construction, the quantities $f'_{N-1}(x_N)$ and $f''_{N-1}(x_N)$ are specified. Indeed, there will be relations of the form

$$f'_{N-1}(x_N) = r(y_0, \cdots, y_N) + s\beta_0 + t\gamma_0, \tag{18.1.27}$$

$$f''_{N-1}(x_N) = \rho(y_0, \cdots, y_N) + \sigma\beta_0 + \tau\gamma_0 \tag{18.1.28}$$

where r and ρ are linear functions of the y_j; and s, t, σ, and τ are proportionality constants. Now adjust both β_0 and γ_0 such that there are the relations

$$f'_{N-1}(x_N) = f'_0(x_0) \tag{18.1.29}$$

and

$$f''_{N-1}(x_N) = f''_0(x_0). \tag{18.1.30}$$

So doing will make the spline fit periodic in that not only will $y_0 = y_N$, also the first and second derivatives will match at the endpoints. In view of (1.4), (1.5), (1.27), and (1.28), these matching relations are equivalent to the conditions

$$\beta_0 = r(y_0, \cdots, y_N) + s\beta_0 + t\gamma_0, \tag{18.1.31}$$

$$\gamma_0 = \rho(y_0, \cdots, y_N) + \sigma\beta_0 + \tau\gamma_0. \tag{18.1.32}$$

These equations have a (unique) solution provided the matrix M defined by

$$M = \begin{pmatrix} s-1 & t \\ \sigma & \tau - 1 \end{pmatrix}. \tag{18.1.33}$$

has a nonzero determinant, which can be shown to be always the case. We conclude that a periodic cubic spline is uniquely specified by the values $y_0, \cdots, y_N$ with $y_0 = y_N$.

Here again we must remark that the pedagogically instructive algorithm we have been describing for computing periodic cubic splines is not numerically stable against roundoff errors. A stable periodic spline routine is also given in Appendix L.

18.1.3 Error Estimate for Spline Approximation

There remains the question of accuracy for a cubic spline approximation. Suppose the function f that is being approximated is known to have a continuous fourth-order derivative. Then it can be shown that the error involved in using its spline approximation f_{sa} has the estimate

$$\text{error}(x) = f(x) - f_{\text{sa}}(x) = (h^4/24)\theta^2(1-\theta)^2 f^{\text{iv}}(x) + O(h^5) \tag{18.1.34}$$

for x in the subinterval $[x_j, x_{j+1}]$ and

$$\theta = (x - x_j)/h. \tag{18.1.35}$$

Note that, according to (1.35), θ lies in the interval $\theta \in [0, 1]$. It is easy to check that in this interval the quantity $\theta^2(1-\theta)^2$ does not exceed 1/16.

As an example, suppose

$$f(x) = 1 - x^4 \tag{18.1.36}$$

and we wish to approximate f over the interval $[-1, 1]$. That is, we set $x_0 = -1$ and $x_N = 1$. Figure 1.3 shows f and its spline fit f_{sa} for $h = .1$ (which corresponds to $N = 20$). They are indistinguishable on the scale shown. Figure 1.4 shows the error that occurs when $h = .1$. By construction the error vanishes at the grid points x_j, and the global error is consistent with the estimate (1.34).

In making the spline fits for Figures 1.3 and 1.4, we have used as input for the end-point derivatives the exact results $f'(-1) = 4$ and $f'(+1) = -4$ based on (1.36). Suppose we instead use (1.24) to estimate the end-point derivatives. Figure 1.5 shows the error that then occurs. We see that there is some error at the endpoints and that, as expected from localization, this error soon damps away so that only the error already seen in Figure 1.4 remains. We remark that the use of (1.24) gives the results $f'(-1) = 3.994$ and $f'(+1) = -3.994$. It is pleasantly surprising that, at its worst, the error in the spline fit is considerably less than might have naively been expected based on the error in the estimated end-point derivatives. Finally we remark that, for our applications, the end points occur in fringe-field regions where both the function being approximated and its derivatives are very small. Thus, we expect that the error made in using (1.22) or (1.24) in this case will be negligible.

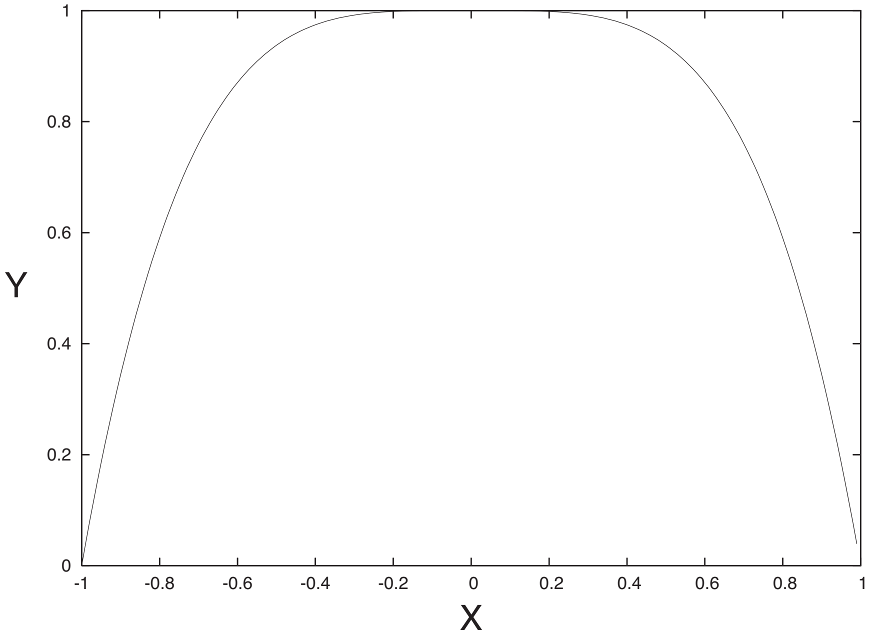


Figure 18.1.3: The function f and its spline fit $f_{\rm sa}$ for $h = .1$. They appear identical.

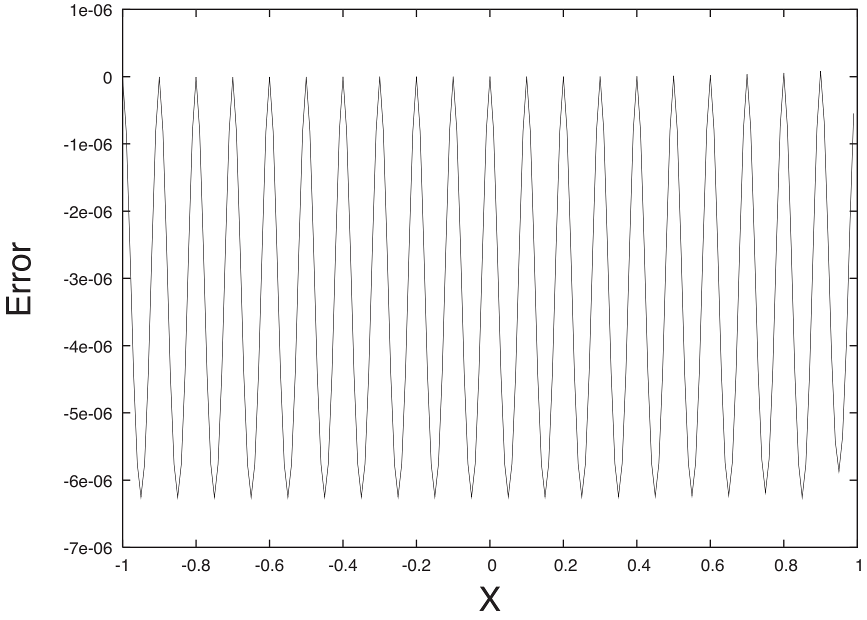


Figure 18.1.4: The difference between the function f and its spline fit $f_{\rm sa}$ for $h = .1$. Here error $= f(x) - f_{\rm sa}(x)$. The spline $f_{\rm sa}$ is constructed using the exact values for the end-point derivatives.

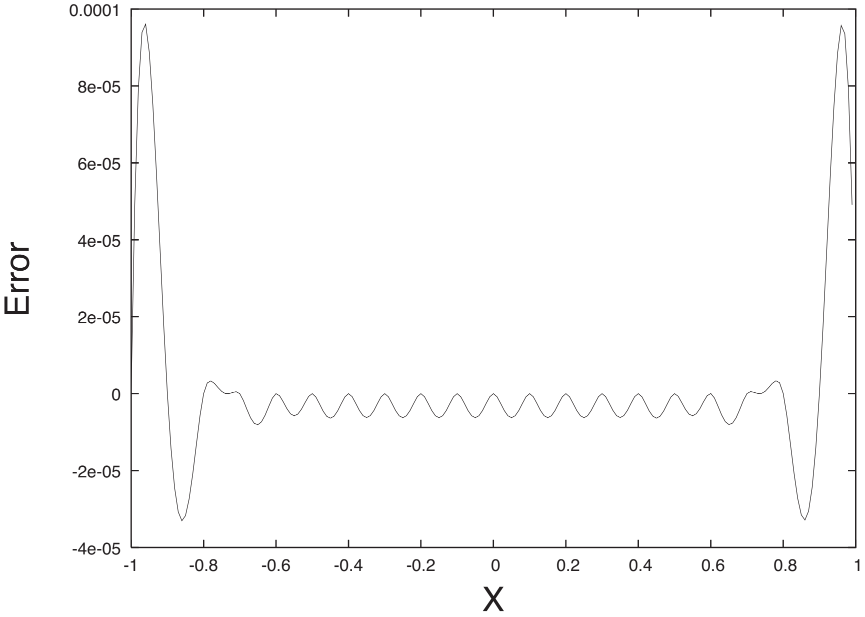


Figure 18.1.5: The difference between the function f and its spline fit $f_{\rm sa}$ for $h = .1$. Here error $= f(x) - f_{\rm sa}(x)$. The spline $f_{\rm sa}$ is constructed using (1.24) to estimate the end-point derivatives.

Exercises

18.1.1. Verify (1.22) through (1.24) using the finite difference calculus of Section 2.4.

18.2 Interpolation

The calculations of Chapter 14 begin with data provided on some regular Cartesian grid. With regard to locations in the z coordinate, we will use those provided, and call them Z_L. However, for the coordinates x and y, interpolation may be required.

In the case where a circular cylinder is employed, we need to interpolate to equi-angular locations given by the relations

$$\bar{x}_i = R\cos(\phi_i), \tag{18.2.1}$$

$$\bar{y}_i = R\sin(\phi_i) \tag{18.2.2}$$

where

$$\phi_0 = 0 \tag{18.2.3}$$

and

$$\phi_N = 2\pi. \tag{18.2.4}$$

See the second frame of Figure 14.1.1. Typically, in the circular cylinder case we take N to have the value $N \approx 50$.

In the case of an elliptical cylinder we write

$$\bar{x}_i = f\cosh(U)\cos(v_i), \tag{18.2.5}$$

$$\bar{y}_i = f \sinh(U) \sin(v_i) \tag{18.2.6}$$

where the v_i are equally spaced with

$$v_0 = 0 \tag{18.2.7}$$

and

$$v_N = 2\pi. \tag{18.2.8}$$

See Figure 14.4.3. Typically, in the elliptic cylinder case, we take N to have the value $N \approx 120$.

18.2.1 Bicubic Interpolation

For each $\bar{x}_i$, $\bar{y}_i$ pair, find the closest 16 points in the regular grid in the x, y plane. See Figure 2.1. Note that the regular grid may be rectangular rather than square. Let X_J and Y_K be the coordinates of the grid point in the lower left corner. That is, we have the following inequalities:

$$X_{J+1} \le \bar{x}_i \le X_{J+2}, \tag{18.2.9}$$

$$Y_{K+1} \le \bar{y}_i \le Y_{K+2}. \tag{18.2.10}$$

Introduce local expansion variables ξ and η by the relations

$$x = X_J + \xi, \tag{18.2.11}$$

$$y = Y_K + \eta; \tag{18.2.12}$$

and also write

$$\bar{x}_i = X_J + \bar{\xi}_i, \tag{18.2.13}$$

$$\bar{y}_i = Y_K + \bar{\eta}_i. \tag{18.2.14}$$

We then interpolate the quantity of interest, be it a potential value or some transverse field component, from the regular grid to the point $\bar{\xi}_i$, $\bar{\eta}_i$ with the aid of a *bicubic* polynomial P in the variables ξ and η.[3] This is a polynomial of the form

$$P(\xi,\eta) = \sum_{m,n=1}^{4} c_{mn}\xi^{m-1}\eta^{n-1} = \sum_{m=1}^{4}(\sum_{n=1}^{4} c_{mn}\eta^{n-1})\xi^{m-1} = \sum_{n=1}^{4}(\sum_{m=1}^{4} c_{mn}\xi^{m-1})\eta^{n-1} \tag{18.2.15}$$

where the coefficients c_{mn} are to be determined. Note that P is cubic in each of the variables ξ and η separately. Hence the name bicubic. Also, note that P is *not* a homogeneous polynomial. For example, it contains the term $\xi^3\eta^3$, but it does not contain the terms ξ^6 or η^6. Finally note that, because c is 4×4, it requires 16 numbers to specify the c_{mn}. This is encouraging, because we have assumed that we have data on the 16 nearest-neighbor grid points.[4]

[3] If the point $\bar{\xi}_i$, $\bar{\eta}_i$ happens to fall on a grid line, then only one-dimensional cubic interpolation is required. If it falls on a grid point, no interpolation is required at all.

[4] This count would not work out so neatly had we attempted what might appear to be more desirable, namely an expansion in homogenous polynomials.

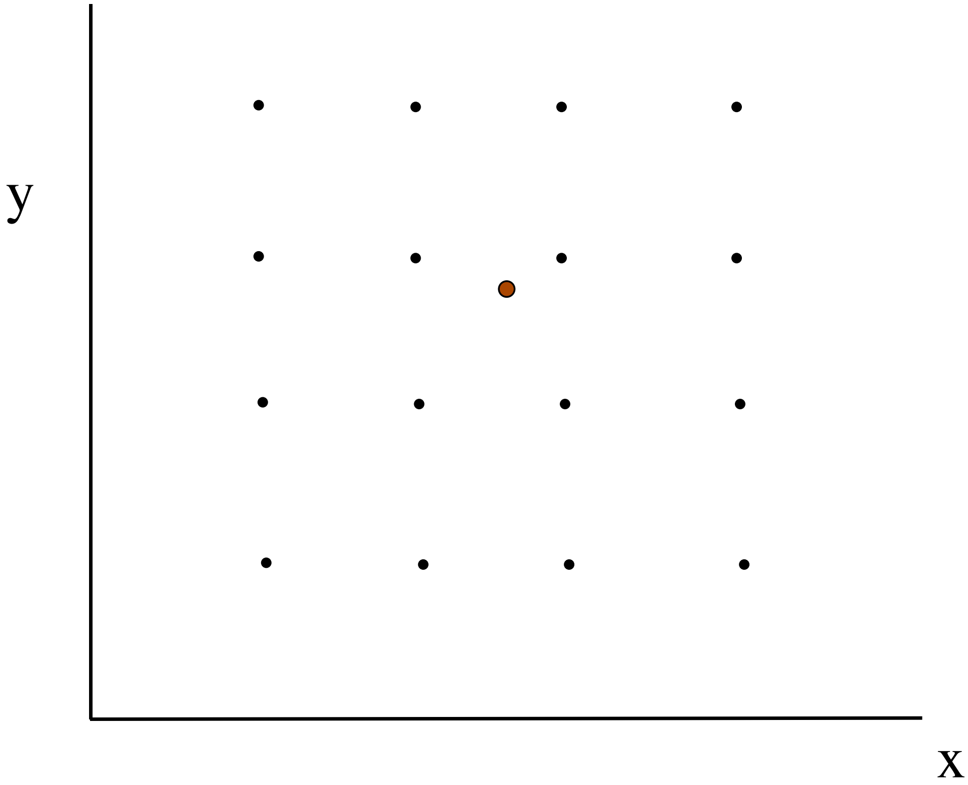


Figure 18.2.1: The point $\bar{x}_i$, $\bar{y}_i$ and its 16 nearest-neighbor grid points. The coordinates of the grid point at the lower left corner are X_J and Y_K.

Let us first verify that P is uniquely defined in terms of values at the the 16 nearest-neighbor grid points. For example, suppose we wish to interpolate a potential function ψ. Let h_x and h_y be the grid spacings in the x and y directions, respectively. Then we know the values Ψ_{jk} given by the relations

$$\Psi_{jk} = \psi[X_J + (j-1)h_x, Y_K + (k-1)h_y, Z_L] \text{ for } j,k \in [1,4]. \tag{18.2.16}$$

In the spirit of (2.11) and (2.12), write

$$\xi_j = (j-1)h_x \tag{18.2.17}$$

and

$$\eta_k = (k-1)h_y. \tag{18.2.18}$$

Then we wish to have the relations

$$\Psi_{jk} = P(\xi_j, \eta_k) = \sum_{m,n=1}^{4} c_{mn} \xi_j^{m-1} \eta_k^{n-1} \ \text{ for } \ j,k \in [1,4], \tag{18.2.19}$$

and hope that these 16 desiderata will determine the 16 c_{mn}.

To explore these relations, define vectors $\hat{\xi}_j$ and $\hat{\eta}_k$ by the rules

$$\hat{\xi}_j = (\xi_j^0, \xi_j^1, \xi_j^2, \xi_j^3)^T = (1, \xi_j, \xi_j^2, \xi_j^3)^T, \tag{18.2.20}$$

and

$$\hat{\eta}_k = (\eta_k^0, \eta_k^1, \eta_k^2, \eta_k^3)^T = (1, \eta_k, \eta_k^2, \eta_k^3)^T. \tag{18.2.21}$$

We will call $\hat{\xi}_j$ and $\hat{\eta}_k$ the *cubic vectors* associated with ξ_j and η_k since they are formed out of the cubic and lower powers of ξ_j and η_k, respectively. With this notation, (2.19) is equivalent to the inner product relations

$$\Psi_{jk} = (\hat{\xi}_j, c\, \hat{\eta}_k) \ \text{ for } \ j,k \in [1,4], \tag{18.2.22}$$

where c is the matrix with entries c_{mn}.

Let $\mathcal{X}$ be the matrix whose columns are the vectors $\hat{\xi}_j$,

$$\mathcal{X} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ \xi_1 & \xi_2 & \xi_3 & \xi_4 \\ \xi_1^2 & \xi_2^2 & \xi_3^2 & \xi_4^2 \\ \xi_1^3 & \xi_2^3 & \xi_3^3 & \xi_4^3 \end{pmatrix}. \tag{18.2.23}$$

Then, by construction, we have the relations

$$\mathcal{X} e^j = \hat{\xi}_j \tag{18.2.24}$$

where the e^j are the standard orthonormal vectors (3.6.4). Similarly, if $\mathcal{Y}$ is the matrix whose columns are the vectors $\hat{\eta}_k$, we have the relations

$$\mathcal{Y} e^k = \hat{\eta}_k. \tag{18.2.25}$$

Insert these relations into (2.22). So doing gives the result

$$\Psi_{jk} = (\mathcal{X}e^j, c\,\mathcal{Y}e^k) = (e^j, \mathcal{X}^T c\,\mathcal{Y}e^k), \tag{18.2.26}$$

which is equivalent to the matrix relation

$$\Psi = \mathcal{X}^T c\,\mathcal{Y}. \tag{18.2.27}$$

We will see shortly that the matrices $\mathcal{X}$ and $\mathcal{Y}$ are invertible. Assuming this to be the case, we may solve (2.27) for c to find the result

$$c = (\mathcal{X}^T)^{-1}\Psi\,\mathcal{Y}^{-1}. \tag{18.2.28}$$

We have found the c_{mn} in terms of the Ψ_{jk}, and therefore P is uniquely specified by the values Ψ_{jk}.

To see that $\mathcal{X}$ is invertible, we examine $\det\mathcal{X}$, which (happily) is a *Vandermonde* determinant. It has the known value

$$\det\mathcal{X} = \prod_{j>k}(\xi_j - \xi_k). \tag{18.2.29}$$

Since the ξ_j are assumed to be *distinct* by construction, $\det\mathcal{X}$ can never vanish. Thus $\mathcal{X}$, and similarly $\mathcal{Y}$, are invertible.

Now that P has been constructed, we find the desired interpolated result $\psi(\bar{x}_i, \bar{y}_i, Z_L)$ by writing

$$\psi(\bar{x}_i, \bar{y}_i, Z_L) \approx P(\bar{\xi}_i, \bar{\eta}_i). \tag{18.2.30}$$

Note that the right side of (2.30) can be written an inner product form involving the two cubic vectors $\hat{\bar{\xi}}_i$ and $\hat{\bar{\eta}}_i$,

$$P(\bar{\xi}_i, \bar{\eta}_i) = (\hat{\bar{\xi}}_i, c\,\hat{\bar{\eta}}_i). \tag{18.2.31}$$

Hence the term *bicubic* interpolation again seems particularly appropriate.

In actual practice, since typically for our purposes any given P would be used only once, it is convenient to proceed somewhat differently. We will interpolate the quantity of interest, be it a potential value or some transverse field component, from the regular grid to the point $\bar{x}_i$, $\bar{y}_i$ with the aid of four plus one cubic polynomials. This approach gives the same result as that obtained by first constructing P and then using (2.30), but employs somewhat more standard tools.

Suppose again that we wish to interpolate a potential function ψ. Using the four values $\psi(X_J, Y_K, Z_L)$, $\psi(X_J, Y_{K+1}, Z_L)$, $\psi(X_J, Y_{K+2}, Z_L)$, $\psi(X_J, Y_{K+3}, Z_L)$, construct the cubic polynomial $g_1(\sigma)$ given by

$$g_1(\sigma) = a_1 + b_1\sigma + c_1\sigma^2 + d_1\sigma^3 \tag{18.2.32}$$

such that

$$g_1[(m-1)h_y] = \psi(X_J, Y_{K+m-1}, Z_L) \text{ for } m = 1, 2, 3, 4. \tag{18.2.33}$$

(This is a standard construction in uniform-spacing *Lagrangian* interpolation.) Also form three more cubic polynomials g_2 through g_4 such that

$$g_k[(m-1)h_y) = \psi(X_{J+k-1}, Y_{K+m-1}, Z_L) \text{ for } m = 1, 2, 3, 4. \tag{18.2.34}$$

Together the four polynomials g_1 through g_4 allow us to interpolate in y for each of the four different x values X_K, X_{K+1}, X_{K+2}, X_{K+3}. (Pictorially, we are interpolating along the four different columns in Figure 2.1.) In particular,using (2.14), we have the four interpolated values

$$\psi_y(X_J, \bar{y}_i, Z_L) = g_1(\bar{\eta}_i), \tag{18.2.35}$$

$$\psi_y(X_{J+1}, \bar{y}_i, Z_L) = g_2(\bar{\eta}_i), \tag{18.2.36}$$

$$\psi_y(X_{J+2}, \bar{y}_i, Z_L) = g_3(\bar{\eta}_i), \tag{18.2.37}$$

$$\psi_y(X_{J+3}, \bar{y}_i, Z_L) = g_4(\bar{\eta}_i). \tag{18.2.38}$$

Next, using the four interpolated potential values $\psi_y(X_{J+k-1}, \bar{y}_i, Z_L)$ with $k \in [1, 4]$, we will interpolate in x. Construct a fifth cubic polynomial $f(\tau)$ of the form

$$f(\tau) = \alpha + \beta\tau + \gamma\tau^2 + \delta\tau^3 \tag{18.2.39}$$

such that

$$f[(k-1)h_x) = \psi_y(X_{J+k-1}, \bar{y}_i, Z_L) \text{ for } k = 1, 2, 3, 4. \tag{18.2.40}$$

Then, interpolating in x using f and (2.13) gives the final desired result

$$\psi(\bar{x}_i, \bar{y}_i, Z_L) \approx \psi_{xy}(\bar{x}_i, \bar{y}_i, Z_L) = f(\bar{\xi}_i). \tag{18.2.41}$$

At this point the observant reader has no doubt noticed that it also possible to interpolate in x first, and then in y, to get a result that we will call $\psi_{yx}(\bar{x}_i, \bar{y}_i, Z_L)$. How are $\psi_{xy}(\bar{x}_i, \bar{y}_i, Z_L)$ and $\psi_{yx}(\bar{x}_i, \bar{y}_i, Z_L)$ related? They are equal. In fact, there are the relations.

$$\psi_{xy}(\bar{x}_i, \bar{y}_i, Z_L) = \psi_{yx}(\bar{x}_i, \bar{y}_i, Z_L) = P(\bar{\xi}_i, \bar{\eta}_i). \tag{18.2.42}$$

That is, all three interpolation results agree. See Exercise 2.1

18.2.2 Bicubic Spline Interpolation

In the discussion so far, we have used cubic polynomials to interpolate in both x and y. An alternate approach is to use the cubic splines of Subsection 15.1.1 to perform interpolations. As before, one could interpolate first in y and then in x, or vice versa. It can be shown that these two results will again be the same. Look again at Figure 2.1, and consider the *central* square/rectangle that contains the point $\bar{x}_i$, $\bar{y}_i$. The coordinates of the grid point at the lower left corner of the central square/rectangle are X_{J+1} and Y_{K+1}. Now introduce *central* expansion variables, again call them ξ and η, by writing

$$\begin{aligned} x &= X_{J+1} + \xi, \\ y &= Y_{K+1} + \eta. \end{aligned} \tag{18.2.43}$$

Then it can be shown that this interpolation procedure is equivalent to using a bicubic polynomial again of the form (2.15), but now in the central expansion variables with the coefficients c_{mn} now determined from the Ψ_{jk} with the aid of cubic splines.

Exercises

18.2.1. Verify that ....

18.3 Fourier Transforms

The work of Sections 14.2 through 14.5 required the computation of Fourier transforms. In this section we will describe numerical methods for this task. We will first define the Fourier transform and find its large $|k|$ behavior. We will then define *discrete* Fourier transforms, and explore their large $|k|$ behavior. Finally, we will define *spline-based* Fourier transforms that have, for our purposes, superior properties.

18.3.1 Exact Fourier Transform and Its Large $|k|$ Behavior

Suppose $f(z)$ is a function that is nonzero (has *support*) only within the interval $[a, b]$, and that we wish to find its linear Fourier transform

$$\tilde{f}(k) = [1/(2\pi)]\int_{-\infty}^{\infty} dz \exp(-ikz)f(z) = [1/(2\pi)]\int_{a}^{b} dz \exp(-ikz)f(z). \tag{18.3.1}$$

Let us examine the behavior of $\tilde{f}$ under the further supposition that f is differentiable and perhaps also has specific properties at the endpoints a and b. Then (3.1) may be integrated by parts to give the relation

$$\tilde{f}(k) = -[1/(2\pi)][1/(ik)]\exp(-ikz)f(z)|_a^b + [1/(2\pi)][1/(ik)]\int_a^b dz \exp(-ikz)f'(z). \tag{18.3.2}$$

The second term on the right in (3.2) may again be integrated by parts to give the result

$$\begin{aligned}&[1/(2\pi)][1/(ik)]\int_a^b dz \exp(-ikz)f'(z)\\ &= -[1/(2\pi)][1/(ik)^2]\exp(-ikz)f'(z)|_a^b + [1/(2\pi)][1/(ik)^2]\int_a^b dz \exp(-ikz)f''(z).\end{aligned} \tag{18.3.3}$$

Evidently, this process of integration by parts may be repeated at will as long as the required higher derivatives of f exist, and each such repetition produces one more power of $1/k$. For future use, we will repeat the process two more times to arrive at the result

$$\begin{aligned}\tilde{f}(k) &= -[1/(2\pi)][1/(ik)]\exp(-ikz)f(z)|_a^b - [1/(2\pi)][1/(ik)^2]\exp(-ikz)f'(z)|_a^b\\ &- [1/(2\pi)][1/(ik)^3]\exp(-ikz)f''(z)|_a^b - [1/(2\pi)][1/(ik)^4]\exp(-ikz)f'''(z)|_a^b\\ &+ [1/(2\pi)][1/(ik)^4]\int_a^b dz \exp(-ikz)f^{\rm iv}(z).\end{aligned} \tag{18.3.4}$$

We see that in general

$$|\tilde{f}(k)| \sim 1/|k| \text{ as } k \to \infty, \tag{18.3.5}$$

and that if f vanishes at the endpoints, $f(a) = f(b) = 0$, as will often be the case, then the first term on the right side of (3.4) will vanish so that

$$|\tilde{f}(k)| \sim 1/|k|^2 \text{ as } k \to \infty, \text{ etc.} \tag{18.3.6}$$

Thus, $\tilde{f}(k)$ must vanish at least as fast as $1/|k|$ for large $|k|$, and often vanishes as $1/|k|^2$.

18.3.2 Inverse Fourier Transform

One of the key features of the linear Fourier transform is the inverse Fourier transform relation

$$f(z) = \int_{-\infty}^{\infty} dk \exp(ikz)\tilde{f}(k). \tag{18.3.7}$$

That is, a function can be *reconstructed* from its linear Fourier transform by using an inverse Fourier transform. Let us further assume that the integral (3.7) can be *cut off* for $|k| > K_c$ where K_c is some suitably large value, say a value where and beyond which the asymptotic behavior (3.5) or (3.6) has effectively driven $\tilde{f}$ to zero. Thus, we write

$$f(z) \approx \int_{-K_c}^{K_c} dk \exp(ikz)\tilde{f}(k). \tag{18.3.8}$$

Eventually we will need to evaluate integrals of the form (3.8) numerically. Therefore, we would like to know something about the properties of $\tilde{f}(k)$ in the interval $[-K_c, K_c]$. In particular, we would like to know how much $\tilde{f}(k)$ oscillates. Suppose that $f(z)$ has support only in the interval $[a, b]$. Then $\tilde{f}(k)$ must encode two pieces of information: it must encode that $f(z)$ is zero outside $[a, b]$ and it must encode the behavior of $f(z)$ within the interval $[a, b]$. We see from (3.1) that $\tilde{f}(k)$ is a generalized sum (integral) of terms of the form $\exp(-i\omega k)$ where $\omega \in [a, b]$. That is, $\tilde{f}(k)$ contains all frequencies $\omega \in [a, b]$ with weights $f(\omega)$. This is a potential disaster, because $|a|$ and/or $|b|$ could be quite large.

For example consider the functions $f_{-1,1}(z)$ and $f_{0,2}(z)$ defined by the relations

$$\begin{aligned} f_{-1,1}(z) &= 1 - z^4 \text{ for } z \in [-1, 1], \\ &= 0 \text{ for } z \text{ outside } [-1, 1]; \end{aligned} \tag{18.3.9}$$

$$\begin{aligned} f_{0,2}(z) &= 1 - (z-1)^4 \text{ for } z \in [0, 2], \\ &= 0 \text{ for } z \text{ outside } [0, 2]. \end{aligned} \tag{18.3.10}$$

Figures 3.1 and 3.2 show their graphs which, evidently and by construction, are simply translations of each other. [See also (1.36) and Figure 1.3.] Calculation shows that their Fourier transforms are given by

$$\Re\tilde{f}_{-1,1}(k) = -[4/(\pi k^5)][k(k^2 - 6)\cos k - 3(k^2 - 2)\sin k], \tag{18.3.11}$$

$$\Im\tilde{f}_{-1,1}(k) = 0; \tag{18.3.12}$$

$$\Re\tilde{f}_{0,2}(k) = [1/(\pi k^5)][-4k(k^2-6)\cos^2 k + 6(k^2-2)\sin 2k], \tag{18.3.13}$$

$$\Im\tilde{f}_{0,2}(k) = [1/(\pi k^5)][12(k^2-2)\sin^2 k - 2k(k^2-6)\sin 2k]. \tag{18.3.14}$$

As expected, $\tilde{f}_{-1,1}(k)$ contains frequency $|\omega| = 1$ from the $\sin k$ and $\cos k$ terms. By contrast, $\tilde{f}_{0,2}(k)$ contains frequency $|\omega| = 2$ from the $\sin 2k$, $\cos^2 k$, and $\sin^2 k$ terms. These different behaviors are also evident in the graphs of these Fourier transforms as shown in Figures 3.3 through 3.5. Clearly the integral (3.8) is more difficult to evaluate for $\tilde{f}_{0,2}(k)$ than for $\tilde{f}_{-1,1}(k)$ because $\tilde{f}_{0,2}(k)$ oscillates twice as often as $\tilde{f}_{-1,1}(k)$. There is an oscillation penalty to be paid for encoding the fact that some f has support in the interval $[0, 2]$ rather than the interval $[-1, 1]$.

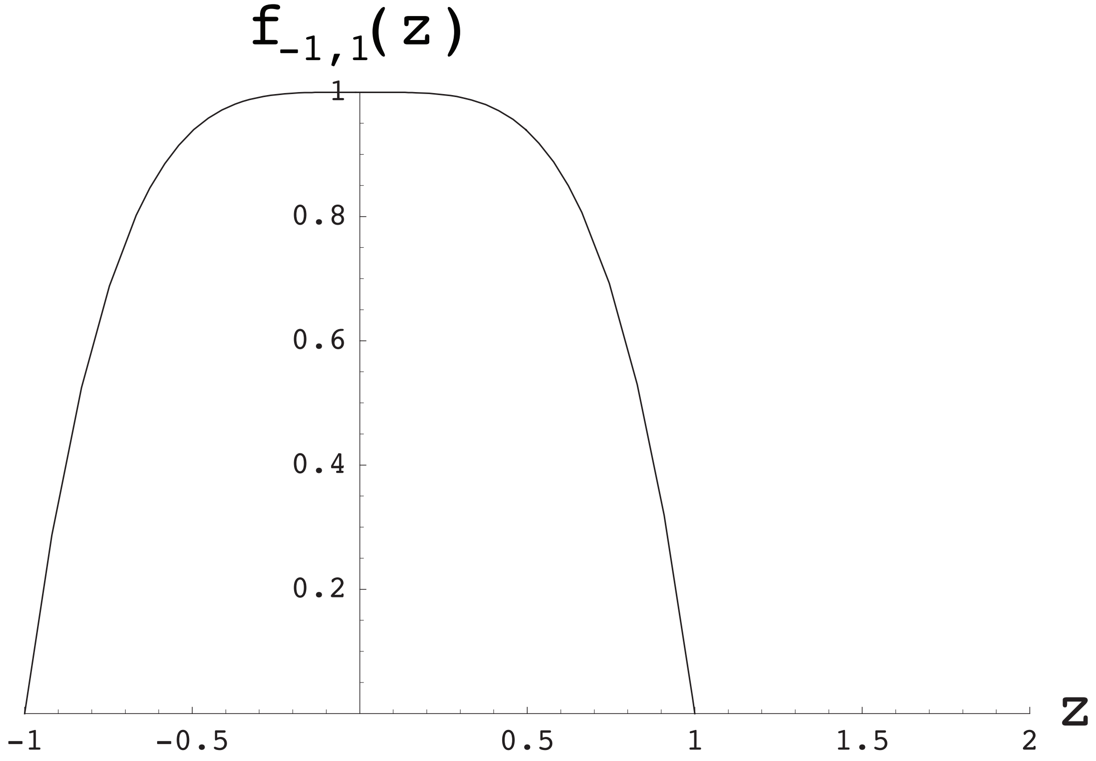


Figure 18.3.1: The function $f_{-1,1}(z)$.

What to do? By simple translation it is always possible to send the interval $[a, b]$ to the interval $[-Z_c, Z_c]$. Therefore. without loss of generality, we may restrict our attention to integrals of the form

$$\tilde{f}(k) = [1/(2\pi)] \int_{-Z_c}^{Z_c} dz \exp(-ikz) f(z). \tag{18.3.15}$$

Then, after all operations have been carried out, we may undo, if we wish, the translation to obtain results in terms of the original coordinates. In this way, we only have to deal with $\tilde{f}(k)$ that contain frequencies satisfying $|\omega| \leq Z_c$.

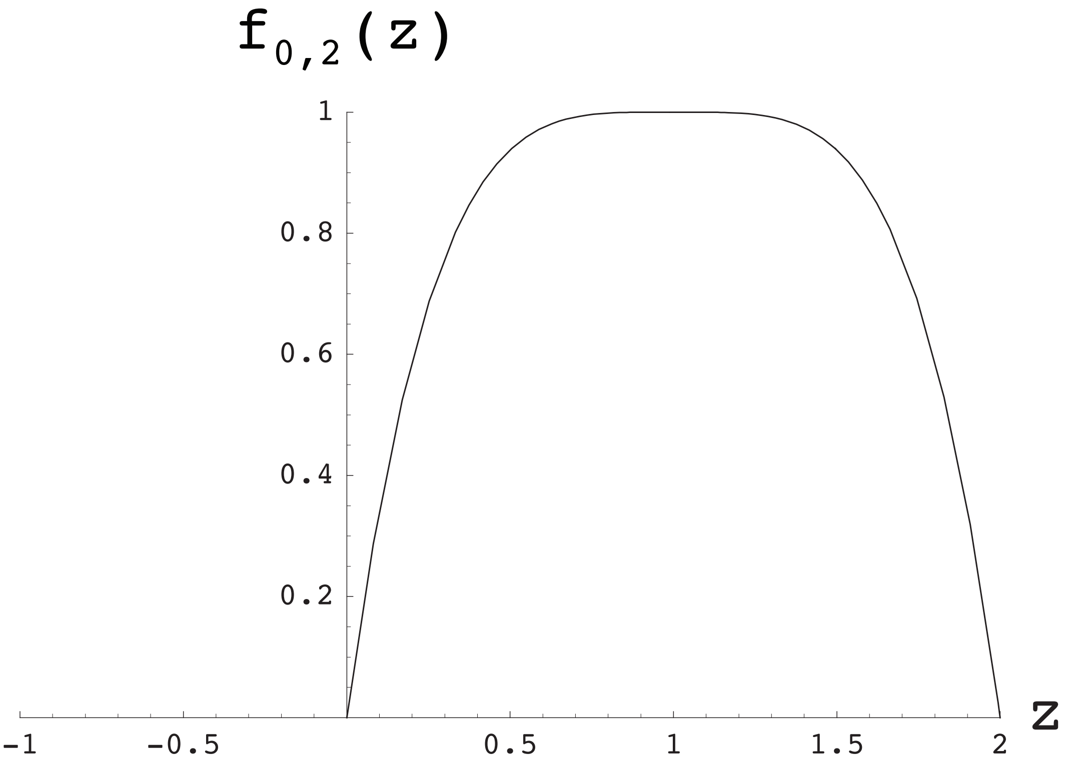


Figure 18.3.2: The function $f_{0,2}(z)$.

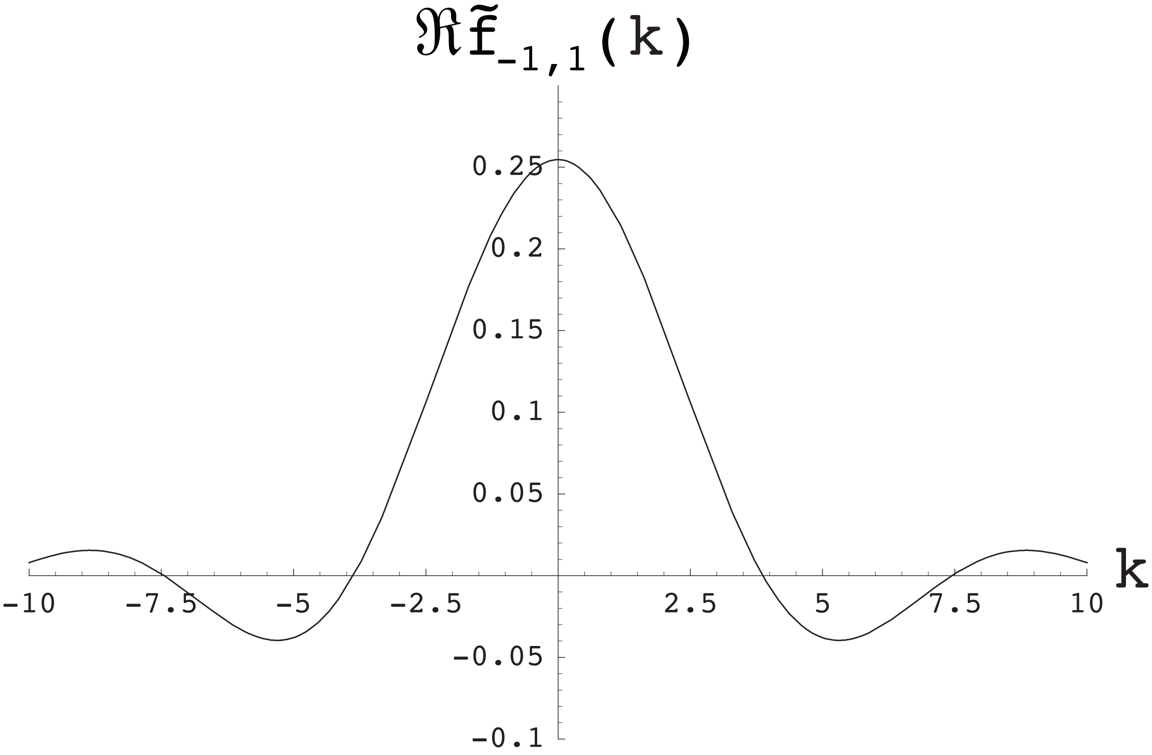


Figure 18.3.3: The function $\Re\tilde{f}_{-1,1}(k)$.

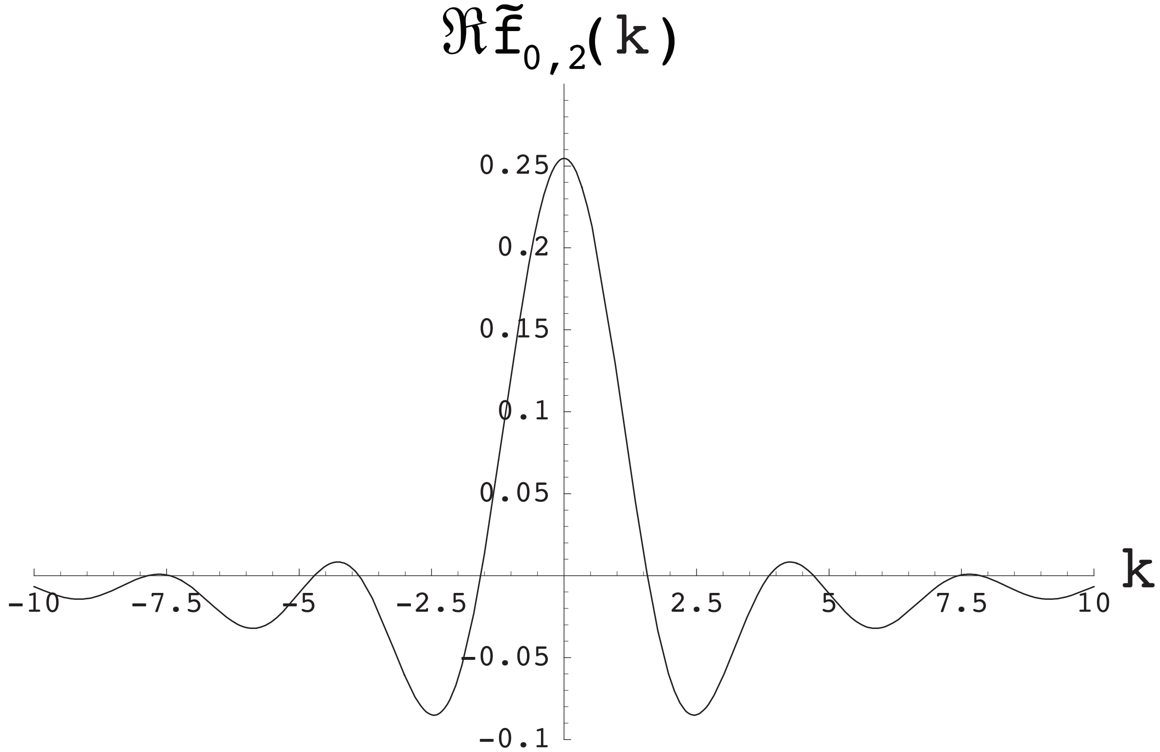


Figure 18.3.4: The function $\Re\tilde{f}_{0,2}(k)$.

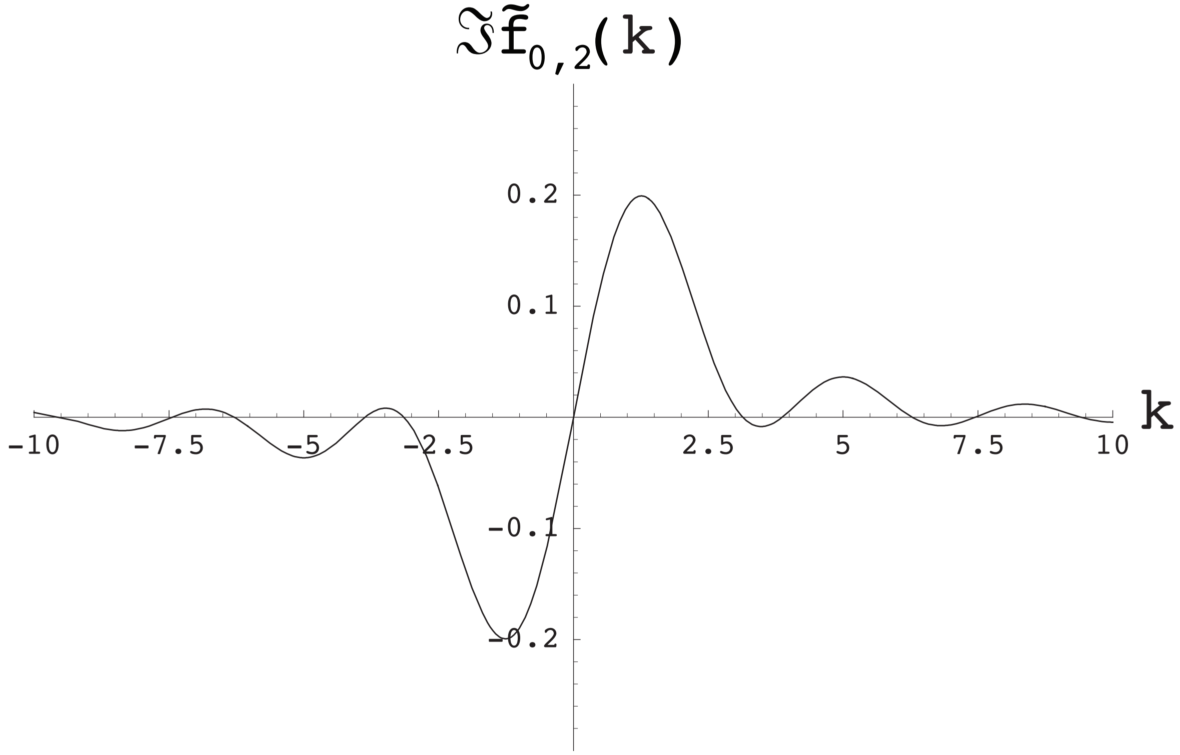


Figure 18.3.5: The function $\Im\tilde{f}_{0,2}(k)$.

18.3.3 Discrete Fourier Transform

Let us now turn to the task of evaluating $\tilde{f}$ numerically. Suppose the interval $[-Z_c, Z_c]$ is subdivided into N subintervals, each of length h, by writing

$$z_n = -Z_c + nh \text{ with } n = 0, 1, \cdots, N \tag{18.3.16}$$

where

$$h = 2Z_c/N \tag{18.3.17}$$

so that

$$z_0 = -Z_c \text{ and } z_N = Z_c. \tag{18.3.18}$$

See Figure 2.1.1 for an earlier analogous construction. Use this subdivision to approximate the integral (3.15) by the Riemann sum

$$\tilde{f}(k) \approx [1/(2\pi)]h\sum_{n=0}^{N-1} \exp(-ikz_n)f(z_n). \tag{18.3.19}$$

The quantity on the right side of (3.19) is called the *discrete* Fourier transform of f.

How accurate is the discrete Fourier transform? Let $g(z)$ be any twice differentiable function. Then, according to the *trapezoidal rule*, there is the result

$$\begin{aligned}\int_{-Z_c}^{Z_c} dz\; g(z) &= h[(1/2)g(z_0) + g(z_1) + g(z_2) + \cdots + g(z_{N-2}) + g(z_{N-1}) + (1/2)g(z_N)] \\ &- (1/6)h^2 g''(\zeta),\end{aligned} \tag{18.3.20}$$

which can be rewritten in the form

$$\begin{aligned}\int_{-Z_c}^{Z_c} dz\; g(z) &= h[g(z_0) + g(z_1) + g(z_2) + \cdots + g(z_{N-2}) + g(z_{N-1})] \\ &- (1/2)hg(z_0) + (1/2)hg(z_N) - (1/6)h^2 g''(\zeta).\end{aligned} \tag{18.3.21}$$

Here ζ is some point in the interval $[-Z_c, Z_c]$. In our case we have

$$g(z) = [1/(2\pi)]\exp(-ikz)f(z) \tag{18.3.22}$$

so that

$$g''(\zeta) = [1/(2\pi)][-k^2 f(\zeta) - 2ikf'(\zeta) + f''(\zeta)]\exp(-ik\zeta). \tag{18.3.23}$$

Thus, we have the result

$$\begin{aligned}\tilde{f}(k) &= [1/(2\pi)]h\sum_{n=0}^{N-1}\exp(-ikz_n)f(z_n) \\ &- (1/2)[1/(2\pi)]hf(-Z_c)\exp(ikZ_c) + (1/2)[1/(2\pi)]hf(Z_c)\exp(-ikZ_c) \\ &- (1/6)h^2[1/(2\pi)][-k^2 f(\zeta) - 2ikf'(\zeta) + f''(\zeta)]\exp(-ik\zeta).\end{aligned} \tag{18.3.24}$$

Upon comparing (3.19) and (3.24), we see that the discrete Fourier transform generally makes errors of order h. However, if $f(-Z_c) = 0$ and $f(Z_c) = 0$, which is often the case, then the discrete transform makes errors of order h^2. But, the errors of order h^2 can be very large if $|k|$ is large.

Let us explore the large $|k|$ behavior of the discrete Fourier transform. If we make use of (3.16), we see that (3.19) can also be written in the form

$$\tilde{f}(k) \approx [1/(2\pi)]h \exp(ikZ_c) \sum_{n=0}^{N-1} \exp(-iknh) f(z_n). \tag{18.3.25}$$

Observe that the function $\exp(-ikh)$ is *periodic* in k with a period K given by

$$K = 2\pi/h = N\pi/Z_c. \tag{18.3.26}$$

Consequently, if we define a function $F(k)$ by writing

$$F(k) = [1/(2\pi)]h \sum_{n=0}^{N-1} \exp(-iknh) f(z_n), \tag{18.3.27}$$

we have the relation

$$F(k + K) = F(k). \tag{18.3.28}$$

But (3.19) can be rewritten in terms of F. We find, using (3.25) and (3.27), the result

$$\tilde{f}(k) \approx \exp(ikZ_c)F(k). \tag{18.3.29}$$

We see that the discrete Fourier transform $\tilde{f}$ is *quasi-periodic* in k. It is a product of the function $\exp(ikZ_c)$, which has the period $2\pi/Z_c$ (so that $|\omega| = Z_c$), and the *envelope* function F which has period K. Thus, the discrete Fourier transform can never satisfy (3.5) or (3.6).[5] In general, the discrete Fourier transform is reliable only in the interval $k \in [-K/2, K/2]$. The quantity $K_{Ny} = K/2$ is called the *Nyquist critical frequency.*

As an example of the behavior of the discrete Fourier transform, consider again the function $f_{-1,1}(z)$ given by (3.9) and shown in Figure 3.1. We have already seen that it has the exact Fourier transform given by (3.11) and (3.12) and shown in Figure 3.3. Figure 3.6 shows both the exact Fourier transform, and the discrete and spline-based Fourier transforms for the case $h = .10$.[6] (The spline-based Fourier transform is discussed in a subsequent subsection.) In this case $K_{Ny} = \pi/h \simeq 31.4$. We observe that, as warned, the discrete Fourier transform results are not reliable for $|k| > K_{Ny}$. Figure 3.7 shows the difference between the exact and discrete Fourier transforms within and somewhat beyond the *Nyquist band* $|k| < K_{Ny}$. Note that, as expected from the quasi-periodicity of the discrete Fourier transform, see Figure 3.6, the error grows as k leaves the Nyquist band.

[5] Observe that the discrete Fourier transform (3.19) can be viewed as the *exact* Fourier transform of the function $h\sum_{n=0}^{N-1} f(z_n)\delta(z - z_n)$, and that this finite sum of *delta* functions must have all frequency components present.

[6] All these Fourier transforms are real because $f_{-1,1}(z)$ as given by (3.9) is an even function.

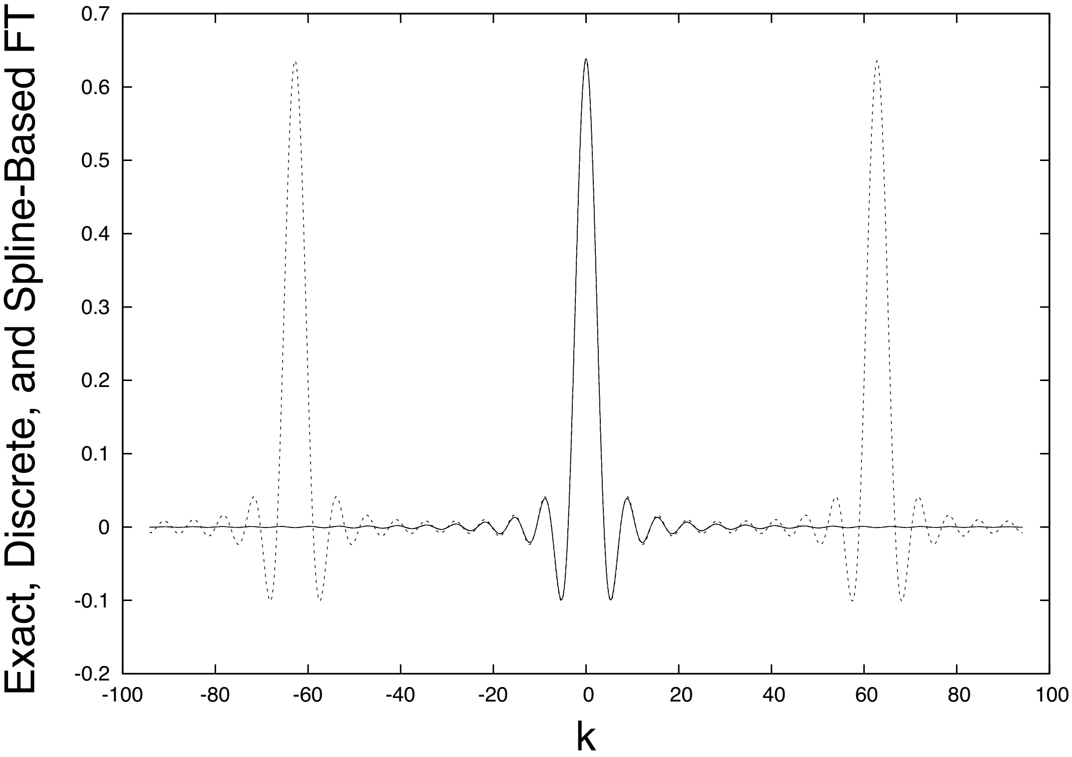


Figure 18.3.6: The exact, discrete, and spline-based Fourier transforms of $f_{-1,1}(z)$ for $h = .10$. On the scale of this figure the exact and spline-based Fourier transforms are indistinguishable. They are both shown as a solid line. The discrete Fourier transform is shown as a dashed line. Note that it is quasi-periodic while the exact and spline-based Fourier transforms fall to zero for large $|k|$.

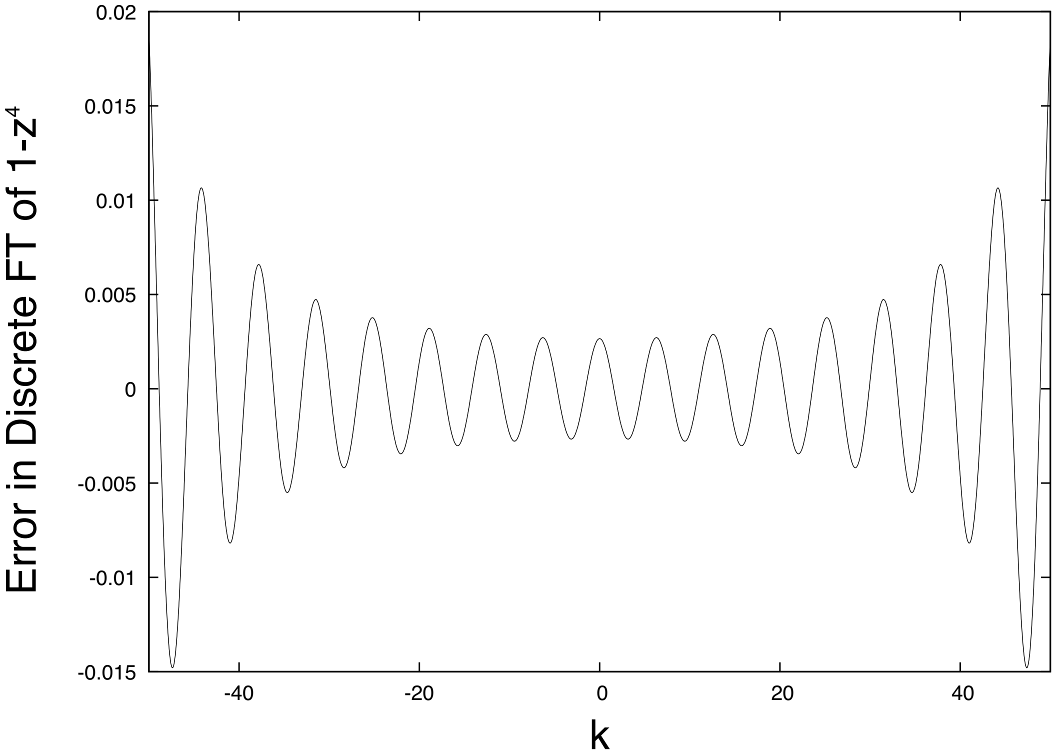


Figure 18.3.7: Difference between the exact and discrete Fourier transforms of $f_{-1,1}(z)$ for $h = .10$.

18.3.4 Discrete Inverse Fourier Transform

Consider again the inverse Fourier transform (3.8). Let us explore to what extent this inverse transform can also be made discrete and to what extent the discrete forward and inverse Fourier transforms are related. Discretize what so far has been the continuous variable k by writing

$$k_m = -K_c + mH \text{ for } m = 0, 1, \cdots, M \tag{18.3.30}$$

with

$$H = 2K_c/M \tag{18.3.31}$$

so that

$$k_0 = -K_c \text{ and } k_M = +K_c. \tag{18.3.32}$$

When this is done, the relation (3.8) can be approximated by the Riemann sum

$$f(z) \approx H \sum_{m=0}^{M-1} \exp(ik_m z)\tilde{f}(k_m). \tag{18.3.33}$$

Evidently both (3.19) and (3.33) are approximate, and presumably they become ever more accurate as $M \to \infty$ and $N \to \infty$. However, if $M = N$ and $K_c = K_{Ny}$, there are the *exact* relations

$$\tilde{f}(k_m) = [1/(2\pi)]h \sum_{n=0}^{N-1} \exp(-ik_m z_n) f(z_n), \tag{18.3.34}$$

and

$$f(z_n) = H \sum_{m=0}^{M-1} \exp(ik_m z_n)\tilde{f}(k_m). \tag{18.3.35}$$

Here (3.34) is taken to be the *definition* of the quantities $\tilde{f}(k_m)$. [They are *not* the exact values that would be found by doing the integral (3.15) exactly for the values $k = k_m$.] And when these approximate values $\tilde{f}(k_m)$ are employed in the (approximate) formula (3.35), the values $f(z_n)$ are recovered *exactly*. This a case where two wrongs do make a right![7] See Exercise 3.1.

18.3.5 Spline-Based Fourier Transforms

Let $f_{\rm sa}$ be a cubic spline approximation to the function f appearing in (3.15). Then we may make the definition

$$\tilde{f}_{\rm sa}(k) = [1/(2\pi)] \int_{-Z_c}^{Z_c} dz \exp(-ikz) f_{\rm sa}(z). \tag{18.3.36}$$

As described in Section 15.1.1, $f_{\rm sa}(z)$ can be constructed from the values $f(z_n)$ with $n \in [0, N]$. Moreover, observe that the definition (3.36) can be evaluated exactly (numerically to machine precision) for any value of k since the Fourier transforms of cubic polynomials

[7] The relation (3.35) does not give the exact values of $f(z)$ when $z \neq z_n$. Also note that (3.35) produces a function of z that is *quasi-periodic* with quasi-period $Z = 2\pi/H = 2$. By contrast, $f(z)$ is supposed to be zero for z outside the interval $[-1, 1]$.

can be found analytically and then evaluated numerically to machine precision. Thus, we have a relation of the form

$$\tilde{f}_{\rm sa}(k) = \sum_{n=0}^{N} g_n^-(k) f(z_n) \tag{18.3.37}$$

where the $g_n^-(k)$ are known functions of k that can be evaluated to machine precision.[8] Suppose we assume that $f(z)$ vanishes at the endpoints $\pm Z_c$. Then $f_{\rm sa}$ will be a differentiable function of z that vanishes at the endpoints. Therefore we expect that $\tilde{f}_{\rm sa}(k)$, unlike the $\tilde{f}$ given by (3.19), will have more nearly appropriate large $|k|$ behavior.[9] See (3.6). The accuracy of $\tilde{f}_{\rm sa}(k)$, namely the difference between $\tilde{f}_{\rm sa}(k)$ and the exact $\tilde{f}(k)$ given by (3.15), depends only on the quality of the spline fit $f_{\rm sa}(z)$, and not on any approximation to the Fourier integral.

As an example of the behavior of $\tilde{f}_{\rm sa}(k)$, consider again the $f_{-1,1}(z)$ given by (3.9). As already described earlier, Figure 3.6 shows the exact Fourier transform, the discrete Fourier transform, and $\tilde{f}_{\rm sa}$ for the case $h = .10$. [Here we have used (1.24) to estimate the end-point derivatives required to construct $f_{\rm sa}(z)$.] We have already noted that the exact and spline-based Fourier transforms appear identical on the scale shown. Figure 3.8 shows their difference. Note that their difference is small as expected from the error estimate (1.34) and the fact that the Fourier transform (3.36) of the spline approximation is evaluated exactly. In summary, as comparison of Figures 3.7 and 3.8 illustrates, for smooth functions the spline-based Fourier transform is much more accurate than the discrete Fourier transform.

What can be said about the inverse Fourier transform? Begin with (3.8). It is approximate because of the cutoff, but this cutoff K_c can in principle be made quite large to assure good accuracy. Next replace $\tilde{f}(k)$ by $\tilde{f}_{\rm sa}(k)$ using (3.36) to get the approximation

$$f(z) \approx \int_{-K_c}^{K_c} dk \exp(ikz)\tilde{f}_{\rm sa}(k). \tag{18.3.38}$$

The quality of this approximation depends on the quality of the spline fit $f_{\rm sa}(z)$. Also suppose we carry out the operation (3.36) for $M+1$ discrete values of k using the k_m values given by (3.30). That is, we compute the quantities $\tilde{f}_{\rm sa}(k_m)$ by the rule

$$\tilde{f}_{\rm sa}(k_m) = \sum_{n=0}^{N} g_n^-(k_m) f(z_n). \tag{18.3.39}$$

We may then use these quantities to try to reconstruct $f(z)$.

In particular, use the $M+1$ values $\tilde{f}_{\rm sa}(k_m)$ to construct a cubic spline approximation to $\tilde{f}_{\rm sa}(k)$ which we will call $\tilde{f}_{\rm sasa}(k)$. [Again we use (1.24), this time applied to the values $\tilde{f}_{\rm sa}(k_m)$, to estimate the required end-point derivatives.] Using this approximation in (3.38) gives the representation

$$f(z) \approx \int_{-K_c}^{K_c} dk \exp(ikz)\tilde{f}_{\rm sasa}(k). \tag{18.3.40}$$

[8] The superscript "$-$" indicates that $\exp(-ikz)$ appears in (3.36).

[9] The function $\tilde{f}_{\rm sa}(k)$ given by (3.36) cannot fall off any faster than $1/|k|^4$ at infinity because a cubic-spline has discontinuous third derivatives. See Exercise 3.2.

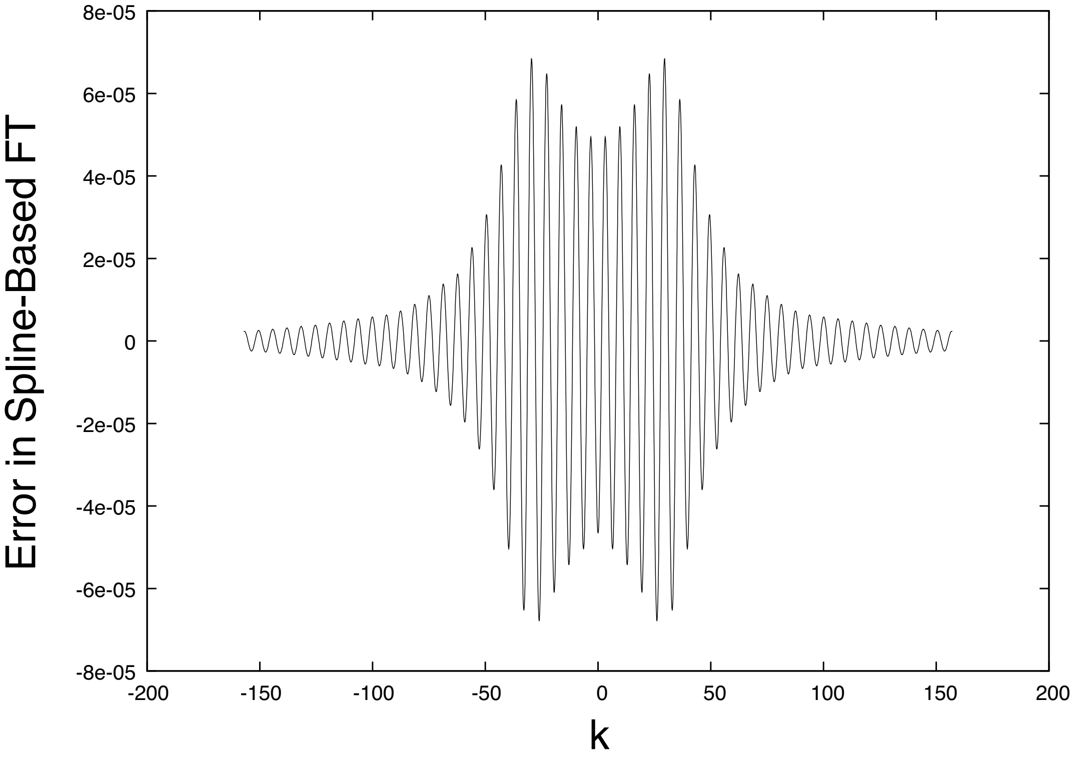


Figure 18.3.8: Difference between the exact and spline-based Fourier transforms of $f_{-1,1}(z)$ for $h = .10$.

Because $\tilde{f}_{\rm sasa}(k)$ is a cubic spline approximation, the integral (3.40) can again be done exactly. That is, there are known functions $g_m^+(z)$ such that[10]

$$f(z) \approx \int_{-K_c}^{K_c} dk \exp(ikz)\tilde{f}_{\rm sasa}(k) = \sum_{m=0}^{M} g_m^+(z)\tilde{f}_{\rm sa}(k_m). \tag{18.3.41}$$

The accuracy of this representation of f depends on how well $\tilde{f}_{\rm sasa}(k)$ approximates $\tilde{f}_{\rm sa}(k)$. In general it will not yield an f that vanishes outside $[-1, 1]$. However, unlike the f produced by (3.33), the f produced by (3.41) will be very small for z outside the interval $[-1, 1]$.

In summary, the accuracy of direct and inverse spline-based Fourier transforms is governed primarily by the quality of spline approximations, and not by how well various Fourier integrals are approximated. Put another way, the use of the discrete Fourier transform does not make any optimistic assumptions about the smoothness properties of $f(z)$. (Indeed, it assumes the worst, a sum of delta functions approximation.) By contrast, the use of spline-based Fourier transforms capitalizes on the assumption that $f(z)$ is not too badly behaved between sampling points z_n.

How well do the discrete and spline-based Fourier transforms work in reconstructing a function? Let us first consider the discrete case. As an example, will again consider the $f_{-1,1}$ given by (3.9). Figure 3.9 shows the reconstructed $f_{-1,1}(z)$ produced by (3.33) with the $\tilde{f}(k_m)$ given by (3.19) or, equivalently, (3.34). Figure 3.10 shows the difference between the exact $f_{-1,1}$ and the reconstructed $f_{-1,1}$. Here again we have used $h = .10$ so that $N = 20$; and we have set $K_c = K_{Ny}$ and $M = N$. We see from Figure 3.9 that the reconstructed $f_{-1,1}$ is quasi-periodic as expected. We see from Figure 3.10 that the error is zero at the sampling points as expected, but rises to as high as 2% elsewhere.

What happens if we instead use spline-based Fourier transforms? We have already seen that, for this example, the *forward* spline-based Fourier transform $\tilde{f}_{\rm sa}(k)$ is more accurate than the discrete Fourier transform. This is because of the high accuracy of the spline approximation $f_{\rm sa}(z)$ to $f(z)$, and the fact that the Fourier transform of the spline approximation is performed exactly. We expect to be able to carry out the *inverse* spline-based Fourier transform with good accuracy provided the spline approximation $\tilde{f}_{\rm sasa}(k)$ to $\tilde{f}_{\rm sa}(k)$ has good accuracy. But now there is a possible problem. Figure 3.11 shows the 21-point spline approximation $\tilde{f}_{\rm sasa}(k)$ over the interval $[-K_{Ny}, K_{Ny}]$ as well as $\tilde{f}_{\rm sa}(k)$ itself. We see that the 21-point spline approximation $\tilde{f}_{\rm sasa}(k)$ is not particularly good because of the oscillatory nature of $\tilde{f}_{\rm sa}(k)$. Figure 3.12 shows the exact $f_{-1,1}(z)$ and the reconstructed $f_{-1,1}(z)$ based on using $\tilde{f}_{\rm sasa}(k)$ in (3.41) with $K_c = K_{Ny}$. Evidently the agreement is not particularly good, reflecting the poor quality of the 21-point spline approximation $\tilde{f}_{\rm sasa}(k)$.

Suppose we instead make K_c somewhat larger than K_{Ny} by setting $K_c = 50$ and also use a 51-point spline approximation to $\tilde{f}_{\rm sasa}(k)$ over this interval $[-K_c, K_c]$. When this is done, it is found that the difference between $\tilde{f}_{\rm sa}(k)$ and its spline fit $\tilde{f}_{\rm sasa}(k)$ is less then 6×10^{-5}. Correspondingly we expect the reconstruction of $f_{-1,1}(z)$ to be much improved. This is indeed the case. Figure 3.13 shows the function $f_{-1,1}(z)$ and its reconstruction using, in (3.41), the 51-point spline approximation $\tilde{f}_{\rm sasa}(k)$ over the interval $k \in [-50, 50]$. The agreement is much improved, and is even good outside the interval $[-1, 1]$ where the discrete

[10] Here the superscript "+" indicates that $\exp(+ikz)$ appears in (3.40) and (3.41).

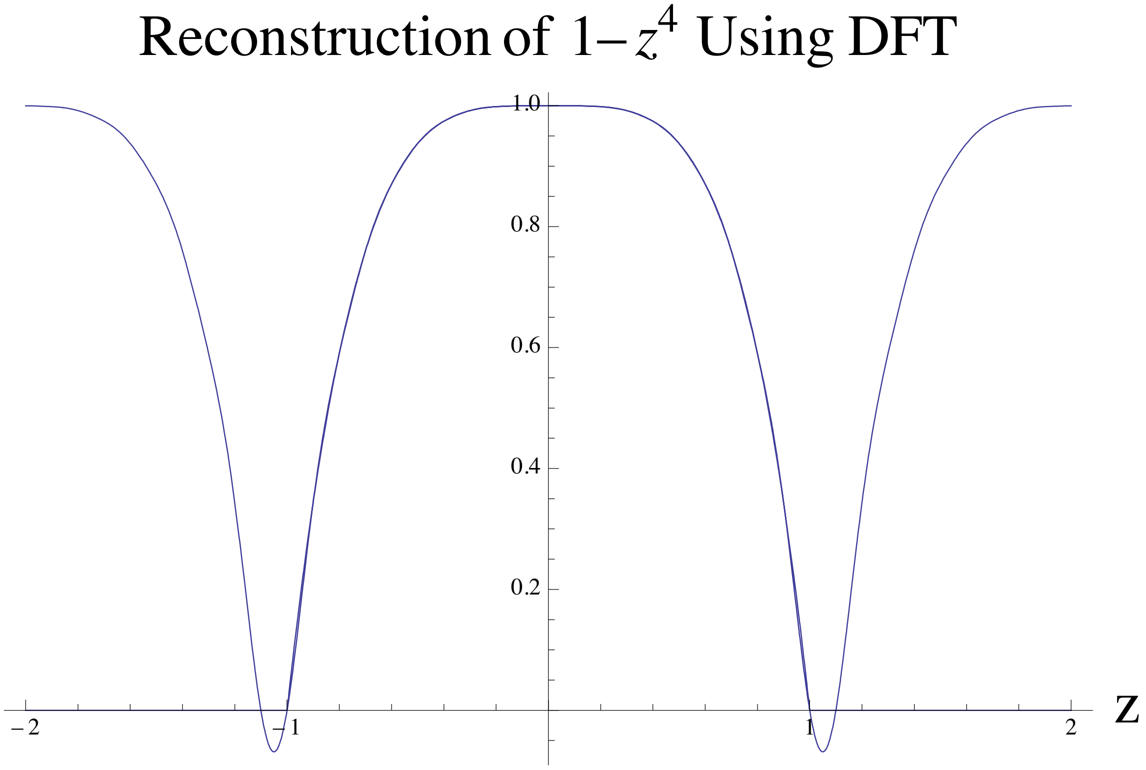


Figure 18.3.9: Reconstruction of $f_{-1,1} = 1 - z^4$ using forward and inverse discrete Fourier transforms.

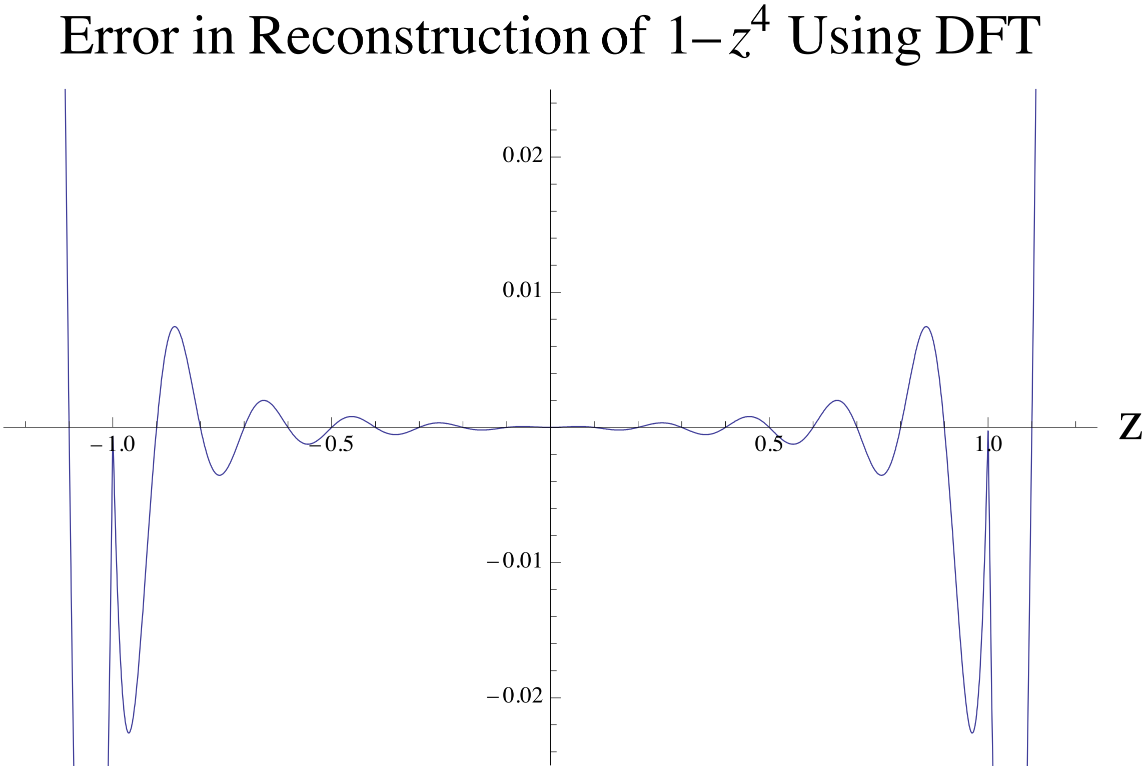


Figure 18.3.10: Error in reconstruction of $f_{-1,1} = 1 - z^4$ using forward and inverse discrete Fourier transforms.

Fourier reconstruction fails. To provide further insight into the error, Figure 3.14 displays the difference between the exact $f_{-1,1}(z)$ and the reconstructed $f_{-1,1}(z)$. Now the error is comparable to that for the discrete case, and, unlike the discrete case, is even small outside the interval $[-1, 1]$. Compare Figures 3.9, 3.10, 3.13, and 3.14. Of course, with more points the discrete-case error also decreases. But it decreases as a smaller power of h than in the spline-based case so that eventually the spline-based method wins.

Moreover, the apparent good performance of the discrete method is misleading. We already know from our previous discussion that its error when performing reconstructions must be zero at the sampling points due to the magic cancellation of errors in the forward and inverse discrete Fourier transformations at these points. But we are ultimately not interested in reconstruction. Rather, we are interested in forward Fourier transformation followed by inverse Fourier transformation with some k-dependent kernel. See (14.2.2), (14.2.6) and (14.3.1), (14.3.6) and (14.4.73), (14.4.74), (14.4.85), (14.4.86). In this context there is no reason to expect cancellation of errors when discrete Fourier transformations are employed. And, when spline-based Fourier transformations are employed, we may expect to see errors that are no worse then those encountered in the case of reconstruction. We conclude that as long as the spline approximations are done with care, the spline-based Fourier transforms should be superior to discrete Fourier transforms.

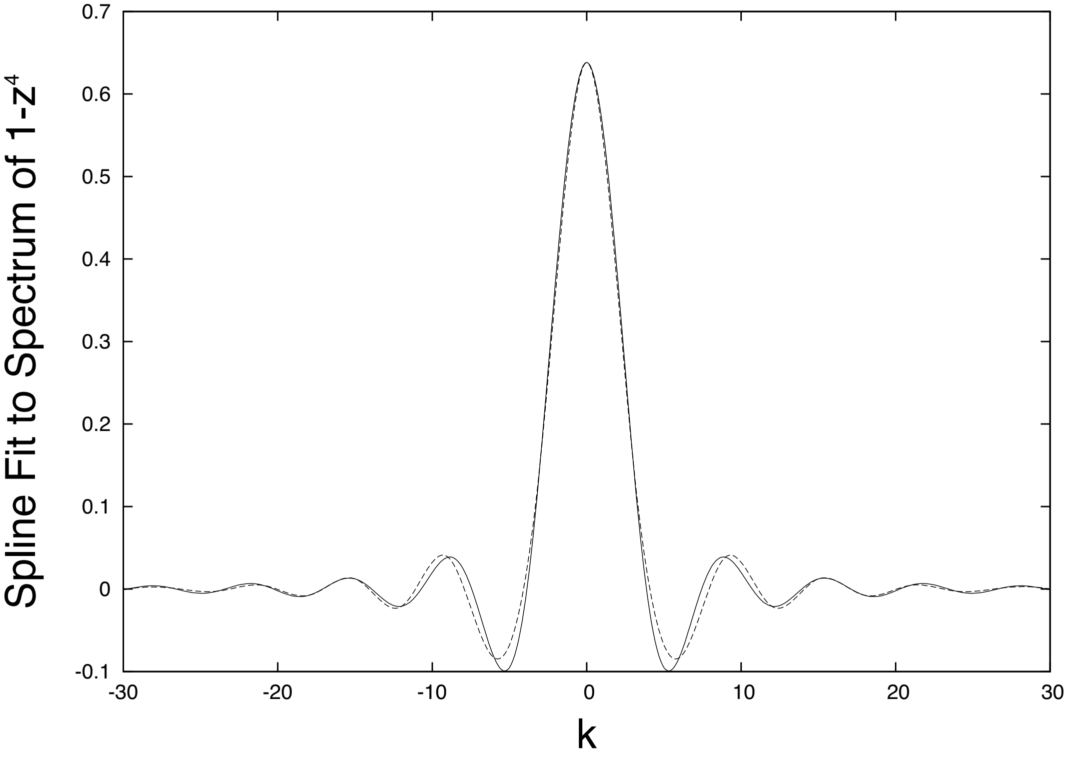


Figure 18.3.11: The function $\tilde{f}_{\text{sa}}(k)$ (solid line) and its 21-point spline approximation $\tilde{f}_{\text{sasa}}(k)$ (dashed line) over the Nyquist band $k \in [-K_{Ny}, K_{Ny}]$.

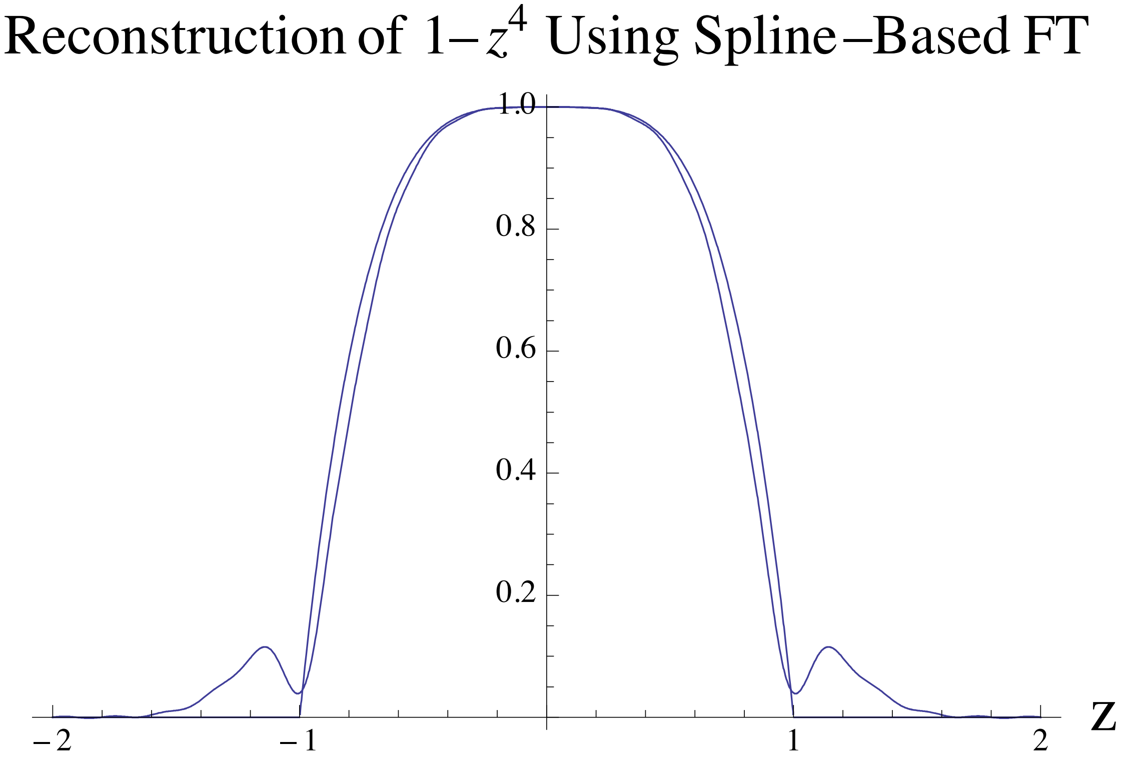


Figure 18.3.12: The function $f_{-1,1}(z) = 1 - z^4$ and its reconstruction using the 21-point spline approximation $\tilde{f}_{\text{sasa}}(k)$ in (3.41).

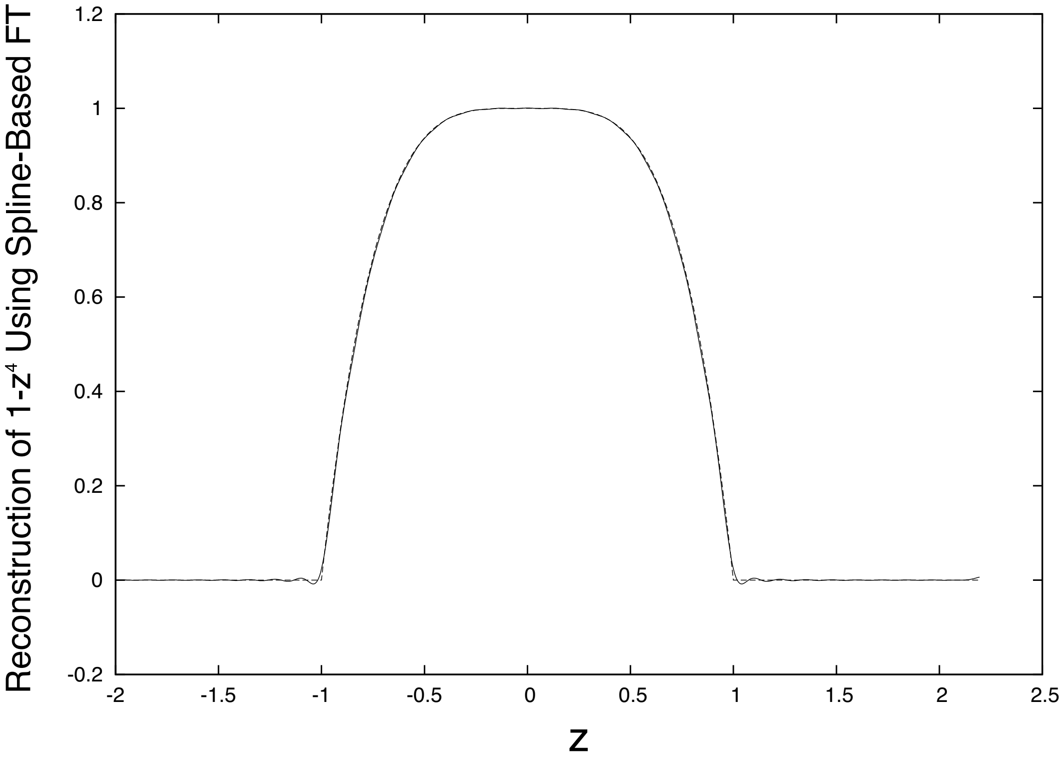


Figure 18.3.13: The function $f_{-1,1}(z) = 1 - z^4$ and its reconstruction using, in (3.41), the 51-point spline approximation $\tilde{f}_{\rm sasa}(k)$ over the interval $k \in [-50, 50]$.

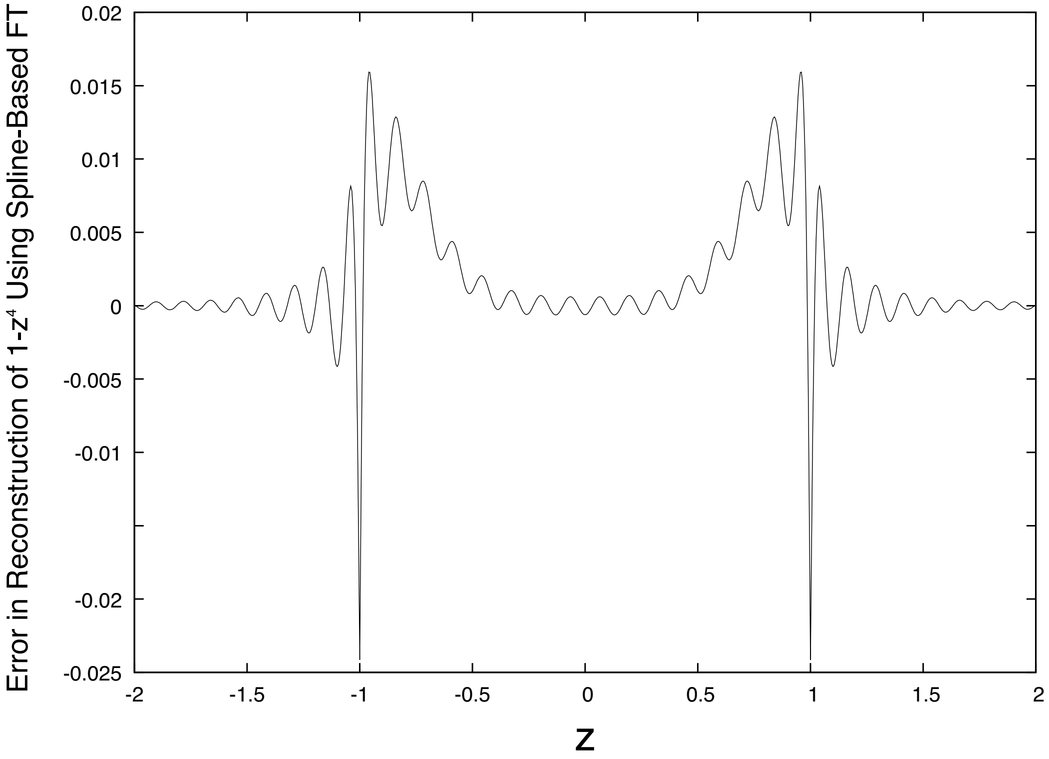


Figure 18.3.14: The difference between the exact function $f_{-1,1}(z) = 1 - z^4$ and its reconstruction using, in (3.41), the 51-point spline approximation $\tilde{f}_{\text{sasa}}(k)$ over the interval $k \in [-50, 50]$.

18.3.6 Fast Spline-Based Fourier Transforms

How much work is involved in computing Fourier transforms? Let M_{mn} be the matrix with entries

$$M_{mn} = [1/(2\pi)]h\exp(-ik_m z_n). \tag{18.3.42}$$

With this notation, the discrete Fourier transformation relation (3.34) can be written in the vector/matrix form

$$\tilde{f}(k_m) = \sum_{n=0}^{N-1} M_{mn} f(z_n). \tag{18.3.43}$$

Suppose we view the matrix M as being *precomputed*, so that we do not count its evaluation toward the work involved, but we do imagine carrying out (3.43) for a collection of k values. Then the major work involved will consist of N^2 multiplications because there are N multiplications in (3.43) for each value of m and there are N such values. Of course, additions are also involved, but they are much less expensive in machine time than multiplications, and so they will be ignored. Thus, it would seem that the work involved in computing discrete Fourier transforms scales as N^2.

One of the celebrated realizations of 20th century computational science is that there are certain *favored* values of N for which there is a *fast* Fourier transform (FFT) algorithm such that the work scales only as $N\log_2 N$ rather than N^2.[11] Although there are other possibilities, these favored values are most commonly taken to be integer powers of 2, $N = 2^\ell$ for some integer ℓ. The use of such N values does not cause any great complication because additional equally-spaced points z_n can added at both ends of the general interval $[a, b]$, and f can be assigned the value 0 at these additional points. (This is called *padding with zeros.*) So doing does not affect the values of the discrete Fourier transform, nor does it affect K_{Ny}. It does affect, however, the location of the sampling points k_m in k space. [See (3.30) with $M = N$ and $K_c = K_{Ny}$.] In fact, it makes the sampling points more finely spaced, which can be viewed as a virtue.

What can be said about the work involved in computing spline-based Fourier transforms? Examination of the logic involved in the construction of discrete FFT algorithms shows that exactly the same considerations apply to spline-based Fourier transforms. Therefore, for every favored value of N, there is also a fast spline-based Fourier transform algorithm for which the work also scales as $N\log_2 N$. Consequently, there is no computational penalty involved in the use of spline-based Fourier transforms.

Exercises

18.3.1. Verify ....

18.3.2. Verify ....

[11] We say *realization* rather than *discovery* because, after the extensive FFT work of Danielson, Lanczos, Cooley, Tukey, and others in the mid 20th century, it became clear in retrospect that their remarkable accomplishments had been anticipated earlier by others including Gauss in 1805.

18.4 Bessel Functions

Look again at Section 14.2. The computation of generalized gradients in terms of potential data on the surface of a circular cylinder required the calculation of Bessel functions $\mathrm{I_m}$, and the equivalent computation in Section 14.3 based on field data required a knowledge of I'_m. For small values of the argument, say $w \leq 1$, these calculations can done using the series (13.2.15). For larger argument values, and in view of the fact that the Bessel functions are needed for many equally-spaced argument values, it is convenient to to compute Bessel functions by integrating the differential equation for $\mathrm{I_m}$ numerically (using the methods of Chapter 2) in the form

$$\mathrm{I}''_\mathrm{m}(w) + (1/w)\mathrm{I}'_\mathrm{m}(w) - (m^2/w^2)\mathrm{I_m}(w) - \mathrm{I_m}(w) = 0. \tag{18.4.1}$$

Here, as initial conditions, we use the series (13.2.15) to evaluate $\mathrm{I_m}$ and I'_m for $w = 1$.

We remark that many of the standard familiar transcendental functions satisfy or are defined by differential equations with good analytic properties. If their values are required at many equally spaced points, these values can often be conveniently and reliably obtained by numerical integration of these differential equations. We will use this method in the next section for Mathieu functions.

18.5 Mathieu Functions

This subsection describes briefly tools needed for the computation of Mathieu functions. In 1914 Whittaker remarked about Mathieu functions that

> their actual analytical determination presents great difficulties.

He could have said the same thing about their numerical computation. Nearly 100 years later there still do not seem to be any open-source algorithms that are fully robust over the required range of the parameter q.

18.5.1 Calculation of Separation Constants $a_n(q)$ and $b_n(q)$

There is a fast algorithm, based on the use of continued fractions, for the computation of the separation constants $a_n(q)$ and $b_n(q)$. There are also algorithms based on matrix diagonalization. Unfortunately, no routines have yet been found that are completely robust. Samples of the existing routines are listed in Appendix M along with a description of their performance. Although much work has been done on this subject by many authors, there is yet more to be done.

18.5.2 Calculation of Mathieu Functions

The calculation of the Mathieu functions themselves is also a delicate matter. As an indication of the difficulty of computing Mathieu functions reliably, we have found that *Mathematica*, useful as it is, also does not compute them accurately for some values of q. However, for our purposes, since we need them only for a relatively few equally-spaced values of the arguments u and v, they can be obtained (with some care and recognition of their symmetry properties) by direct numerical integration (using the methods of Chapter 2) of their defining differential equations. Thus, for us, the major problem is accurate computation of the separation constants.

Specifically, for the $\mathrm{ce}_n(v,q)$, we integrate the equation (14.4.22) over the interval $[0,2\pi]$ with the initial conditions

$$Q_n^c(0,q)=1, \tag{18.5.1}$$

$$Q_n^{c\,\prime}(0,q)=0. \tag{18.5.2}$$

Here, and in what follows, a $\prime$ denotes differentiation with respect to v. Moreover, the notation $Q_n^c(v,q)$ indicates that q is to be computed using (14.4.23), that this resulting q value is next used to compute $a=a_n(q)$, and that these values of q and a are then used in (14.4.22). Simultaneously, we integrate the first-order differential equation

$$N_c'(v)=[Q_n^c(v,q)]^2, \tag{18.5.3}$$

again over the interval $[0,2\pi]$, with the initial condition

$$N_c(0)=0. \tag{18.5.4}$$

Finally, we find $\mathrm{ce}_n(v,q)$ from the relation

$$\mathrm{ce}_n(v,q)=[\sqrt{\pi/N_c(2\pi)}]Q_n^c(v,q). \tag{18.5.5}$$

In this way we generate a solution of (14.4.22) that is even in v, and also satisfies the normalization requirement (14.4.38) for $m=n$. Finally, we check numerically the periodicity requirements

$$Q_n^c(2\pi,q)=1,$$
$$Q_n^{c\,\prime}(2\pi,q)=0. \tag{18.5.6}$$

We require that the relations (5.6) are always satisfied to high precision. So doing provides a check on both the accuracy of the $a_n(q)$ and the numerical integration procedure. We remark that because of the symmetry conditions described at the end of Section 14.4.4, it is really only necessary to integrate over the interval $[0,\pi/2]$ and then verify that (4.54) or (4.57) are satisfied. Moreover, if the $a_n(q)$ are known to be accurate and there are strongly forbidden regions in v, it is only necessary to integrate over the still smaller interval $[0,v_{\mathrm{deep}}]$.

The calculation of the $\mathrm{se}_n(v,q)$ is done in a similar way. Now we set $a=b_n(q)$ and integrate (14.4.22) with the initial conditions

$$Q_n^s(0,q)=0, \tag{18.5.7}$$

$$Q_n^{s\,\prime}(0,q)=1. \tag{18.5.8}$$

At the same time we again integrate the differential equation

$$N_s'(v)=[Q_n^s(v,q)]^2 \tag{18.5.9}$$

with the initial condition

$$N_s(0)=0. \tag{18.5.10}$$

Then we define $\mathrm{se}_n(v,q)$ by the relation

$$\mathrm{se}_n(v,q) = [\sqrt{\pi/N_s(2\pi)}]Q_n^s(v,q). \tag{18.5.11}$$

Finally, we check numerically the periodicity requirements that now

$$Q_n^s(2\pi,q) = 0,$$

$$Q_n^s{}'(2\pi,q) = 1. \tag{18.5.12}$$

We require that (5.12) be satisfied to high precision thereby providing a check on the accuracy of both the $b_n(q)$ and, as before, the numerical integrator. Again, using symmetry it is really only necessary to integrate over the interval $[0,\pi/2]$. And if the $b_n(q)$ are known to be accurate and there are strongly forbidden regions in v, it is only necessary to integrate over the still smaller interval $[0,v_{\text{deep}}]$.

We still have to describe the computation of $\mathrm{Ce}_n(u,q)$ and $\mathrm{Se}_n(u,q)$. Now we will integrate (14.4.21) numerically. For the case of $\mathrm{Ce}_n(u,q)$ we find from (14.4.56) the initial condition

$$\mathrm{Ce}_n(0,q) = \mathrm{ce}_n(0,q) = \sqrt{\pi/N_c(2\pi)}. \tag{18.5.13}$$

Here we have also used (5.1) and (5.5). And, since $\mathrm{Ce}_n(u,q)$ is even in u, we have the second initial condition

$$\mathrm{Ce}_n'(0,q) = 0. \tag{18.5.14}$$

Thus, given k, we find q and $a = a_n(q)$. Then, having selected U, we integrate (14.4.21) over the interval $u \in [0,U]$ with the initial conditions

$$P_n^c(0,q) = \sqrt{\pi/N_c(2\pi)} \tag{18.5.15}$$

and

$$P_n^c{}'(0,q) = 0. \tag{18.5.16}$$

The result of this process is the value $\mathrm{Ce}_n(U,q) = P_n^c(U,q)$.

The computation of $\mathrm{Se}_n(u,q)$ proceeds similarly. From (14.4.57) we see that there are the initial conditions

$$\mathrm{Se}_n(0,q) = -i\mathrm{se}_n(0,q) = 0, \tag{18.5.17}$$

$$\mathrm{Se}_n'(u,q)|_{u=0} = -i\mathrm{se}_n'(iu,q)|_{u=0}(i) = \mathrm{se}_n'(0,q) = \sqrt{\pi/N_s(2\pi)}. \tag{18.5.18}$$

Here we have used (5.8) and (5.11). Thus, given k, we find q and $a = b_n(q)$. Then, having selected U, we integrate (14.4.21) over the interval $u \in [0,U]$ with the initial conditions

$$P_n^s(0,q) = 0 \tag{18.5.19}$$

and

$$P_n^s{}'(0,q) = \sqrt{\pi/N_s(2\pi)}. \tag{18.5.20}$$

The result of this process is the value $\mathrm{Se}_n(U,q) = P_n^s(U,q)$.

18.5.3 Calculation of Fourier and Mathieu-Bessel Connection Coefficients

The functions $\mathrm{ce}_r(v,q)$ and $\mathrm{se}_r(v,q)$ are periodic with period 2π and therefore have Fourier expansions in terms of the functions $\cos(mv)$ and $\sin(mv)$. See (14.4.52) through (14.4.55). As shown in Appendix N, the Fourier coefficients that appear in these expansions, which depend on q, are key to computing the Mathieu-Bessel connection coefficients $\alpha^r_m(k)$ and $\beta^r_m(k)$. See (14.4.78) and (14.4.79). In this subsection we describe how these Fourier coefficients can be computed numerically.

Let us begin with the functions $\mathrm{ce}_r(v,q)$. Since they are periodic and even, they have Fourier expansions of the form

$$\mathrm{ce}_r(v,q) = \sum_{m=0}^{\infty} A^r_m(q)\cos(mv). \tag{18.5.21}$$

There are known algorithms for the computation of the Fourier coefficients $A^r_m(q)$ but, as was the case with those for the computation of the $\mathrm{ce}_r(v,q)$, we have found that they are not robust. However, by the orthogonality property of the trigonometric functions, it follows that

$$A^r_0(q) = [1/(2\pi)]\int_0^{2\pi} dv\ \mathrm{ce}_r(v,q),$$

$$A^r_m(q) = (1/\pi)\int_0^{2\pi} dv\ \mathrm{ce}_r(v,q)\cos(mv) \text{ for } m \ge 1. \tag{18.5.22}$$

By (5.5) we may also write

$$A^r_0(q) = [1/(2\pi)][\sqrt{\pi/N_c(2\pi)}]\int_0^{2\pi} dv\ Q^c_r(v,q),$$

$$A^r_m(q) = (1/\pi)[\sqrt{\pi/N_c(2\pi)}]\int_0^{2\pi} dv\ Q^c_r(v,q)\cos(mv) \text{ for } m \ge 1. \tag{18.5.23}$$

Let $\hat{A}^r_m(v,q)$ be the functions defined for various values of m and r by the differential equations

$$\hat{A}^{r\,\prime}_0(v,q) = (1/2)Q^c_r(v,q),$$

$$\hat{A}^{r\,\prime}_m(v,q) = Q^c_r(v,q)\cos(mv) \text{ for } m \ge 1 \tag{18.5.24}$$

with the common initial conditions

$$\hat{A}^r_m(0,q) = 0. \tag{18.5.25}$$

The differential equations (5.24) can be integrated numerically over the interval $[0,2\pi]$ simultaneously with those for the $Q^c_r(v,q)$ and (5.3). Then we find that the $A^r_m(q)$ are given by the relations

$$A^r_m(q) = (1/\pi)[\sqrt{\pi/N_c(2\pi)}]\hat{A}^r_m(2\pi,q). \tag{18.5.26}$$

Similarly, for the functions $\mathrm{se}_r(v,q)$, there are Fourier expansions of the form

$$\mathrm{se}_r(v,q)=\sum_{m=1}^{\infty} B_m^r(q)\sin(mv). \tag{18.5.27}$$

And, again by the orthogonality property of the trigonometric functions, it follows that

$$B_m^r(q)=(1/\pi)\int_0^{2\pi} dv\ \mathrm{se}_r(v,q)\sin(mv). \tag{18.5.28}$$

With the aid of (5.11) this relation can also be written in the form

$$B_m^r(q)=(1/\pi)[\sqrt{\pi/N_s(2\pi)}]\int_0^{2\pi} dv\ Q_r^s(v,q)\sin(mv). \tag{18.5.29}$$

Now let $\hat{B}_m^r(v,q)$ be the functions defined for various values of m and r by the differential equations

$$\hat{B}_m^{r\,\prime}(v,q)=Q_r^s(v,q)\sin(mv) \tag{18.5.30}$$

with the common initial conditions

$$\hat{B}_m^r(0,q)=0. \tag{18.5.31}$$

Now we find that the $B_m^r(q)$ are given by the relations

$$B_m^r(q)=(1/\pi)[\sqrt{\pi/N_s(2\pi)}]\hat{B}_m^r(2\pi,q). \tag{18.5.32}$$

At this point we remark that the functions $\cos(mv)$ and $\sin(mv)$ required to integrate the differential equations (5.24) and (5.30), as well as the equations of the form (14.4.22) for the $Q_r^c(v,q)$ and the $Q_r^s(v,q)$, can also be computed on the fly by simultaneously integrating numerically the differential equations for the trigonometric functions. The needed hyperbolic functions in (14.4.21) can be calculated analogously. So doing is faster than using the built-in Fortran or C functions for the trigonometric functions.

Bibliography

Splines

[1] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Third Edition, Springer (2002).

[2] F.B. Hildebrand, *Introduction to Numerical Analysis*, Second Edition, Dover (1987).

[3] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in Fortran 77*, Second Edition, Cambridge (2003).

Discrete Fourier Transforms

[4] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in Fortran 77*, Second Edition, Cambridge (2003).

[5] C.F. Van Loan, *Computational Frameworks for the Fast Fourier Transform*, SIAM (1992).

[6] A. Zonst, *Understanding the FFT*, Citrus Press (1995).

[7] R.E. Crandall, *Projects in Scientific Computation*, Springer-Verlag (2000).

[8] R.E. Crandall, *Topics in Advanced Scientific Computation*, Springer-Telos (1996).

[9] H. Joseph Weaver, *Theory of Discrete and Continuous Fourier Analysis*, John Wiley (1989).

[10] E. Stein and R. Shakarchi, *Fourier Analysis: An Introduction*, Princeton University Press (2003).

[11] B. P. Lathi, *Signal Processing & Linear Systems*, Berkeley-Cambridge Press (1998). See also `http://web.itu.edu.tr/hulyayalcin/Signal_Processing_Books/Lathi_Signal_Processing_and_Linear_Systems.pdf`.

Chapter 19

Numerical Benchmarks

How well do the surface methods of Chapter 14 work? To test them we need a problem that is sufficiently complex and challenging to fully exercise the numerical algorithms while at the same time being exactly soluble. In that way we will be able to gauge the accuracy of the numerical methods by comparing numerical results with exact analytic results. Such a test problem is that of the magnetic monopole doublet described in Section 13.7. For this problem we will assume that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2. See Section 13.7.1.

19.1 Circular Cylinder Numerical Results for Monopole Doublet

In this section we will apply the circular cylinder numerical method of Section 14.3 to the monopole doublet problem to investigate how accurately this method is able to reproduce the exact analytic results for the on-axis gradients found in Section 13.7.2.[1] For our benchmark calculation we will employ a cylinder with radius $R = 2$ cm. See Figure 13.7.1. We will work up to the desired numerical comparison by stages. In this way we will be able to judge the accuracy of various intermediate steps.

Observe that the integrands of (14.3.8) and (14.3.23), apart from multiplicative constants, consist of the *product* of a *kernel* $[k^{n+|m|-1}/I'_m(kR)]$ and the *Fourier coefficients* $[\tilde{\tilde{B}}_\rho(R, m', k')]$ or $[\tilde{\tilde{B}}^\alpha_\rho(R, m', k')]$. The kernels are *universal* (the same for all problems) and the Fourier coefficients are specific to each problem. In what follows we will be examining both. For convenience, we will use the Fourier coefficients $[\tilde{\tilde{B}}^\alpha_\rho(R, m', k')]$.

19.1.1 Testing the Spline-Based Inverse ($k \rightarrow z$) Fourier Transform

Suppose we know exactly the Fourier coefficients $\tilde{\tilde{B}}^\alpha_\rho(R, m', k')$ as given by (14.3.14). We will see in the next two paragraphs that, for the case of the monopole doublet, the $\tilde{\tilde{B}}^\alpha_\rho(R, m', k')$ can indeed be found exactly. We can insert these exact quantities into (14.3.23), and then

[1] A similar study could be made of the accuracy of the related method of Section 14.2, with analogous results.

perform the required integration numerically using the methods described in Section 15.3.5. In this way we will be able to test the accuracy of the numerical routines for $I_m'(w)$ and for spline-based inverse Fourier transforms. This test will be performed in later paragraphs.

Exact Fourier coefficients

To find the $\tilde{\tilde{B}}_\rho^\alpha(R, m', k')$ exactly, suppose the $C_{m,\alpha}^{[0]}(z)$ have the Fourier representation

$$C_{m,\alpha}^{[0]}(z) = \int_{-\infty}^{\infty} dk\ \tilde{C}_{m,\alpha}^{[0]}(k) \exp(ikz). \tag{19.1.1}$$

From (14.3.23) evaluated with $n = 0$ we have the result

$$C_{m,\alpha}^{[0]}(z) = (1/2)^m (1/m!) \int_{-\infty}^{\infty} dk [k^{m-1}/I_m'(kR)] \tilde{\tilde{B}}_\rho^\alpha(R, m, k) \exp(ikz). \tag{19.1.2}$$

It follows from the uniqueness of the Fourier representation that there are the relations

$$\tilde{C}_{m,\alpha}^{[0]}(k) = (1/2)^m (1/m!) [k^{m-1}/I_m'(kR)] \tilde{\tilde{B}}_\rho^\alpha(R, m, k), \tag{19.1.3}$$

which can be solved for the $\tilde{\tilde{B}}_\rho^\alpha(R, m, k)$ to give the relations

$$\tilde{\tilde{B}}_\rho^\alpha(R, m, k) = 2^m (m!) [I_m'(kR)/(k)^{m-1}] \tilde{C}_{m,\alpha}^{[0]}(k). \tag{19.1.4}$$

Of course, in view of (1.1), the $\tilde{C}_{m,\alpha}^{[0]}(k)$ are also given by the inverse Fourier transform

$$\tilde{C}_{m,\alpha}^{[0]}(k) = [1/(2\pi)] \int_{-\infty}^{\infty} dz\ C_{m,\alpha}^{[0]}(z) \exp(-ikz). \tag{19.1.5}$$

Therefore, from (13.7.9) and (13.7.33), we conclude that for the monopole doublet there are the results

$$\tilde{C}_{m=0}^{[0]}(k) = 0, \tag{19.1.6}$$

$$\tilde{C}_{m,c}^{[0]}(k) = 0, \tag{19.1.7}$$

$$\tilde{C}_{m,s}^{[0]}(k) = 0 \text{ for } m \text{ even}, \tag{19.1.8}$$

and the only nonzero integrals on the right side of (1.5) are of the form

$$\int_{-\infty}^{\infty} dz\ C_{m,s}^{[0]}(z) \exp(-ikz) \text{ with } m \text{ odd}. \tag{19.1.9}$$

From (13.7.33) we see that we need the integrals

$$\int_{-\infty}^{\infty} dz\ \exp(-ikz) \beta^{2m+1}(z) \text{ for } m \text{ odd}. \tag{19.1.10}$$

These integrals can be done analytically, and have the values

$$\int_{-\infty}^{\infty} dz\ \exp(-ikz) \beta^{2m+1}(z) = \{2/[1 \cdot 3 \cdot 5 \cdot 7 \cdots (2m-1)]\} a^{m+1} |k|^m K_m(a|k|). \tag{19.1.11}$$

Putting everything together gives the final result that the only nonzero $\tilde{\tilde{B}}_\rho^\alpha(R,m,k)$ are those given by the relations

$$\tilde{\tilde{B}}_\rho^s(R,m,k) = (-4g/\pi)(-1)^{(m-1)/2}|k|I'_m(kR)K_m(a|k|) \text{ for } m \text{ odd.} \tag{19.1.12}$$

We observe that the $\tilde{\tilde{B}}_\rho^s(R,m,k)$ for the monopole doublet are pure real and are even functions of k. The imaginary part vanishes because, for the monopole doublet, $B_\rho(R,\phi,z)$ is an even function of z. See (13.7.8) and (14.3.14) through (14.3.16).

As examples of the behavior of the $\tilde{\tilde{B}}_\rho^s(R,m,k)$, Figures 1.1 and 1.2 show the functions $\Re\tilde{\tilde{B}}_\rho^s(R,1,k)$ and $\Re\tilde{\tilde{B}}_\rho^s(R,7,k)$. As already described, $I'_m(kR)$ grows exponentially at infinity. See (14.3.7). By contrast, $K_m(a|k|)$ decays exponentially to zero at infinity,

$$|K_m(a|k|)| \sim \exp(-|k|a)(\pi)^{1/2}/\sqrt{2|k|a} \text{ as } |k| \to \infty. \tag{19.1.13}$$

Since $a > R$, it follows that the $\tilde{\tilde{B}}_\rho^s(R,m,k)$ are exponentially damped at infinity,

$$|\tilde{\tilde{B}}_\rho^s(R,m,k)| \sim \exp[-|k|(a-R)] \text{ as } |k| \to \infty. \tag{19.1.14}$$

The function $I'_m(kR)$ is an entire function of k. The function $K_m(a|k|)$ is singular at the origin. Analysis reveals that the product $|k|I'_m(kR)K_m(a|k|)$ is finite at the origin and has additionally a logarithmic singularity at the origin of the form $|k|^{2m}\log|k|$.[2] Thus the $\tilde{\tilde{B}}_\rho^s(R,m,k)$, with $m \geq 1$, are finite for all k and vanish exponentially at infinity.

Kernels

Since, as emphasized earlier, the integrand in (14.3.23) is the product of a kernel and a Fourier coefficient, we should also examine the kernels. Figures 1.3 through 1.5 display the kernels $[k^{n+m-1}/I'_m(kR)]$ for the representative cases (m,n)=(1,0), (1,6), and (7,0) when $R = 2$ cm. We see, as expected, that they fall off rapidly for large $|k|$. The intermediate cases give analogous results.

Exact Integrands

We are now ready to evaluate the integrals (14.3.23) to find the $C_{m,s}^{[n]}(z)$. Upon inserting (1.12) into (14.3.23), we find the results

$$C_{m,s}^{[n]}(z) = \left\{\frac{-4g(-1)^{(m-1)/2}}{\pi 2^m(m!)}\right\}\int_{-\infty}^{\infty} dk\ \exp(ikz)(ik)^n|k|^m K_m(|k|a). \tag{19.1.15}$$

Or, more directly and as an algebraic check, we may use (1.5) to yield the result

$$\begin{aligned}\tilde{C}_{m,s}^{[0]}(k) &= [1/2\pi]\int_{-\infty}^{\infty} dz\ C_{m,s}^{[0]}(z)\exp(-ikz) \\ &= \{-4g/[\pi 2^m(m!)]\}(-1)^{(m-1)/2}|k|^m K_m(|k|a).\end{aligned} \tag{19.1.16}$$

[2] It can be shown that this singularity in $\tilde{\tilde{B}}_\rho^s(R,m,k)$ is related to the $\sim 1/|z|^3$ behavior of $B_\rho(R,\phi,z)$ and the $\sim 1/|z|^{2m+1}$ behavior of the $C_{m,s}^{[0]}(z)$ for large $|z|$. See (13.7.8) and (13.7.34). In general, the faster the falloff at $|z| = \infty$, the milder the singularity at $k = 0$.

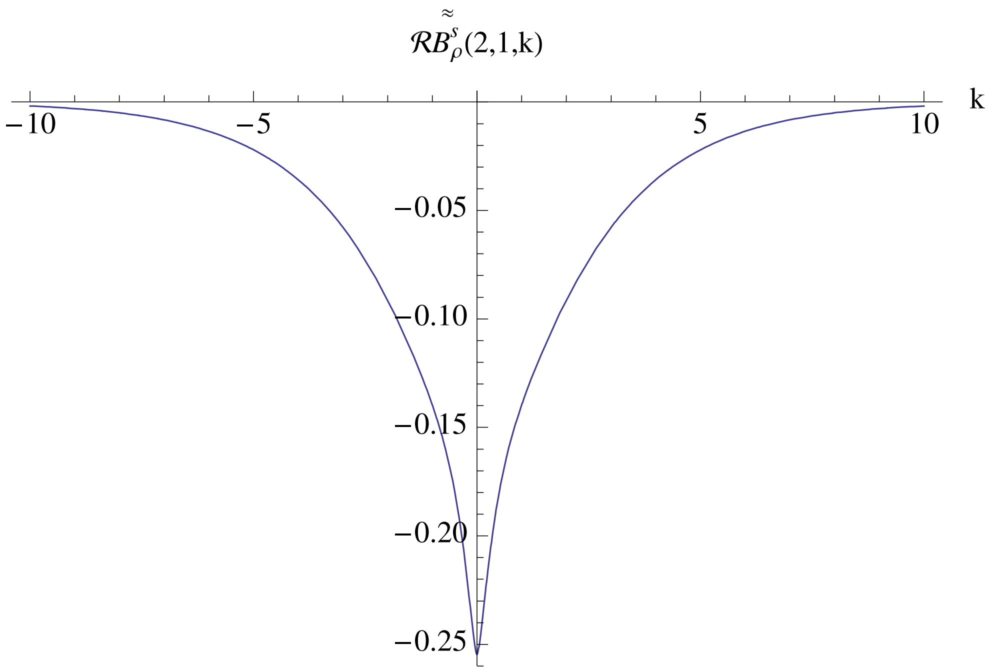


Figure 19.1.1: The real part of $\tilde{\tilde{B}}^s_\rho(R,1,k)$ as a function of k for the monopole doublet in the case that $R = 2$ cm and $a = 2.5$ cm. The imaginary part vanishes.

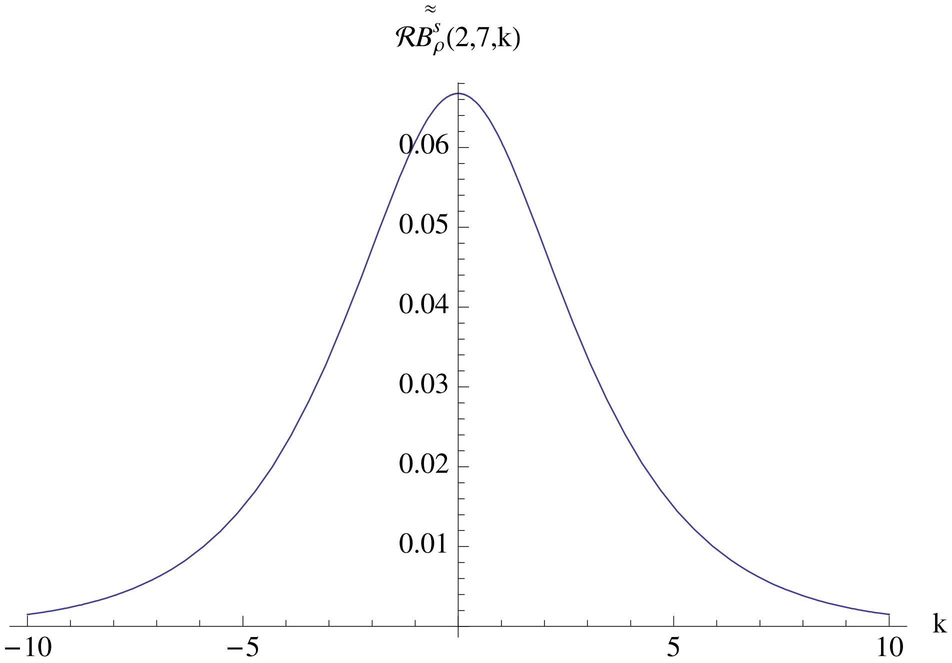


Figure 19.1.2: The real part of $\Re\tilde{\tilde{B}}^s_\rho(R,7,k)$ as a function of k for the monopole doublet in the case that $R = 2$ cm and $a = 2.5$ cm. The imaginary part vanishes.

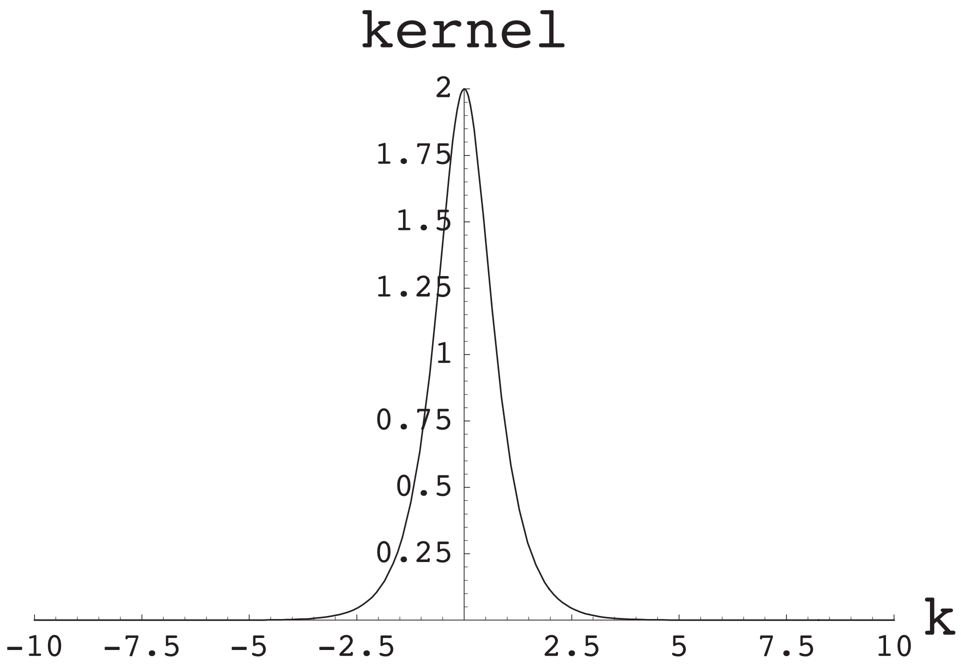


Figure 19.1.3: The kernel $[k^{n+m-1}/I'_m(kR)]$ as a function of k in the case that $m = 1$, $n = 0$, and $R = 2$ cm.

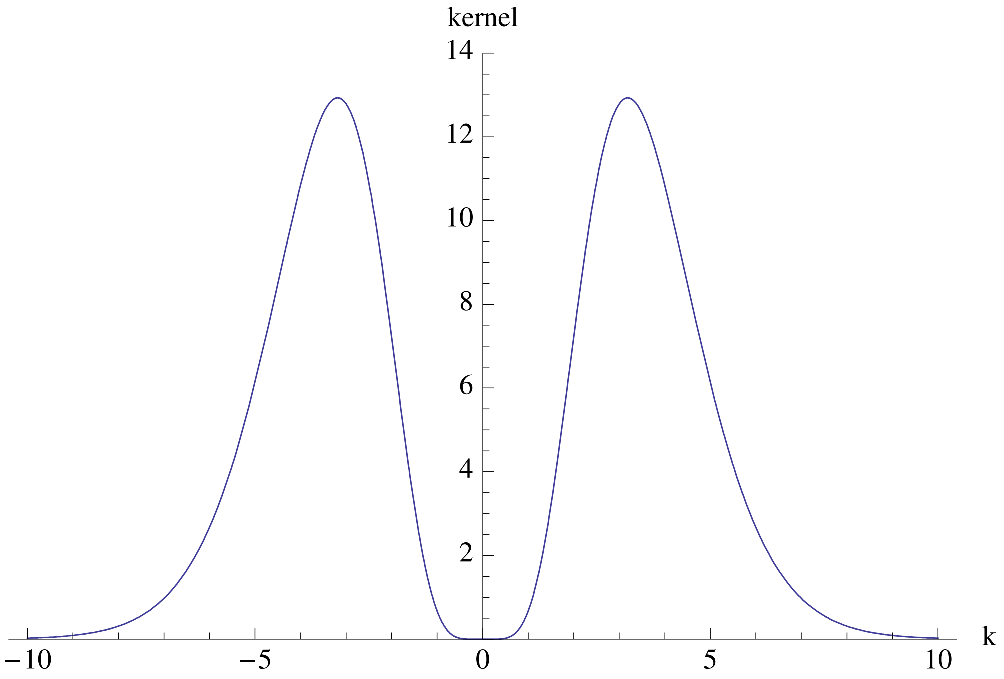


Figure 19.1.4: The kernel $[k^{n+m-1}/I'_m(kR)]$ as a function of k in the case that $m = 1$, $n = 6$, and $R = 2$ cm.

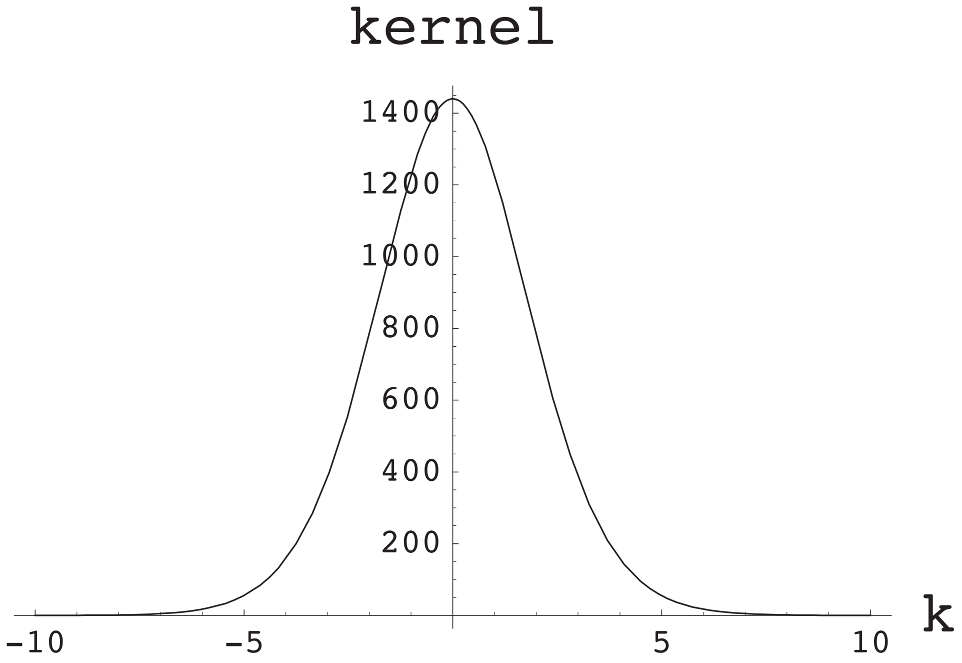


Figure 19.1.5: The kernel $[k^{n+m-1}/I'_m(kR)]$ as a function of k in the case that $m = 7$, $n = 0$, and $R = 2$ cm.

Note that the nonzero $\tilde{C}^{[0]}_{m,s}(k)$ for the monopole doublet are pure real because $C^{[0]}_{m,\alpha}(z)$ is real and, for the monopole doublet case, is an even function of z. Correspondingly, the $\tilde{C}^{[0]}_{m,s}(k)$ are even functions of k. Now, more compactly, we may write

$$C^{[n]}_{m,s}(z) = \int_{-\infty}^{\infty} dk\ (ik)^n \tilde{C}^{[0]}_{m,s}(k) \exp(ikz). \tag{19.1.17}$$

It is these integrals that we want to evaluate using spline-based inverse Fourier transforms in order to illustrate and verify the accuracy of this numerical method.

The integrands that are expected to be the hardest to integrate accurately are those for the extreme case $\{m = 1, n = 0\}$ because of its $|k|^2 \log |k|$ singularity at the origin and the extreme case $\{m = 1, n = 6\}$ because of its oscillatory behavior. These integrands, which are required to compute $C^{[0]}_{1,s}(z)$ and $C^{[6]}_{1,s}(z)$, respectively, are shown in Figures 1.6 and 1.7.

As expected, the integrands fall of rapidly for large $|k|$ because, as seen earlier, both the Fourier coefficients and the kernels fall of rapidly for large $|k|$. Indeed, inspection of Figures 1.6 and 1.7 shows that the integrands have effectively fallen to zero when $|k| > 10$. Therefore, and in order to be conservative, we will evaluate the integrals (14.3.23) using the somewhat larger cutoff $K_c = 20$. Moreover, we will use 401-point spline fits to the integrands over the interval $k \in [-K_c, K_c]$ so that $H = .1$. See (15.3.30).

Spline-Based Inverse Fourier Transform Results

Figures 1.8 and 1.9 show the functions $C^{[0]}_{1,s}(z)$ and $C^{[6]}_{1,s}(z)$ obtained in this way as well as the exact results given by (13.7.33) and its derivatives. Figures 1.10 and 1.11 show for these

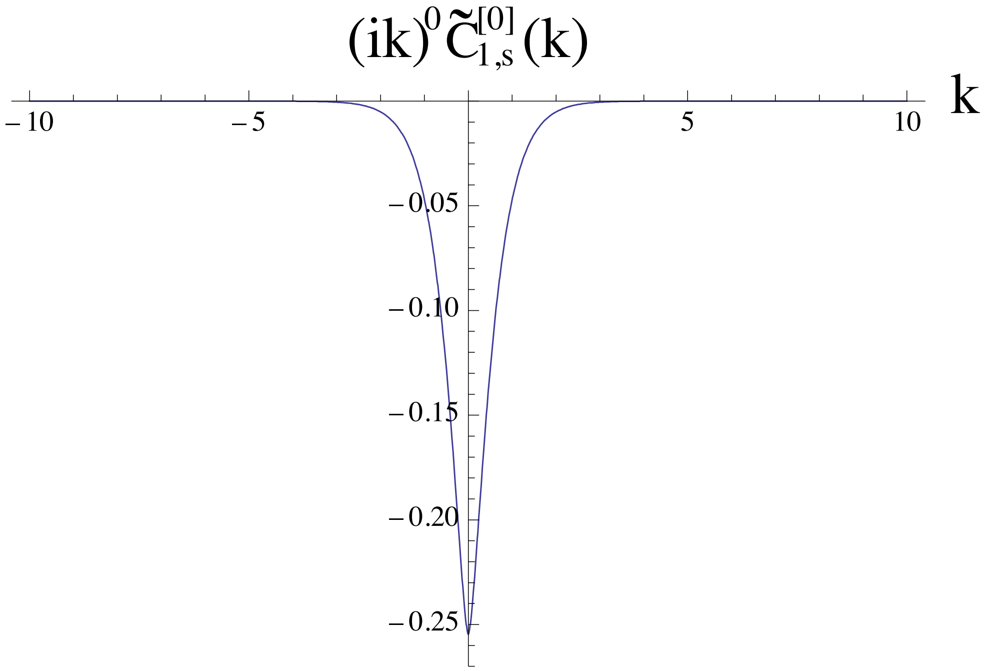


Figure 19.1.6: The integrand $(ik)^n\tilde{C}^{[0]}_{m,s}(k)$ for $m = 1$ and $n = 0$ as a function of k in the case that $R = 2$ cm. It is required to compute $C^{[0]}_{1,s}(z)$.

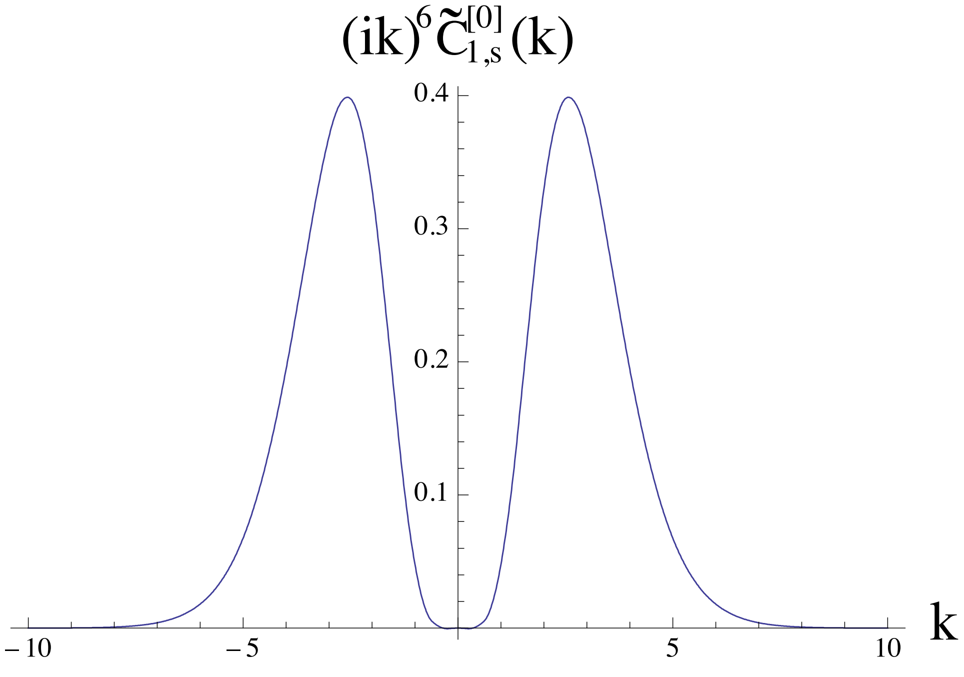


Figure 19.1.7: The integrand $(ik)^n\tilde{C}^{[0]}_{m,s}(k)$ for $m = 1$ and $n = 6$ as a function of k in the case that $R = 2$ cm. It is required to compute $C^{[6]}_{1,s}(z)$.

cases the differences between the exact and numerical results. Evidently the worst case is that of $C^{[0]}_{1,s}(z)$, and the error in this case is approximately 1.6 parts in 10^4. Further numerical study shows that the error is smaller for the other $C^{[n]}_{m,s}(z)$ listed in (13.7.35). Finally, further numerical study shows that the error can be made even smaller by increasing the number of points in the spline fit.[3] In this context we remark that the error in $C^{[0]}_{1,s}(z)$ decreases only as H^3 while the error in the other $C^{[n]}_{m,s}(z)$ decreases as H^4. This behavior is due to the $|k|^2 \log |k|$ singularity in $\tilde{C}^{[0]}_{1,s}(k)$ so that $\tilde{C}^{[0]}_{1,s}(k)$ does not have a finite fourth derivative at the origin. [See (15.1.34).] The integrands $(ik)^n \tilde{C}^{[0]}_{m,s}(k)$ for other values of m, n are less singular at the origin.

We conclude that if the integrand is computed to high precision (exactly, in this case), the spline-based inverse Fourier transform gives accurate results.

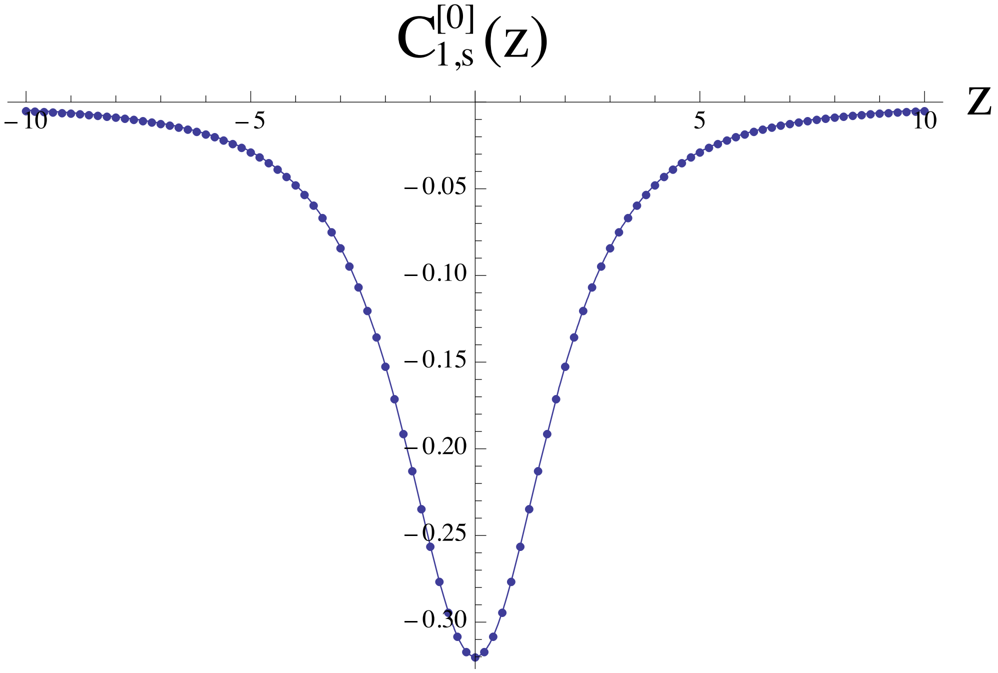


Figure 19.1.8: Exact and numerical results for $C^{[0]}_{1,s}(z)$. Exact results are shown as a solid line (see Figure 13.7.8), and numerical results are shown as dots.

19.1.2 Testing the Forward ($z \rightarrow k$) and ($\phi \rightarrow m$) Fourier Transforms

Having verified the accuracy of the spline-based inverse Fourier transform for carrying out the integration (14.3.23), let us also use splines to find $\tilde{\tilde{B}}_\rho(R, m', k')$. The radial component $B_\rho(\rho, \phi, z)$ of the magnetic field $\boldsymbol{B}$ is given by (13.7.8) and its behavior is displayed in Figures 13.7.6 and 13.7.7. According to (14.3.1), calculating $\tilde{\tilde{B}}_\rho(R, m', k')$ requires both a Fourier

[3] The value of K_c can also be increased. However, this does not seem to be necessary unless much higher accuracy is required.

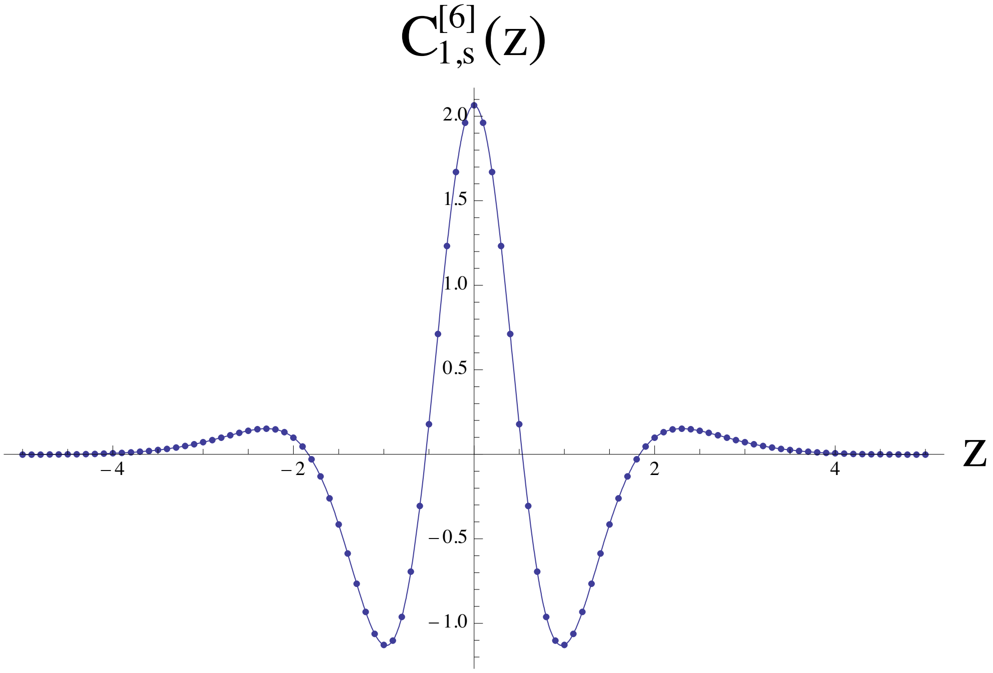


Figure 19.1.9: Exact and numerical results for $C_{1,s}^{[6]}(z)$. Exact results are shown as a solid line (see Figure 13.7.9), and numerical results are shown as dots.

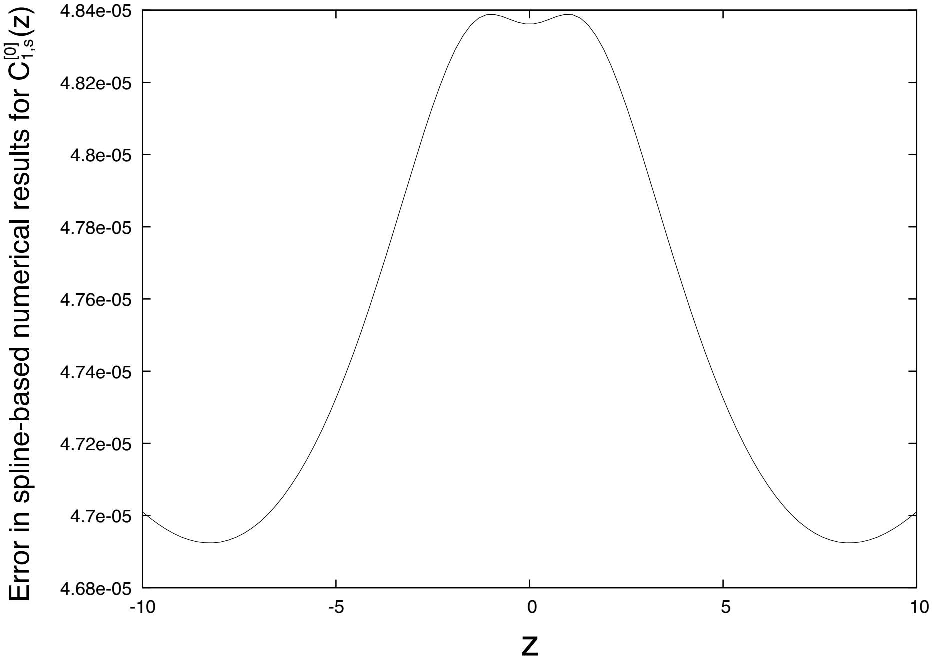


Figure 19.1.10: Difference between exact and spline-based numerical results for $C_{1,s}^{[0]}(z)$ using an exact integrand in (14.3.23).

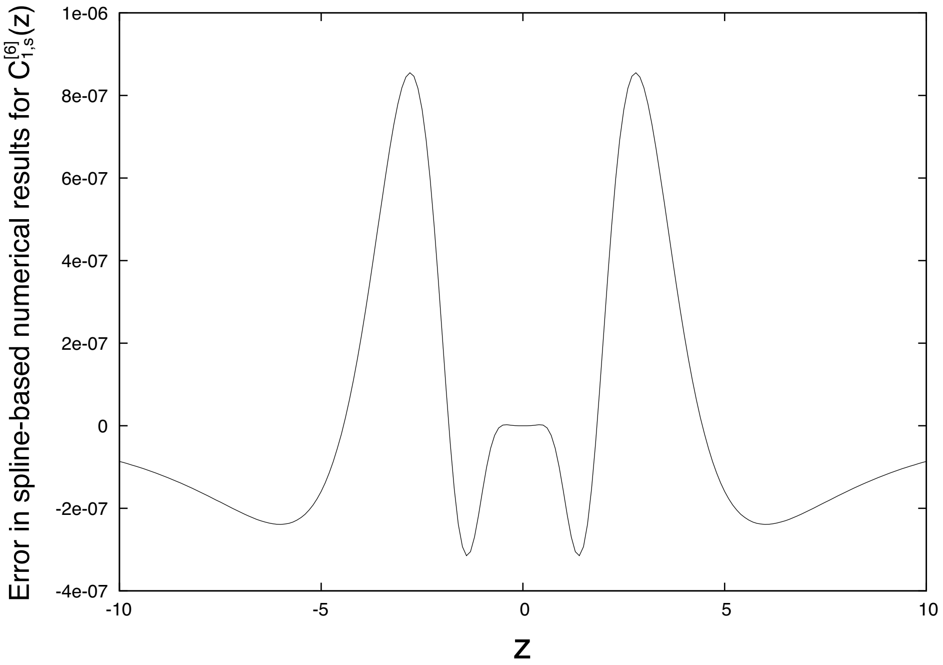


Figure 19.1.11: Difference between exact and spline-based numerical results for $C^{[6]}_{1,s}(z)$ using an exact integrand in (14.3.23).

transform involving integration over z, and the computation of Fourier coefficients involving integration over ϕ. The integrations may be performed in either order. Thus, we may define $\tilde{B}_\rho(R,\phi,k')$ by writing

$$\tilde{B}_\rho(R,\phi,k') = [1/(2\pi)] \int_{-\infty}^{\infty} dz\ \exp(-ik'z) B_\rho(R,\phi,z), \tag{19.1.18}$$

and then obtain $\tilde{\tilde{B}}_\rho(R,m',k')$ from the relation

$$\tilde{\tilde{B}}_\rho(R,m',k') = [1/(2\pi)] \int_0^{2\pi} d\phi\ \exp(-im'\phi) \tilde{B}_\rho(R,\phi,k'). \tag{19.1.19}$$

Performing the Forward ($z \to k$) Fourier Transform

Let us first attack the problem of evaluating (1.18). Based on inspection of Figure 13.7.7, we might imagine that we could safely cut off the Fourier transform integral (1.18) by setting the integrand to zero for $|z| > Z_c$ with $Z_c = 10$. However, in this case looks are deceiving. Examine (13.7.8) for the case $\phi = \pi/2$. The two terms within the first set of curly brackets tend to cancel for large $|z|$. The two terms within the second set of curly brackets do not. The term within the second set of curly brackets that falls of most slowly for large $|z|$ is given by the relation

$$\text{term with slowest falloff} = -ga[z^2 + (R+a)^2]^{-3/2} = -ga[z^2 + b^2]^{-3/2} \tag{19.1.20}$$

where

$$b = R + a. \tag{19.1.21}$$

If we are interested in evaluating $\tilde{B}_\rho(R,\phi,k)$ accurately for $\phi=\pi/2$ and $k=0$, then we must make the comparison

$$\int_{-\infty}^{\infty} dz\ [z^2+b^2]^{-3/2} \text{ versus} \int_{-Z_c}^{Z_c} dz\ [z^2+b^2]^{-3/2}. \tag{19.1.22}$$

These two integrals have the values

$$\int_{-\infty}^{\infty} dz\ [z^2+b^2]^{-3/2}\ = 2/b^2, \tag{19.1.23}$$

$$\int_{-Z_c}^{Z_c} dz\ [z^2+b^2]^{-3/2}\ = (2/b^2)\{Z_c/[Z_c^2+b^2]^{-1/2}\}. \tag{19.1.24}$$

Therefore the *fractional error* involved in imposing a cutoff is given by the relation

$$\text{fractional error} = 1 - Z_c/[Z_c^2+b^2]^{-1/2} \approx (1/2)(b/Z_c)^2. \tag{19.1.25}$$

Suppose we are willing to accept a fractional error on the order of 10^{-4}. Then we must have

$$Z_c \approx 100b/\sqrt{2}. \tag{19.1.26}$$

For $a=2.5$ and $R=2$, which yields $b=4.5$, this means that in reality we must have

$$Z_c \approx 300. \tag{19.1.27}$$

We remark that this is a worst case, where fringe fields fall off only as $1/|z|^3$. For cases where the fringe fields fall off more rapidly (e.g. quadrupoles, higher-order multipoles, dipoles with field clamps, etc.) the cutoff in z can be smaller.

Performing the Forward ($\phi \to m$) Fourier Transform

We are ready to carry out the spline-based calculation of $\tilde{\tilde{B}}_\rho(R,m',k')$. Let us select 4801 equally-spaced points z_j in the interval $z\in[-Z_c,Z_c]$ with $Z_c=300$, and let us select 49 equally-spaced values ϕ_ℓ in the interval $\phi\in[0,2\pi]$. Then, for each ϕ_ℓ, we carry out a 4801-point (in z) spline-based Fourier transform to find (for $R=2$) the quantities $\tilde{B}_\rho(R,\phi_\ell,k')$. Next we evaluate (1.19) using a 49-point (in ϕ) Riemann sum discrete angular Fourier transform to obtain $\tilde{\tilde{B}}_\rho(R,m',k')$. See Exercise 1.2 for an explanation of why 49 points should be adequate and, indeed, give good accuracy.

Spline-Based Forward and Inverse Fourier Transform Results

As a last step in this part of our exercise, let us use the integrands based on this $\tilde{\tilde{B}}_\rho(R,m',k')$ to carry out the same spline-based inverse Fourier transform described earlier. That is, we no longer work with exact integrands in (14.3.6), but rather use approximate integrands based on spline-based and discrete integrations over z and ϕ. However, we still do use exact values of $B_\rho(R,\phi,z)$ on the cylinder. The result of this process is an *almost completely* numerically calculated set of functions $C_{m,\alpha}^{[n]}(z)$. Examination of these numerically calculated functions shows that they also well approximate the exact functions $C^{[n]}_{m,\alpha}(z)$. For example, Figures 1.12 and 1.13 show the differences between the exact and numerically calculated $C^{[0]}_{1,s}(z)$ and $C^{[6]}_{1,s}(z)$. We see that Figure 1.12 resembles Figure 1.10. Surprisingly, the error in $C^{[0]}_{1,s}(z)$ is now slightly less than before, but remains approximately 1.6 parts in 10^4. Apparently in this case the errors involved in the approximate computation of $\tilde{\tilde{B}}_\rho(R, m', k')$ cancel to some extent the errors involved in the spline-based inverse Fourier transform. We also see that Figure 1.13 somewhat resembles Figure 1.11, but now the error in $C^{[6]}_{1,s}(z)$ is approximately 7 parts on 10^5 whereas it was 5 parts in 10^7 in Figure 1.11. In this case the errors involved in the approximate computation of $\tilde{\tilde{B}}_\rho(R, m', k')$ add to the overall error. (However the overall error is still acceptably small.) Indeed, we find that if we compute $\tilde{\tilde{B}}_\rho(R, m', k')$ more exactly by by increasing the number of points in ϕ beyond 49, increasing Z_c beyond 300, and increasing the number of points in z beyond 4801, then Figure 1.12 morphs into Figure 1.10, and Figure 1.13 morphs into Figure 1.11. With regard to the errors for the other nonzero $C^{[n]}_{m,s}(z)$, we find that they are also acceptably small. Finally, we find that the $C^{[0]}_{m,\alpha}(z)$ that should vanish are, in fact, numerically very small.

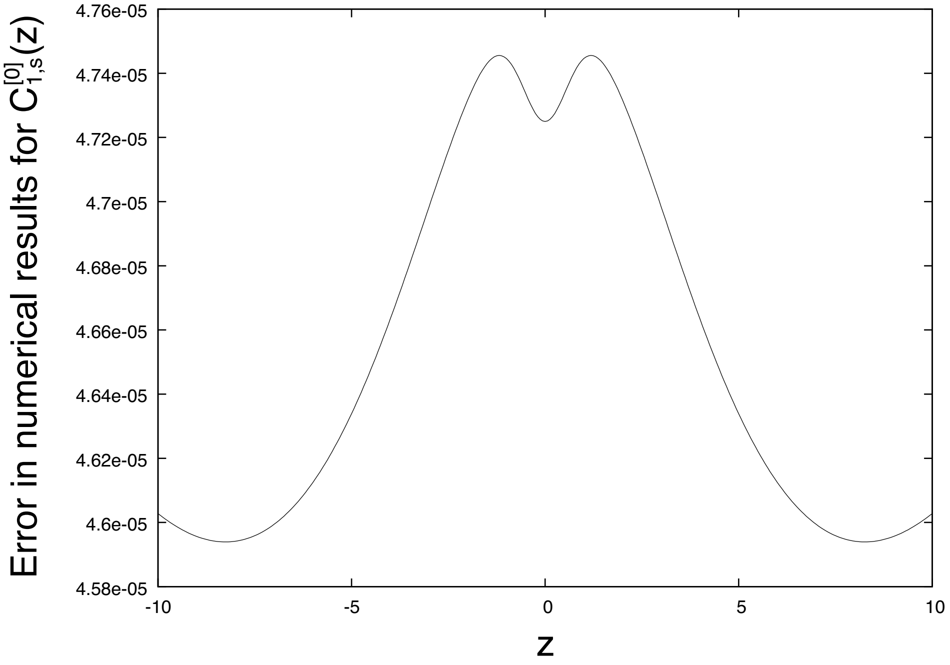


Figure 19.1.12: Difference between exact and numerical results for $C^{[0]}_{1,s}(z)$ using a spline-based integrand in (14.3.23) and exact values of $B_\rho(R, \phi, z)$ on the cylinder.

19.1.3 Test of Interpolation off a Grid

To complete our test, let us no longer use exact values of $B_\rho(R, \phi, z)$ on the cylinder. Rather, suppose we set up a regular grid in x, y, z space centered on the origin $(0, 0, 0)$. Let x and y range over the values $x \in [-2.4, 2.4]$ and $y \in [-2.4, 2.4]$, and (as before) let z range over the values $z \in [-300, 300]$. Use 49 grid points each in x and y so that $h_x = h_y = .1$,

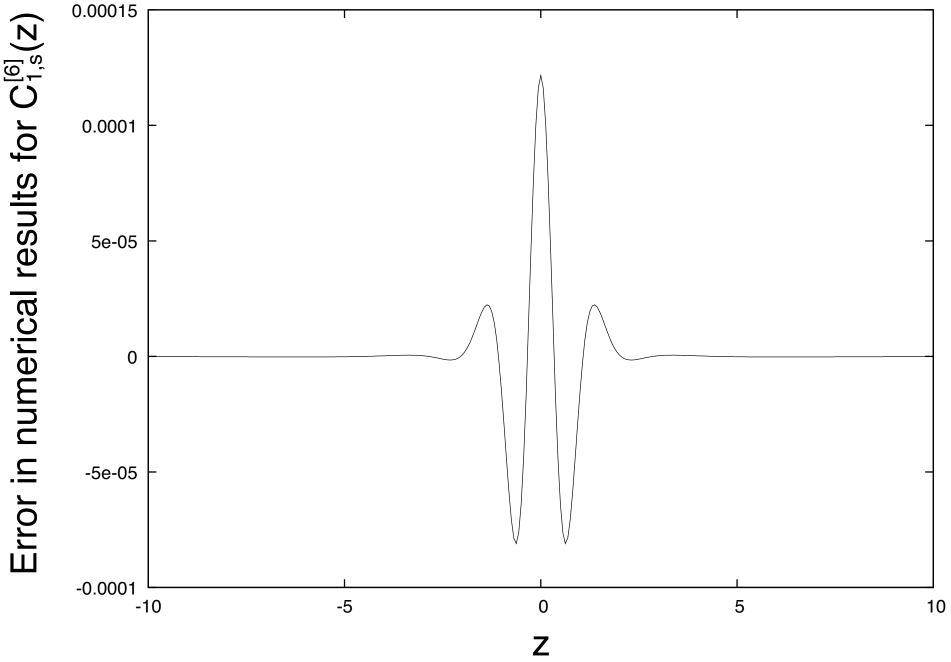


Figure 19.1.13: Difference between exact and numerical results for $C^{[6]}_{1,s}(z)$ using a spline-based integrand in (14.3.23) and exact values of $B_\rho(R, \phi, z)$ on the cylinder.

and (again as before) use 4801 grid points in z so that $h_z = .125$. Thus, use a total of $49 \times 49 \times 4801 = 11,527,201$ grid points. For each grid point specify the three components B_x, B_y, and B_z using (13.7.4) through (13.7.6) evaluated at these grid points. Employ bicubic interpolation (see Section 15.2.1) to interpolate $\boldsymbol{B}$ at these grid points onto the selected angular points on the cylinder $R = 2$, and then compute $B_\rho(R, \phi, z)$ at these angular points.[4] Finally, proceed as before using these approximate values of $B_\rho(R, \phi, z)$ on the cylinder. In particular, evaluate the angular Fourier transforms with a Riemann sum using 49 angular points and evaluate the forward linear transforms for 401 k values in the range $k \in [-K_c, K_c]$ with $K_c = 20$. Use these same points in k space to evaluate the inverse Fourier transforms. The result of this process is a *completely numerically* calculated set of functions $C^{[n]}_m(z)$.

Examination of these completely numerically calculated functions shows that they also well approximate the exact functions $C^{[n]}_{m,\alpha}(z)$. For example, Figures 1.14 and 1.15 show the differences between the exact and the completely numerically calculated $C^{[0]}_{1,s}(z)$ and $C^{[6]}_{1,s}(z)$. We see that Figure 1.14 is very similar to Figure 1.12, and Figure 1.15 is very similar to Figure 1.13. Consequently the error is little changed, and we conclude that interpolation from the grid onto the cylinder introduces little additional error. The errors for the other nonzero $C^{[n]}_{m,s}(z)$ listed in (13.7.35) are comparable. For example, Figure 1.16 shows the

[4] Alternatively, one could use bicubic spline interpolation. See Section 15.2.2. Note that the B_z component does not, in fact, contribute to $B_\rho(R, \phi, z)$. Note also that only a relatively small number of the 11,527,201 points are actually used because only values at those points relatively near the circular cylinder are needed to interpolate onto the cylinder. And, after interpolation onto the cylinder, only $49 \times 4801 = 235,249$ surface values of B_ρ are used in the remainder of the calculation.

exact and completely numerical results for $C^{[0]}_{7,s}(z)$, and Figure 1.17 shows the difference between the exact and completely numerical results for $C^{[0]}_{7,s}(z)$. We see that the error is approximately 4 parts in 10^4. Finally, the $C^{[0]}_{m,\alpha}(z)$ that should vanish are, in fact, again numerically very small.

We have demonstrated, for the monopole-doublet problem, that the steps in the first three boxes shown in Figure 14.1.1 can be carried out to yield results having good numerical accuracy. As remarked earlier, again see Figures 13.7.6 and 13.7.7, the surface field we have been working with is quite singular, more singular than fields likely to be encountered in practice. Thus the fact that the circular cylinder surface method has succeeded in this rather extreme case indicates that it is likely to work even better in actual physical applications.

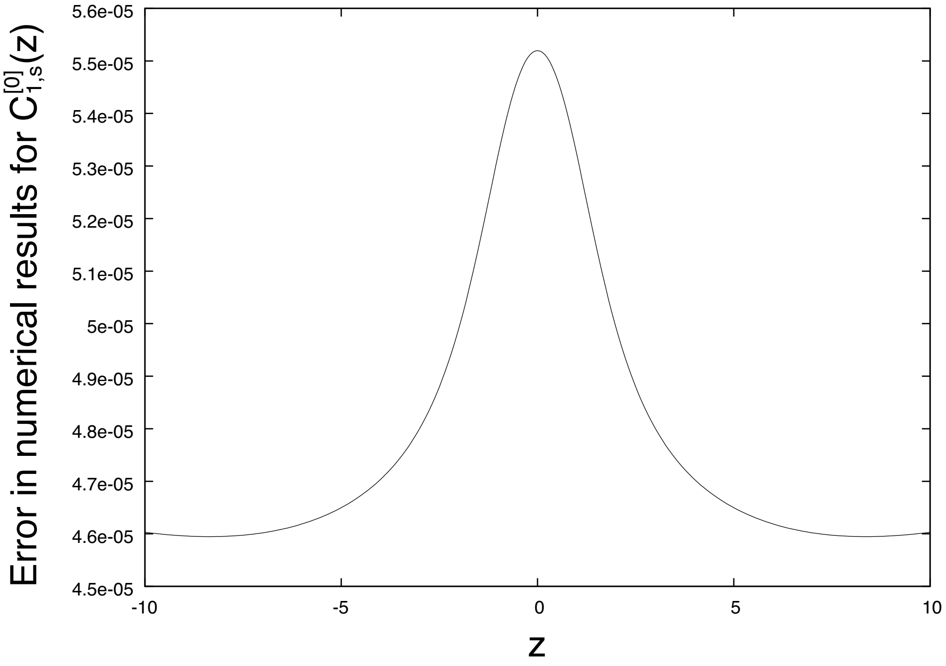


Figure 19.1.14: Difference between exact and completely numerical results for $C^{[0]}_{1,s}(z)$ using a spline-based integrand in (14.3.23) and interpolated values of $B_\rho(R, \phi, z)$ on the cylinder based on field data provided on a grid.

19.1.4 Reproduction of Interior Field Values

Another, but less stringent, test of accuracy is to use the completely numerically obtained on-axis gradients to compute $\boldsymbol{B}$ at the *interior* grid points with the aid of (13.2.69) through (13.2.71). These computed values can be compared with the known values of $\boldsymbol{B}$ at the interior grid points.[5] Before making such a comparison, some discussion is required.

[5] This test is less stringent because it does not compare derivatives.

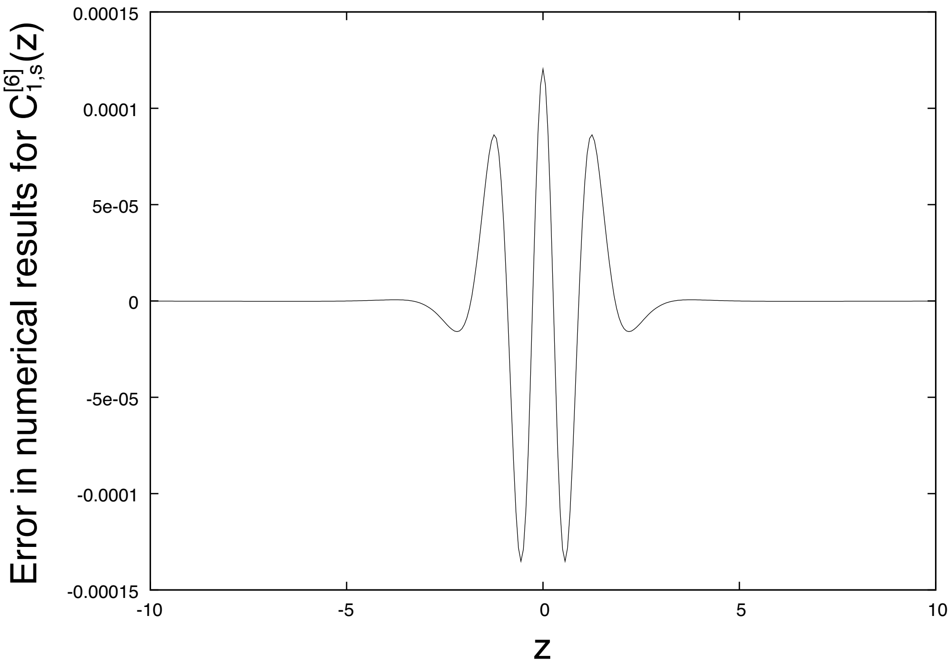


Figure 19.1.15: Difference between exact and completely numerical results for $C^{[6]}_{1,s}(z)$ using a spline-based integrand in (14.3.23) and interpolated values of $B_\rho(R, \phi, z)$ on the cylinder based on field data provided on a grid.

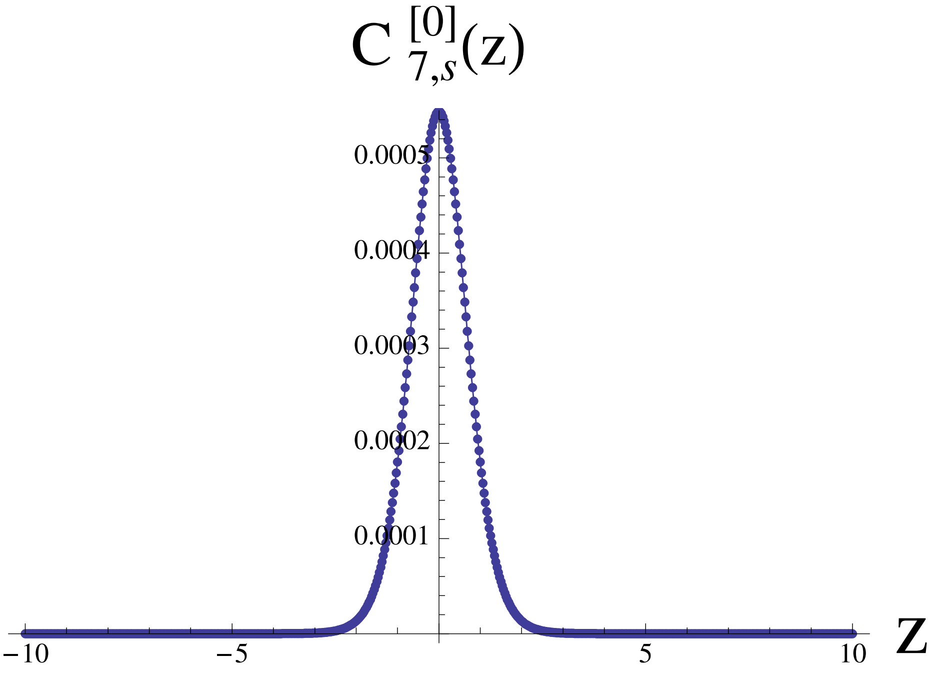


Figure 19.1.16: Exact and completely numerical results for $C^{[0]}_{7,s}(z)$. Exact results are shown as a solid line (see Figure 13.7.15), and numerical results are shown as dots.

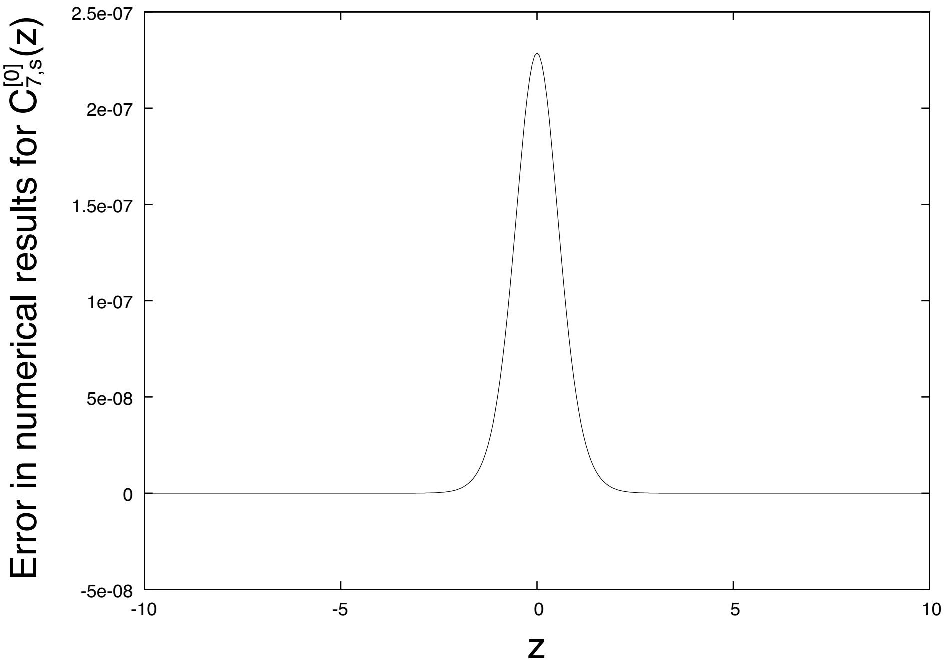


Figure 19.1.17: Difference between exact and completely numerical results for $C_{7,s}^{[0]}(z)$ using a spline-based integrand in (14.3.23) and interpolated values of $B_\rho(R, \phi, z)$ on the cylinder based on field data provided on a grid.

What to Hope for

First, since the surface fields we have been working with are quite singular, let us investigate the region over which the $C_{m,s}^{[n]}(z)$ that we have decided to employ, see (13.7.35), can be expected to give a good representation of $\boldsymbol{B}(x, y, z)$. As an initial exploration, let us consider the behavior of the Fourier series representation

$$B_\rho(R = 2, \phi, z = 0) = \sum_{m=-\infty}^{\infty} a_m \exp(im\phi) \tag{19.1.28}$$

for the function $B_\rho(R = 2, \phi, z = 0)$ shown in Figure 13.7.6. Suppose this series is *truncated* so that only terms for which $|m| \leq 7$ are retained. See Exercise 1.3. Call the resulting function $B_\rho^{\rm Tr}(R = 2, \phi, z = 0)$. Figure 1.18 displays $B_\rho^{\rm Tr}(R = 2, \phi, z = 0)$ as a function of ϕ, and Figure 1.19 shows the difference between $B_\rho(R = 2, \phi, z = 0)$ and $B_\rho^{\rm Tr}(R = 2, \phi, z = 0)$. Evidently terms well beyond $|m| = 7$ must be retained in (1.28) to adequately represent the surface field. See also the table of Fourier coefficients in Exercise 1.3. It follows that many $C_{m,s}^{[n]}(z)$ beyond those listed in (13.7.35) are required to represent the field near the surface $R = 2$.

As an illustration of this conclusion, let $\boldsymbol{B}^{\rm TrA}$ denote the field computed using the series (13.2.69) through (13.2.71) *truncated* so that only the $C_{m,s}^{[n]}(z)$ listed in (13.7.35) are retained, and using the *analytic* expressions (13.7.33) and their z derivatives for these $C_{m,s}^{[n]}(z)$.[6] Also,

[6] Note that the use of only the $C_{m,s}^{[n]}(z)$ listed in (13.7.35) to compute $\boldsymbol{B}^{\rm TrA}$ amounts to using 6th-order

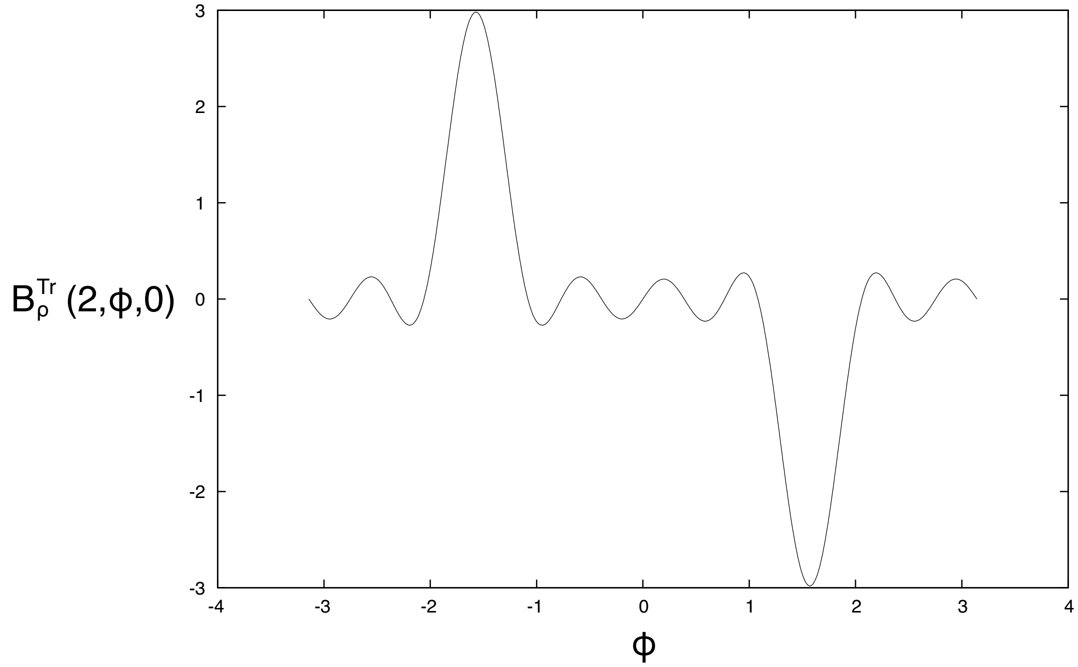


Figure 19.1.18: The quantity $B_\rho^{\mathrm{Tr}}(R = 2, \phi, z = 0)$ for the monopole doublet in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2.

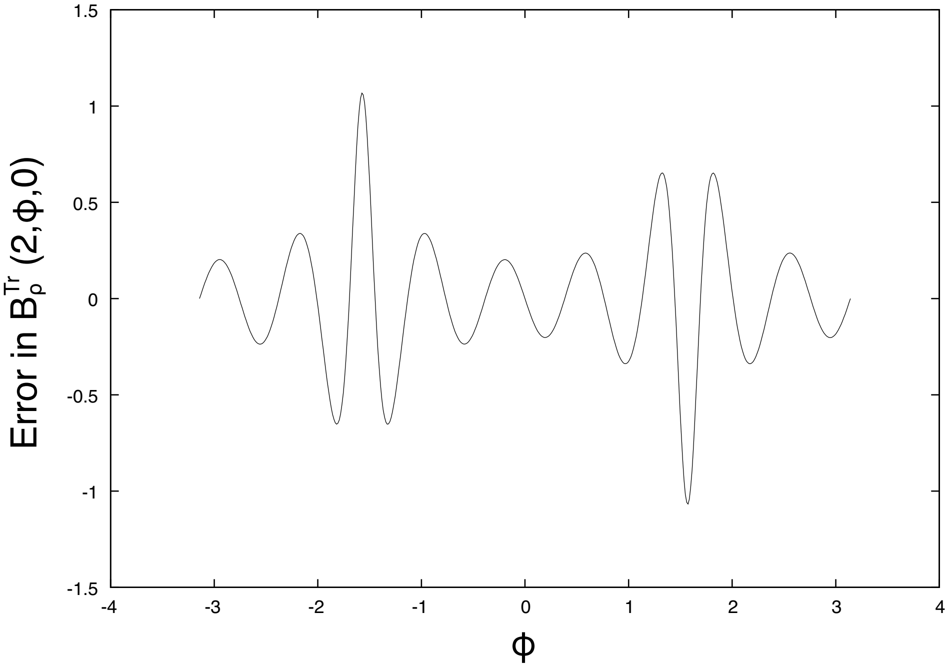


Figure 19.1.19: Difference between $B_\rho(R = 2, \phi, z = 0)$ and $B_\rho^{\mathrm{Tr}}(R = 2, \phi, z = 0)$ for the monopole doublet in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2.

let $\boldsymbol{B}^{\rm Exact}$ denote the known *exact* values of $\boldsymbol{B}$ computed using (13.7.4) through (13.7.6). Figures 1.20 through 1.22 show, as a function of ϕ, the quantity $||\boldsymbol{B}^{\rm TrA} - \boldsymbol{B}^{\rm Exact}||/||\boldsymbol{B}||^{\rm Max}$ for various values of ρ and z. Here $||\boldsymbol{B}||^{\rm Max}$ is the *maximum* value of $||\boldsymbol{B}||$ within the cylinder with radius ρ. We conclude that within the cylinder $\rho \leq 1/2$ the relative error in the field due to truncating the cylindrical multipole expansion is less than a few parts in 10^5. Note also that the domain of good approximation opens up as z leaves the plane $z = 0$. This behavior is to be expected based on the analytic properties of $\psi(x, y, z)$ when x, y, z are treated as three *complex* variables. See Examples 2.1 and 2.2 in Section 31.2; and Examples 3.1 and 3.2 in Section 31.3.

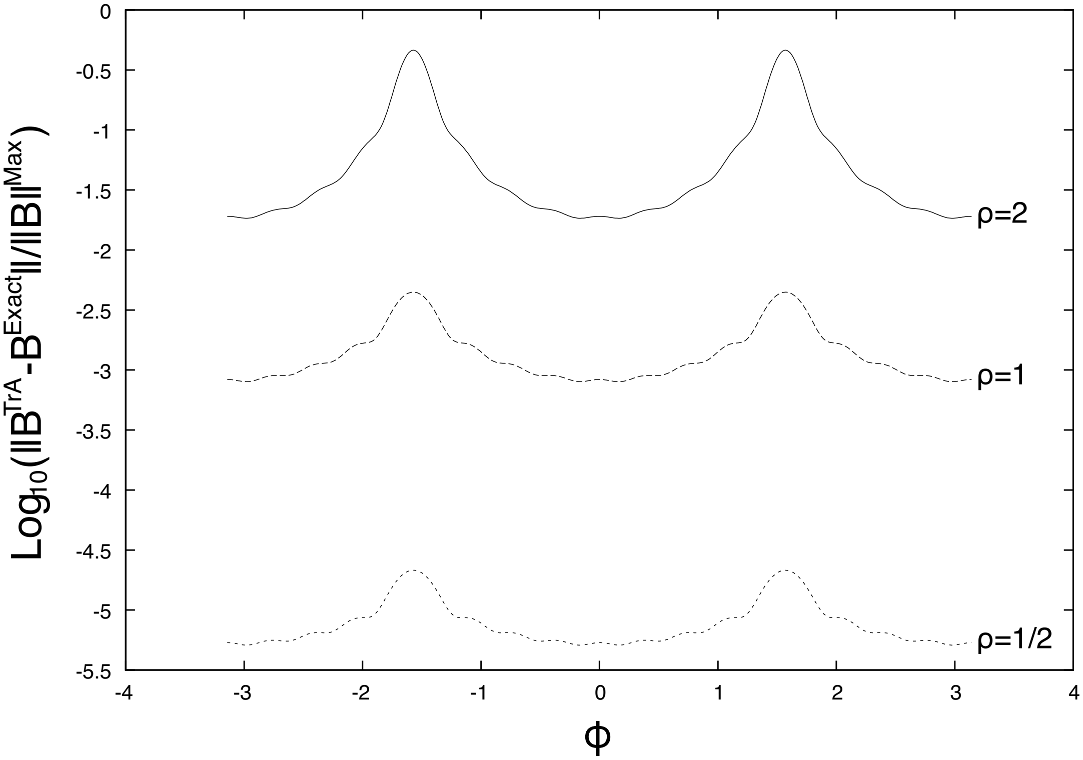


Figure 19.1.20: The logarithm base 10 of the quantity $||\boldsymbol{B}^{\rm TrA} - \boldsymbol{B}^{\rm Exact}||/||\boldsymbol{B}||^{\rm Max}$ as a function of ϕ for three ρ values and $z = 0$, for the monopole doublet, in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2. The solid line corresponds to $\rho = 2$, the dashed line to $\rho = 1$, and the dotted line to $\rho = 1/2$.

Test of Field Reproduction within the Cylinder $\rho \leq 1/2$

The previous discussion has determined (for the case of a monopole doublet) the region where the truncated cylindrical multipole expansion, which retains only the terms listed in (13.7.35), is expected to be accurate assuming the listed $C_{m,s}^{[n]}(z)$ are known exactly. For real

polynomials in the x, y variables (with z-dependent coefficients) for the components B_x and B_y, and a 5th-order polynomial in x, y for B_z.

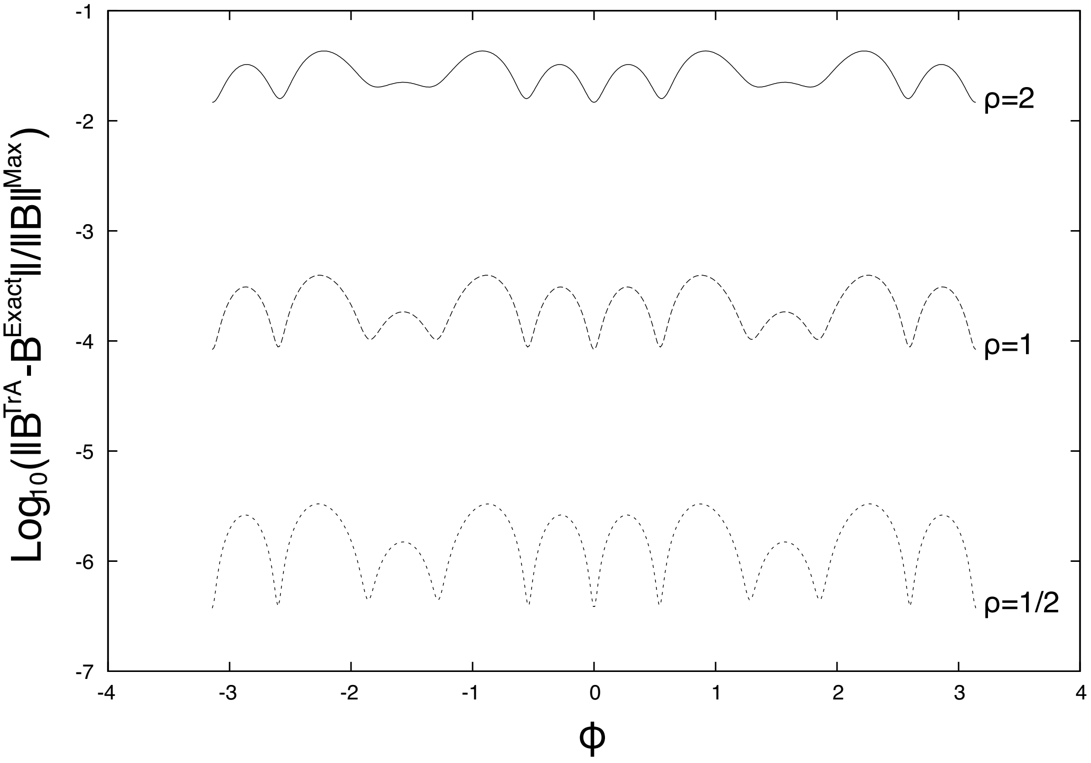


Figure 19.1.21: The logarithm base 10 of the quantity $||\boldsymbol{B}^{\mathrm{TrA}} - \boldsymbol{B}^{\mathrm{Exact}}||/||\boldsymbol{B}||^{\mathrm{Max}}$ as a function of ϕ for three ρ values and $z = 2.5$, for the monopole doublet, in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2. The solid line corresponds to $\rho = 2$, the dashed line to $\rho = 1$, and the dotted line to $\rho = 1/2$.

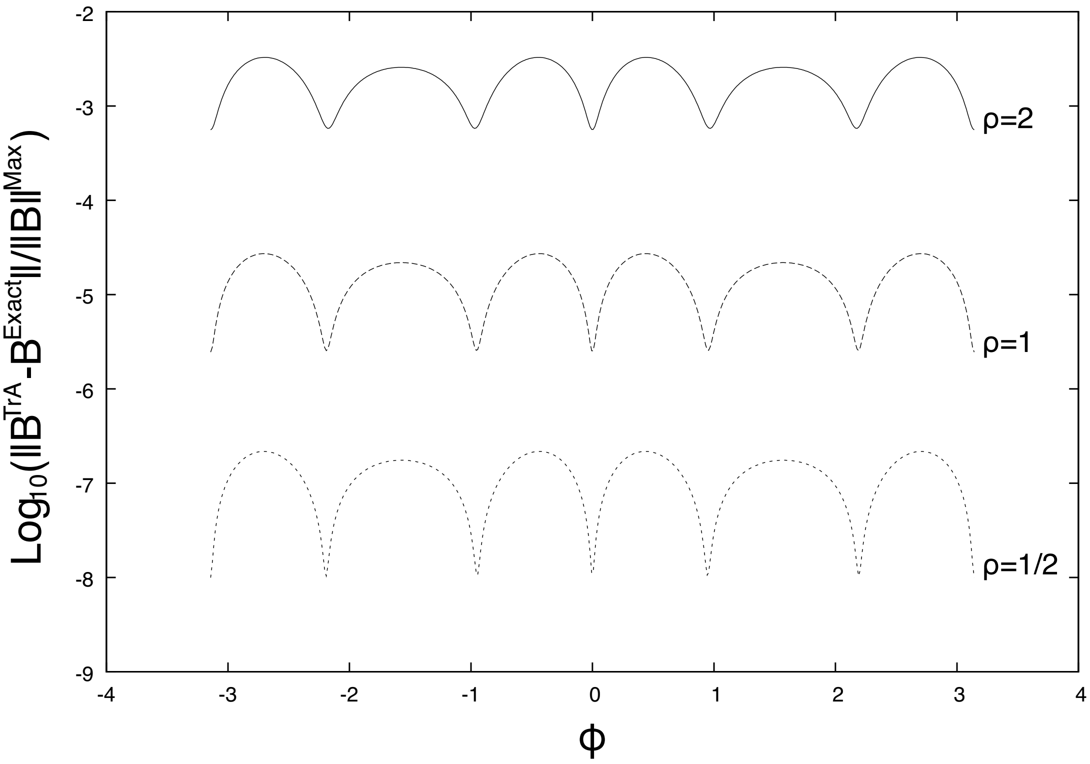


Figure 19.1.22: The logarithm base 10 of the quantity $||\boldsymbol{B}^{\rm TrA} - \boldsymbol{B}^{\rm Exact}||/||\boldsymbol{B}||^{\rm Max}$ as a function of ϕ for three ρ values and $z = 5$, for the monopole doublet, in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2. The solid line corresponds to $\rho = 2$, the dashed line to $\rho = 1$, and the dotted line to $\rho = 1/2$.

problems we only know the $C^{[n]}_{m,\alpha}(z)$ as calculated numerically based on grid data interpolated onto a surface, and only have the values of $\boldsymbol{B}$ at the grid points. Let $\boldsymbol{B}^{\rm Num}$ denote the completely *numerically* calculated values of $\boldsymbol{B}$ based on the completely numerically calculated $C^{[n]}_{m,\alpha}(z)$ with $(m+n) \leq 7$, and let $\boldsymbol{B}^{\rm Grid}$ now denote the values of $\boldsymbol{B}$ at the *grid* points obtained using (13.7.4) through (13.7.6). Examination of the completely numerically calculated values and the corresponding grid values shows that

$$||\boldsymbol{B}^{\rm Num} - \boldsymbol{B}^{\rm Grid}|| < 6 \times 10^{-5} \tag{19.1.29}$$

for all grid points within the cylinder $\rho \leq (1/2)$. Observe that, according to Figure 13.7.3, the magnitude of the maximum on-axis field is .32. Thus, the maximum error within the cylinder $\rho = 1/2$ compared to the maximum on-axis field is approximately 2 parts in 10^4. The smallness of this error again illustrates the accuracy of the method.[7] It also is a data-based indication that the cylindrical harmonic expansion is converging well for the interior $\rho \leq 1/2$ and that the Maxwell equations are well satisfied by the interior values of $\boldsymbol{B}^{\rm Grid}$.[8]

Exercises

19.1.1. Verify (1.11), (1.12), (1.15), and (1.16). You may need the identity

$$(2m)! = 2^m m![1 \cdot 3 \cdot 5 \cdot 7 \cdots (2m-1)]. \tag{19.1.30}$$

Prove this identity, say, by induction.

19.1.2. This exercise studies the discrete Fourier transform of a periodic function. Suppose $f(\phi)$ is a 2π periodic function and therefore has a Fourier expansion of the form

$$f(\phi) = \sum_{m=-\infty}^{\infty} a_m \exp(im\phi). \tag{19.1.31}$$

We know that the coefficients a_j are given by the relation

$$a_j = [1/(2\pi)] \int_0^{2\pi} d\phi \; f(\phi) \exp(-ij\phi). \tag{19.1.32}$$

Select N discrete phi values ϕ_n according to the rule

$$\phi_n = n(2\pi/N) \text{ for } n = 0, 1, \cdots, N-1. \tag{19.1.33}$$

[7] This small error means, among other things, that Figures 13.7.3 through 13.7.5 are well reproduced by $\boldsymbol{B}^{\rm Num}$. Note that *only* the values of $B^{\rm Grid}_x$ and $B^{\rm Grid}_y$ on grid points near the surface of the cylinder $R = 2$ were used to determine all three components of $\boldsymbol{B}^{\rm Num}$ at all grid points within the cylinder.

[8] By construction, the $\boldsymbol{B}^{\rm Num}$ satisfy the Maxwell equations to good precision. And, in this case, the $\boldsymbol{B}^{\rm Grid}$ also satisfy the Maxwell equations since they were obtained by evaluating the Maxwell solution (13.7.4) through (13.7.6) at the grid points. However, in the general case, the $\boldsymbol{B}^{\rm Grid}$ are supplied by some 3-D electromagnetic code, and the 'Maxwellian goodness' of these quantities depends on the quality of the 3-D electromagnetic code.

Next define quantities A_j by the rule

$$A_j = (1/N)\sum_{n=0}^{N-1} f(\phi_n)\exp(-ij\phi_n). \tag{19.1.34}$$

Verify that (1.34) is the discrete version of (1.32).

Show that combining (1.31), (1.33), and (1.34) gives the result

$$\begin{aligned} A_j &= \sum_{m=-\infty}^{\infty} a_m\{(1/N)\sum_{n=0}^{N-1}\exp[inm(2\pi/N)]\exp[-ijn(2\pi/N)]\} \\ &= \sum_{m=-\infty}^{\infty} a_m\{(1/N)\sum_{n=0}^{N-1}\exp[in(m-j)(2\pi/N)]\}. \end{aligned} \tag{19.1.35}$$

However, there is the identity

$$\sum_{n=0}^{N-1} x^n = (1-x^N)/(1-x). \tag{19.1.36}$$

Using this identity, show that

$$(1/N)\sum_{n=0}^{N-1}\exp[in(m-j)(2\pi/N)] = (1/N)\{1-\exp[i(m-j)2\pi]\}/\{1-\exp[i(m-j)(2\pi/N)]\}. \tag{19.1.37}$$

From (1.37), show that

$$(1/N)\sum_{n=0}^{N-1}\exp[in(m-j)(2\pi/N)] = 0 \text{ if } (m-j)/N \text{ is not an integer,} \tag{19.1.38}$$

and

$$(1/N)\sum_{n=0}^{N-1}\exp[in(m-j)(2\pi/N)] = 1 \text{ if } (m-j)/N \text{ is an integer.} \tag{19.1.39}$$

[Note that $(m-j)$ is always an integer.] Verify that these results can be written more compactly in the form

$$(1/N)\sum_{n=0}^{N-1}\exp[in(m-j)(2\pi/N)] = \sum_{\ell=-\infty}^{\infty}\delta_{m,j+\ell N}. \tag{19.1.40}$$

Employ (1.35) and (1.40) to get the final result

$$A_j = \sum_{\ell=-\infty}^{\infty} a_{j+\ell N}. \tag{19.1.41}$$

Verify that (1.41) implies the periodicity relation

$$A_{j+N} = A_j. \tag{19.1.42}$$

The relation (1.41) can be rewritten in the form

$$A_j = a_j + \sum_{\ell \neq 0}^{\infty} a_{j+\ell N}. \tag{19.1.43}$$

Thus, A_j is a good approximation to a_j provided N is large enough that the $a_{j+\ell N}$ are small for $\ell = \pm 1, \pm 2, \cdots$ in such a way that the sum in (1.43) is also small.

Note that under the assumptions made, namely periodicity and rapid falloff of the Fourier coefficients, the discrete angular Fourier transform is much more accurate than the estimate (15.3.24) would promise.

19.1.3. This exercise is a continuation of Exercise 1.2. It shows that good performance of the discrete Fourier transform for periodic functions can be assured under the the assumption of suitable *analyticity*. As an application, it justifies the use of the $N = 49$ discrete Fourier transform to obtain $\tilde{B}_\rho(R, m', k')$ in the case of the monopole doublet.

As in (1.28), let $f(\phi)$ be the function $B_\rho(R = 2, \phi, z = 0)$ shown in Figure 13.7.6. Its first few nonzero Fourier coefficients, obtained by accurate numerical integration of (1.32) using *Mathematica*, are listed below in Table 1.1. As shown in Figure 1.23, they fall of exponentially with increasing j in the fashion

$$|a_j| \sim (.8)^j. \tag{19.1.44}$$

Consequently, the contribution of the sum in (1.43) will be small provided N is reasonably large and j is considerably smaller than N. Table 1.1 also lists the first few nonzero values of A_j obtained from (1.34) with $N = 49$. We see from the table that (when $N = 49$) the A_j are good approximations to the a_j for $j \leq 7$, as advertised.

Table 19.1.1: The exact and discrete (with $N = 49$) Fourier coefficients of $f(\phi)$.

j	$2\Im a_j$	$2\Im A_j$
1	0.986833050662540	0.9868330457500823
3	-0.859714503908633	-0.8597145139462157
5	0.656703778334560	0.6567037866425125
7	-0.477523703835796	-0.47752370010754797
9	0.337961137907473	0.33796112868607286
11	-0.235072232641152	-0.235072209887016
13	0.161532491767950	0.16153245497610078
15	-0.110003634817578	-0.11000357184645376
17	0.074393215827366	0.0743931186747128
19	-0.050032679038838	-0.05003252910493426
21	0.033497429399684	0.0334971980691082
23	-0.022342757613886	-0.022342400791792692
25	0.0148553386831777	0.01485478841656739

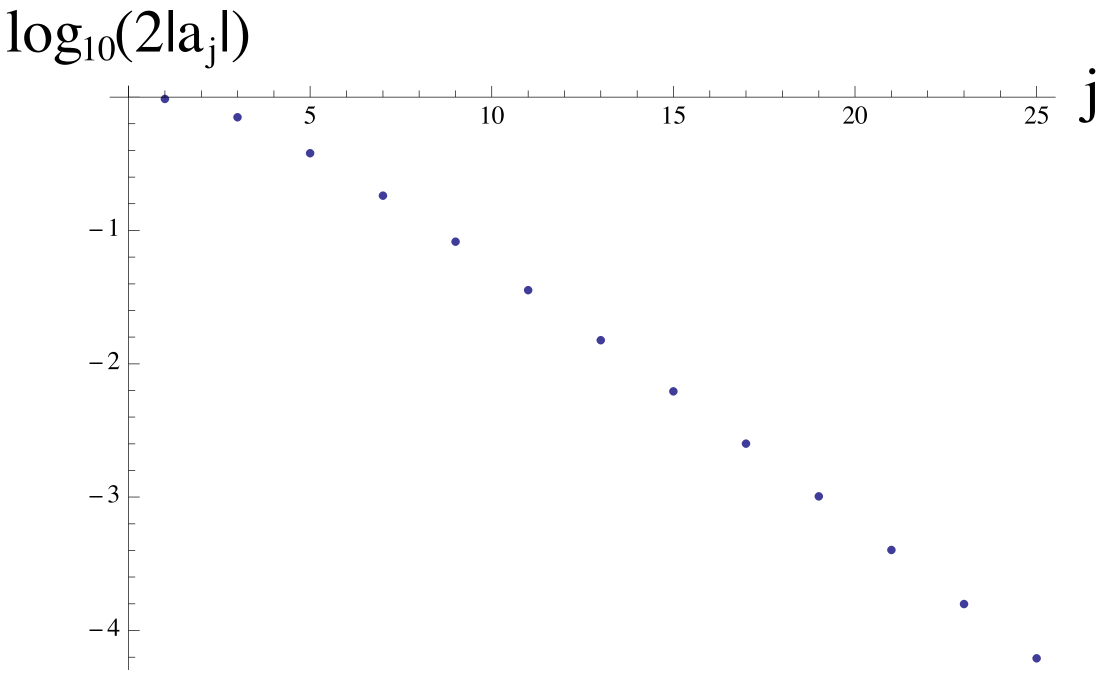


Figure 19.1.23: The quantities $\log_{10}(2|a_j|)$ as a function of j. For large j the points fall on a straight line having slope $\log_{10}(.8)$.

Similar considerations apply to the function $\tilde{B}_\rho(R,\phi,k)$. Figures 1.24 and 1.25 illustrate the cases $\tilde{B}_\rho(R=2,\phi,k=0)$ and $\tilde{B}_\rho(R=2,\phi,k=20)$, and Figures 1.26 and 1.27 display their Fourier coefficients. Again the Fourier coefficients fall of as $(.8)^j$. This behavior can be understood as follows: It can be shown that the Fourier coefficients of $B_\rho(R,\phi,z)$ for $z \neq 0$ fall off even more rapidly than (1.44). We conclude that since the Fourier coefficients of $B_\rho(R,\phi,z)$ fall off at least as rapidly as (1.44) for all z, then the Fourier coefficients of $\tilde{B}_\rho(R,\phi,k)$ must also fall off in this fashion, because $\tilde{B}_\rho(R,\phi,k)$ may be viewed as a linear combination of the $B_\rho(R,\phi,z)$. See (1.22).

How might one anticipate the relation (1.44)? Introduce the complex variable λ by writing

$$\lambda = \exp i\phi. \tag{19.1.45}$$

With this change of variable the integral (1.32) becomes

$$a_j = [-i/(2\pi)] \oint_C d\lambda\, f(-i\log\lambda)\lambda^{-(j+1)} \tag{19.1.46}$$

where the contour C is the unit circle. Now deform the contour to make C as large a circle as possible without encountering singularities of $f(-i\log\lambda)$. Suppose this circle has radius Λ. Then the integral (1.46) has the asymptotic behavior

$$|a_j| \sim (1/\Lambda)^j. \tag{19.1.47}$$

Evidently (13.7.8), when evaluated at $z=0$, is singular when

$$\sin\phi = \pm[(R^2+a^2)/(2aR)]. \tag{19.1.48}$$

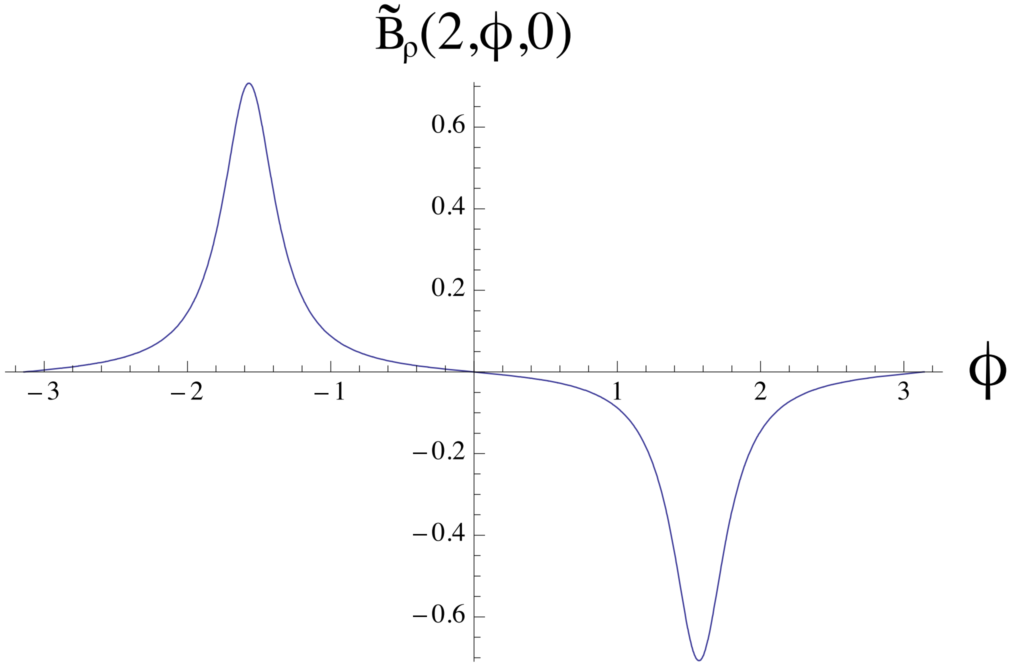


Figure 19.1.24: The real part of $\tilde{B}_\rho(R = 2, \phi, k = 0)$ for the monopole doublet in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2. The imaginary part vanishes.

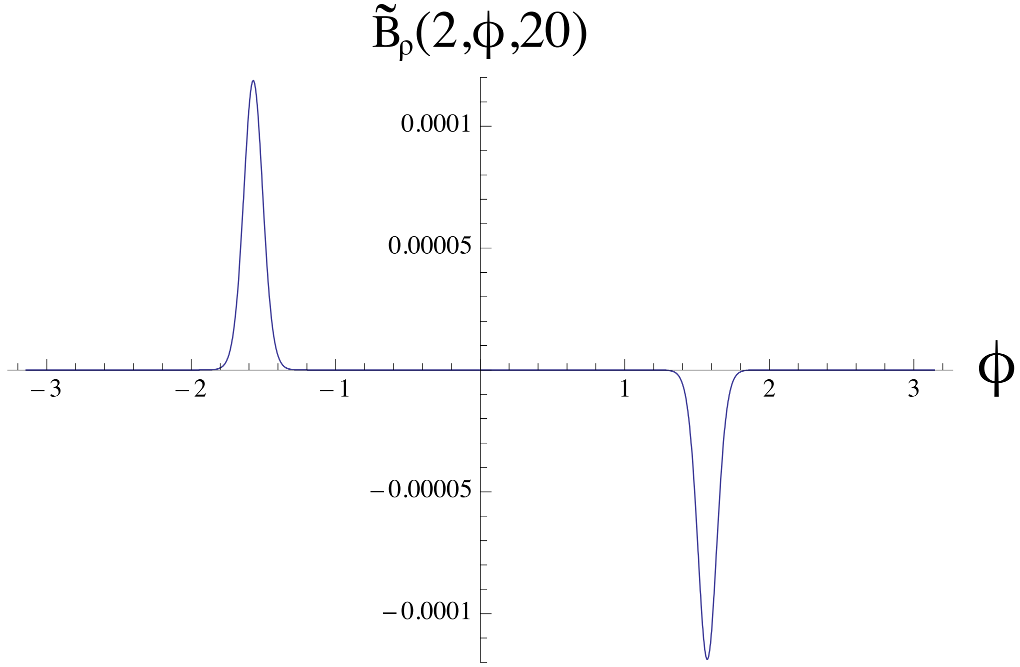


Figure 19.1.25: The real part of $\tilde{B}_\rho(R = 2, \phi, k = 20)$ for the monopole doublet in the case that $a = 2.5$ cm and $g = 1$ Tesla-(cm)^2. The imaginary part vanishes.

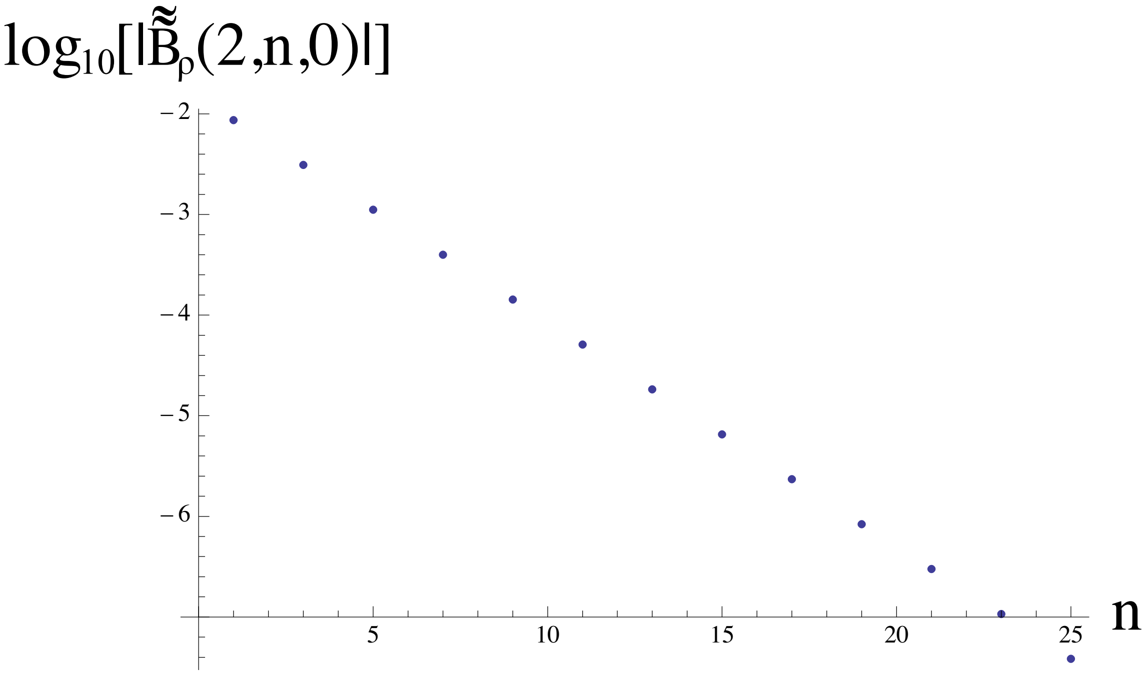


Figure 19.1.26: The quantities $\log_{10}[|\tilde{\tilde{B}}_\rho(R = 2, n, k = 0)|]$ as a function of n. For large n the points fall on a straight line having slope $\log_{10}(.8)$.

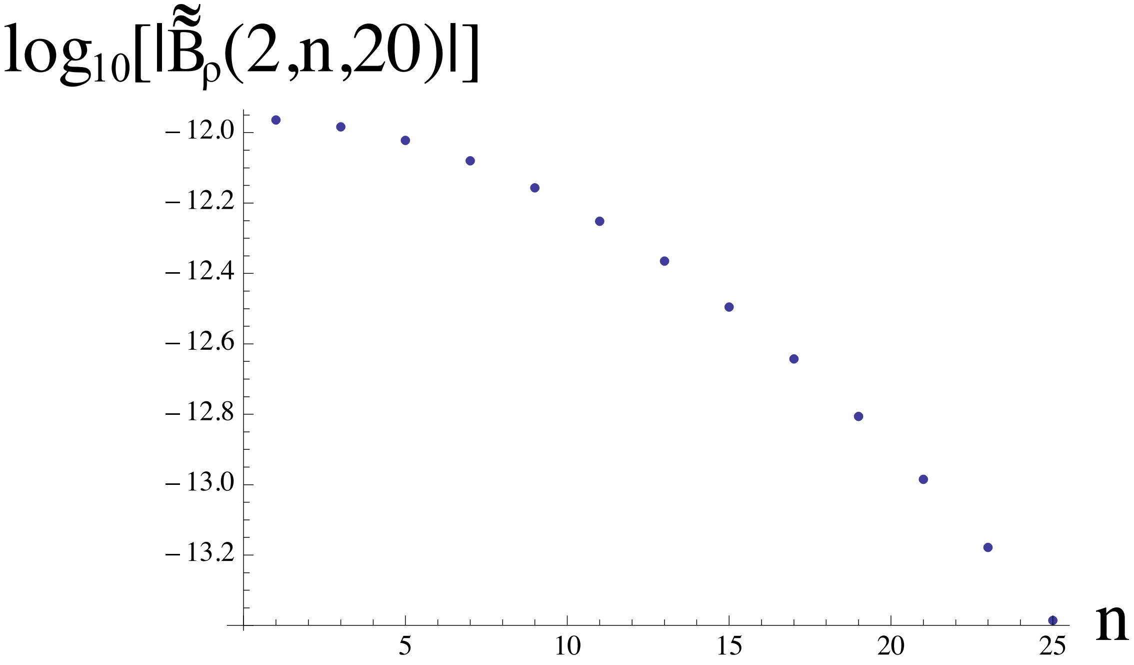


Figure 19.1.27: The quantities $\log_{10}[|\tilde{\tilde{B}}_\rho(R = 2, n, k = 20)|]$ as a function of n. For large n the points fall on a straight line having slope $\log_{10}(.8)$.

Show that (1.48) yields the relation

$$1/\Lambda = R/a = 2/2.5 = .8. \tag{19.1.49}$$

Suppose $z \neq 0$ in (13.7.8). Show that then Λ is larger than the value given in (1.49). Thus, we have analyzed the worst case, the case with the slowest falloff with increasing j.

19.1.4. Review Exercise 1.3. Look at the relation (14.3.2), the expansion (13.2.37), and the monopole doublet results (13.7.19), (13.7.29), and (13.7.33). Use these quantities to find explicit expressions for the Fourier Coefficients $\tilde{B}_\rho(R, m, z)$, and find their large m behavior.

19.2 Elliptical Cylinder Numerical Results for Monopole Doublet

In this subsection we will benchmark the numerical method of Section 14.4 to demonstrate the use of an elliptical cylinder. Here we have two goals: First, as a practical matter, the infinite sums over r that occur in (14.4.83) through (14.4.86) must be truncated, and we must establish that this can be done while still achieving a desired accuracy. Second, we must demonstrate that all our numerical machinery actually works.

As just done for the case of a circular cylinder, we will again try to reproduce the exact results for the on-axis gradients of the same monopole doublet. But now, as an example, we will use as our surface that of an elliptical cylinder for which the ellipse has a semi-major axis ($x^{\max}$) of 4 cm in the x direction and a semi-minor axis ($y^{\max}$) of 2 cm in the y direction. See Figure 14.4.2. This is achieved by setting

$$u = U = \tanh^{-1}(y^{\max}/x^{\max}) = \tanh^{-1}(2/4) = .549306144 \tag{19.2.1}$$

and

$$f = 4/\cosh(U) = \sqrt{12} = 3.464101615 \text{ cm} \tag{19.2.2}$$

in equations (14.4.1) and (14.4.2) so that we have the relations

$$x^{\max} = f \cosh U = 4 \text{ cm}, \tag{19.2.3}$$

$$y^{\max} = f \sinh U = 2 \text{ cm}. \tag{19.2.4}$$

19.2.1 Finding the Mathieu Coefficients

In the elliptical cylinder case there are fewer calculations we can carry out exactly compared to the circular cylinder case. However, some of the routines we will be using in the elliptic case will be the same as in the circular case, and we have already benchmarked them in the circular case.

Exact Results for the Forward ($z \rightarrow k$) Fourier Transform

There is a quantity we can still compute exactly for the elliptic cylinder when applied to the monopole doublet case, and that is the function $\tilde{F}(v, k)$. See (14.4.72).[9] Upon combining (14.4.67) and (14.4.72) we see that

$$\tilde{F}(v, k) = (f \sinh U \cos v)\tilde{B}_x(U, v, k) + (f \cosh U \sin v)\tilde{B}_y(U, v, k) \tag{19.2.5}$$

where

$$\tilde{B}_x(U, v, k) = [1/(2\pi)] \int_{-\infty}^{\infty} dz \exp(-ikz) B_x(U, v, z), \tag{19.2.6}$$

$$\tilde{B}_y(U, v, k) = [1/(2\pi)] \int_{-\infty}^{\infty} dz \exp(-ikz) B_y(U, v, z). \tag{19.2.7}$$

Examine $B_x(U, v, z)$ and $B_y(U, v, z)$ as given by (14.4.1), (14.4.2), (13.7.4), and (13.7.5). Define quantities $b_\pm(v)$ by the rule

$$b_\pm(v) = [x^2 + (y \pm a)^2]^{1/2} = \{[f \cosh(U) \cos(v)]^2 + [f \sinh(U) \sin(v) \pm a]^2\}^{1/2}. \tag{19.2.8}$$

Using this definition we may write

$$[x^2 + (y \pm a)^2 + z^2] = [z^2 + b_\pm^2(v)], \tag{19.2.9}$$

and $B_x(U, v, z)$ and $B_y(U, v, z)$ take the form

$$\begin{aligned} B_x(U, v, z) \quad &= \quad gf \cosh(U) \cos(v)[z^2 + b_-^2(v)]^{-3/2} \\ &\quad -gf \cosh(U) \cos(v)[z^2 + b_+^2(v)]^{-3/2}, \end{aligned} \tag{19.2.10}$$

$$\begin{aligned} B_y(U, v, z) \quad &= \quad g[f \sinh(U) \sin(v) - a][z^2 + b_-^2(v)]^{-3/2} \\ &\quad -g[f \sinh(U) \sin(v) + a][z^2 + b_+^2(v)]^{-3/2}. \end{aligned} \tag{19.2.11}$$

Next recall the Fourier transform relation

$$[1/(2\pi)] \int_{-\infty}^{\infty} dz \;\; \exp(-ikz)[z^2 + b_\pm^2(v)]^{-3/2} = (1/\pi)[|k|/b_\pm(v)]K_1[|k|b_\pm(v)]. \tag{19.2.12}$$

For convenience, define the functions $F_\pm(v, k)$ by the rule

$$F_\pm(v, k) = (1/\pi)[|k|/b_\pm(v)]K_1[|k|b_\pm(v)]. \tag{19.2.13}$$

Then, in terms of these functions, we have the relations

$$\begin{aligned} \tilde{B}_x(U, v, k) \quad &= \quad gf \cosh(U) \cos(v)F_-(v, k) \\ &\quad -gf \cosh(U) \cos(v)F_+(v, k), \end{aligned} \tag{19.2.14}$$

[9] Of course, we will eventually want to demonstrate that we can compute this function numerically with high accuracy using field data on grid points.

$$\begin{aligned}\tilde{B}_y(U,v,k) &= g[f\sinh(U)\sin(v)-a]F_-(v,k) \\ &\quad -g[f\sinh(U)\sin(v)+a]F_+(v,k). \qquad (19.2.15)\end{aligned}$$

Combining (2.5), (2.8), and (2.13) through (2.15) gives a final expression for $\tilde{F}(v,k)$,

$$\tilde{F}(v,k) = . \qquad (19.2.16)$$

Note that for the magnetic monopole doublet the functions $F_\pm(v,k)$ and $\tilde{F}(v,k)$ are pure real and are even functions of k.

Let us pause to study the functions $F_\pm(v,k)$. The function $K_1(w)$ has, at the origin, the behavior

$$K_1(w) \approx (1/w) + (w/2)\log(w/2); \qquad (19.2.17)$$

and therefore

$$wK_1(w) \approx 1 + (w^2/2)\log(w/2) \qquad (19.2.18)$$

at the origin. At $w = +\infty$, $K_1(w)$ has the behavior

$$K_1(w) \approx [\pi/(2w)]^{1/2}\exp(-w). \qquad (19.2.19)$$

Consequently $F_\pm(v,k)$, and therefore also $\tilde{F}(v,k)$, are well behaved for all k, but are not analytic at the origin because of the log term.[10]

Exact Results for the Mathieu Coefficients

According to (14.4.73) and (14.4.74), the next step is to compute the Mathieu coefficients by performing the angular integrals[11]

$$\tilde{\tilde{F}}_r^c(k) = (1/\pi)\int_0^{2\pi} dv\ \mathrm{ce}_r(v,q)\tilde{F}(v,k), \qquad (19.2.20)$$

$$\tilde{\tilde{F}}_r^s(k) = (1/\pi)\int_0^{2\pi} dv\ \mathrm{se}_r(v,q)\tilde{F}(v,k). \qquad (19.2.21)$$

For the monopole doublet, because the transverse field components are even in z, the functions $\tilde{\tilde{F}}_r^\alpha(k)$ will be real and even in k. Unfortunately, unlike the circular cylinder case for which we were able to find the $\tilde{\tilde{B}}_\rho(R,m,k)$ analytically, see (1.14) through (1.16), we do not have analytic results for the $\tilde{\tilde{F}}_r^\alpha(k)$. However, since we do know $\tilde{F}(v,k)$ analytically, we can compute the $\tilde{\tilde{F}}_r^\alpha(k)$ numerically.[12] With these functions in hand, we will be able to explore how many of them need to be retained in the sums (14.4.85) and (14.4.86).

[10] As emphasized earlier, this lack of analyticity at the origin is related to the fact that $\boldsymbol{B}$ for the monopole doublet falls off only as $|z|^{-3}$ at $\pm\infty$.

[11] Note that, unlike the circular cylinder case where the $(z \to k)$ and $(\phi \to m)$ Fourier transforms can be performed in either order, see Section 16.1.2, in the elliptic cylinder case we must first perform the $(z \to k)$ transform and then the $(v \to r)$ transform. This is because $\mathrm{ce}_r(v,q)$ and $\mathrm{se}_r(v,q)$ depend on k through (14.4.23).

[12] Of course, we will eventually want to demonstrate that we can also compute these functions to high accuracy using only data on grid points.

To see what we might expect for the Mathieu coefficients, $\tilde{\tilde{F}}_r^\alpha(k)$, it is useful to examine the angular dependence of $\tilde{F}(v,k)$. As examples, Figures 2.1 and 2.2 display the functions $\Re\tilde{F}(v,k=0)$ and $\Re\tilde{F}(v,k=20)$. Also shown, in Figure 2.3, is the quantity $\Re\tilde{F}(v=\pi/2,k)$. Note that the maximum amplitude of $\tilde{F}(v,k)$ decreases dramatically as k increases, in accord with (2.19), and the peaks become sharper. As is evident from Figures 2.1 and 2.2, and can also be seen from (2.16), $\tilde{F}(v,k)$ is an odd function of v. Therefore we immediately know, in this monopole doublet case, that

$$\tilde{\tilde{F}}_r^c(k) = 0 \tag{19.2.22}$$

since the $\mathrm{ce}_r(v,q)$ are even functions of v. Moreover, these figures and (2.16) show that $\tilde{F}(v,k)$ is symmetric about the values $v=\pm\pi/2$. Since the $\mathrm{se}_r(v,q)$ for even r are antisymmetric about $v=\pm\pi/2$, we conclude, again in this monopole doublet case, that the only nonvanishing angular integrals will be

$$\tilde{\tilde{F}}_r^s(k) = (1/\pi)\int_0^{2\pi} dv\ \mathrm{se}_r(v,q)\tilde{F}(v,k) \tag{19.2.23}$$

with $r = 1,3,5,7\cdots$.

Finally, what range of q values is of interest? According to (14.4.23) and (2.2), q and k are connected in this instance by the relation

$$q = -k^2f^2/4 = -3k^2. \tag{19.2.24}$$

If we use k values in the range $k\in[-K_c,K_c]$, q will lie in the range $q\in[q_{\min},0]$ with

$$q_{\min} = -3K_c^2. \tag{19.2.25}$$

Since the fields on the surface of the elliptical cylinder are no more singular than those on the surface of the circular cylinder, we could set $K_c=20$ as done before for the circular cylinder case. Doing so yields $q_{\min}=-1200$. Or, being less conservative, we might use $K_c=10$, in which case $q_{\min}=-300$. This is the extreme q value used in making Figures 14.4.13 through 14.4.21.

With this background in mind, examine Figures 2.4 and 2.5. They show the Mathieu coefficients $\tilde{\tilde{F}}_r^s(k)$ as a function of k for the cases $r=1$ through $r=11$ and $r=17$ through $r=25$. The curves for the intervening values $r=13$ and $r=15$ behave analogously. We see that all the $\tilde{\tilde{F}}_r^s(k)$ tend to zero with increasing $|k|$.

This large $|k|$ behavior can be understood as follows: Suppose $f(v)$ and $g(v)$ are any two 2π periodic functions. Into the vector space of such functions introduce a scalar product by the usual rule

$$(f,g) = [1/(2\pi)]\int_0^{2\pi} dv\ \bar{f}(v)g(v) \tag{19.2.26}$$

where a bar denotes complex conjugation. Also define a norm by the usual rule

$$(||f||)^2 = (f,f). \tag{19.2.27}$$

Then, by the Schwarz inequaltiy, there is the result

$$|(f,g)| \le ||f||\ ||g||. \tag{19.2.28}$$

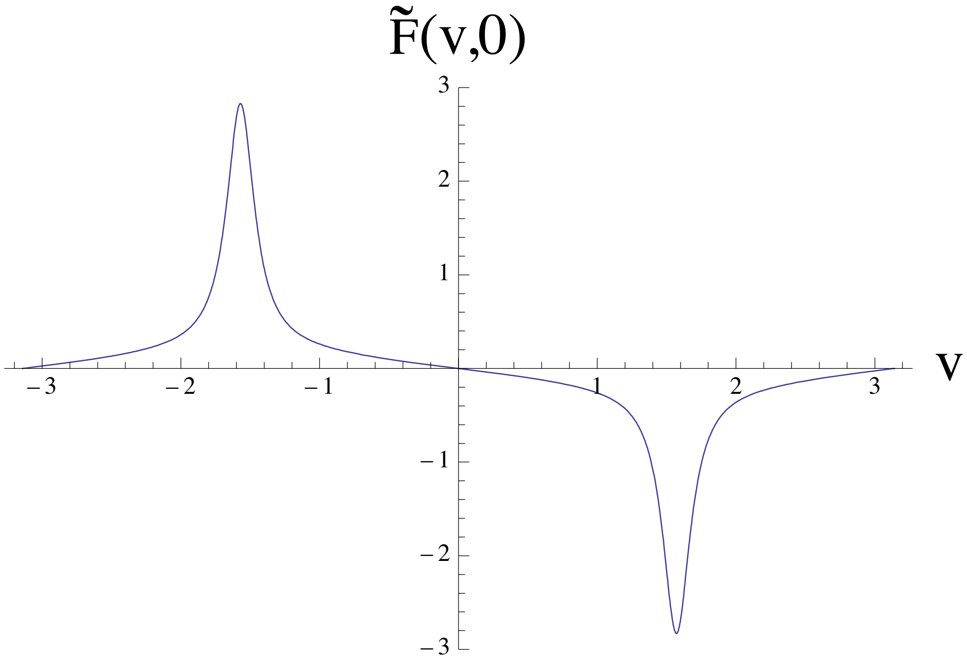


Figure 19.2.1: The real part of $\tilde{F}(v, k = 0)$ for the monopole doublet in the case that $x^{\max} = 4$ cm, $y^{\max} = 2$ cm, $a = 2.5$ cm, and $g = 1$ Tesla-(cm)^2. The imaginary part vanishes.

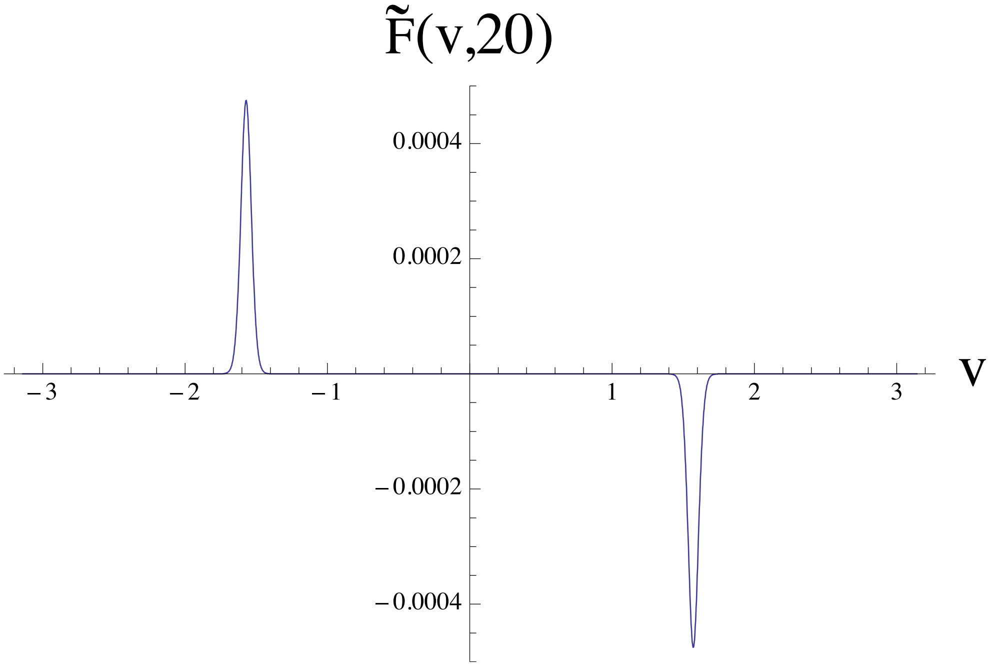


Figure 19.2.2: The real part of $\tilde{F}(v, k = 20)$ for the monopole doublet in the case that $x^{\max} = 4$ cm, $y^{\max} = 2$ cm, $a = 2.5$ cm, and $g = 1$ Tesla-(cm)^2. The imaginary part vanishes.

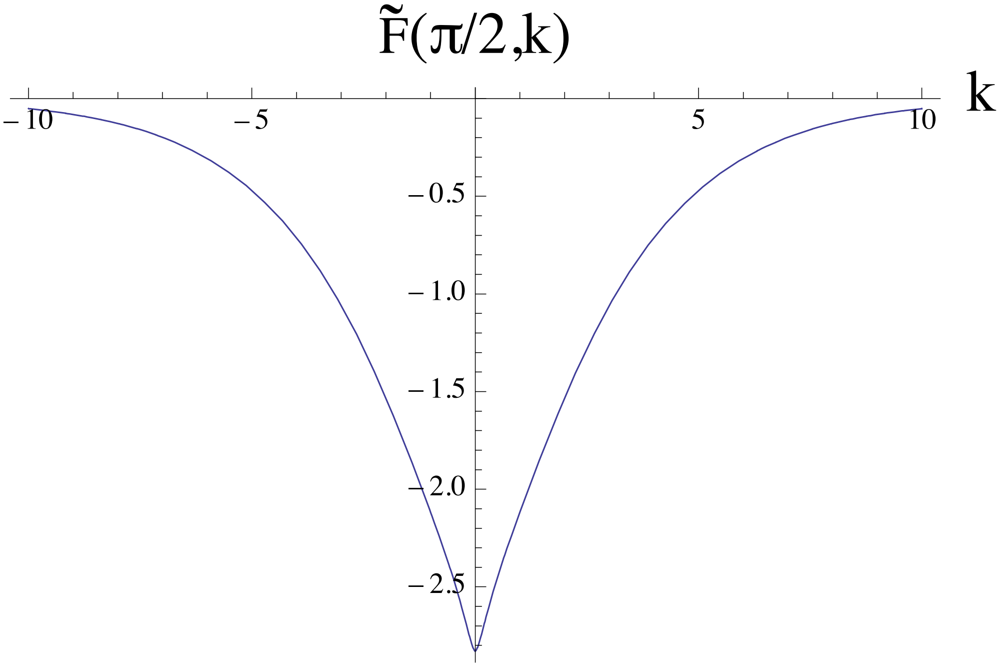


Figure 19.2.3: The real part of $\tilde{F}(v=\pi/2,k)$ for the monopole doublet in the case that $x^{\max}=4$ cm, $y^{\max}=2$ cm, $a=2.5$ cm, and $g=1$ Tesla-(cm)^2. The imaginary part vanishes.

With this notation in mind, we see that (2.23) can be rewritten in the form

$$\tilde{\tilde{F}}_r^s(k)=2(\mathrm{se}_r,\tilde{F}). \tag{19.2.29}$$

Therefore, using (2.28), we have the inequality

$$|\tilde{\tilde{F}}_r^s(k)|\leq 2||\mathrm{se}_r||\ ||\tilde{F}||. \tag{19.2.30}$$

Also, from the normalization (14.4.39), we see that

$$||\mathrm{se}_r||=1/\sqrt{2}. \tag{19.2.31}$$

We conclude that there is the r *independent* bound

$$|\tilde{\tilde{F}}_r^s(k)|\leq\sqrt{2}\ ||\tilde{F}|| \tag{19.2.32}$$

where

$$(||\tilde{F}||)^2=[1/(2\pi)]\int_0^{2\pi}dv\ [\tilde{F}(v,k)]^2. \tag{19.2.33}$$

(See also Exercise 2.1.) Figure 2.6 displays the quantity $\sqrt{2}\ ||\tilde{F}||$ as a function of k. Evidently it also decreases with increasing $|k|$. In fact, we can get a loose bound on $||\tilde{F}||$ from its definition (2.33) by estimating the integral,

$$(||\tilde{F}||)^2=[1/(2\pi)]\int_0^{2\pi}dv\ [\tilde{F}(v,k)]^2\leq[\tilde{F}(\pi/2,k)]^2 \tag{19.2.34}$$

from which it follows that

$$||\tilde{F}|| \leq |\tilde{F}(\pi/2, k)|. \tag{19.2.35}$$

Here we have used the fact that $[\tilde{F}(v, k)]^2$ takes its maxima at $v = \pm\pi/2$. See Figures 2.1 and 2.2. We already know that $|\tilde{F}(\pi/2, k)|$ decreases exponentially with increasing $|k|$. Again see Figure 2.3 and (2.16) through (2.19). We conclude that the $|\tilde{\tilde{F}}_r^s(k)|$ must all decrease at least this rapidly as well.

In fact there are two reasons why, for fixed r, the $|\tilde{\tilde{F}}_r^s(k)|$ must eventually decrease even more rapidly as $|k| \to \infty$. First, as comparison of Figures 2.1 and 2.2 indicates, the peaks in $\tilde{F}(v, k)$ at $v = \pm\pi/2$ become more narrow with increasing $|k|$. Second, look at (2.21). We expect that the greatest contribution to this integral will come from $v = \pm\pi/2$ because $\tilde{F}(v, k)$ is peaked there. But it is precisely at these v values that the $\mathrm{se}_r(v, q)$ vanish rapidly as $q \to -\infty$ since these v values are in the middle of the forbidden region. See, for example, Figures 14.4.19 and 14.4.20. To observe these considerations in action for the case of $|\tilde{\tilde{F}}_r^s(k)|$ with $r = 29$, see Figure 2.30 in Exercise 2.4.

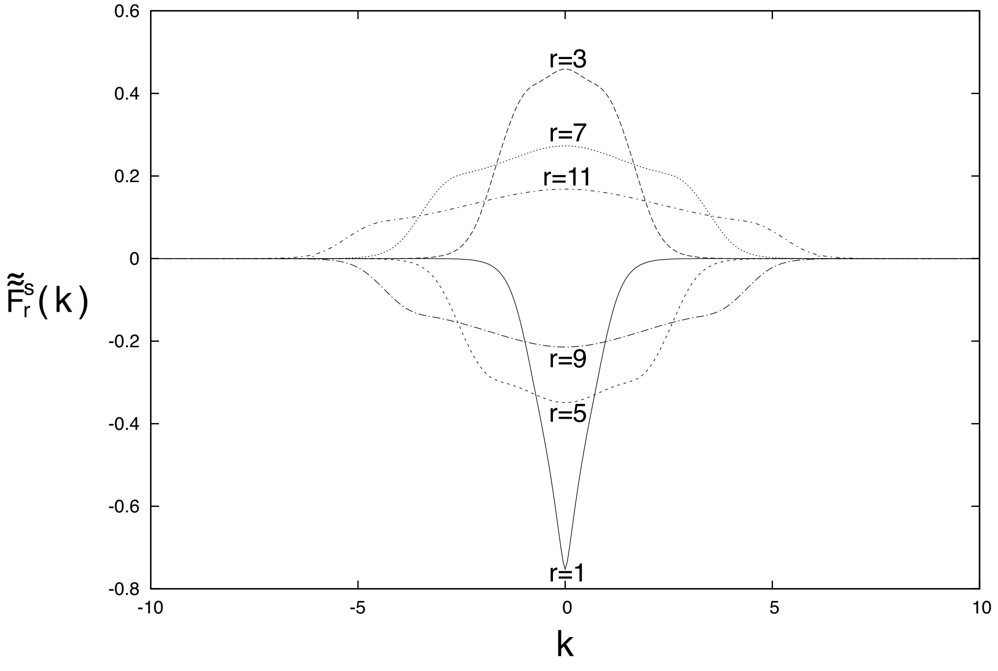


Figure 19.2.4: The real parts of the Mathieu coefficients $\tilde{\tilde{F}}_r^s(k)$ as a function of k, with $r = 1, 3, 5, 7, 9, 11$, for the monopole doublet in the case that $x^{\max} = 4$ cm, $y^{\max} = 2$ cm, $a = 2.5$ cm, and $g = 1$ Tesla-(cm)^2. The imaginary parts vanish. The solid curve, the one with the largest negative excursion at $k = 0$, is that for $r = 1$. The curves alternate in sign, and the magnitudes of their values at $k = 0$ decrease, for each successive value of r. For example, the curve with the largest positive excursion at $k = 0$ is that for $r = 3$.

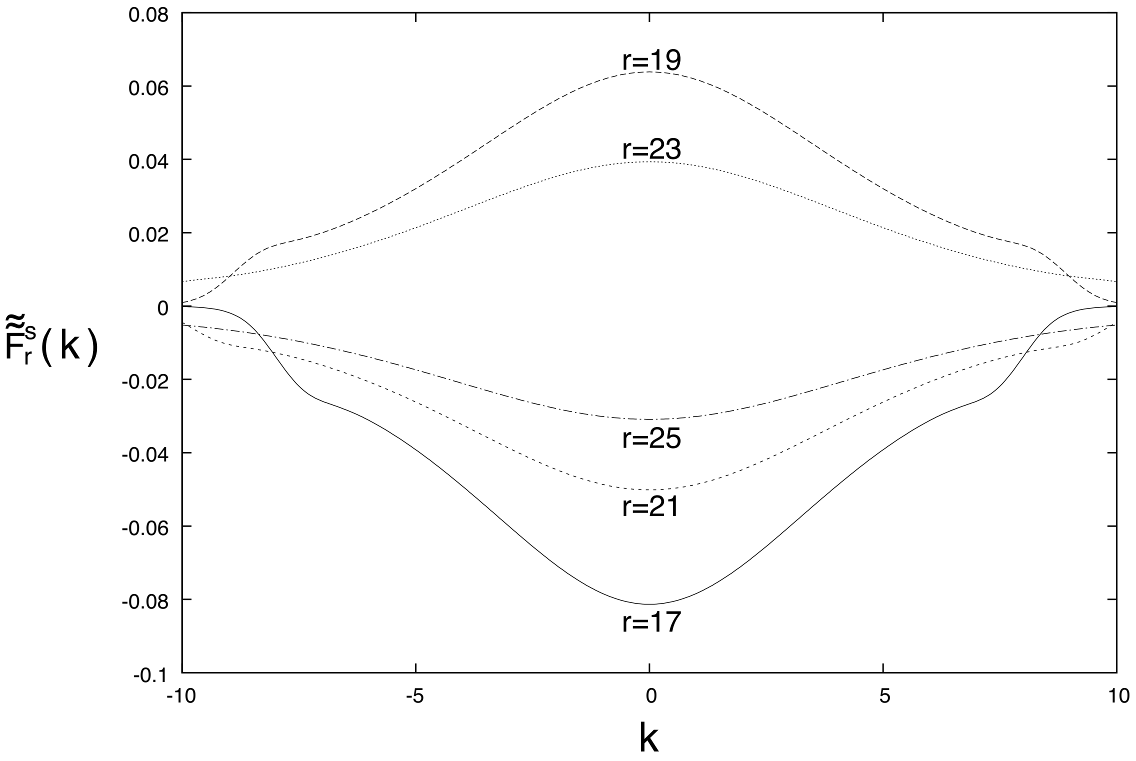


Figure 19.2.5: The real parts of the Mathieu coefficients $\tilde{\tilde{F}}^s_r(k)$ as a function of k, with $r = 17, 19, 21, 23, 25$, for the monopole doublet in the case that $x^{\max} = 4$ cm, $y^{\max} = 2$ cm, $a = 2.5$ cm, and $g = 1$ Tesla-(cm)^2. The imaginary parts vanish. The solid curve, the one with the largest negative excursion at $k = 0$, is that for $r = 17$. The curves alternate in sign, and the magnitudes of their values at $k = 0$ decrease for each successive value of r. For example, the curve with the largest positive excursion at $k = 0$ is that for $r = 19$.

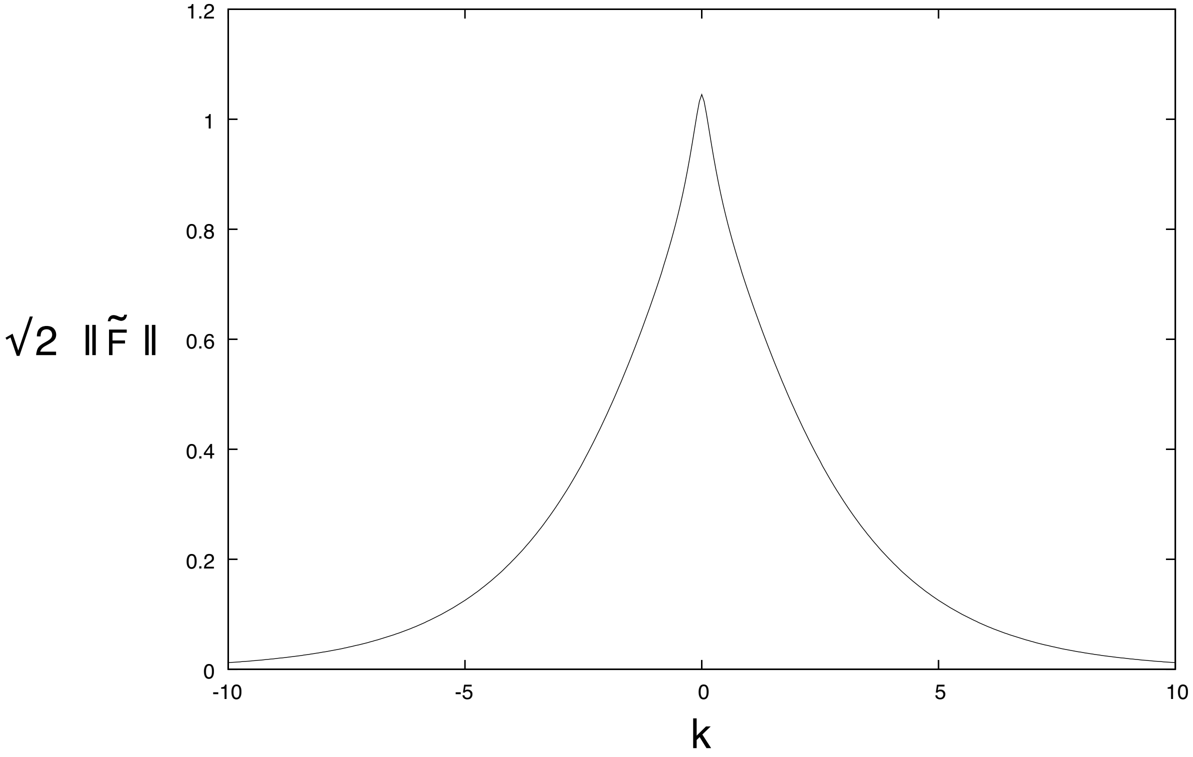


Figure 19.2.6: The quantity $\sqrt{2}\ ||\tilde{F}||$ as a function of k for the monopole doublet in the case that $x^{\max} = 4$ cm, $y^{\max} = 2$ cm, $a = 2.5$ cm, and $g = 1$ Tesla-(cm)^2.

19.2.2 Behavior of Kernels

We have determined the behavior of the Mathieu coefficients $\tilde{\tilde{F}}_r^s(k)$ for the case of the monopole doublet. In analogy with our previous discussion of the circular cylinder case in Section 16.1.1, let us next examine the kernels $k^m\beta_m^r(k)/\mathrm{Se}_r'(U,q)$ that appear in (14.4.86). Figure 2.7 shows the kernels $k^m\beta_m^r(k)/\mathrm{Se}_r'(U,q)$ for the case $m=1$ and $r=1,3,5,7,9,11$. We see that each kernel has constant sign and the absolute value of each goes monotonically to 0 as $|k|\to\infty$. Also, as Figure 2.8 shows, their absolute values at $k=0$ go monotonically to zero as r increases.

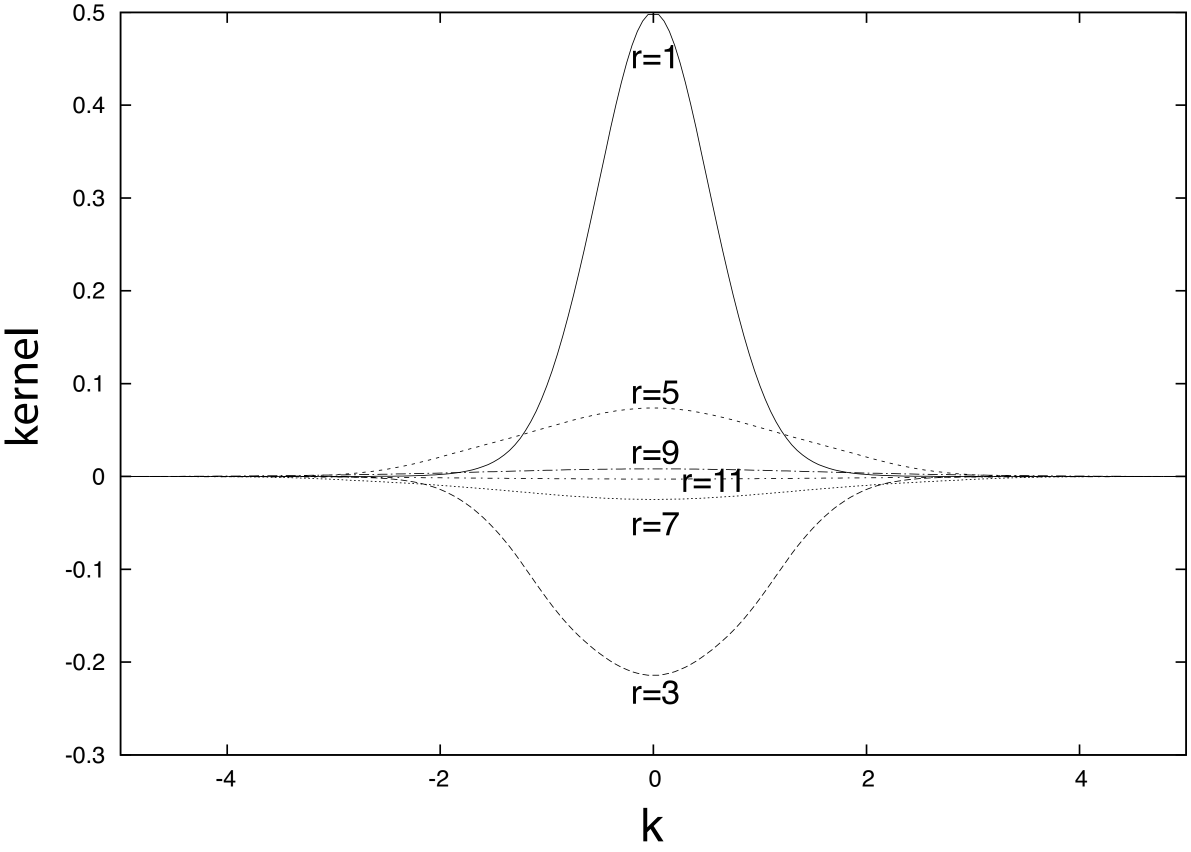


Figure 19.2.7: The kernels $k^m\beta_m^r(k)/\mathrm{Se}_r'(U,q)$ for the case $m=1$ and $r=1,3,5,7,9,11$, as a function of k, with q and k related by (2.24) and U given by (2.1). The kernel for $r=1$ is the one with the largest positive value at $k=0$. Kernels for successive values of r alternate in sign. Their absolute values at $k=0$ decrease monotonically with increasing r.

Figure 2.9 shows the kernels $k^m\beta_m^r(k)/\mathrm{Se}_r'(U,q)$ for the case $m=7$ and $r=1,3,5$. Figure 2.10 shows the kernels $k^m\beta_m^r(k)/\mathrm{Se}_r'(U,q)$ for the case $m=7$ and $r=7,9,11$. Figure 2.11 shows the kernels $k^m\beta_m^r(k)/\mathrm{Se}_r'(U,q)$ for the case $m=7$ and $r=13,15,17,21,23$. We see that the kernels for $r<m$ vanish at $k=0$. Those with $r\geq m$ are finite at $k=0$, but ultimately go to 0 as $r\to\infty$. See Figure 2.12. Finally, all kernels go rapidly to zero as $|k|\to\infty$. We have documented the extreme cases $m=1$ and $m=7$. The intermediate cases $m=3$ and $m=5$ are similar to the $m=7$ case: the kernels for $r<m$ vanish at

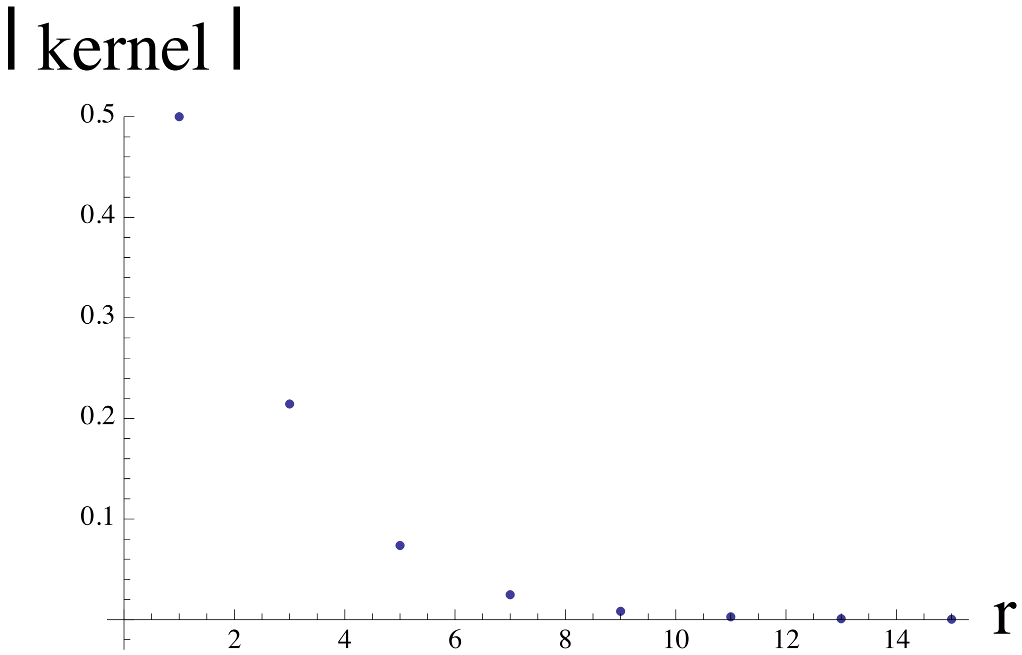


Figure 19.2.8: Absolute values of the kernels $k^m\beta^r_m(k)/\mathrm{Se}'_r(U,q)$ evaluated at $k=0$ for the case $m=1$ and $r\in[1,15]$ with U given by (2.1).

$k=0$; those with $r\geq m$ are finite at $k=0$, but ultimately go to 0 as $r\to\infty$; all kernels go rapidly to zero as $|k|\to\infty$.

19.2.3 Truncation of Series

We have studied the Mathieu coefficients $\tilde{\tilde{F}}^s_r(k)$ and the kernels $k^m\beta^r_m(k)/\mathrm{Se}'_r(U,q)$. Next, we need to study their combinations as they occur in (14.4.84) and (14.4.86). In particular, let us look at the quantities $(1/2)^m(1/m!)k^mG_{m,s}(k)$. Of course, in the monopole doublet case, we also know exactly what the result should be. Comparison of (14.4.86) and (1.1) gives the relation

$$(1/2)^m(1/m!)k^mG_{m,s}(k)=\tilde{C}^{[0]}_{m,s}(k), \tag{19.2.36}$$

and we know the right side of (2.36) from (1.20).

Figure 2.13 shows their values for the cases $m=1,3,5,7$. Figures 2.14 through 2.16 show their values for the cases $m=3,5,7$ separately. These are the cases that we need for our magnet monopole doublet example.

As described in the beginning of this subsection, we must truncate the infinite sums over r that occur in (14.4.83) through (14.4.86) in order to obtain practical results. We will do this by assuming that the truncation error is comparable to the size of the last retained term.[13] By this criterion, we retained all terms with values of r through $r=r_{\max}(m)$ with $r_{\max}(m)=11,19,25,29$ for the cases $m=1,3,5,7$, respectively. Figures 2.17 through

[13] This assumption is justified because both the Mathieu coefficients $\tilde{\tilde{F}}^s_r(k)$ and the kernels $k^m\beta^r_m(k)/\mathrm{Se}'_r(U,q)$ fall off exponentially in r for large r.

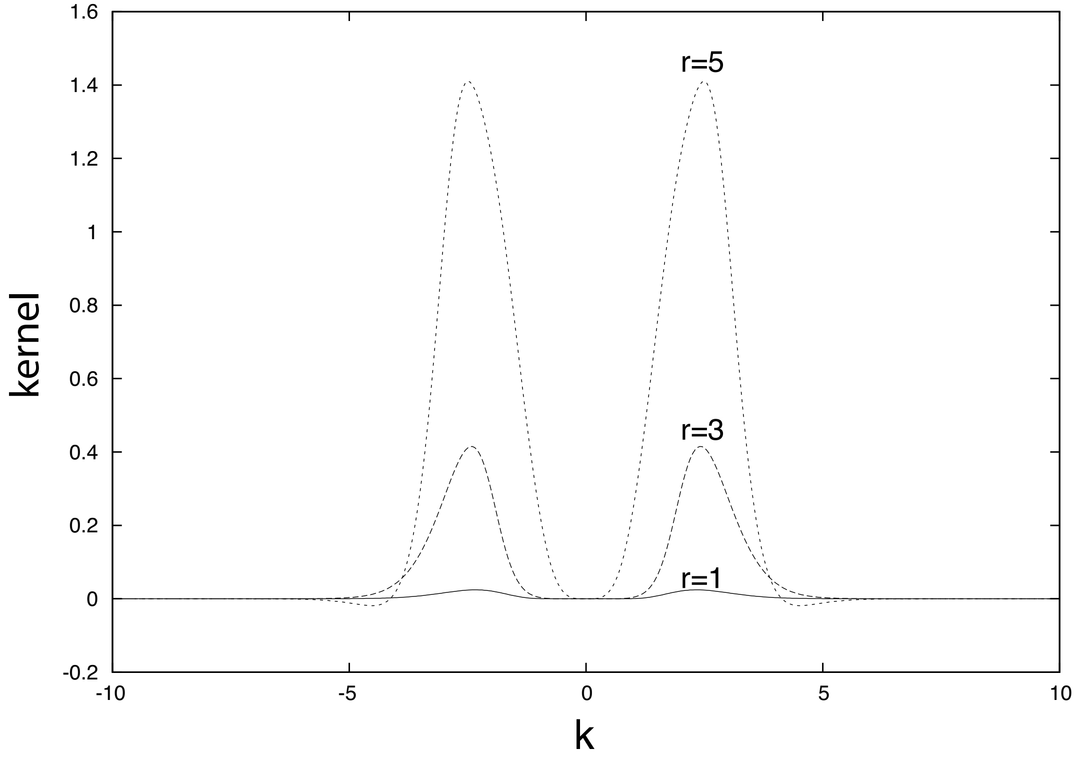


Figure 19.2.9: The kernels $k^m \beta_m^r(k)/\mathrm{Se}_r'(U,q)$ for the case $m = 7$ and $r = 1, 3, 5$, as a function of k, with q and k related by (2.24) and U given by (2.1). The kernel that has the largest positive value is that for $r = 5$. The kernel with the next largest positive value is that for $r = 3$. The remaining kernel is that for $r = 1$.

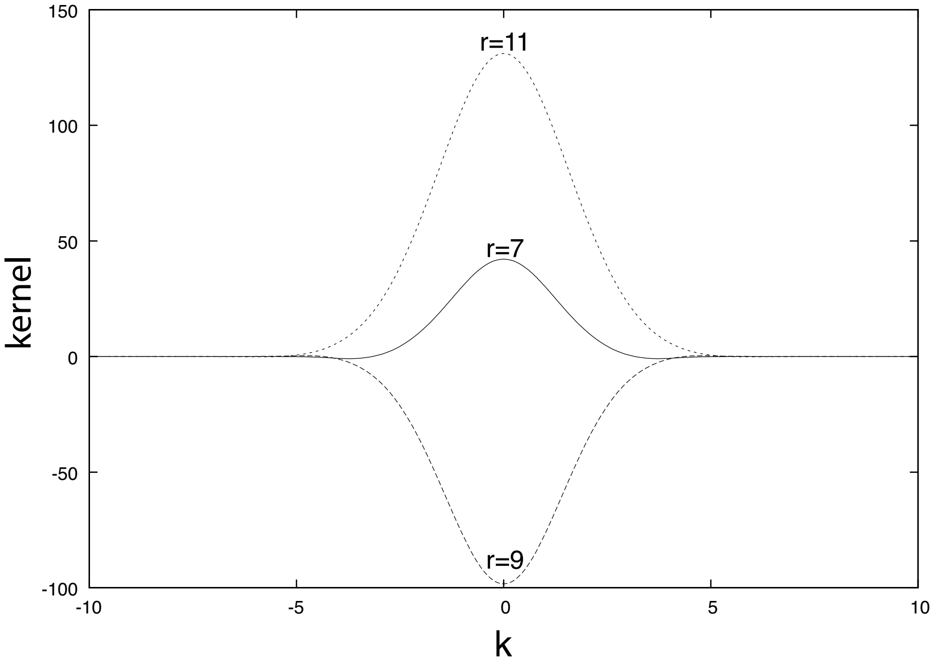


Figure 19.2.10: The kernels $k^m \beta_m^r(k)/\mathrm{Se}_r'(U,q)$ for the case $m = 7$ and $r = 7, 9, 11$, as a function of k, with q and k related by (2.24) and U given by (2.1). The kernel for $r = 7$ is the one with the smallest positive value at $k = 0$. Kernels for successive values of r alternate in sign. Their magnitudes at $k = 0$ increase monotonically with increasing r in the range $r \in [7, 11]$.

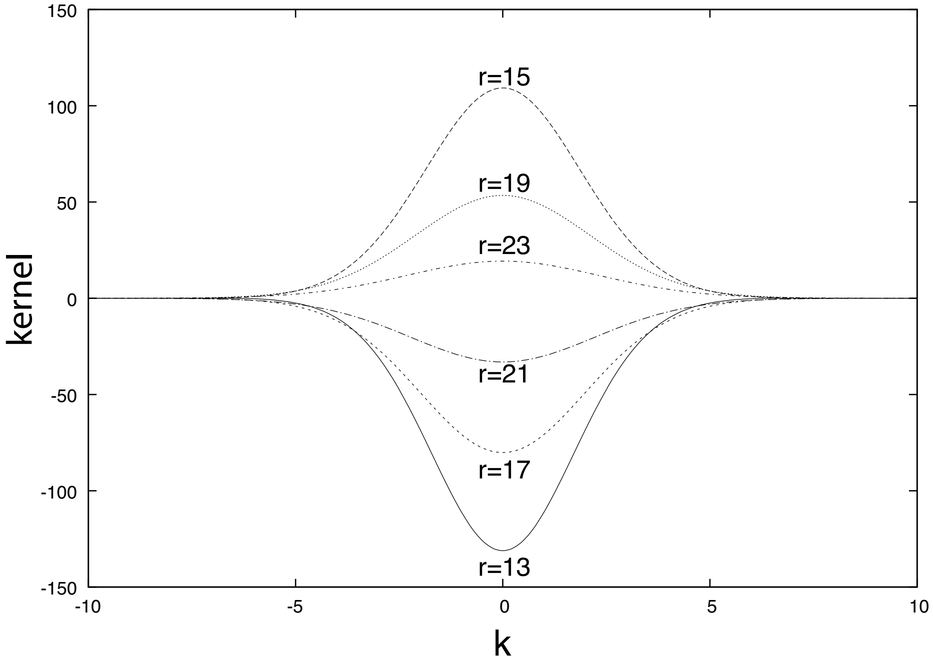


Figure 19.2.11: The kernels $k^m\beta^r_m(k)/\mathrm{Se}'_r(U,q)$ for the case $m = 7$ and $r = 13, 15, 17, 19, 21, 23$, as a function of k, with q and k related by (2.24) and U given by (2.1). The kernel for $r = 13$ is the one with the largest negative value at $k = 0$. Kernels for successive values of r alternate in sign. Their magnitudes at $k = 0$ decrease monotonically with increasing r in the range $r \in [13, 23]$.

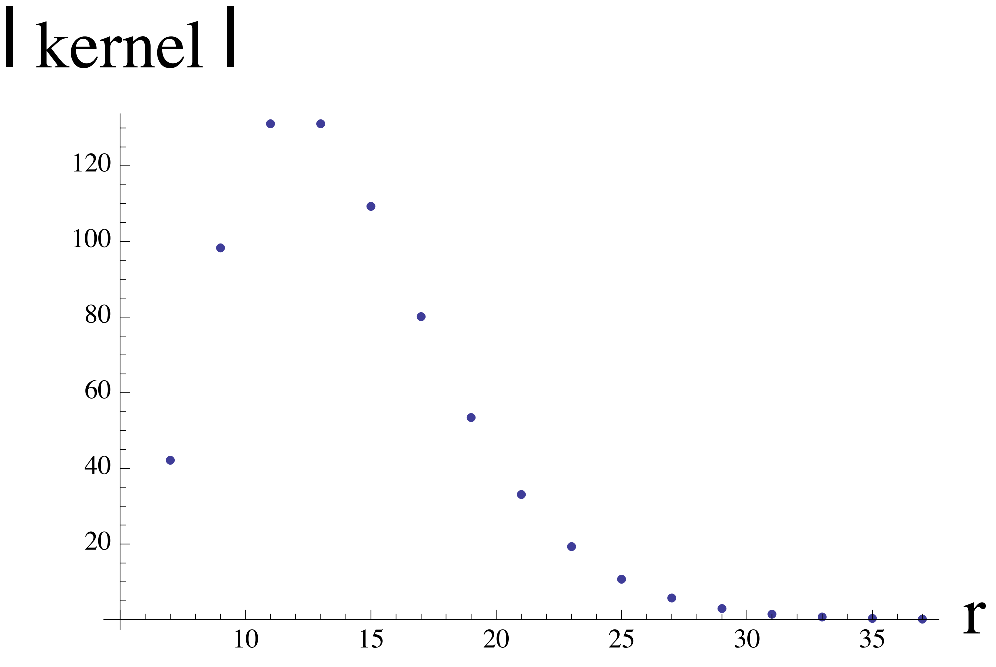


Figure 19.2.12: Absolute values of the kernels $k^m\beta^r_m(k)/\mathrm{Se}'_r(U,q)$ evaluated at $k = 0$ for the case $m = 7$ and $r \in [7, 37]$ with U given by (2.1).

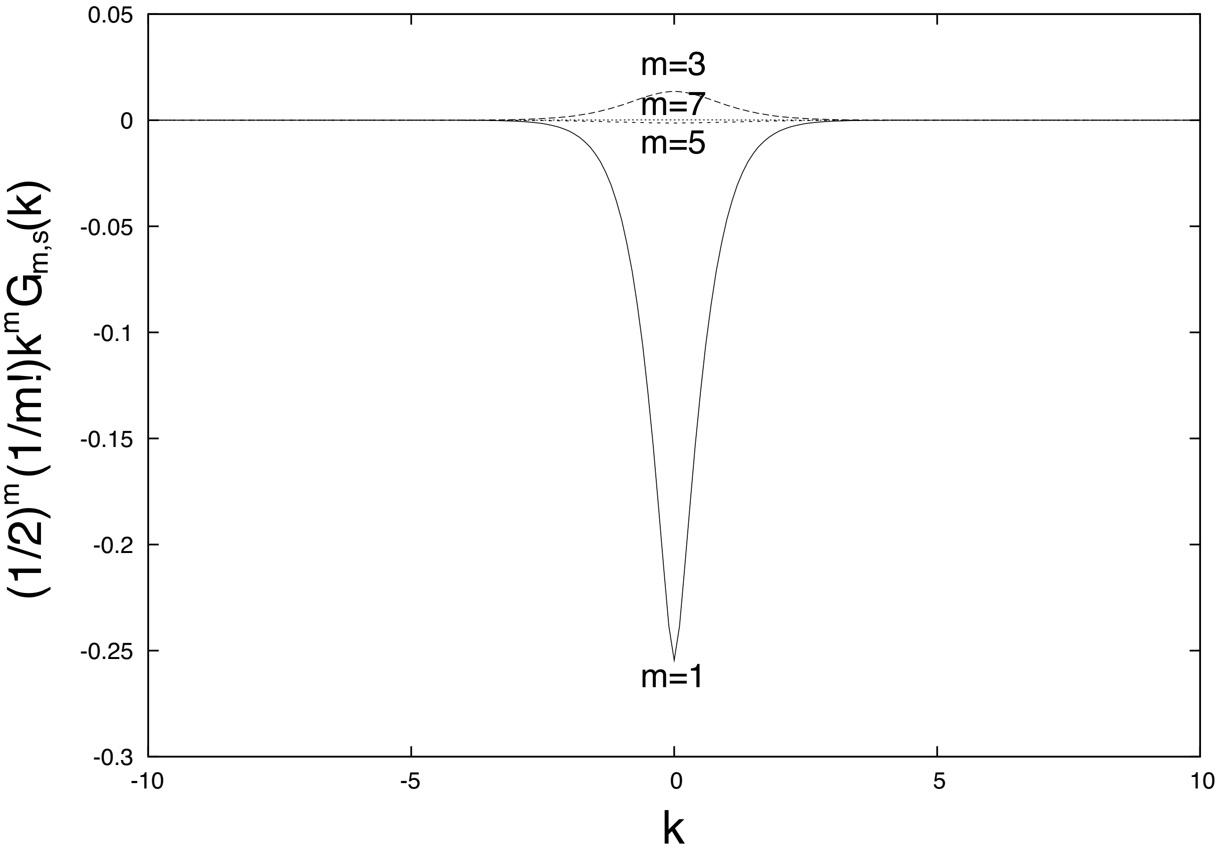


Figure 19.2.13: The real parts of $(1/2)^m(1/m!)k^m G_{m,s}(k)$ for the monopole doublet when $m = 1, 3, 5, 7$. The imaginary parts vanish. The quantities decrease in magnitude with increasing m. For example, the curve with the largest negative value at $k = 0$ is that for $m = 1$.

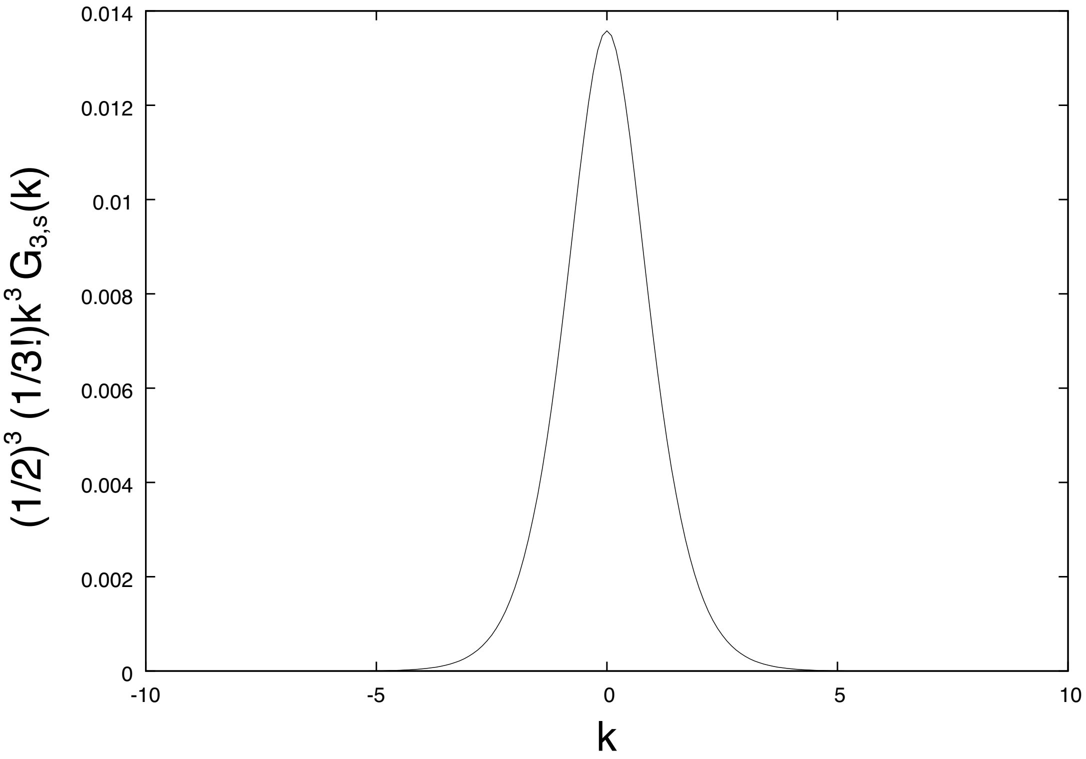


Figure 19.2.14: The real part of $(1/2)^m(1/m!)k^m G_{m,s}(k)$ for $m = 3$.

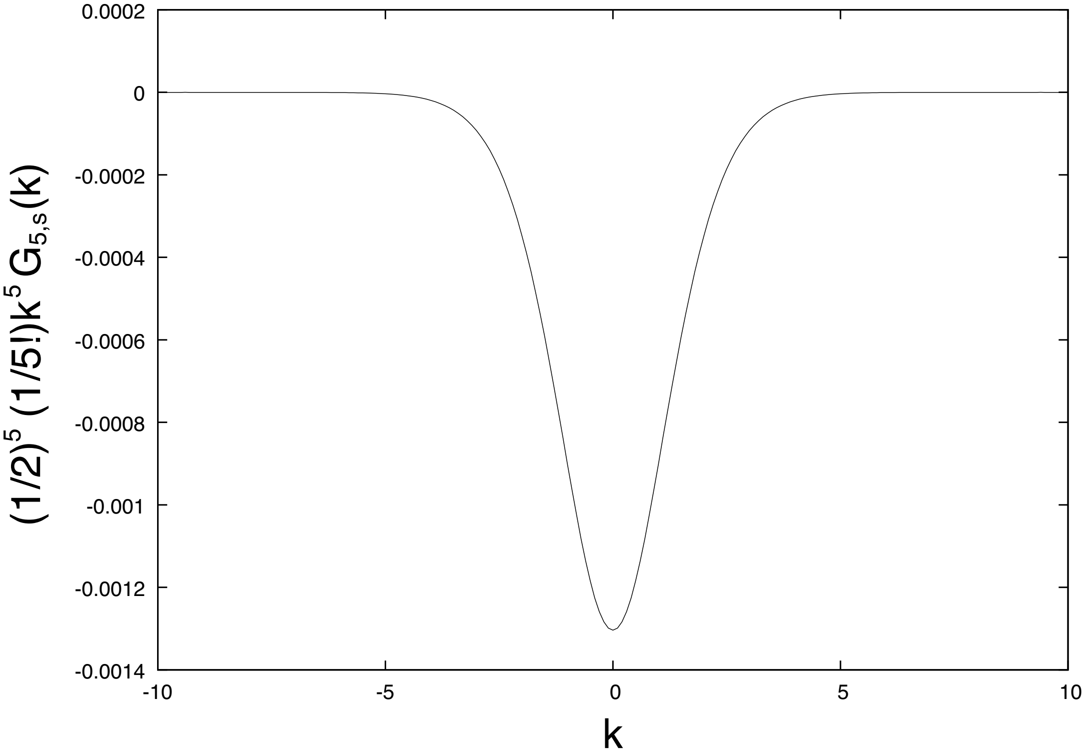


Figure 19.2.15: The real part of $(1/2)^m(1/m!)k^m G_{m,s}(k)$ for $m = 5$.

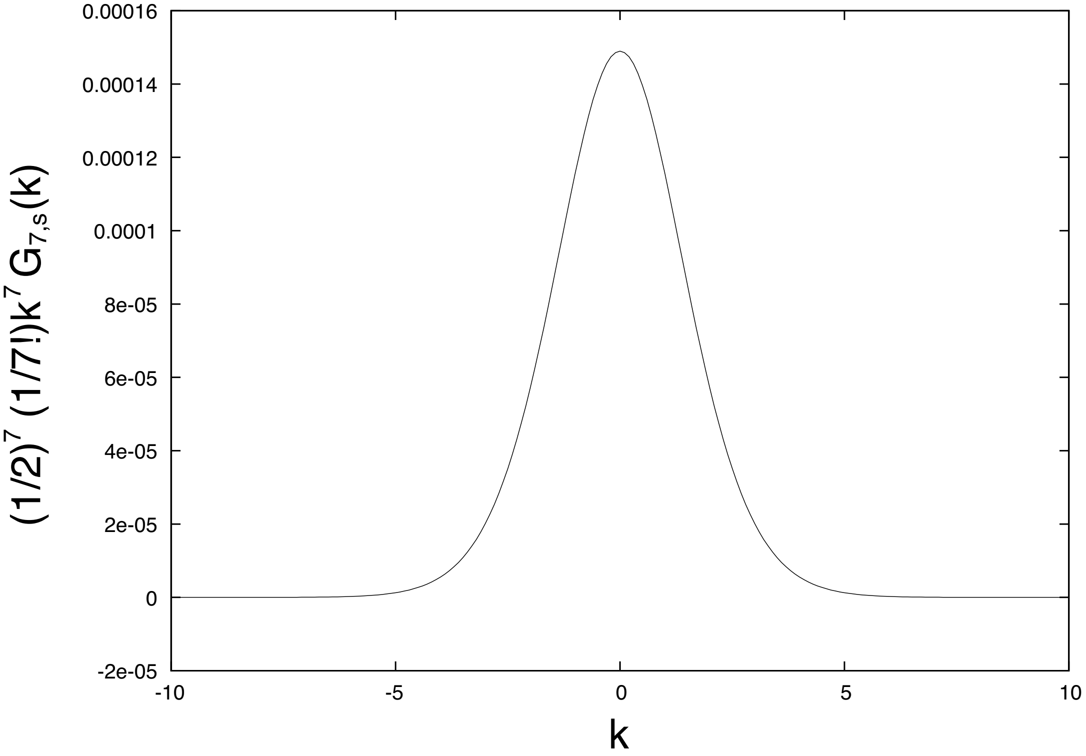


Figure 19.2.16: The real part of $(1/2)^m(1/m!)k^mG_{m,s}(k)$ for $m = 7$.

2.20 show the last retained term in each case. Judging from Figures 2.13 through 2.16, we estimate the errors to be 10,4,3,6 parts in 10^4 for the cases $m = 1, 3, 5, 7$ respectively.

Figures 2.21 through 2.24 show the actual truncation error, the difference between the truncated result and the exact result (2.36). Comparison of Figures 2.13 through 2.16 and Figures 2.21 through 2.24 shows that the actual truncation errors are 3,2,2,4 parts in 10^4 for $m = 1, 3, 5, 7$, respectively.

19.2.4 Approximation of Angular Integrals by Riemann Sums

At this point we have verified that the truncation criterion is adequate and that the routines used to compute the Mathieu functions and the Mathieu-Bessel function connection coefficients are working properly. The next step is to explore the replacement of angular integrals by Riemann sums.

Select N discrete v values v_n according to the rule

$$v_n = n(2\pi/N) \text{ for } n = 0, 1, \cdots, N-1. \tag{19.2.37}$$

What we want to do is to approximate the integrals (2.20) and (2.21) by the Riemann sums

$$\tilde{\tilde{F}}_r^c(k) \approx (2/N)\sum_{n=0}^{N-1} \mathrm{ce}_r(v_n, q)\tilde{F}(v_n, k), \tag{19.2.38}$$

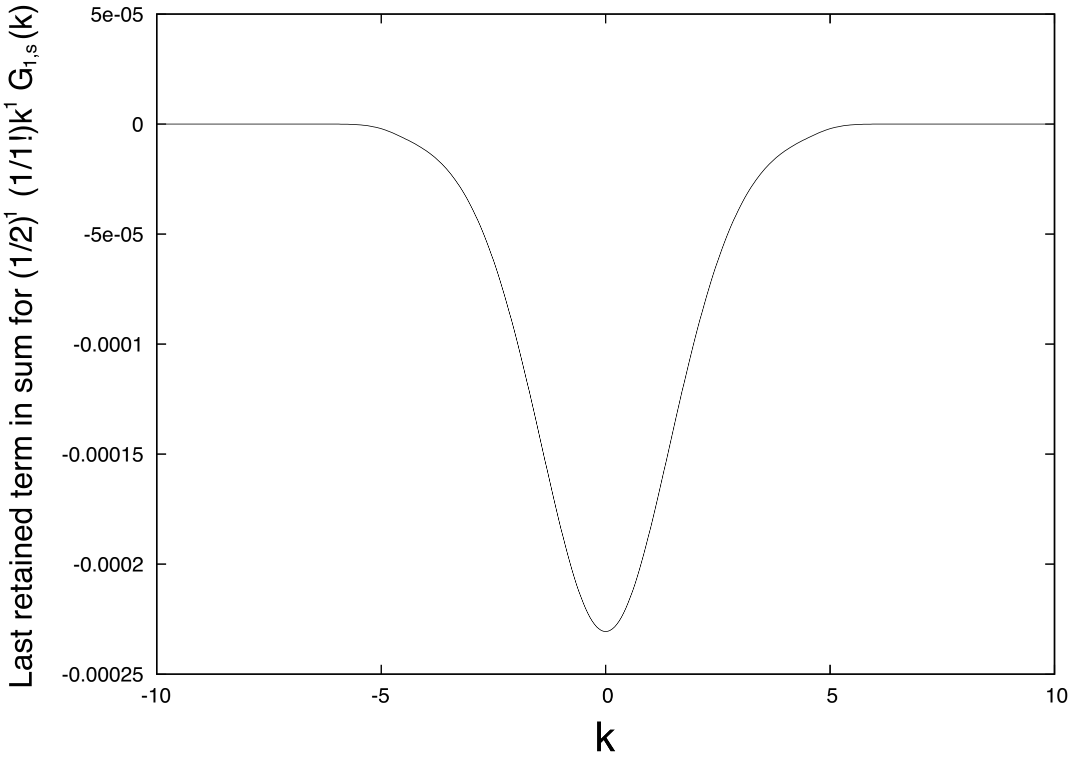


Figure 19.2.17: The real part of the last retained term in the sum for $(1/2)^m(1/m!)k^mG_{m,s}(k)$ with $m = 1$ based on truncating the series (14.4.84) beyond $r = r_{\max}(1) = 11$. The imaginary part vanishes.

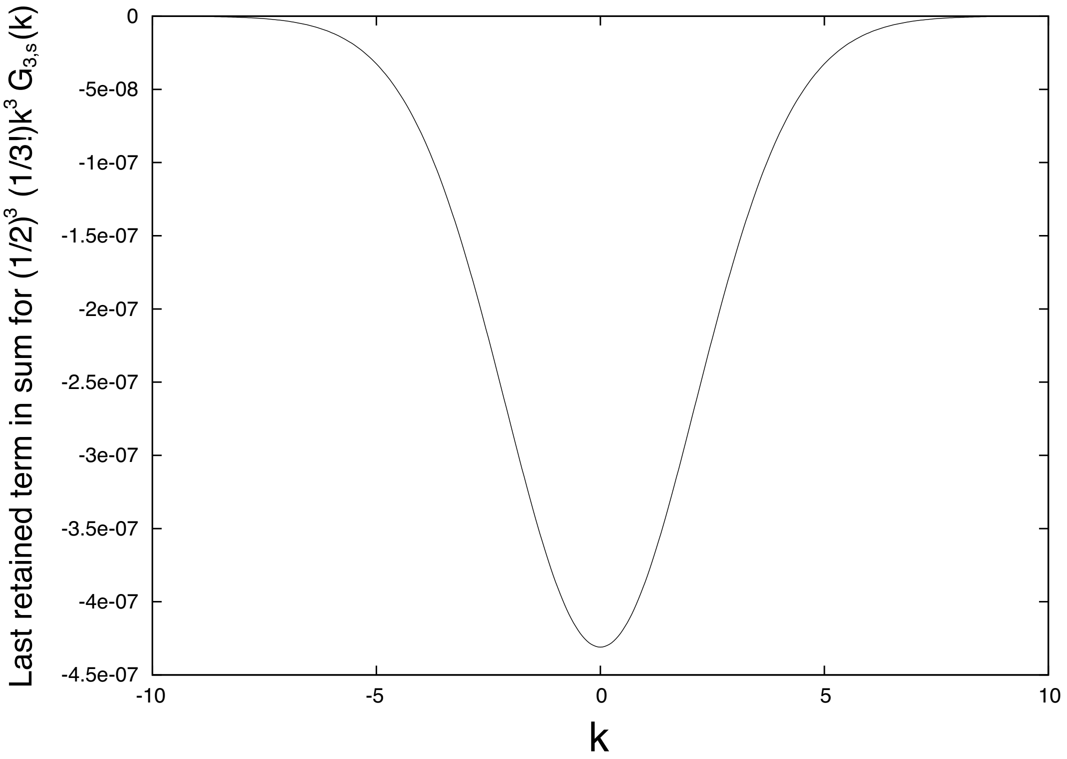


Figure 19.2.18: The real part of the last retained term in the sum for $(1/2)^m(1/m!)k^mG_{m,s}(k)$ with $m = 3$ based on truncating the series (14.4.84) beyond $r = r_{\max}(3) = 19$. The imaginary part vanishes.

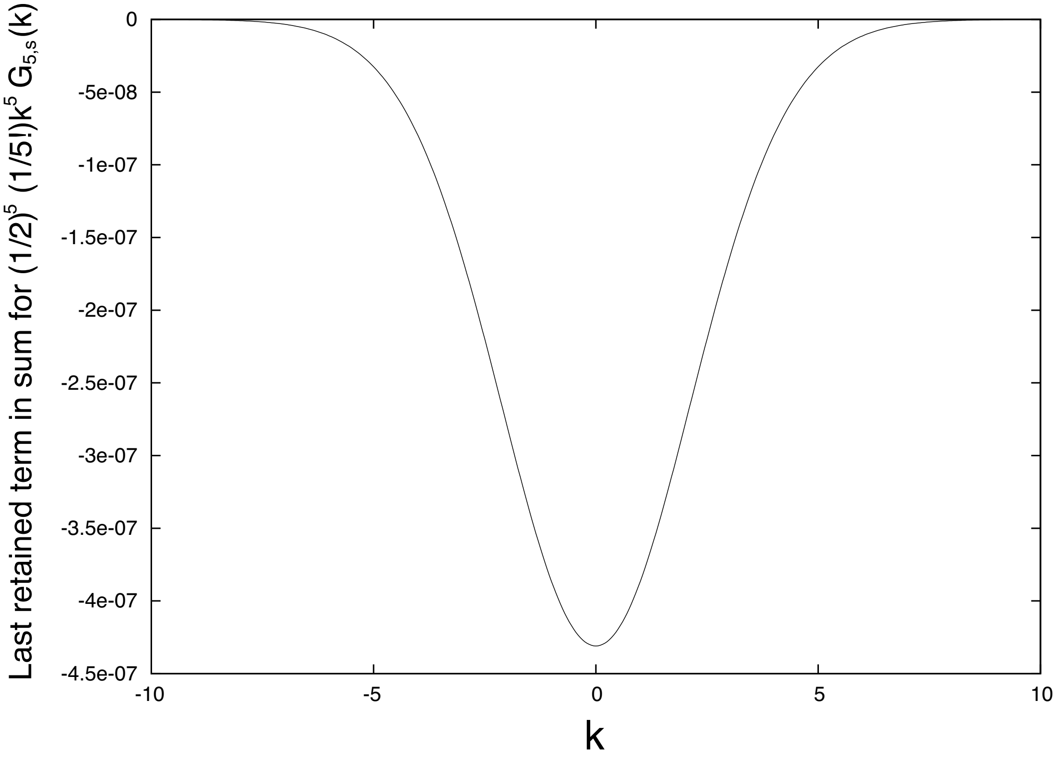


Figure 19.2.19: The real part of the last retained term in the sum for $(1/2)^m(1/m!)k^mG_{m,s}(k)$ with $m = 5$ based on truncating the series (14.4.84) beyond $r = r_{\max}(5) = 25$. The imaginary part vanishes.

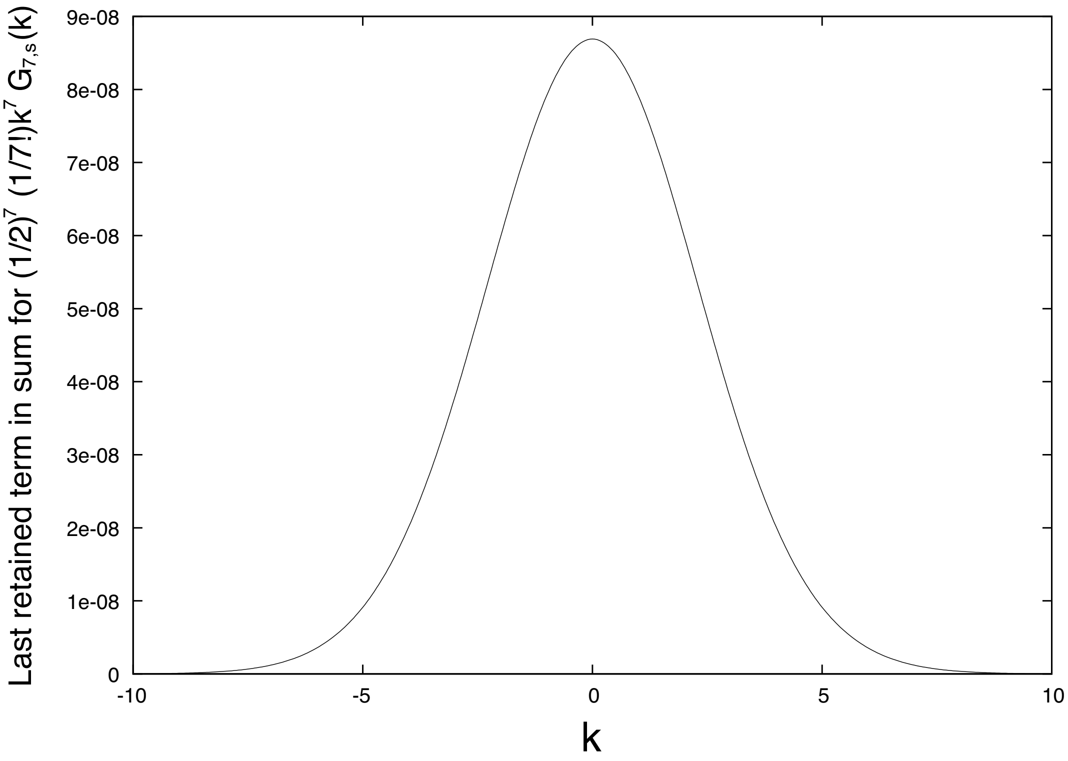


Figure 19.2.20: The real part of the last retained term in the sum for $(1/2)^m(1/m!)k^m G_{m,s}(k)$ with $m = 7$ based on truncating the series (14.4.84) beyond $r = r_{\max}(7) = 29$. The imaginary part vanishes.

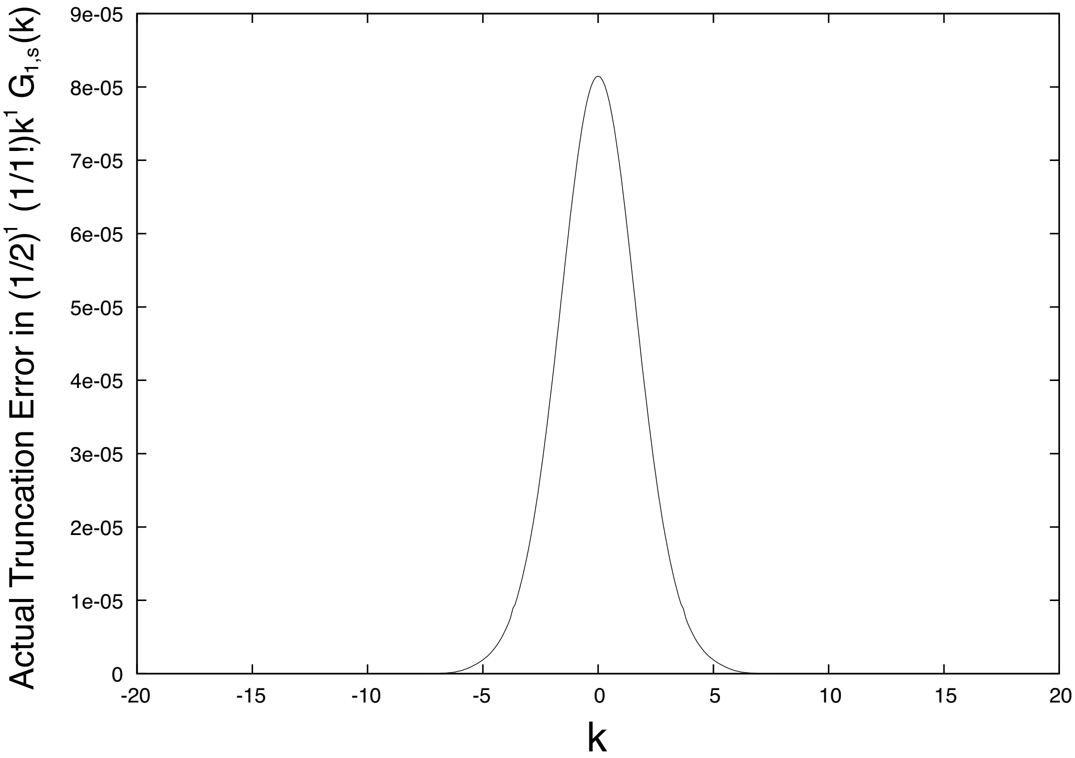


Figure 19.2.21: Real part of actual truncation error in $(1/2)^m(1/m!)k^m G_{m,s}(k)$ for $m = 1$ produced by truncating the series (14.4.84) beyond $r = r_{\max}(1) = 11$. The imaginary part vanishes.

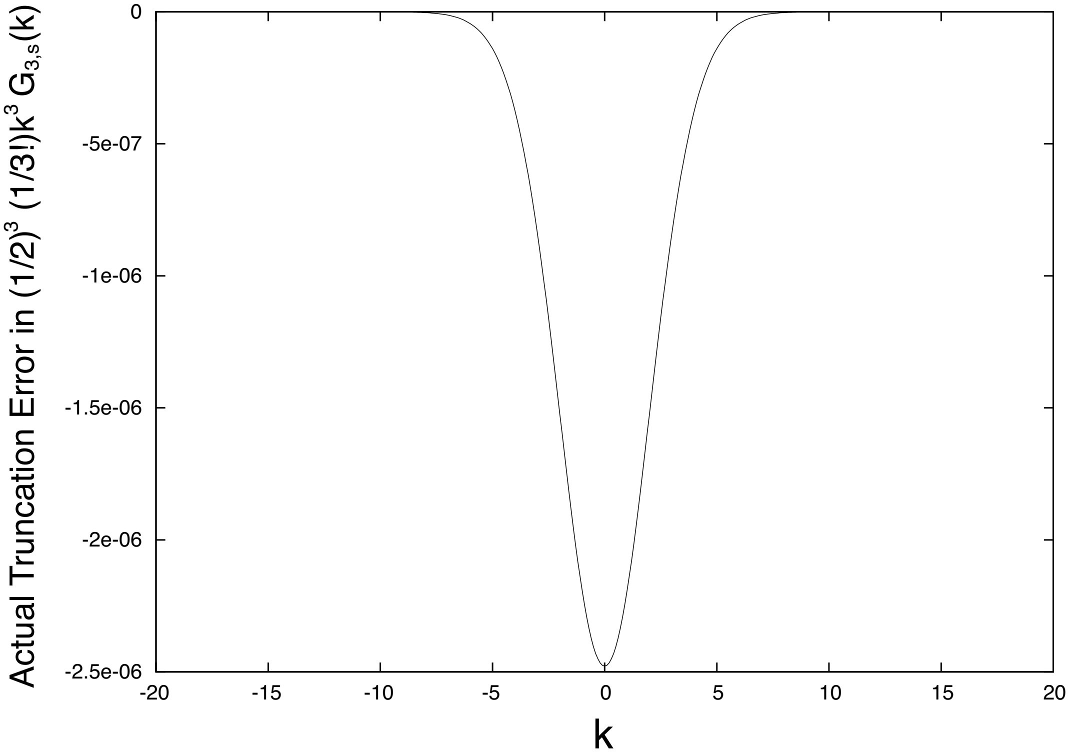


Figure 19.2.22: Real part of actual truncation error in $(1/2)^m(1/m!)k^mG_{m,s}(k)$ for $m = 3$ produced by truncating the series (14.4.84) beyond $r = r_{\max}(3) = 19$. The imaginary part vanishes.

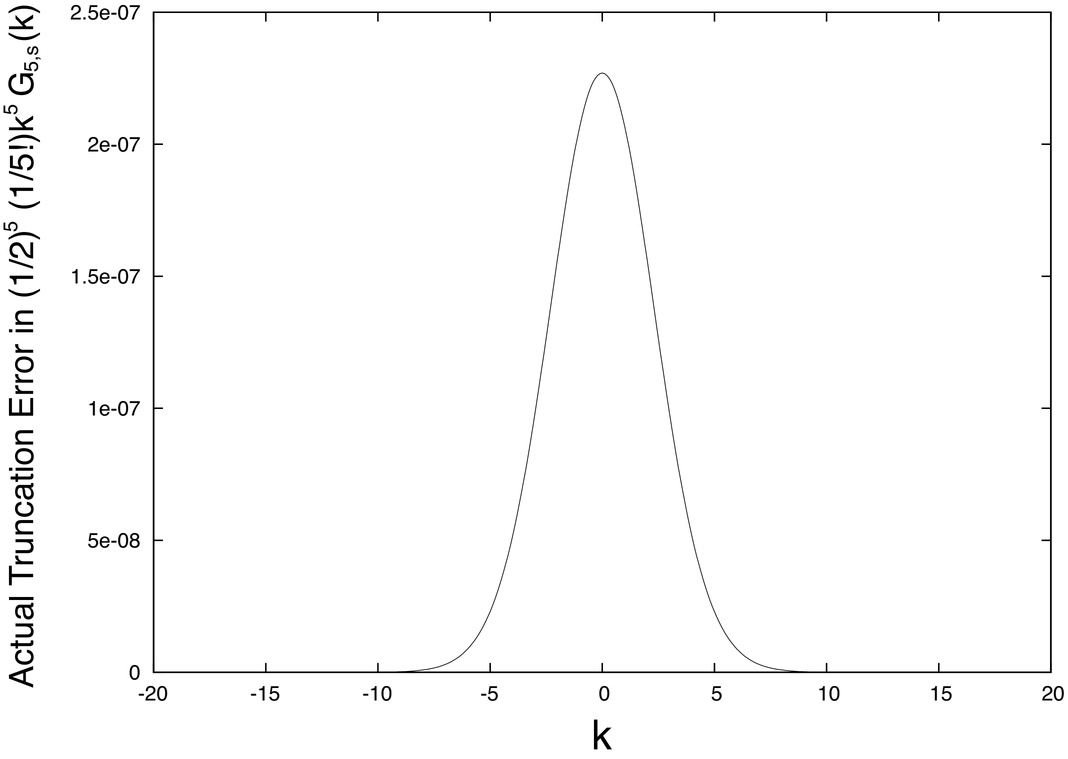


Figure 19.2.23: Real part of actual truncation error in $(1/2)^m(1/m!)k^m G_{m,s}(k)$ for $m = 5$ produced by truncating the series (14.4.84) beyond $r = r_{\max}(5) = 25$. The imaginary part vanishes.

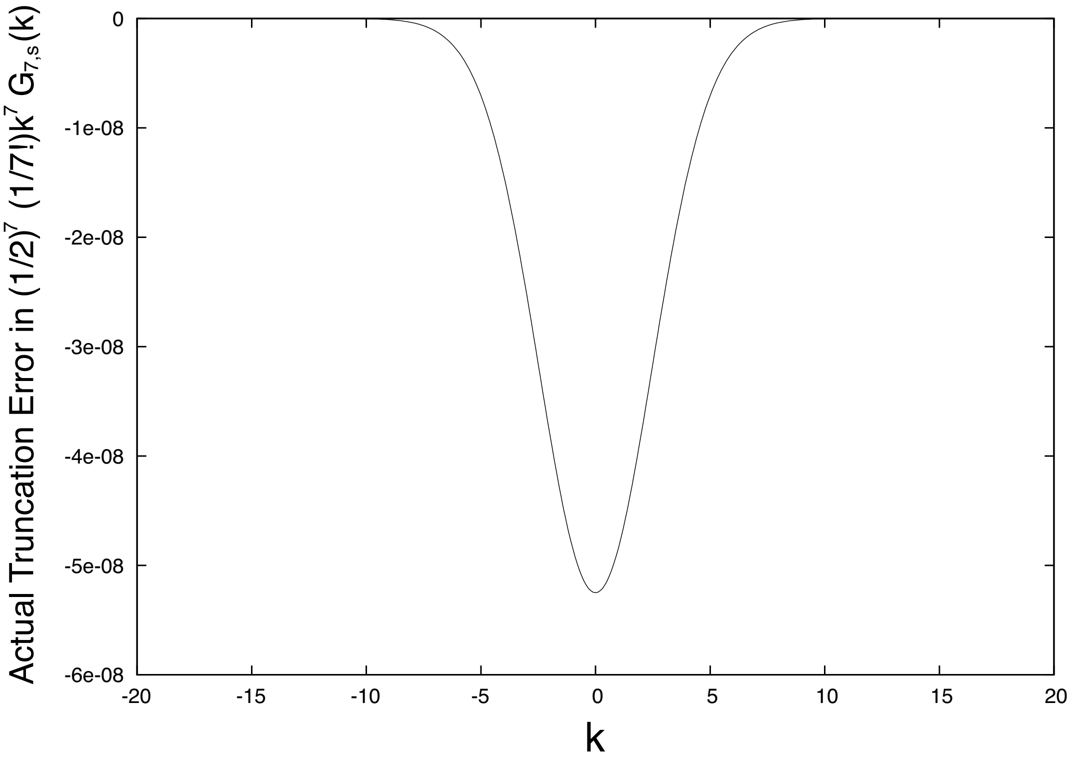


Figure 19.2.24: Real part of actual truncation error in $(1/2)^m(1/m!)k^mG_{m,s}(k)$ for $m = 7$ produced by truncating the series (14.4.84) beyond $r = r_{\max}(7) = 29$. The imaginary part vanishes.

$$\tilde{\tilde{F}}_r^s(k) \approx (2/N) \sum_{n=0}^{N-1} \mathrm{se}_r(v_n, q)\tilde{F}(v_n, k). \tag{19.2.39}$$

The accuracy of these approximations depends on the large m behavior of the Fourier coefficients for $\tilde{F}(v, k)$ and the Fourier coefficients for the Mathieu functions $\mathrm{ce}_r(v, q)$ and $\mathrm{se}_r(v, q)$. See Exercise 2.2. There it is shown that (2.38) and (2.39) are good approximations providing the Fourier coefficients are sufficiently small for $m \geq N$. For the monopole doublet case this proves to be true for $N \geq 120$. See Exercises 2.3 and 2.4. In that case the use of (2.38) and (2.39) with $N = 120$ is accurate to 5 parts in 10^4 for all k and r values of interest. Indeed, the $\tilde{\tilde{F}}_r^s(k)$ used in making Figures 2.4 and 2.5 were obtained using (2.39) with $N = 120$.

19.2.5 Further Tests

Following the pattern established in the circular cylinder case, the next items we should verify are these:

- So far, we have been using exact results for $\tilde{F}(v_n, k)$. We should verify that sufficiently accurate results for $\tilde{F}(v_n, k)$ can be obtained by evaluating (14.4.72) numerically using splines.

- With the use of this spline-computed $\tilde{F}(v_n, k)$ and the related Riemann sum $\tilde{\tilde{F}}_r^s(k)$ derived from it employing (2.39), we should verify that (2.36) is still well satisfied.

- Assuming that (2.36) is well satisfied, we should verify that inserting these results into (14.4.86) and then carrying out the indicated inverse Fourier transform, again numerically using splines, yields sufficiently accurate approximations to the functions $C_{m,s}^{[n]}(z)$.

- Finally, all these procedures should also yield satisfactory results when data is taken from a grid and interpolated onto the elliptic cylinder. See Figures 13.7.2 and 14.4.3.

The first three items above depend on the accuracy of the numerical spline-based Fourier transform routines. Since this accuracy has already been established in the circular cylinder case, and the demands made of the spline-based routines are no more stringent in the elliptic cylinder case, it seems sensible to proceed directly to verifying the last item. Moreover, its verification also provides added evidence of the veracity of the first three.

19.2.6 Completion of Test

Therefore, to complete our test, we again set up a regular grid in x, y, z space. We again let y and z range over the intervals $y \in [-2.4, 2.4]$ with $h_y = .1$ and $z \in [-300, 300]$ with $h_z = .125$. However, for x we will use the interval $x \in [-4.4, 4.4]$ with $h_x = .1$ in order to ensure that all the ellipse given by (2.1) and (2.2) lies within the grid. Thus, we use 89 grid points in x, 49 grid points in y, and 4801 grid points in z for a total of $89 \times 49 \times 4801 = 20,937,161$ grid points.[14] As before, we evaluate the angular integrals using a Riemann sum with $N = 120$. Doing so requires interpolation off the grid onto the elliptical cylinder at 120 angular points for each of the 4801 selected values of z. And we evaluate the forward linear transforms for 401 k values in the range $k \in [-K_c, K_c]$ with $K_c = 20$. We also use these same points in k space to evaluate the inverse Fourier transforms. That is, $H = .1$. Finally we use the same values for $r_{\max}(m)$ as before. The result of this process is a *completely numerically* calculated set of functions $C_m^{[n]}(z)$.

We find that the results so obtained for the $C_m^{[n]}(z)$ are to the eye indistinguishable, for example, from Figures 1.7, 1.8, and 1.15. As more precise indicators of accuracy, Figures 2.25 through 2.27 show differences between exact and grid-value based completely numerically computed results for $C_{1,s}^{[0]}(z)$, $C_{1,s}^{[6]}(z)$, and $C_{7,s}^{[0]}(z)$. We see that the error is between 4 and 5 parts in 10^4, which is just slightly larger than the circular cylinder error result.

We have demonstrated, for the monopole-doublet problem, that the steps in the first three boxes shown in Figure 14.1.1 can also be carried out to yield results having good numerical accuracy for the case of a cylinder with elliptical cross section. As remarked earlier, again see Figure 13.7.7 and now also Figure 2.28, the surface field we have been working with is quite singular, more singular than fields likely to be encountered in practice. Thus the fact that the elliptical cylinder surface method has succeeded in this rather extreme case indicates that it is likely to work even better in actual physical applications.

Exercises

19.2.1. From (2.20) through (2.22) we know that $\tilde{F}(v, k)$ must have an expansion of the form

$$\tilde{F}(v,k) = \sum_r \tilde{\tilde{F}}_r^s(k)\mathrm{se}_r(v,q). \tag{19.2.40}$$

Use this expansion to show that there is the relation

$$||\tilde{F}||^2 = (1/2)\sum_n |\tilde{\tilde{F}}_n^s(k)|^2, \tag{19.2.41}$$

which is a form of *Parseval's* theorem applied to Mathieu expansions. From (2.41) conclude that

$$|\tilde{\tilde{F}}_r^s(k)| = \sqrt{2}[||\tilde{F}||^2 - \sum_{n \neq r} |\tilde{\tilde{F}}_n^s(k)|^2]^{1/2} \leq \sqrt{2}||\tilde{F}||]. \tag{19.2.42}$$

19.2.2. This exercise is a generalization of Exercise 1.2. Suppose $f(\phi)$ and $g(\phi)$ are 2π periodic functions and therefore have Fourier expansions of the form

$$f(\phi) = \sum_{m=-\infty}^{\infty} f_m \exp(im\phi), \tag{19.2.43}$$

[14] Of course, as in the circular cylinder case, most of these grid points are actually unused since only relatively few are sufficiently near the surface of the elliptic cylinder to be employed in the interpolation procedure. After interpolation onto the surface of the elliptic cylinder, $120 \times 4801 = 576,120$ surface values of $F(U, v, z)$ are used in the remainder of the calculation.

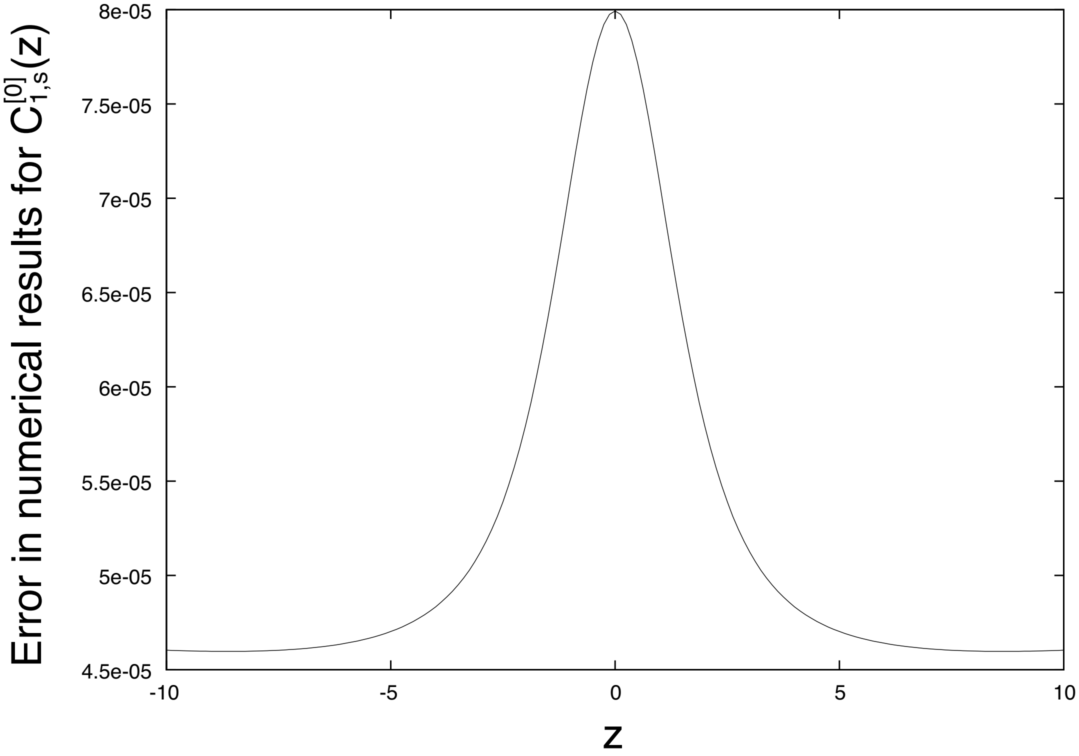


Figure 19.2.25: Difference between exact and completely numerically computed results for $C_{1,s}^{[0]}(z)$ based on field data provided on a grid and interpolated onto an elliptic cylinder with $x_{\max} = 4$ cm and $y_{\max} = 2$ cm.

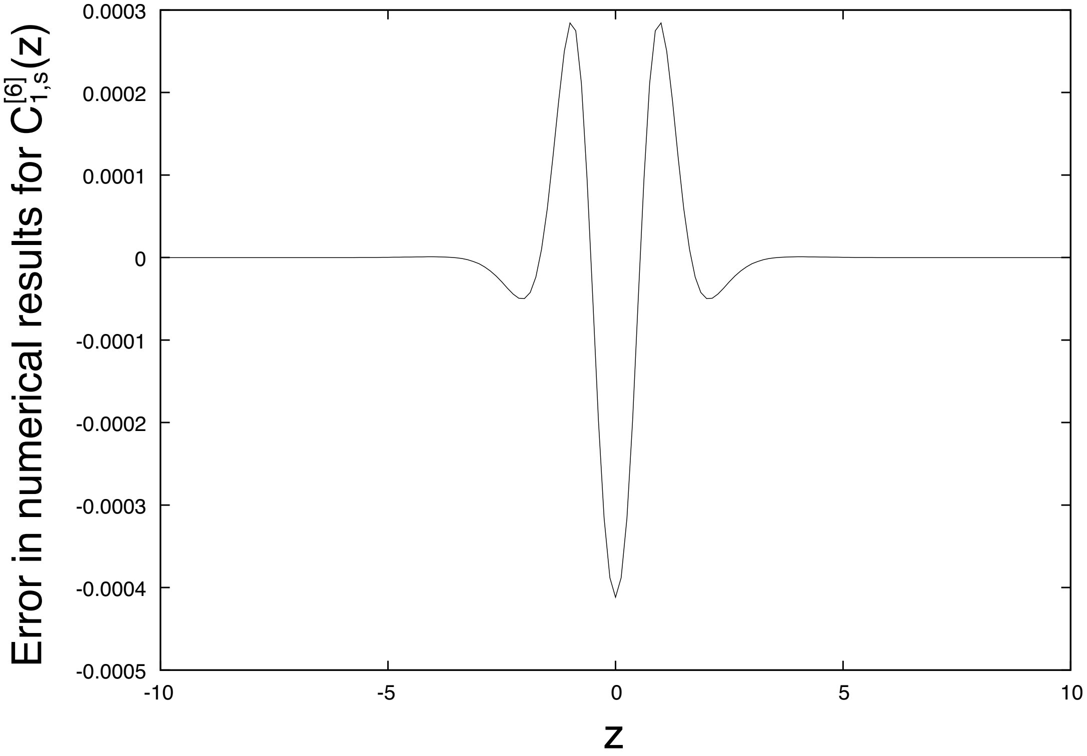


Figure 19.2.26: Difference between exact and completely numerically computed results for $C^{[6]}_{1,s}(z)$ based on field data provided on a grid and interpolated onto an elliptic cylinder with $x_{\rm max} = 4$ cm and $y_{\rm max} = 2$ cm.

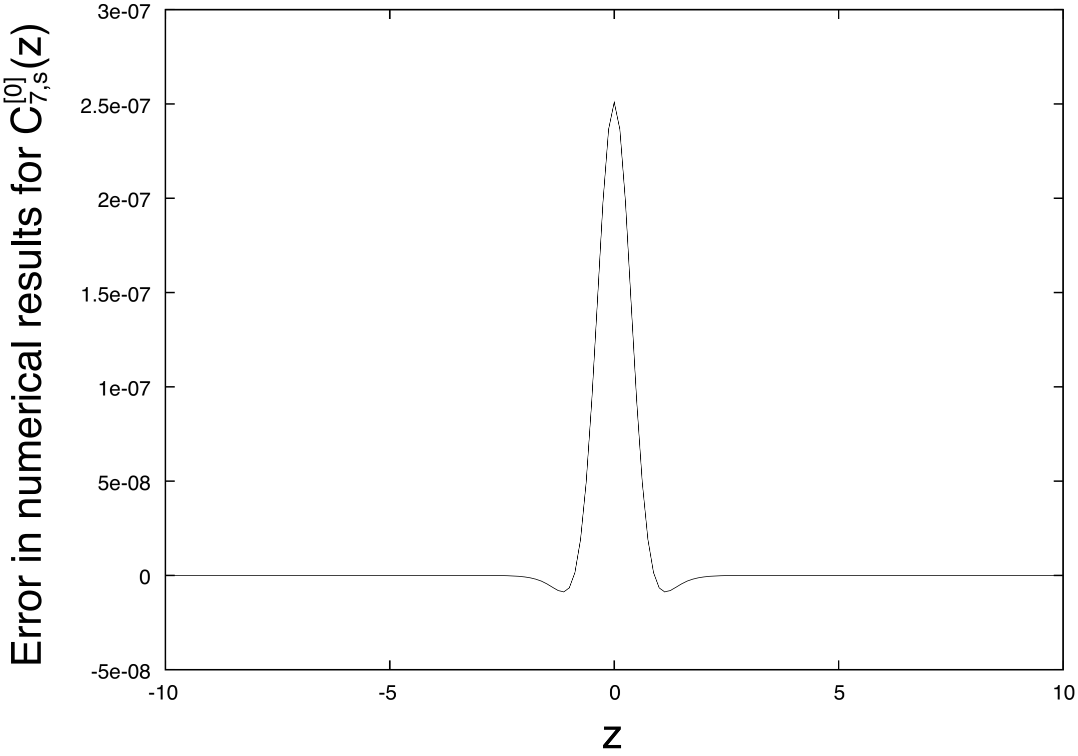


Figure 19.2.27: Difference between exact and completely numerically computed results for $C_{7,s}^{[0]}(z)$ based on field data provided on a grid and interpolated onto an elliptic cylinder with $x_{\max} = 4$ cm and $y_{\max} = 2$ cm..

$$g(\phi) = \sum_{m=-\infty}^{\infty} g_m \exp(im\phi). \tag{19.2.44}$$

Into the vector space of such functions introduce a scalar product by the usual rule

$$(f, g) = [1/(2\pi)] \int_0^{2\pi} d\phi \ \bar{f}(\phi) g(\phi) \tag{19.2.45}$$

where a bar denotes complex conjugation. Show that

$$(f, g) = \sum_{m=-\infty}^{\infty} \bar{f}_m g_m, \tag{19.2.46}$$

which is *Parseval's* theorem for Fourier expansions. Again select N discrete phi values ϕ_n according to the rule (6.1.37), and use these discrete values to define a *discrete* scalar product by the rule

$$(f, g)_{\rm d} = (1/N) \sum_{n=0}^{N-1} \bar{f}(\phi_n) g(\phi_n). \tag{19.2.47}$$

Verify that (2.47) is the discrete version of (2.45). Show that

$$(f, g)_{\rm d} = \sum_{m=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} \bar{f}_m g_{m+\ell N}, \tag{19.2.48}$$

which can be rewritten in the form

$$\begin{aligned} (f, g)_{\rm d} &= \sum_{m=-\infty}^{\infty} \bar{f}_m g_m + \sum_{m=-\infty}^{\infty} \sum_{\ell \neq 0} \bar{f}_m g_{m+\ell N} \\ &= (f, g) + \sum_{m=-\infty}^{\infty} \sum_{\ell \neq 0} \bar{f}_m g_{m+\ell N}. \end{aligned} \tag{19.2.49}$$

Thus, $(f, g)_{\rm d}$ is a good approximation to (f, g) provided the f_m and g_m fall off sufficiently rapidly for large $|m|$, and N is sufficiently large, so that the sum in (2.49) can be neglected.

Let us apply these results to the computation of the $\tilde{\tilde{F}}{}_r^s(k)$ as given by (2.21). Employing scalar product notation, we may write (2.21) in the form

$$\tilde{\tilde{F}}{}_r^s(k) = 2({\rm se}_r, \tilde{F}). \tag{19.2.50}$$

We know that the ${\rm se}_r$ have the Fourier expansion(15.5.27), which we rewrite in the complex form

$${\rm se}_r(v, q) = \sum_{m=-\infty}^{\infty} \hat{B}_m^r(q) \exp(imv). \tag{19.2.51}$$

Also, let $\tilde{\tilde{F}}{}^F(m, k)$ denote the *Fourier* coefficients of $\tilde{F}(v, k)$ so that we may write

$$\tilde{F}(v, k) = \sum_{m=-\infty}^{\infty} \tilde{\tilde{F}}{}^F(m, k) \exp(imv). \tag{19.2.52}$$

Then we have the result

$$(\mathrm{se}_r, \tilde{F})_{\mathrm{d}} = (\mathrm{se}_r, \tilde{F}) + \sum_{m=-\infty}^{\infty} \sum_{\ell \neq 0} \bar{\hat{B}}_m^r(q) \tilde{\tilde{F}}^F(m + \ell N, k). \tag{19.2.53}$$

In the next two exercises we will study the large $|m|$ behavior of the Fourier coeficients $\tilde{\tilde{F}}^F(m,k)$ and $\hat{B}_m^r(q)$.

19.2.3. This exercise is a continuation of Exercise 2.2. Here we will examine the falloff of the *Fourier* coefficients of $F(U, v, z)$ and $\tilde{F}(v, k)$ for the monopole doublet example. If you have not already done so, read Exercise 1.3. Here in analogy, we want to view $F(U, v, z)$ as a function of λ, with

$$\lambda = \exp iv, \tag{19.2.54}$$

and locate its singularities in λ.

From (14.4.67) we see that F is analytic in v save for those points where $B_x(U, v, z)$ and $B_y(U, v, z)$ have singularities in v. And from (13.7.4) and (13.7.5) we see that these singularities occur where $\psi(x, y, z)$ as given by (13.7.3) is singular, namely the points where

$$x^2 + (y \pm a)^2 + z^2 = 0. \tag{19.2.55}$$

Use (14.4.1), (14.4.2), and (2.1) through (2.4) to show that (2.25) is equivalent to the condition

$$\sin^2(v) \pm (2a/f)\sinh(U)\sin(v) - \{[f\cosh(U)]^2 + a^2 + z^2\}/f^2 = 0, \tag{19.2.56}$$

which can also be written in the form

$$\sin^2(v) \pm (2ay^{\max}/f^2)\sin(v) - [(x^{\max})^2 + a^2 + z^2]/f^2 = 0. \tag{19.2.57}$$

For the given values of f, $x^{\max}$, $y^{\max}$, and $a = 2.5$, and setting $z = 0$ as before, verify that (2.57) has the solutions

$$\sin(v) = \pm 1.0073, \ \pm 1.84067 \tag{19.2.58}$$

Correspondingly, verify that λ has the values

$$\lambda = \pm 1.12863i, \ \pm .88603i, \ \pm .29533i, \ \pm 3.38601i, \tag{19.2.59}$$

and therefore Λ has the value

$$\Lambda = 1.12863 \tag{19.2.60}$$

and $1/\Lambda$ has the value

$$1/\Lambda = 1/1.12863 = .88603. \tag{19.2.61}$$

Upon comparing (2.61) with (1.53), we see that the Fourier coefficients of $F(U, v, z = 0)$ are expected to fall off less rapidly that those of $B_\rho(R = 2, \phi, z = 0)$. Figure 2.28 displays the function $F(U, v, z = 0)$, with U given by (2.1). Observe that the peaks at $v = \pm\pi/2$ are sharper than those in Figure 13.7.6 for the corresponding case of a circular cylinder. Sharper peaks imply the existence of greater high 'frequency' content, which is consistent with the slower falloff of the Fourier coefficients.

Finally, it can be shown that, like the circular case, Λ is larger when $z \neq 0$. It follows, because the $\tilde{F}(v,k)$ may be viewed as linear combinations of the $F(U,v,z)$, that the $\tilde{\tilde{F}}^F(m,k)$ must also have the asymptotic behavior

$$|\tilde{\tilde{F}}^F(m,k)| \sim (.88603)^{|m|}. \tag{19.2.62}$$

This slower falloff is also consistent with the properties of $\tilde{F}(v,k)$. Figures 2.1 and 2.2 display the functions $\tilde{F}(v,k=0)$ and $\tilde{F}(v,k=20)$. For comparison, Figures 1.23 and 1.24 show $\tilde{B}_\rho(R,\phi,k=0)$ and $\tilde{B}_\rho(R,\phi,k=20)$ for the case of the circular cylinder. We see that the angular behavior is similar, but more peaked in the case of the elliptic cylinder. Since $\tilde{F}(v,k)$ and $\tilde{B}_\rho(R,\phi,k)$ have similar angular behavior, we conclude that the large m behavior of $\tilde{\tilde{F}}^F(m,k)$ and $\tilde{\tilde{B}}_\rho(R,m,k)$ should be similar. However, since the elliptic case is more peaked, we expect the falloff with m will be slower, as observed.

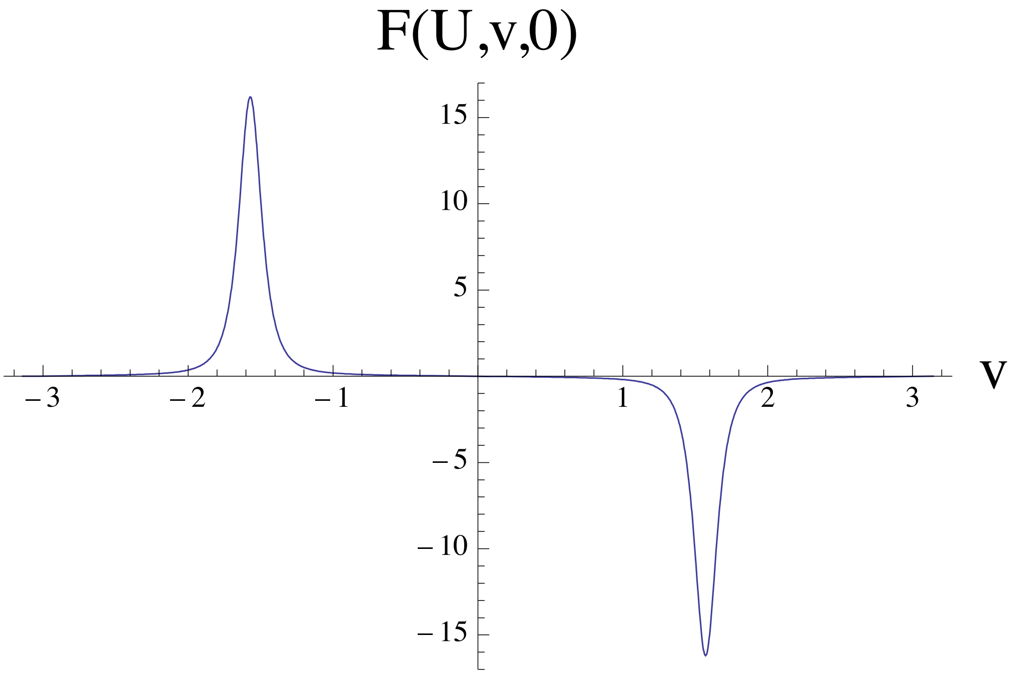


Figure 19.2.28: The quantity $F(U,v,z=0)$ for the monopole doublet in the case that $x^{\max}=4$ cm, $y^{\max}=2$ cm, $a=2.5$ cm, and $g=1$ Tesla-(cm)^2.

19.2.4. This exercise is a continuation of Exercise 2.2. Here we will examine the Fourier coefficients for the Mathieu functions $\mathrm{ce}_r(v,q)$ and $\mathrm{se}_r(v,q)$. That is, we want to study the functions $A^r_m(q)$ and $B^r_m(q)$ given in (15.5.21) and (15.5.27).

When $q=0$ the Mathieu functions become the trigonometric functions. See (14.4.35) through (14.4.37). Consequently, we have the results

$$A^r_m(0) = \delta_{mr} \text{ when } m \neq 0, \tag{19.2.63}$$

$$A^r_m(0) = \sqrt{2}\delta_{mr} \text{ when } m = 0, \tag{19.2.64}$$

$$B_m^r(0) = \delta_{mr} \text{ with } m, r \geq 1. \tag{19.2.65}$$

For $q \neq 0$ the $A_m^r(q)$ and $B_m^r(q)$ need to be computed numerically, and the methods of Section 15.5.3 are convenient for doing so.

From (15.5.22) and (15.5.28), and applying the logic of Exercise 1.3, we see that as $m \to \infty$ the $A_m^r(q)$ and $B_m^r(q)$, for fixed q and r, must fall off faster than $(1/\Lambda)^m$ for any $\Lambda > 1$ because the Mathieu functions $\mathrm{ce}_r(v, q)$ and $\mathrm{se}_r(v, q)$ are entire functions of v. This is good news, but we still would like to know how large m must be for this asymptotic behavior to set in. We will see that the answer to this question depends on q.

For simplicity, let us study the behavior of $B_m^r(q)$ as a function of m and q for the case $r = 7$, which is relevant to the case of the magnetic monopole doublet. Table 2.1 lists the values of $B_m^7(q)$ for various values of m and q. We see that, as q becomes more negative, the m value for which $|B_m^7(q)|$ peaks becomes ever larger. This is because, as q becomes ever more negative, more and more of the v axis is forbidden. See Section 14.4.4. However, we also know that $\mathrm{se}_7(v, q)$ must have 7 zeroes in the half-open interval $v \in [0, \pi)$. Thus, these oscillations are crowded into an ever smaller regions about 0 and $\pm\pi$, therefore leading to ever higher effective frequencies of oscillation.

Table 19.2.1: The coefficients $B_m^7(q)$.

$m\backslash q$	0	-50	-100	-150	-200	-250	-300
1	0	-.564487	-.375636	-.311321	-.274811	-.250197	-.232052
3	0	.293386	-.123551	-.237834	-.284830	-.307162	-.318251
5	0	.165556	.448665	.354776	.253034	.170433	.105346
7	1	-.465523	-.069863	.210855	.324419	.362328	.365811
9	0	.095015	-.435315	-.337074	-.163346	-.018371	.089069
11	0	.473149	.013745	-.311063	-.395713	-.372390	-.308973
13	0	.321478	.442586	.161649	-.088508	-.245162	-.327431
15	0	.118454	.432784	.453953	.318139	.151822	.003982
17	0	.029177	.236414	.409101	.456423	.415762	.331884
19	0	.005265	.089434	.234953	.355749	.421796	.437909
21	0	.000734	.025531	.099327	.196896	.287659	.356275
23	0	.000082	.005770	.032960	.084599	.149978	.216970
25	0	.000007	.001065	.008913	.029531	.063243	.106189
27	0	5.81E-7	.000164	.002014	.008615	.022275	.043350
29	0	3.81E-8	.000021	.000387	.002142	.006694	.015110

We saw on Exercise 2.2 that the discrete scalar product (discrete angular Riemann sum) is a good approximation to the true scalar product if N is sufficiently large such that

$$\tilde{\tilde{F}}^F(m, k) \approx 0 \text{ for } m \geq N \tag{19.2.66}$$

and

$$B_m^r(q) \approx 0 \text{ for } m \geq N \tag{19.2.67}$$

with $k \in [-K_c, K_c]$ and r being within its required range. See (2.53). In our benchmarking we have set $N = 120$. Verify that $(.88603)^{120} = 4.9 \times 10^{-7}$ so that, in view of (2.62), (2.66) is well satisfied. We next have to worry about the condition (2.67). From Table 2.1 we infer that $B_m^7(q = -50)$ begins to fall off rapidly with increasing m for m somewhat greater than 17, and $B_m^7(q = -300)$ begins to fall off rapidly with increasing m for m somewhat greater than 29. Assuming this trend continues as q becomes ever more negative, estimate by linear extrapolation that $B_m^7(q = -1200)$ begins to fall off rapidly with increasing m for m somewhat greater than 72. Also, we know that the largest r value we are interested in is $r_{\max}(7) = 29$. Therefore we might infer that $B_m^{29}(q = -1200)$ begins to fall of rapidly for m somewhat larger than $72 + (29 - 7) = 94$. Since 120 is significantly larger than 94, we infer that (2.67) should also be well satisfied for the choice $N = 120$.

These expectations are verified by the following calculations: Let ${}^N\tilde{\tilde{F}}_r^s(k)$ denote the result of the Riemann sum (2.93), and let

$$ {}^\infty\tilde{\tilde{F}}_r^s(k) = \tilde{\tilde{F}}_r^s(k) \tag{19.2.68}$$

denote its $N \to \infty$ limit given by the integral (2.21). Figure 2.29 shows the quantity ${}^\infty\tilde{\tilde{F}}_{29}^s(k)$ obtained by careful numerical evaluation of the integral in (2.21) when $r = 29$. En passant, we also take this opportunity to show in Figure 2.30 the base 10 logarithm of the three quantities $[-{}^\infty\tilde{\tilde{F}}_{29}^s(k)]$, $[\sqrt{2}||\tilde{F}||]$, and $[\sqrt{2}|\tilde{F}(\pi/2, k)|]$ to illustrate the inequalities (2.32) and (2.35). More to the point of this exercise, Figures 2.31 through 2.33 show the error quantities

$$ {}^N\tilde{\tilde{F}}_{29}^s(k) \; - \; {}^\infty\tilde{\tilde{F}}_{29}^s(k) $$

for $N = 40, 80,$ and 120. We see that the error decreases rapidly with increasing N, and is less than 6 parts in 10^4 when $N = 120$. Indeed, further calculation shows that the error is approximately 4 parts in 10^6 when $N = 160$.

19.2.5. Show that the quantities B_m^n comprise the entries of an infinite orthogonal matrix.

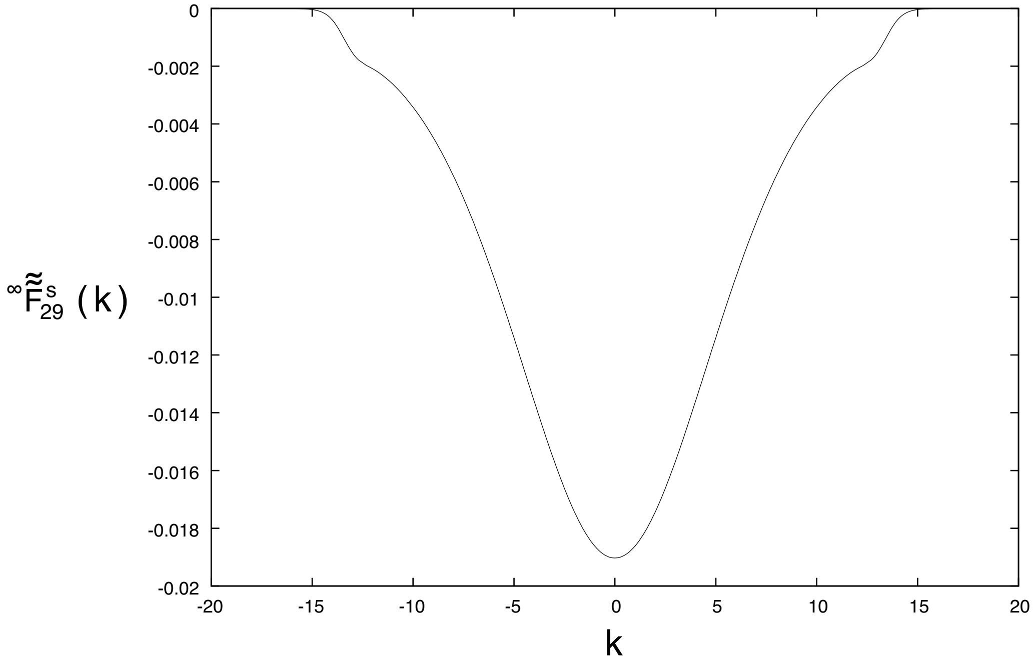


Figure 19.2.29: Real part of $^{\infty}\tilde{\tilde{F}}^{s}_{29}(k)$. The imaginary part vanishes.

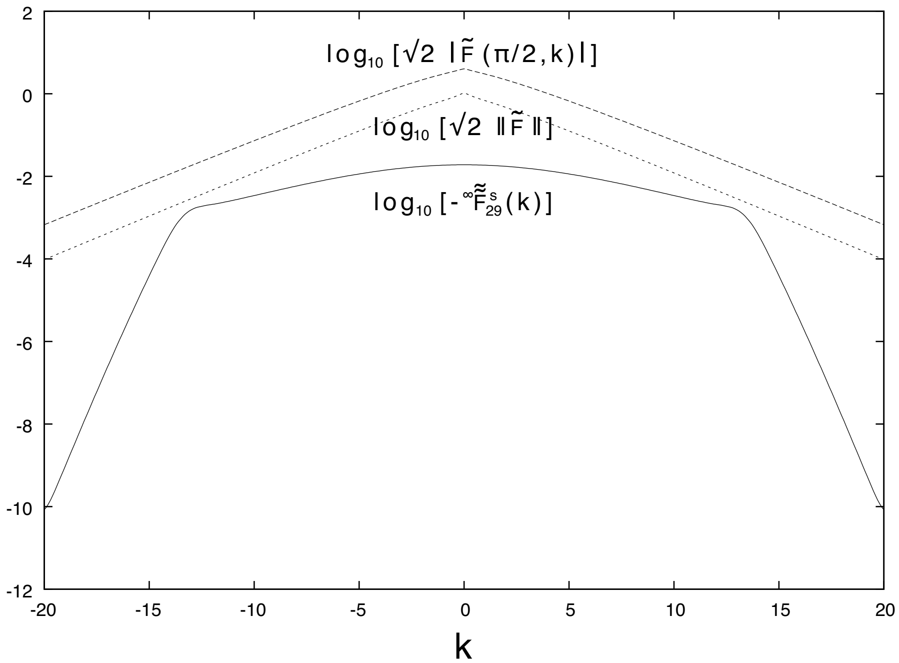


Figure 19.2.30: The base 10 logarithm of three quantities. The quantity $\log_{10}[\sqrt{2}|\tilde{F}(\pi/2,k)|]$ is the top curve. The middle curve is the quantity $\log_{10}[\sqrt{2}||\tilde{F}||]$, and the bottom curve is $\log_{10}[-^{\infty}\tilde{\tilde{F}}^{s}_{29}(k)]$. Together they illustrate the inequalities (2.32) and (2.35).

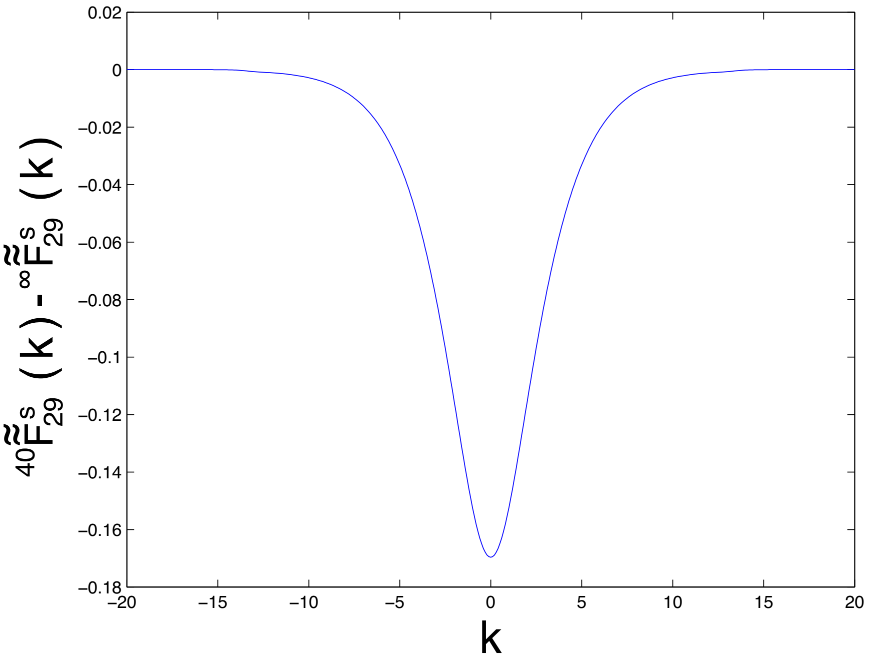


Figure 19.2.31: Real part of the error quantity ${}^{N}\tilde{\tilde{F}}^{s}_{29}(k) - {}^{\infty}\tilde{\tilde{F}}^{s}_{29}(k)$ for $N = 40$. The imaginary part vanishes.

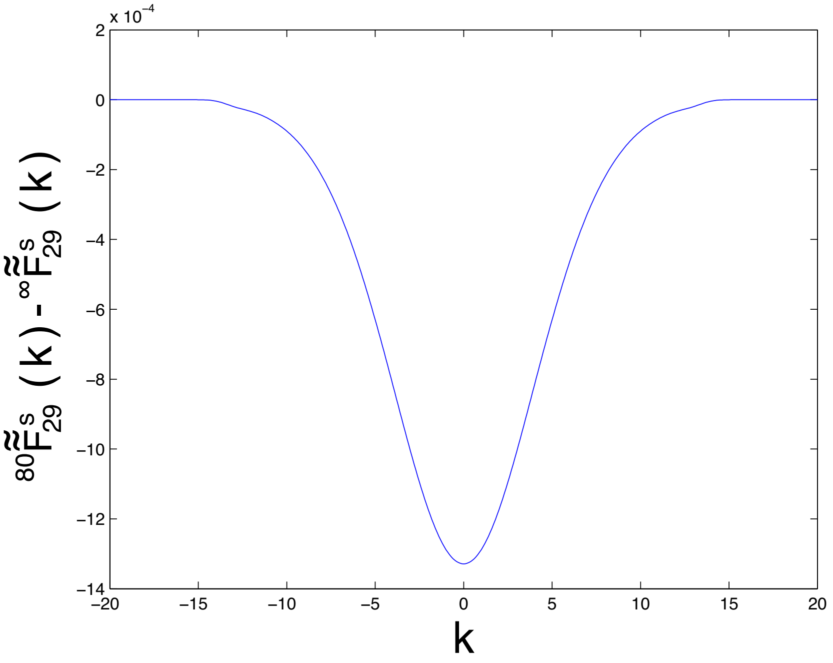


Figure 19.2.32: Real part of the error quantity ${}^{N}\tilde{\tilde{F}}^{s}_{29}(k) - {}^{\infty}\tilde{\tilde{F}}^{s}_{29}(k)$ for $N = 80$. The imaginary part vanishes.

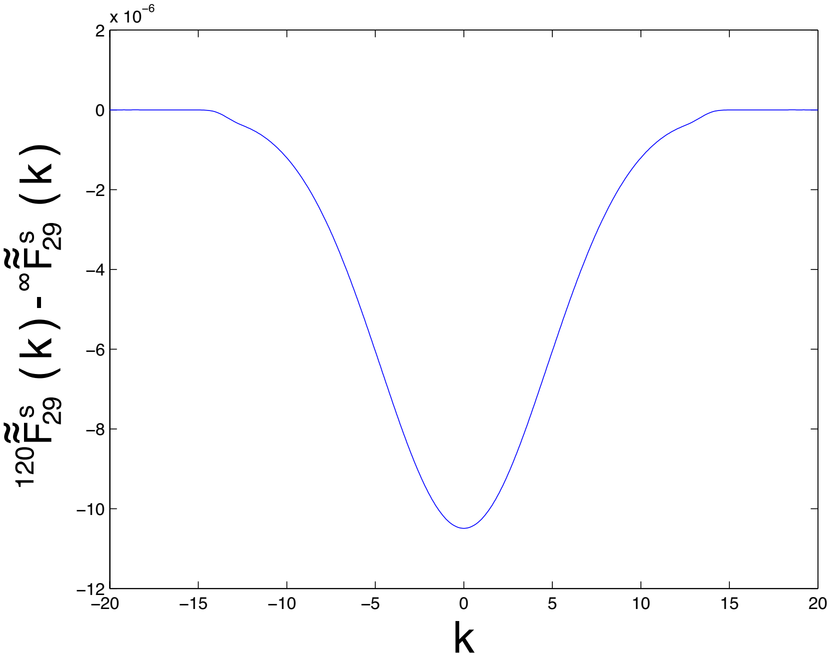


Figure 19.2.33: Real part of the error quantity ${}^{N}\tilde{\tilde{F}}^{s}_{29}(k) - {}^{\infty}\tilde{\tilde{F}}^{s}_{29}(k)$ for $N = 120$. The imaginary part vanishes.

19.3 Rectangular Cylinder Numerical Results for Monopole Doublet

Bibliography

General References

[1] C. Mitchell, "Calculation of Realistic Charged-Particle Transfer Maps", University of Maryland Physics Department Ph.D. Thesis (2007).

[2] C. Mitchell and A. Dragt, "Accurate transfer maps for realistic beam-line elements: Part I, straight elements", *Physical Review Special Topics - Accelerators and Beams* **13** 064001 (2010).

Bessel Functions

[3] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Chapter 9, Dover (1972). Also available on the Web by Googling "abramowitz and stegun 1972".

[4] F. Olver, D. Lozier, R. Boisvert, and C. Clark, Editors, *NIST Handbook of Mathematical Functions*, Cambridge (2010). See also the Web site http://dlmf.nist.gov/

Fourier Integrals

[5] A. Erdélyi, Edit., *Tables of Integral Transforms*, McGraw-Hill (1954).

Angular Integrals

[6] M. Javed and L. Trefethen, "A trapezoidal rule error bound unifying the Euler-Maclaurin formula and geometric convergence for periodic functions", *Proc. R. Soc. A Math. Phys. Eng. Sci.* **470**, 20130571 (2013). See also the Web site http://people.maths.ox.ac.uk/trefethen/publication/PDF/2014_150.pdf

[7] L. Trefethen and J. Weideman, "The exponentially convergent trapezoidal rule", (2014). *SIAM Rev.* **56**, 385–458 (2014). See the Web site http://people.maths.ox.ac.uk/trefethen/publication/PDF/2014_149.pdf

Chapter 20

Smoothing and Insensitivity to Errors

20.1 Introduction

In the previous Chapters 13 and 14 mention was made of the smoothing feature of surface methods. In this chapter we will explore the smoothing behavior of surface methods in more detail.

By way of introduction, imagine for simplicity that we initially use the surface of a circular cylinder. Suppose there are measurement or computational errors in the radial surface field values $B_\rho(\rho = R, \phi, z)$. What effect do these errors have on the determination of the generalized on-axis gradients and their derivatives? We will see that due to smoothing the effect of these errors is rather mild.

20.1.1 Preliminary Considerations

The relative insensitivity of surface methods to errors arises from a basic property of solutions to Laplace's equation: the value of ψ at some interior point is an appropriately weighted average of its values over any surrounding boundary. Consequently, ψ is smoother in the interior of a region than it may be on a boundary of this region. Correspondingly, errors in boundary values are averaged.

Something can also be said about surface methods and fitting errors. Suppose ψ is a *harmonic* function (satisfies $\nabla^2\psi = 0$) in some domain $\mathcal{D}$. Then, it can be shown that ψ assumes its maxima and minima on the *boundary* of $\mathcal{D}$. Imagine that $\psi^{\rm exact}$ is the true scalar potential and $\psi^{\rm approx}$ is some approximation to it. We know that $\psi^{\rm exact}$ is harmonic, and suppose that $\psi^{\rm approx}$ has been constructed to be harmonic. Then we know that $\psi^{\rm error} = \psi^{\rm approx} - \psi^{\rm exact}$ is harmonic. Therefore the magnitude of the error must be largest on the boundary of $\mathcal{D}$. However, if we do a good job of fitting $\psi^{\rm exact}$ by $\psi^{\rm approx}$ on the boundary of $\mathcal{D}$, then the error on the boundary will be small. And, thanks to $\psi^{\rm error}$ being harmonic, the error in the interior of $\mathcal{D}$ will be even smaller.

20.1.2 Analyticity

For a preliminary exploration of smoothing, suppose, for example, that the magnetic field is produced by an iron-dominated magnet, and is therefore localized in space. In this case the integrals (14.3.2) can be considered to have, in practice, finite limits of integration. With some care, an effective cut-off can also be found even if the fields extend to infinity since they fall off sufficiently rapidly at infinity. Also, since the generalized Bessel function I_m' increases exponentially as described by (14.3.9), there is also, in effect, a cut-off in k for the integrals (14.3.8) defining the generalized gradients.

Next suppose that the $\tilde{B}_\rho(R, m', z)$ are absolutely integrable,

$$\int_{-\infty}^{\infty} dz |\tilde{B}_\rho(R, m', z)| < \infty. \tag{20.1.1}$$

This will certainly be the case if $B_\rho(R, \phi, z)$ and hence the $\tilde{B}_\rho(R, m', z)$ are localized in z space. It follows from (14.3.2) that the Fourier transforms $\tilde{\tilde{B}}_\rho(R, m', k')$ are then bounded,

$$|\tilde{\tilde{B}}_\rho(R, m', k')| < [1/(2\pi)] \int_{-\infty}^{\infty} dz |\tilde{B}_\rho(R, m', z)| < \infty. \tag{20.1.2}$$

Now look at the integral representations (14.3.8) for the generalized gradients. We see that, due to the bounds (1.2) and the fall off in k at infinity produced by the $I_m'(kR)$ denominators, the integrals (14.3.8) are absolutely convergent in the domain

$$\Re(z) \in (-\infty, \infty)\ ,\ \Im(z) \in (-R, R). \tag{20.1.3}$$

Thus, under very mild assumptions about the surface data $B_\rho(\rho = R, \phi, z)$, including the possibility of errors, we conclude that the generalized gradients are *analytic* in the strip (1.3). Note that commonly used fringe-field models, those that assume constant (z independent) fields for z within the body of a magnet, zero fields outside beyond the fringe-field regions, and interpolating linear ramps in the fringe-field regions, violate this analyticity property because of singularities in the first derivative at the beginnings and ends of the ramps. These models are therefore unphysical, and their use could lead to erroneous conclusions.

20.1.3 Equivalent Spatial Kernel

In Section 14.3 the on-axis gradients were related to fields on the surface of a circular cylinder by various Fourier transform operations. We will now see that, for the circular cylinder case, they can also be viewed as being related by integration against a spatial kernel. Later, analogous results will be found for the cases of elliptic and rectangular cylinders.

Begin by relabeling variables so that (14.3.14) through (14.3.16) take (for $m > 0$) the form

$$\tilde{\tilde{B}}_\rho^\alpha(R, m, k) = [1/(2\pi)] \int_{-\infty}^{\infty} dz' \exp(-ikz') \tilde{B}_\rho^\alpha(R, m, z') \tag{20.1.4}$$

with

$$\tilde{B}_\rho^s(R, m, z') = (1/\pi) \int_0^{2\pi} d\phi \sin(m\phi) B_\rho(R, \phi, z'), \tag{20.1.5}$$

$$\tilde{B}_\rho^c(R, m, z') = (1/\pi) \int_0^{2\pi} d\phi \cos(m\phi) B_\rho(R, \phi, z'), \tag{20.1.6}$$

and (again for $m > 0$) (14.3.23) takes the form

$$C_{m,\alpha}^{[n]}(z) = i^n(1/2)^m(1/m!)\int_{-\infty}^{\infty} dk[k^{n+m-1}/I_m'(kR)]\tilde{\tilde{B}}_\rho^\alpha(R,m,k)\exp(ikz). \quad (20.1.7)$$

Our goal is to re-express the relations (1.4) and (1.7) in terms of a spatial kernel.

To do this, insert (1.4) into (1.7) to find the relation

$$\begin{aligned}
&C_{m,\alpha}^{[n]}(z) = [1/(2\pi)]i^n(1/2)^m(1/m!) \times \\
&\int_{-\infty}^{\infty} dk[k^{n+m-1}/I_m'(kR)]\exp(ikz)\int_{-\infty}^{\infty} dz'\exp(-ikz')\tilde{B}_\rho^\alpha(R,m,z') \\
&= [1/(2\pi)]i^n(1/2)^m(1/m!) \times \\
&\int_{-\infty}^{\infty} dz'\tilde{B}_\rho^\alpha(R,m,z')\int_{-\infty}^{\infty} dk[k^{n+m-1}/I_m'(kR)]\exp(ikz)\exp(-ikz') \\
&= [1/(2\pi)]i^n(1/2)^m(1/m!) \times \\
&\int_{-\infty}^{\infty} dz'\tilde{B}_\rho^\alpha(R,m,z')\int_{-\infty}^{\infty} dk[k^{n+m-1}/I_m'(kR)]\exp[ik(z-z')]. \quad (20.1.8)
\end{aligned}$$

Define the kernel $K_m^{[n]}$ by the rule

$$K_m^{[n]}(z,z') = [1/(2\pi)]i^n(1/2)^m(1/m!)\int_{-\infty}^{\infty} dk[k^{n+m-1}/I_m'(kR)]\exp[ik(z-z')]. \quad (20.1.9)$$

With this definition, we may rewrite (1.8) in the form

$$C_{m,\alpha}^{[n]}(z) = \int_{-\infty}^{\infty} dz' K_m^{[n]}(z,z')\tilde{B}_\rho^\alpha(R,m,z'). \quad (20.1.10)$$

That is, $C_{m,\alpha}^{[n]}(z)$ has been expressed as the result of integrating $\tilde{B}_\rho^\alpha(R,m,z')$ against the spatial kernel $K_m^{[n]}(z,z')$.

Let us now explore the properties of $K_m^{[n]}$. In the definition (1.9) make the substitution

$$\lambda = kR \text{ or } k = \lambda/R. \quad (20.1.11)$$

So doing gives the result

$$K_m^{[n]}(z,z') = [1/(2\pi)]i^n(1/2)^m(1/m!)(1/R)^{n+m}\int_{-\infty}^{\infty} d\lambda[\lambda^{n+m-1}/I_m'(\lambda)]\exp[i\lambda(z-z')/R]. \quad (20.1.12)$$

Staring at (1.12) suggests writing

$$K_m^{[n]}(z,z') = (1/R)^{n+m}L_m^{[n]}(\Delta) \quad (20.1.13)$$

where

$$\Delta = (z-z')/R \quad (20.1.14)$$

and

$$L_m^{[n]}(\Delta) = [1/(2\pi)]i^n(1/2)^m(1/m!)\int_{-\infty}^{\infty} d\lambda[\lambda^{n+m-1}/I'_m(\lambda)]\exp(i\lambda\Delta). \tag{20.1.15}$$

We observe, consistent with the notation being employed, that there is the relation

$$L_m^{[n]}(\Delta) = (\partial_\Delta)^n L_m^{[0]}(\Delta). \tag{20.1.16}$$

We also observe that the integrand factor $[\lambda^{m-1}/I'_m(\lambda)]$ appearing in

$$L_m^{[0]}(\Delta) = [1/(2\pi)](1/2)^m(1/m!)\int_{-\infty}^{\infty} d\lambda[\lambda^{m-1}/I'_m(\lambda)]\exp(i\lambda\Delta) \tag{20.1.17}$$

is *even* in λ. It follows that the $L_m^{[n]}(\Delta)$, and hence also the $K_m^{[n]}(z,z')$, are purely real. Moreover, the $L_m^{[0]}(\Delta)$ are even in Δ.

Graphs of some of the functions $L_m^{[n]}(\Delta)$ are shown in Figures 1.1 through 1.3 below. Before commenting on them, it is also useful to examine the integrands $\tilde{L}_m^{[0]}(\lambda)$, which are the Fourier transforms of the $L_m^{[0]}(\Delta)$, defined by the relations

$$\tilde{L}_m^{[0]}(\lambda) = [1/(2\pi)](1/2)^m(1/m!)[\lambda^{m-1}/I'_m(\lambda)]. \tag{20.1.18}$$

Some of them are displayed in Figures 1.4 and 1.5. Again we recall the asymptotic behavior

$$|I'_m(\lambda)| \sim \exp(|\lambda|)/\sqrt{2\pi|\lambda|} \text{ as } |\lambda| \to \infty, \tag{20.1.19}$$

from which we conclude that the $\tilde{L}_m^{[0]}(\lambda)$ essentially vanish exponentially at infinity.

What insights can be gained from examining Figures 1.1 through 1.5? First, we observe from Figure 1.3 that the $L_m^{[0]}$ become ever narrower with increasing m. Correspondingly, in accord with the uncertainty principle relating Fourier transform pairs, Figure 1.5 shows that the $\tilde{L}_m^{[0]}$ become ever broader with increasing m.

Next, from (1.10), we see that it is desirable that the $K_m^{[0]}(z,z')$ be slowly varying in z', because then noise in $\tilde{B}_\rho^\alpha(R,m,z')$ will be averaged over a large interval in z'. From (1.14) and Figure 1.3 we see that the $K_m^{[0]}(z,z')$ will be more nearly constant in z' the larger the value of R, and therefore there is ever more smoothing in z' as R is increased. We have already observed that the $L_m^{[0]}$ become somewhat narrower as m increases. Therefore there is somewhat less smoothing in z' for larger m. However, there is still ever more smoothing in z' as R is increased. Also, inspection of (1.13) shows that there is a $(1/R)^{n+m}$ factor relating $K_m^{[n]}$ and $L_m^{[n]}$. When the associated $C_{m,\alpha}^{[n]}$ are used in (13.2.37) or analogous expansions for $\boldsymbol{A}$, it is evident that the effective dimensionless expansion factor is $(\rho/R)^{n+m}$. Thus, although there is somewhat less smoothing in z' for larger m, there is increased suppression of high angular harmonic noise, and moreover this suppression is enhanced as R is increased. We also observe that there is increased suppression of high angular harmonic noise as n is increased.

Finally, let us examine smoothing from the perspective of k space. In the integral (1.7) make the change of variables

$$\lambda = kR \tag{20.1.20}$$

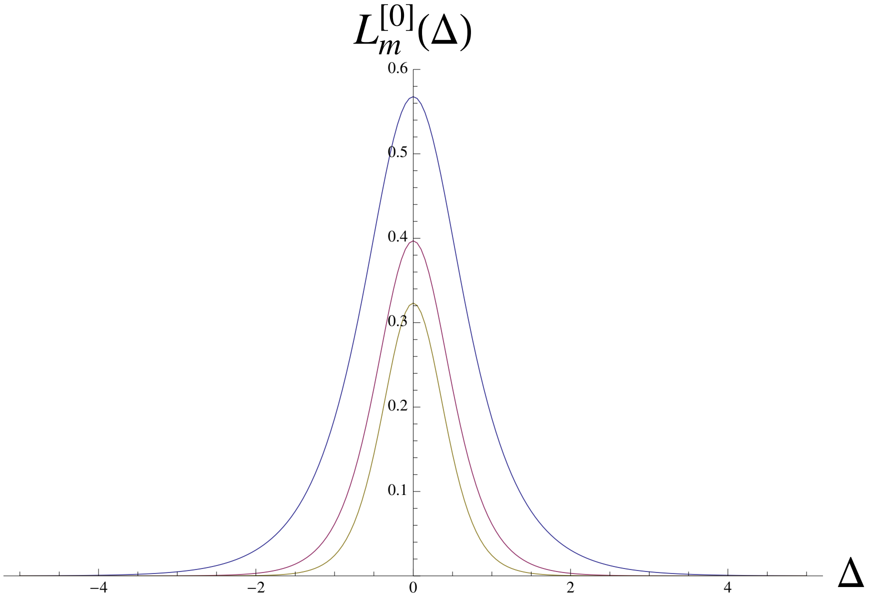


Figure 20.1.1: The spatial kernels $L_1^{[0]}(\Delta)$ through $L_3^{[0]}(\Delta)$. For $\Delta = 0$, the kernels $L_m^{[0]}(\Delta)$ decrease with increasing m.

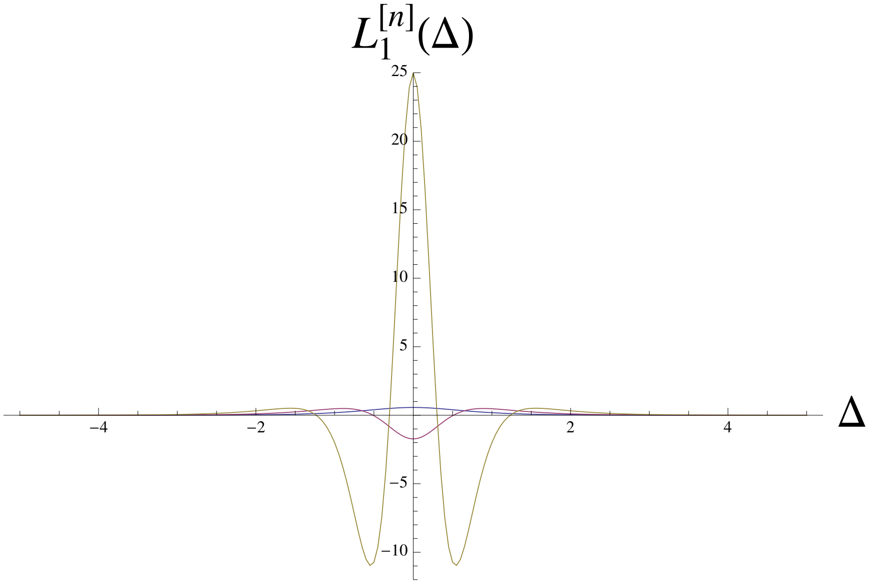


Figure 20.1.2: The spatial kernels $L_1^{[n]}(\Delta)$ for $n = 0, 2, 4$. Note that they satisfy (1.16). In particular, they have n zeroes.

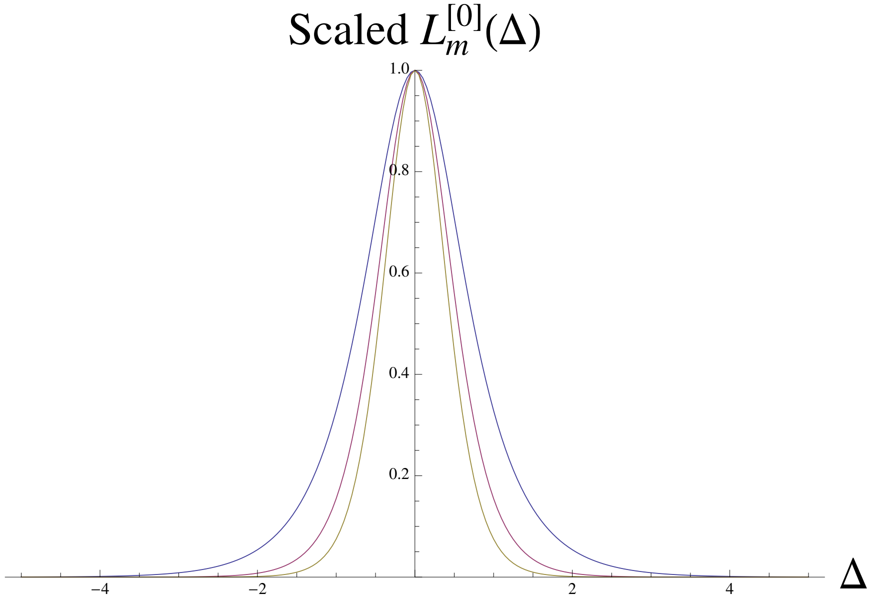


Figure 20.1.3: The *scaled* spatial kernels $L_1^{[0]}(\Delta)$ through $L_3^{[0]}(\Delta)$, all normalized to 1 at $\Delta = 0$. The scaled $L_m^{[0]}$ become ever narrower with increasing m.

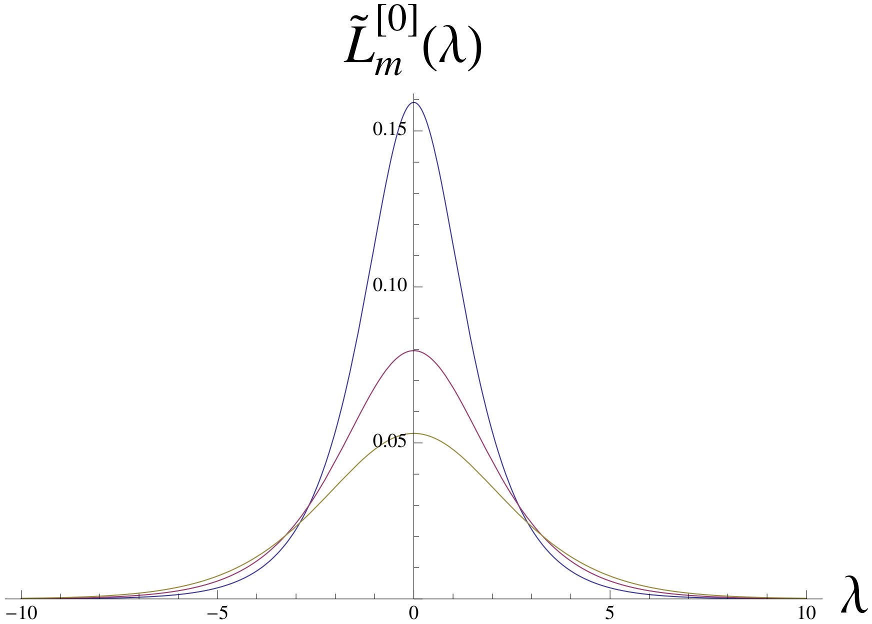


Figure 20.1.4: The integrands $\tilde{L}_m^{[0]}(\lambda)$ for $m = 1, 2, 3$. For $\lambda = 0$, the integrands decrease with increasing m.

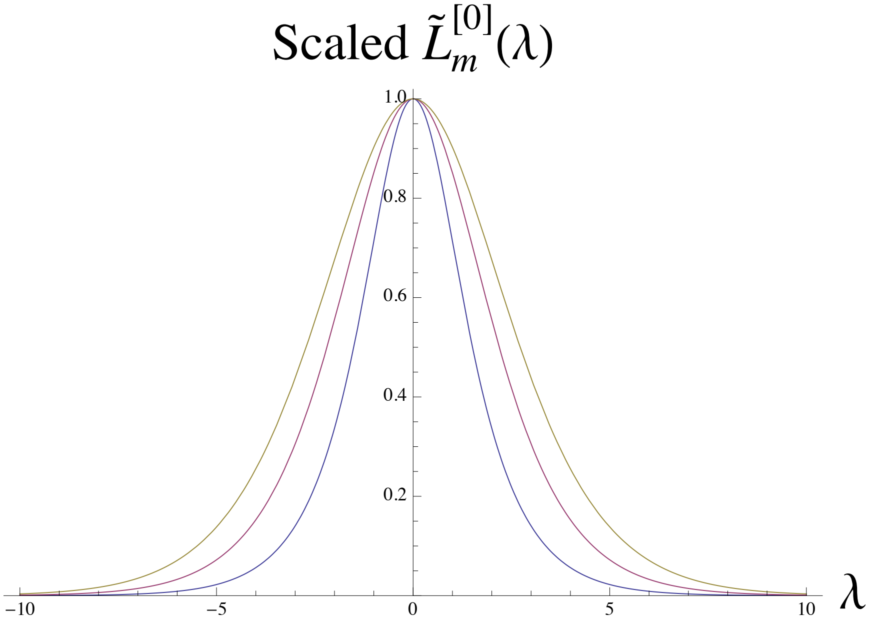


Figure 20.1.5: The *scaled* integrands $\tilde{L}_m^{[0]}(\lambda)$ for $m = 1, 2, 3$, all normalized to 1 at $\lambda = 0$. The scaled integrands become ever broader with increasing m.

to find the result

$$C_{m,\alpha}^{[n]}(z) = (1/R)^{n+m} i^n (1/2)^m (1/m!) \int_{-\infty}^{\infty} d\lambda [\lambda^{n+m-1}/I'_m(\lambda)] \tilde{\tilde{B}}_\rho^\alpha(R, m, \lambda/R) \exp(i\lambda z/R). \tag{20.1.21}$$

With the aid of the definition (1.18) this result can be rewritten in the form

$$C_{m,\alpha}^{[n]}(z) = 2\pi i^n (1/R)^{n+m} \int_{-\infty}^{\infty} d\lambda \tilde{L}_m^{[n]}(\lambda) \tilde{\tilde{B}}_\rho^\alpha(R, m, \lambda/R) \exp(i\lambda z/R). \tag{20.1.22}$$

Consider, for example, the case $n = 0$. Since the $\tilde{L}_m^{[0]}(\lambda)$ are peaked about $\lambda = 0$, we have the approximate result

$$\tilde{L}_m^{[0]}(\lambda) \tilde{\tilde{B}}_\rho^\alpha(R, m, \lambda/R) \approx \tilde{L}_m^{[0]}(\lambda) \tilde{\tilde{B}}_\rho^\alpha(R, m, 0), \tag{20.1.23}$$

and this result becomes ever more exact the larger the value of R. From (1.4) we see that evaluating $\tilde{\tilde{B}}_\rho^\alpha(R, m, k)$ for $k \approx 0$ essentially amounts to averaging $\tilde{B}_\rho^\alpha(R, m, z')$ over z', thereby suppressing the effect of noise. From Figure 1.5 we see that this smoothing in z' becomes somewhat less effective with increasing m because then the $\tilde{L}_m^{[0]}(\lambda)$ are less peaked about $\lambda = 0$. But again there is a $(1/R)^{n+m}$ factor that comes into play so that the effective dimensionless expansion factor is again $(\rho/R)^{n+m}$. Thus, although there is somewhat less smoothing in z' for larger m, there is increased suppression of high angular harmonic noise as n and m are increased, and again this suppression is further enhanced as R is increased.

20.1.4 What Work Lies Ahead

We will now study smoothing and insensitivity to errors in more detail depending on what surface is used. We will do so for the monopole-doublet examples treated in Sections 16.1 through 16.3, but our conclusions will be general. Section 17.2 treats the use of circular cylinders, and Sections 17.3 and 17.4 treat the use of elliptic and rectangular cylinders.

Exercises

20.1.1. Show that the on-axis gradients associated with the magnetic monopole doublet and given by (13.7.33) are analytic in the domain (1.3), and have singularities on the boundary. Show that the same is true for the on-axis gradient associated with the air-core solenoid of Section 11.11.

20.1.2. Verify that that the integrand factor $[\lambda^{m-1}/I'_m(\lambda)]$ is *even* in λ, and therefore the $L_m^{[n]}(\Delta)$, and also the $K_m^{[n]}(z,z')$, are purely real. Verify that the $L_m^{[0]}(\Delta)$ are even in Δ.

20.2 Circular Cylinders

Review Section 16.1.3 that described, for the monopole-doublet test case and the use of the surface of a circular cylinder, the calculation of on-axis gradients based on field data provided on a grid. Suppose we add to the field data at each grid point small random field components in the x and y directions.[1] What will be the effect of this noise on the on-axis gradients? We could use the noisy data to compute the on-axis gradients, and compare the results with those obtained in the absence of noise (and which we know agree very well with exact results). However, observe that the on-axis gradients are *linear* functions of the input field values on the grid points. Therefore, we can just as well compute the on-axis gradients that arise from *pure* noise without any background field. Doing so will give us better insight. If these purely noise-generated on-axis gradients are small compared to those for the noise-free data, then we will know that the effect of noise is small.

How shall we assign random field values to each grid point? Suppose the grid points are numbered from 1 to N. For example, in the calculation described in Section 16.1.3, $N = 11,527,201$. Let (x_j, y_j, z_j) be the coordinates of the j^{th} grid point. Let $B_y(0,0,z)$ be the vertical on-axis field arising from the monopole doublet and displayed in Figure 13.7.3. To model noise we make the Ansätze

$$B_x^{\text{noise}}(x_j, y_j, z_j) = \epsilon B_y(0,0,z_j)\delta_x(j), \tag{20.2.1}$$

$$B_y^{\text{noise}}(x_j, y_j, z_j) = \epsilon B_y(0,0,z_j)\delta_y(j). \tag{20.2.2}$$

Here the $\delta_x(j)$ and $\delta_y(j)$ are random numbers uniformly distributed in the interval $[-1, 1]$, and we set $\epsilon = .01$. By this prescription we produce a random field that is proportional, at

[1] Note, as observed earlier, that the z component of the field makes no contributions to the on-axis gradients.

the 1% level, to the strength of the monopole-doublet on-axis vertical field.[2]

The first step in the purely numerical calculation is to interpolate the field onto the surface of the cylinder and find its normal component at each of the cylinder sampling points. Figures 2.1 and 2.2 show the resulting $B_\rho(R, \phi, z = 0)$ for two different random number seeds. Compare with Figure 13.7.6. Correspondingly, Figures 2.3 and 2.4 show $B_\rho(R, \phi = \pi/2, z)$. Compare with Figure 13.7.7. We see that the surface field is noisy as expected, and the noise field falls to zero as $z \to \pm\infty$ because $B_y(0, 0, z)$ does so.

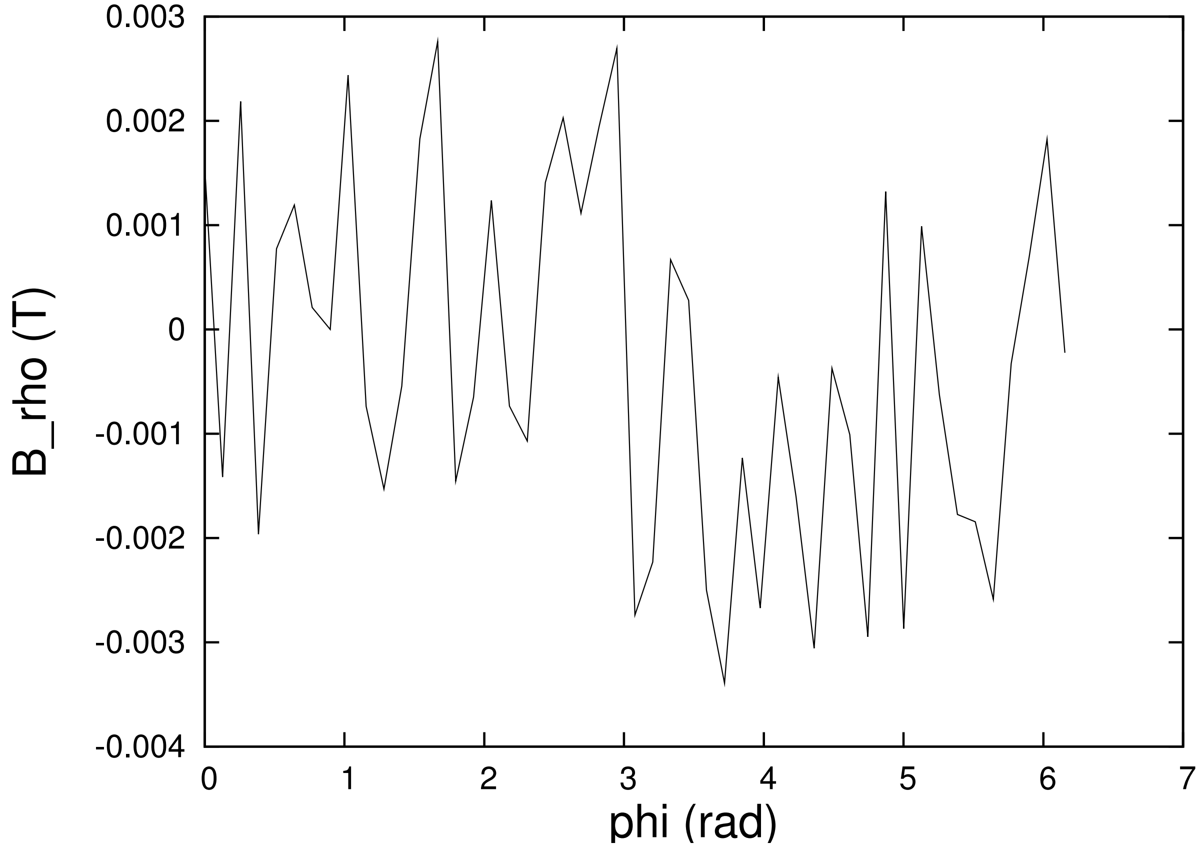


Figure 20.2.1: The function $B_\rho(R, \phi, z = 0)$ produced by a pure noise field.

Suppose, for example, that we now wish to compute the $C^{[0]}_{1,s}(z)$ produced by the pure noise field. Then, according to (14.3.23), we first need to compute $\tilde{\tilde{B}}^s_\rho(R, m = 1, k)$. And, because no particular symmetry is assumed for the noise, $\tilde{\tilde{B}}^s_\rho(R, m = 1, k)$ will have both real and imaginary parts. Figures 2.5 and 2.6 display the real parts for the two different choices of random number seed. [The imaginary parts behave analogously. They no longer vanish because $\tilde{B}^s_\rho(R, m = 1, z)$ for the noise is not assumed to be even in z.] In both cases

[2] This would be the general procedure. Actually, for this study, we generate a random field at all the $N = 20, 937, 161$ points described in Section 16.2.6; but, of course, only those required for interpolation onto the surface of the circular cylinder are actually used. We do this because in Section 17.3 we want to compare the use of circular and elliptic cylinders, and for this purpose we want to have a common data base. Note that the grid points described in Section 16.1.3 are a subset of those described in Section 16.2.6.

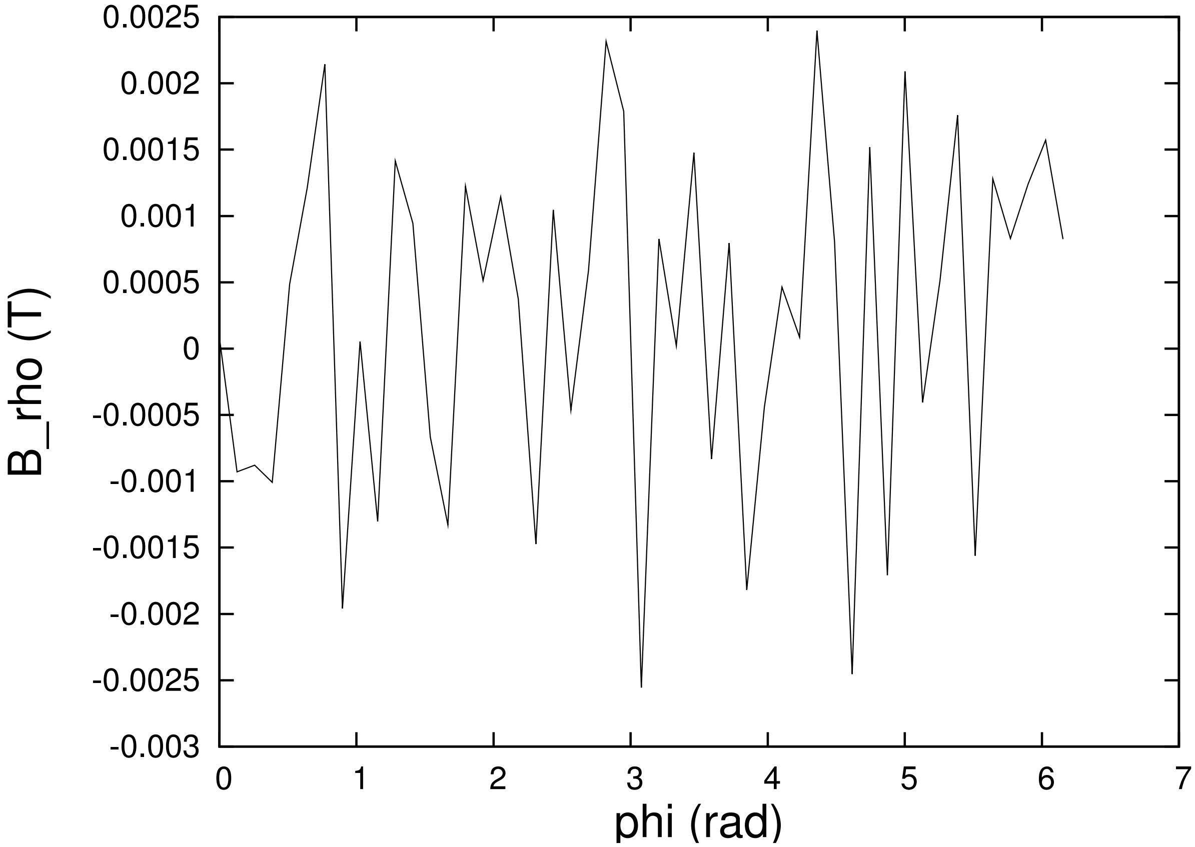


Figure 20.2.2: The function $B_\rho(R, \phi, z = 0)$ produced by a pure noise field arising from a second different random number seed.

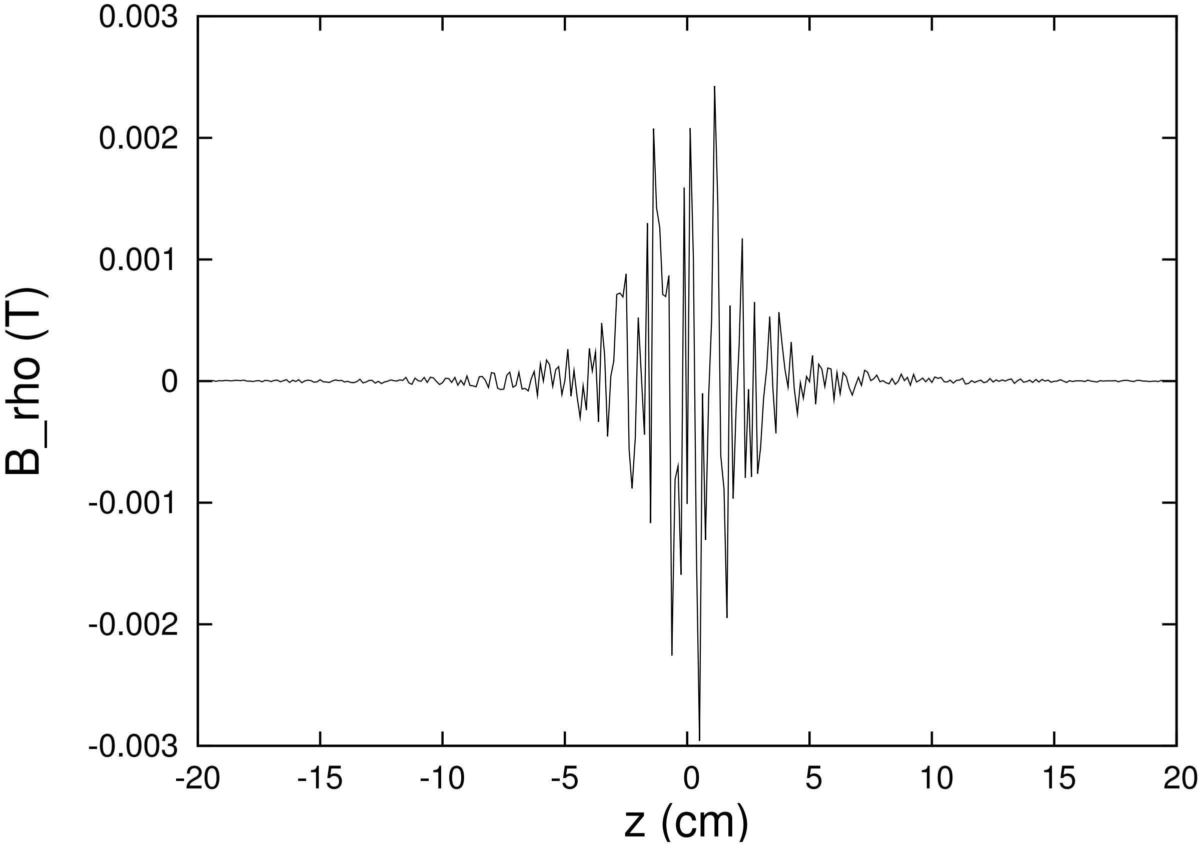


Figure 20.2.3: The function $B_\rho(R, \phi = \pi/2, z)$ produced by a pure noise field.

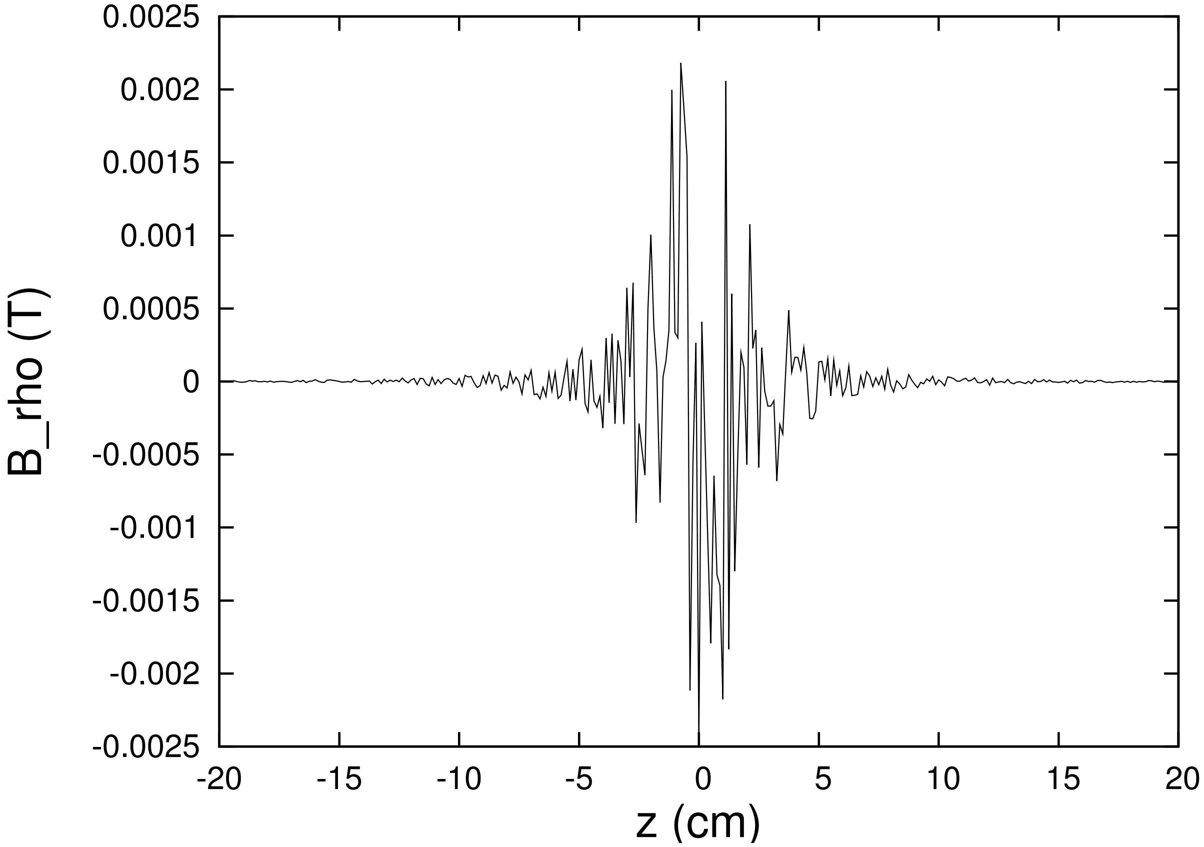


Figure 20.2.4: The function $B_\rho(R, \phi = \pi/2, z)$ produced by a pure noise field arising from a second different random number seed.

they have support for large values of $|k|$ as expected due to noise. Compare with Figure 16.1.1. In fact, because $h_z = .125$, we expect the noise to have Fourier contributions out to $K_{Ny} = \pi/h_z = 8\pi$, which is consistent with what is displayed.

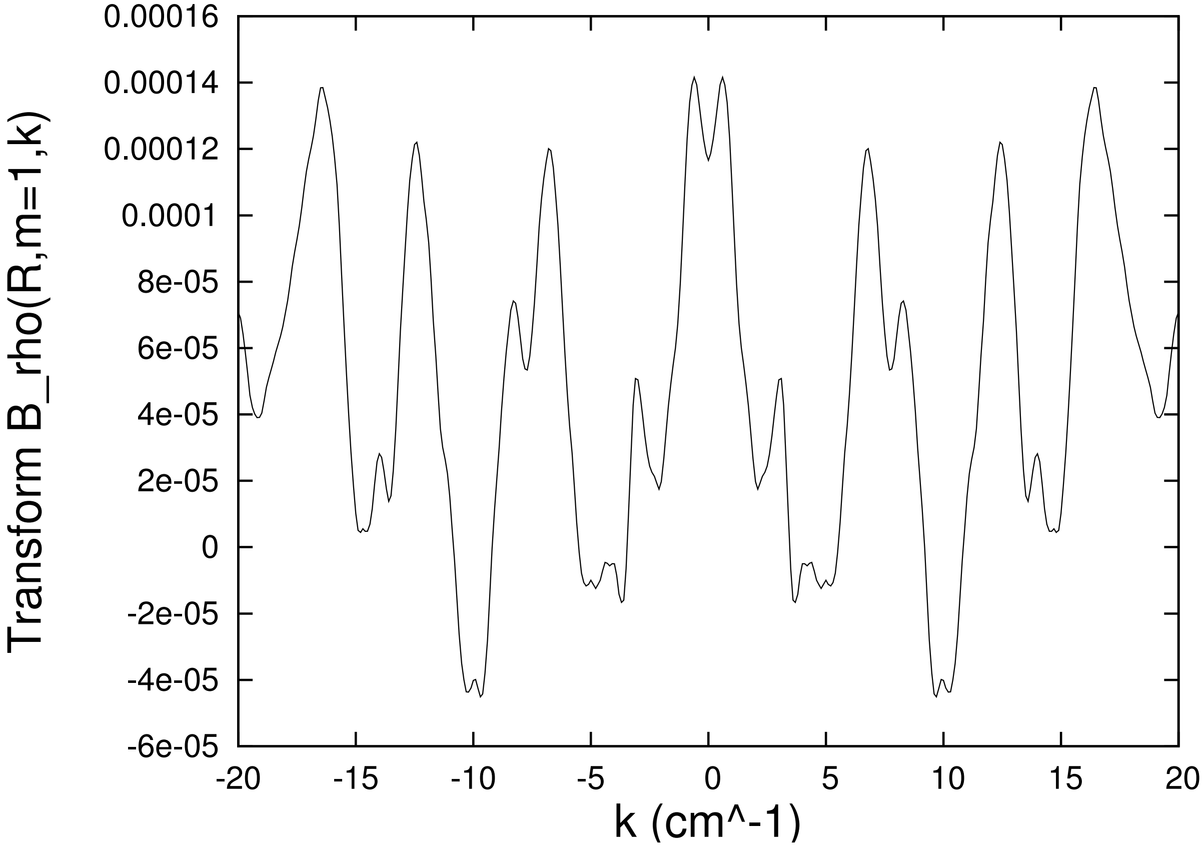


Figure 20.2.5: Real part of $\tilde{\tilde{B}}^s_\rho(R, m = 1, k)$ produced by a pure noise field. The imaginary part is comparable.

Next, again according to (14.3.23), we need to multiply $\tilde{\tilde{B}}^s_\rho(R, m = 1, k)$ by the kernel shown in Figure 16.1.3. Figures 2.7 and 2.8 show (for the real part) the results of this multiplication for the two different seed cases. We see, in both cases, that high spatial frequency noise is filtered out by the kernel.

Finally, we need to carry out the integration in (14.3.23). Figures 2.9 and 2.10 show the $C^{[0]}_{1,s}(z)$ so obtained for each noise realization. Comparison of these figures with Figure 16.1.7 shows that in this study a 1% noise in field data produces at most a .03% error in $C^{[0]}_{1,s}(z)$. Note that, unlike the case of Figure 16.1.7, $C^{[0]}_{1,s}(z)$ in Figures 2.9 and 2.10 is not symmetric about $z = 0$. There is no assumed symmetry for the noise.

What can be said about the other $C^{[n]}_{m,s}(z)$? They too are small. For example, Figures 2.11 through 2.14 show the functions $C^{[6]}_{1,s}(z)$ and $C^{[0]}_{7,s}(z)$. Comparison with Figures 16.1.8 and 16.1.15 shows that in this case a 1% noise in field data produces at most a .02% error in $C^{[6]}_{1,s}(z)$ and a .08% error in $C^{[0]}_{7,s}(z)$. It is remarkable that the error in the on-axis gradients is considerably smaller than that in the field data. It seems particularly remarkable that

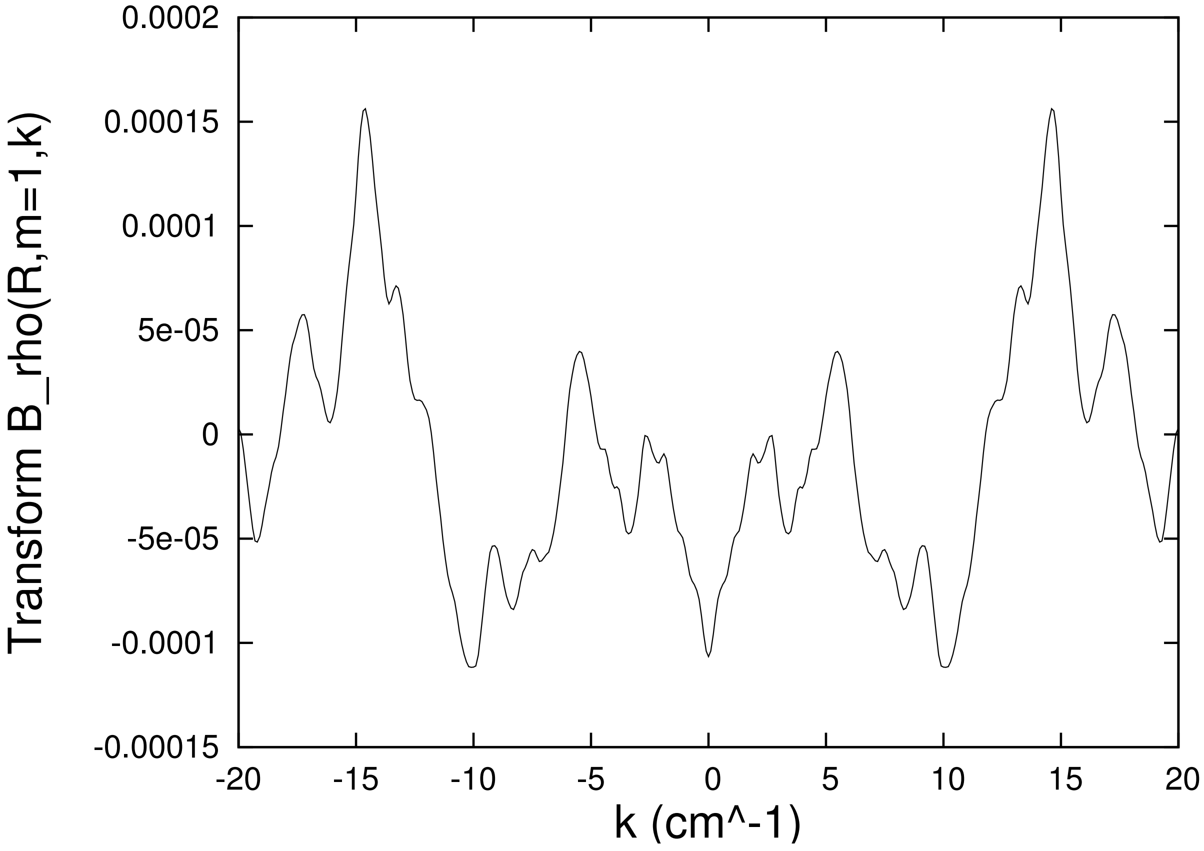


Figure 20.2.6: Real part of $\tilde{\tilde{B}}_\rho(R, m = 1, k)$ produced by a pure noise field arising from a second different random number seed. The imaginary part is comparable.

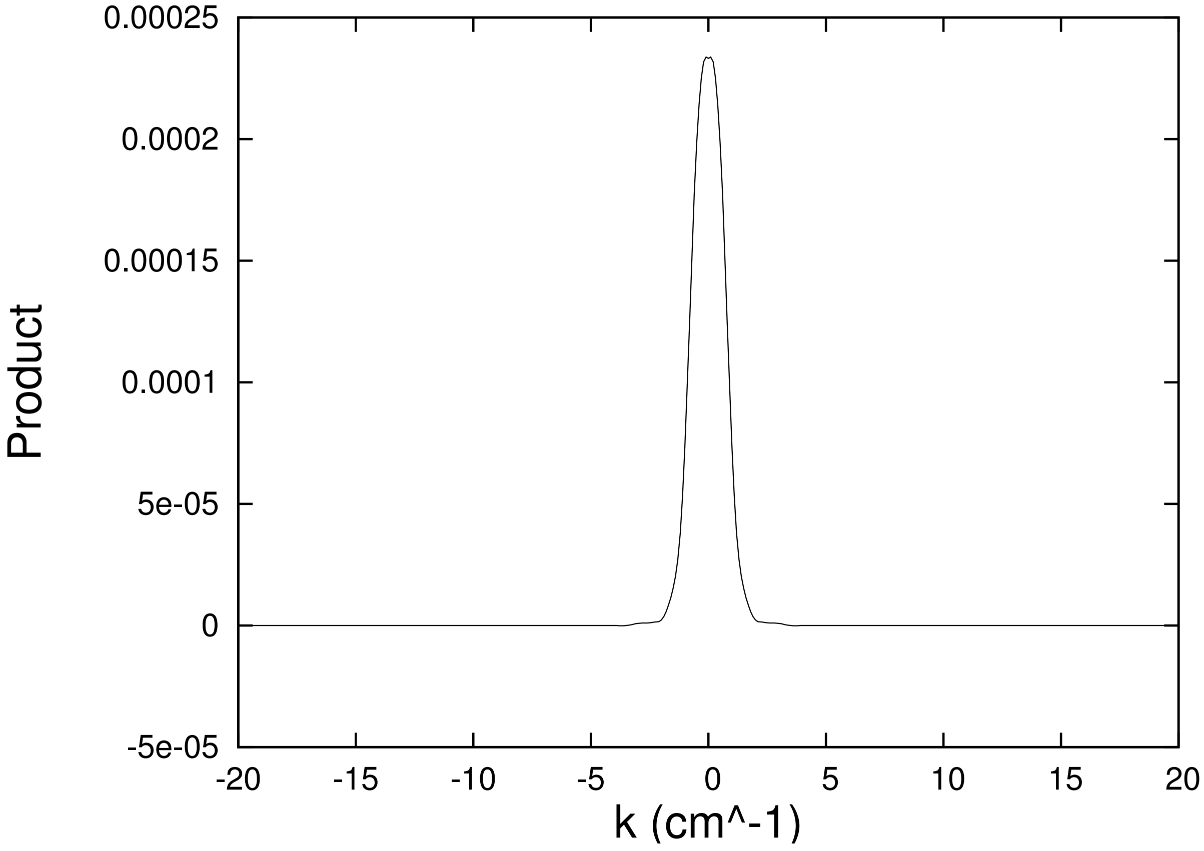


Figure 20.2.7: The product of $\Re\tilde{\tilde{B}}_\rho(R, m = 1, k)$ for the first random number seed and the kernel of Figure 16.1.3.

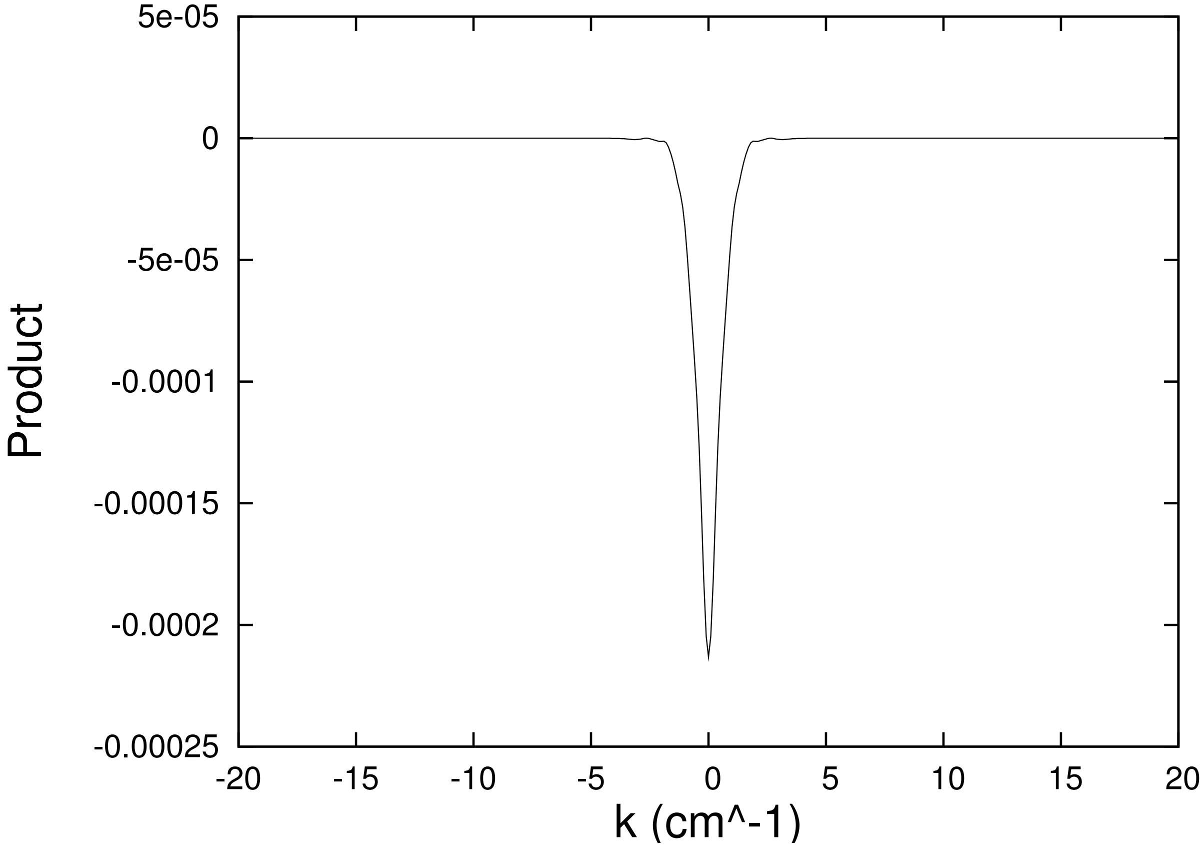


Figure 20.2.8: The product of $\Re\tilde{\tilde{B}}_\rho(R, m = 1, k)$ for the second different random number seed and the kernel of Figure 16.1.3.

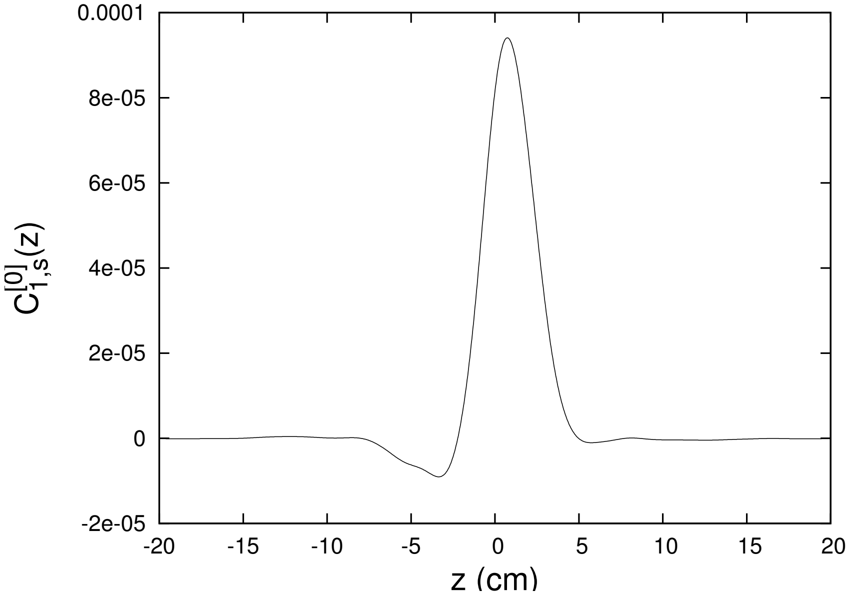


Figure 20.2.9: The function $C^{[0]}_{1,s}(z)$ produced by a pure noise field.

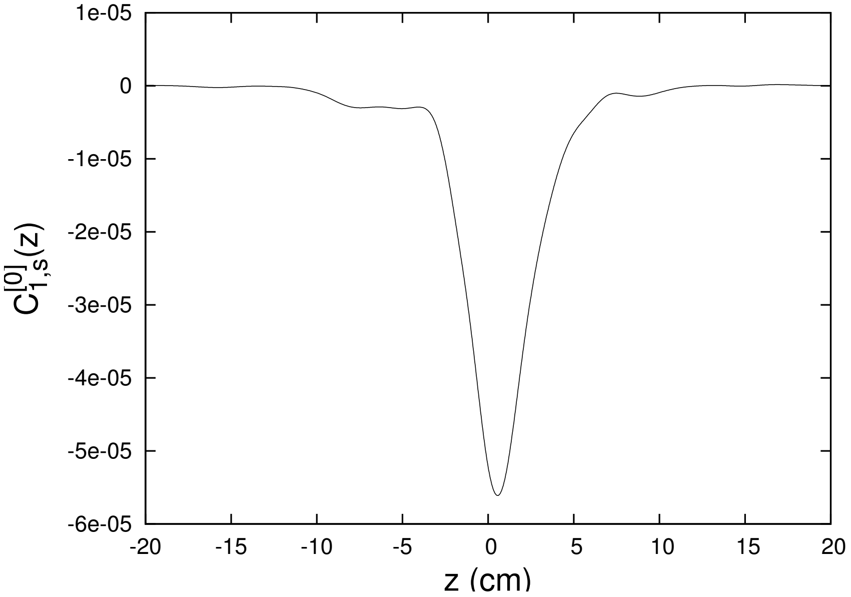


Figure 20.2.10: The function $C^{[0]}_{1,s}(z)$ produced by a pure noise field arising from a second different random number seed.

the error in $C^{[6]}_{1,s}(z)$ is so small because it involves 6 derivatives and the interpolated surface data, as evidenced by Figures 2.3 through 2.6, essentially has no derivatives.

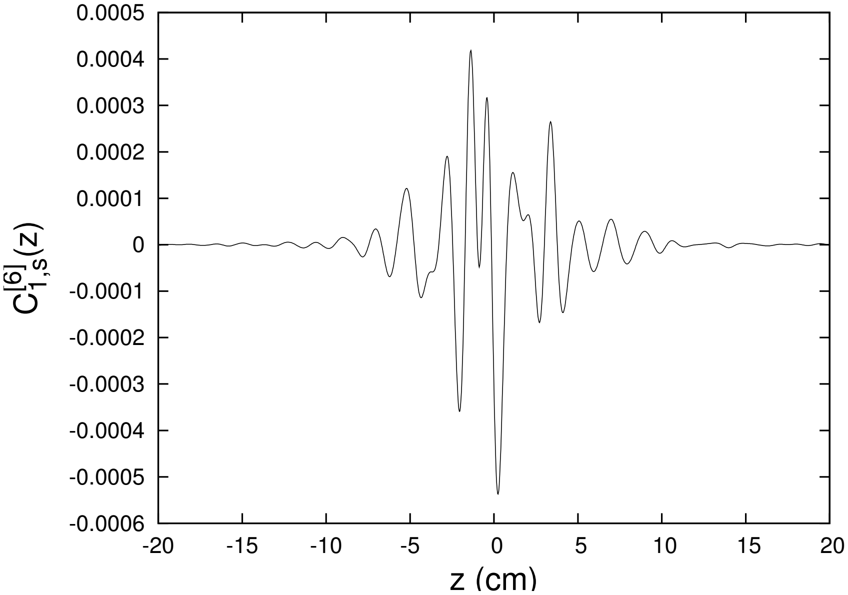


Figure 20.2.11: The function $C^{[6]}_{1,s}(z)$ produced by a pure noise field.

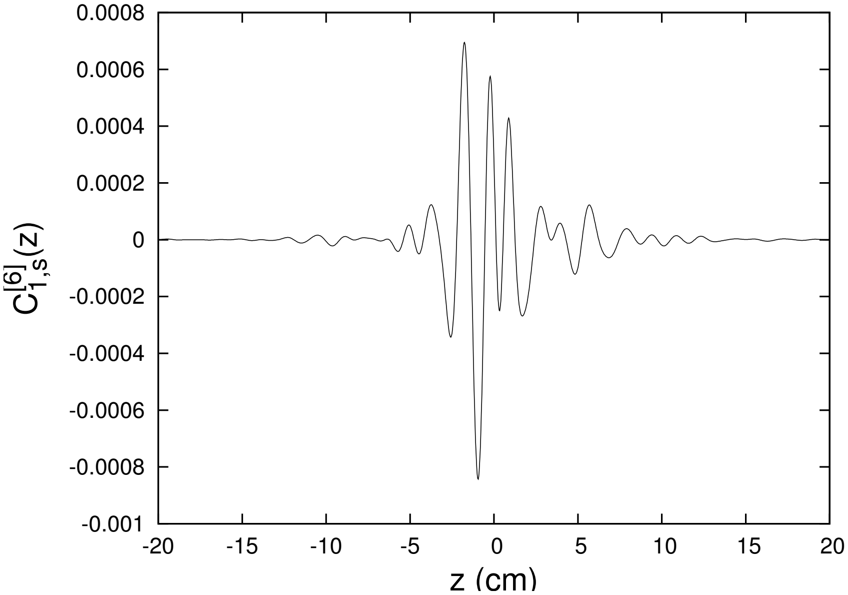


Figure 20.2.12: The function $C^{[6]}_{1,s}(z)$ produced by a pure noise field arising from a second different random number seed.

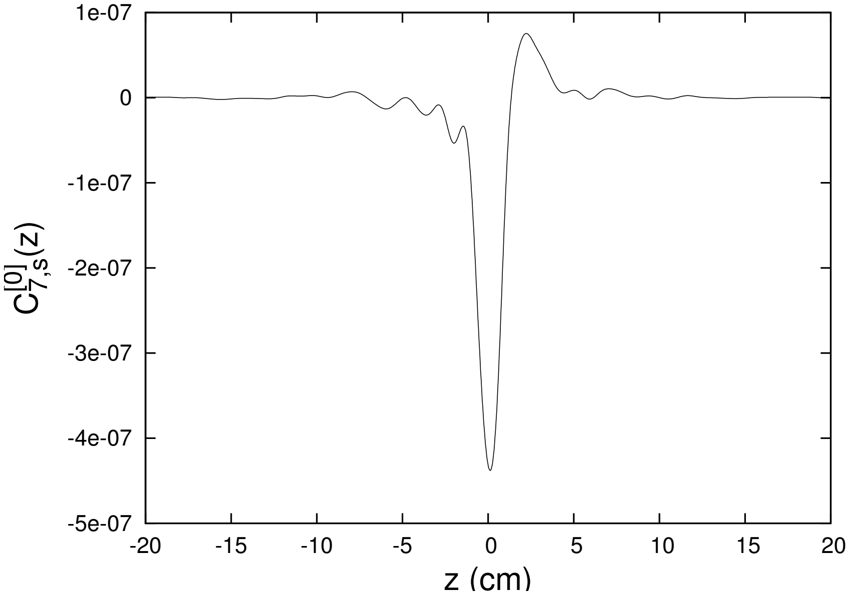


Figure 20.2.13: The function $C^{[0]}_{7,s}(z)$ produced by a pure noise field.

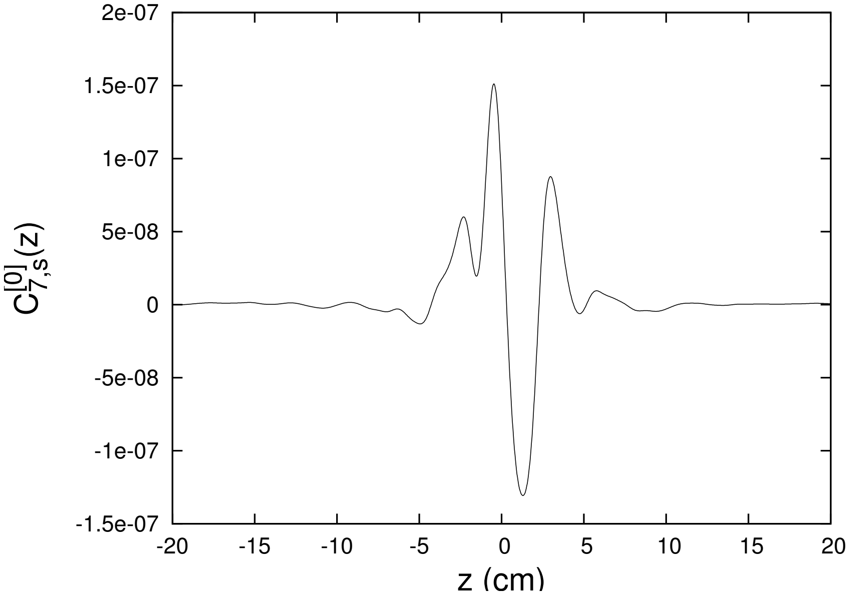


Figure 20.2.14: The function $C^{[0]}_{7,s}(z)$ produced by a pure noise field arising from a second different random number seed.

We end this section by exploring how smoothing depends on the radius of the circular cylinder. Figures 2.9 through 2.14 presented results for the case of a circular cylinder having radius $R = 2$. What if we had instead used a circular cylinder with $R = 1$? Presumably the effect of noise will be larger because there will then be less filtering. See (14.2.7).

Let us make a simple model of what to expect. Look at (14.3.23). Consistent with arising from a noise source, suppose the associated $\tilde{\tilde{B}}_\rho(R, m, k)$ is essentially independent of k. Then we have the bound

$$|C_{m,\alpha}^{[n]}(z)| \leq |\tilde{\tilde{B}}_\rho^\alpha(R, m, k \approx 0)|[(1/2)^m(1/m!)] \int_{-\infty}^{\infty} dk[|k|^{n+m-1}/|I_m'(kR)|]. \tag{20.2.3}$$

By the change of variables (1.20), the integral appearing in (2.3) can be brought to the form

$$\int_{-\infty}^{\infty} dk[|k|^{n+m-1}/|I_m'(kR)|] = (1/R)^{n+m} \int_{-\infty}^{\infty} d\lambda[|\lambda|^{n+m-1}/|I_m'(\lambda)|]. \tag{20.2.4}$$

Correspondingly, the bound (2.3) now takes the form

$$|C_{m,\alpha}^{[n]}(z)| \leq |\tilde{\tilde{B}}_\rho^\alpha(R, m, k \approx 0)|[(1/2)^m(1/m!)](1/R)^{n+m} \int_{-\infty}^{\infty} d\lambda[|\lambda|^{n+m-1}/|I_m'(\lambda)|]. \tag{20.2.5}$$

On the assumption that the noise itself is independent of R, we see that because of smoothing the $C_{m,\alpha}^{[n]}(z)$ due to noise may be expected to decrease with increasing R as $(1/R)^{n+m}$.

What actually happens? Figures 2.15 through 2.20 compare the results for $R = 1$ and $R = 2$. We see that, as a general trend, noise indeed has a larger effect when the smaller cylinder is employed. This is particularly true, as expected, for large values of $n + m$. But, in the case of Figure 2.15, the $C_{1,s}^{[0]}(z)$ computed from noise on the $R = 1$ cylinder is smaller than that computed from noise on the $R = 2$ cylinder. How can this be? As explained in the beginning of this section, see (2.1) and (2.2), in our model the noise values at the various grid points are independent. It can happen, through statistical fluctuations, that the net noise on the $R = 1$ cylinder is considerably less than on the $R = 2$ cylinder, so much so that this fluctuation effect more than compensates the poorer smoothing supplied by the smaller cylinder.

As a check on this explanation, suppose we attempt to make the noise on the $R = 1$ cylinder nearly the same as that on the $R = 2$ cylinder. One way to do this is the following: Suppose a noise value is required at some point on the $R = 1$ cylinder having angle ϕ. Instead of interpolating off nearby grid points, we may find the point on the $R = 2$ cylinder having the same ϕ value, and then interpolate off grid points near this $R = 2$ point. Figure 2.21 shows what happens when this done. Evidently the effect of noise on the $R = 1$ cylinder is now larger than the effect of essentially the same noise on the $R = 2$ cylinder.

We conclude that, as hoped, expected, and advertised, the use of surface methods (in this case the surface of a circular cylinder) does indeed yield results that are relatively insensitive to noise, and that this insensitivity is improved by placing the surface farther from the axis.

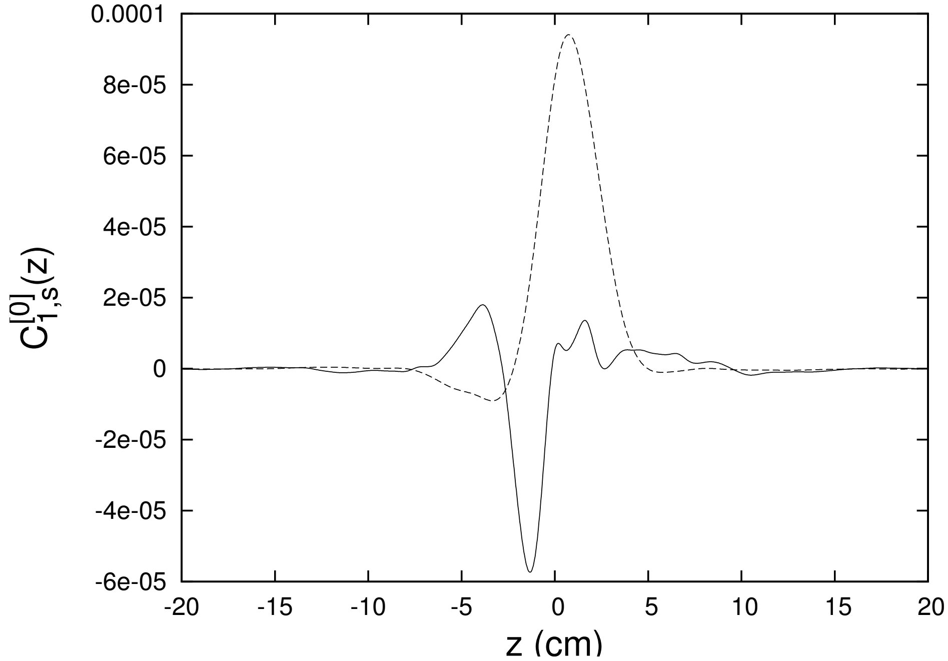


Figure 20.2.15: The functions $C_{1,s}^{[0]}(z)$ produced by a pure noise field on circular cylinders having $R = 1$ (solid line) and $R = 2$ (dashed line).

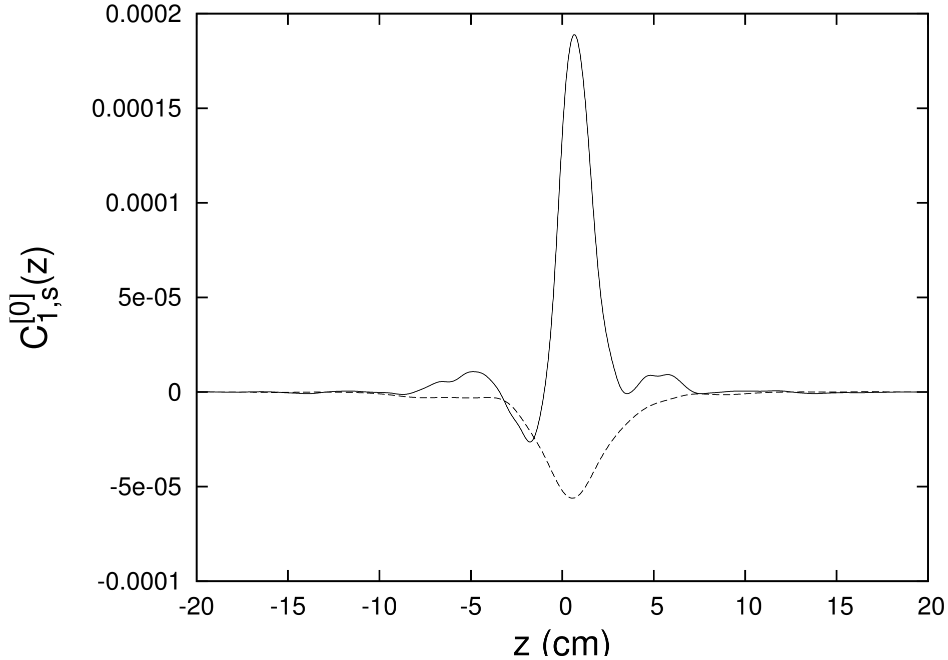


Figure 20.2.16: The functions $C_{1,s}^{[0]}(z)$ produced by a pure noise field on circular cylinders having $R = 1$ (solid line) and $R = 2$ (dashed line) and arising from a second different random number seed.

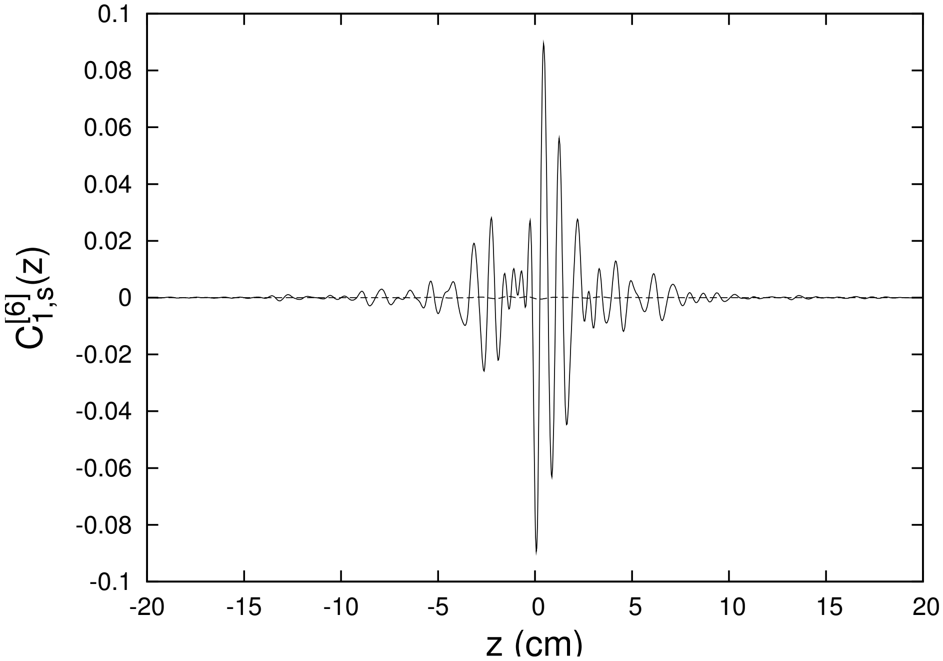


Figure 20.2.17: The functions $C^{[6]}_{1,s}(z)$ produced by a pure noise field on circular cylinders having $R = 1$ (solid line) and $R = 2$ (dashed line).

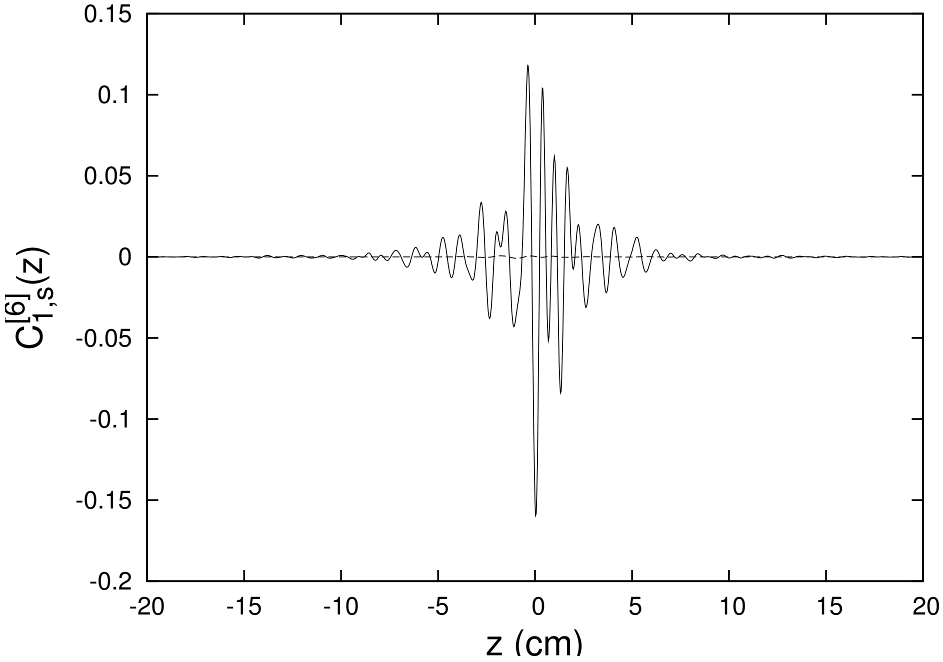


Figure 20.2.18: The functions $C^{[6]}_{1,s}(z)$ produced by a pure noise field on circular cylinders having $R = 1$ (solid line) and $R = 2$ (dashed line) and arising from a second different random number seed.

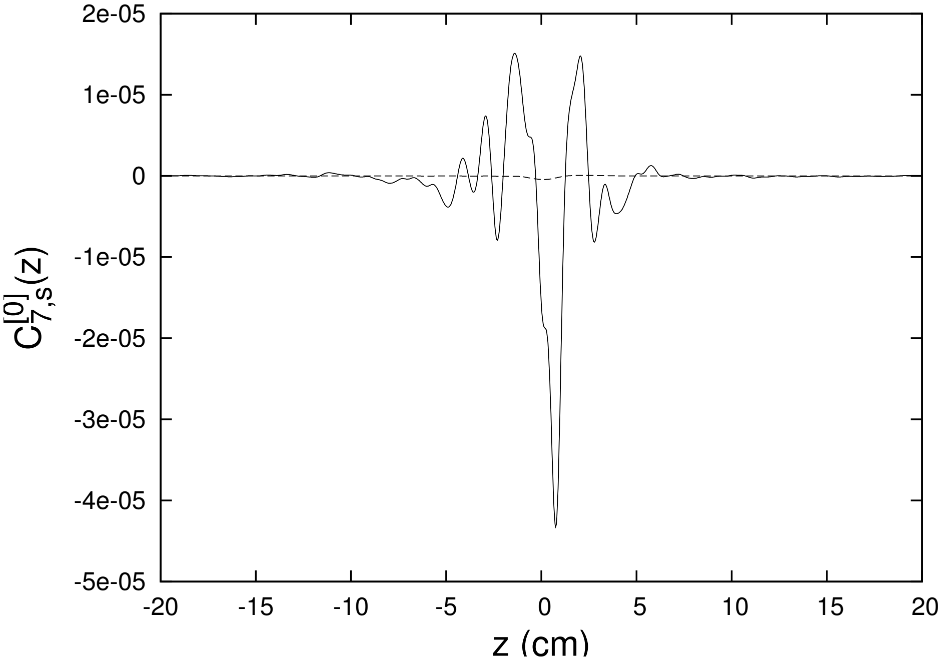


Figure 20.2.19: The functions $C^{[0]}_{7,s}(z)$ produced by a pure noise field on circular cylinders having $R = 1$ (solid line) and $R = 2$ (dashed line).

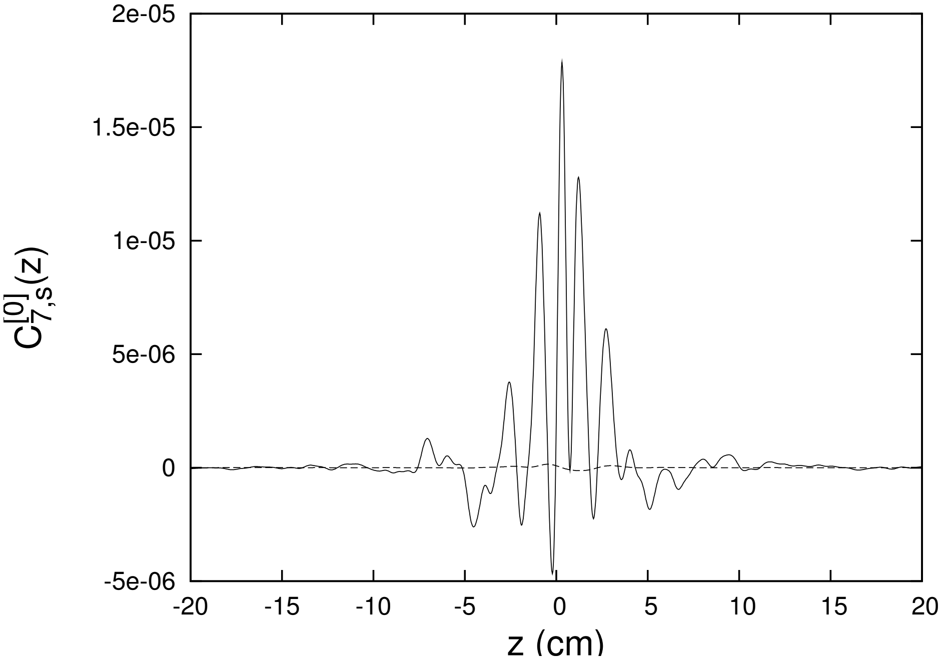


Figure 20.2.20: The functions $C^{[0]}_{7,s}(z)$ produced by a pure noise field on circular cylinders having $R = 1$ (solid line) and $R = 2$ (dashed line) and arising from a second different random number seed.

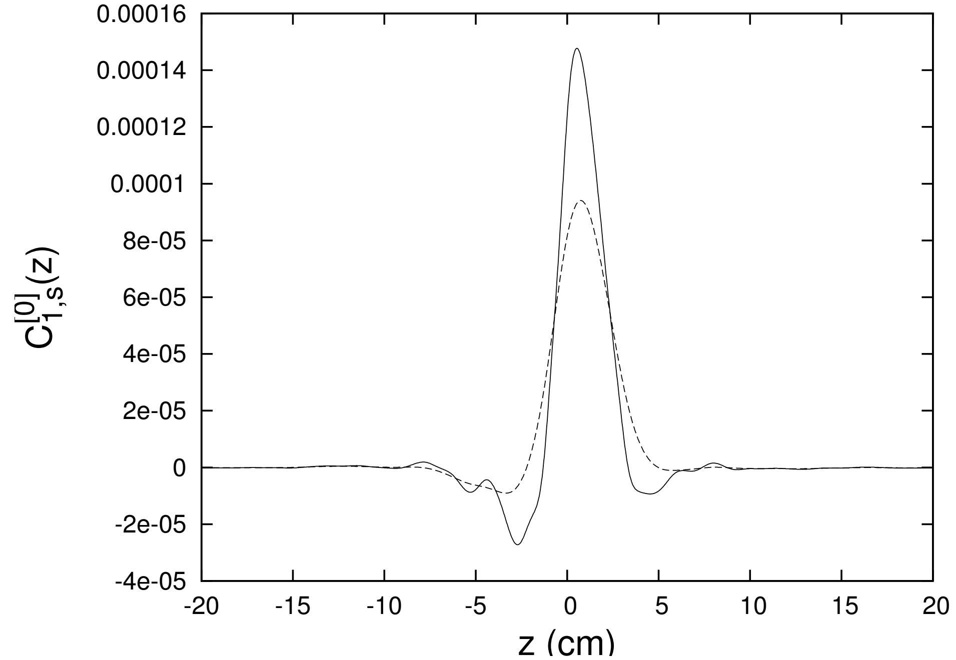


Figure 20.2.21: The functions $C^{[0]}_{1,s}(z)$ produced by nearly the same pure noise fields on circular cylinders having $R = 1$ (solid line) and $R = 2$ (dashed line).

20.3 Elliptic Cylinders

The discussion of this section parallels that of the previous section, but now deals with the monopole-doublet test case and the use of the surface of an elliptic cylinder as described in Section 16.2.6. We use the same noise model as that of the previous section.

Now the first step in the purely numerical calculation is to interpolate the field onto the surface of the elliptic cylinder and to find the function $F(U, v, z)$ given by (14.4.67). Figures 3.1 and 3.2 show the resulting $F(U, v, z = 0)$ for two different random number seeds. Compare with Figure 16.2.28. Correspondingly, Figures 3.3 and 3.4 show $F(U, v = \pi/2, z)$. In view of (14.4.66), this quantity is proportional to the normal component of B when $\phi = \pi/2$ so that these figures should be compared with Figure 13.7.7. Observe that the surface field is noisy as expected.

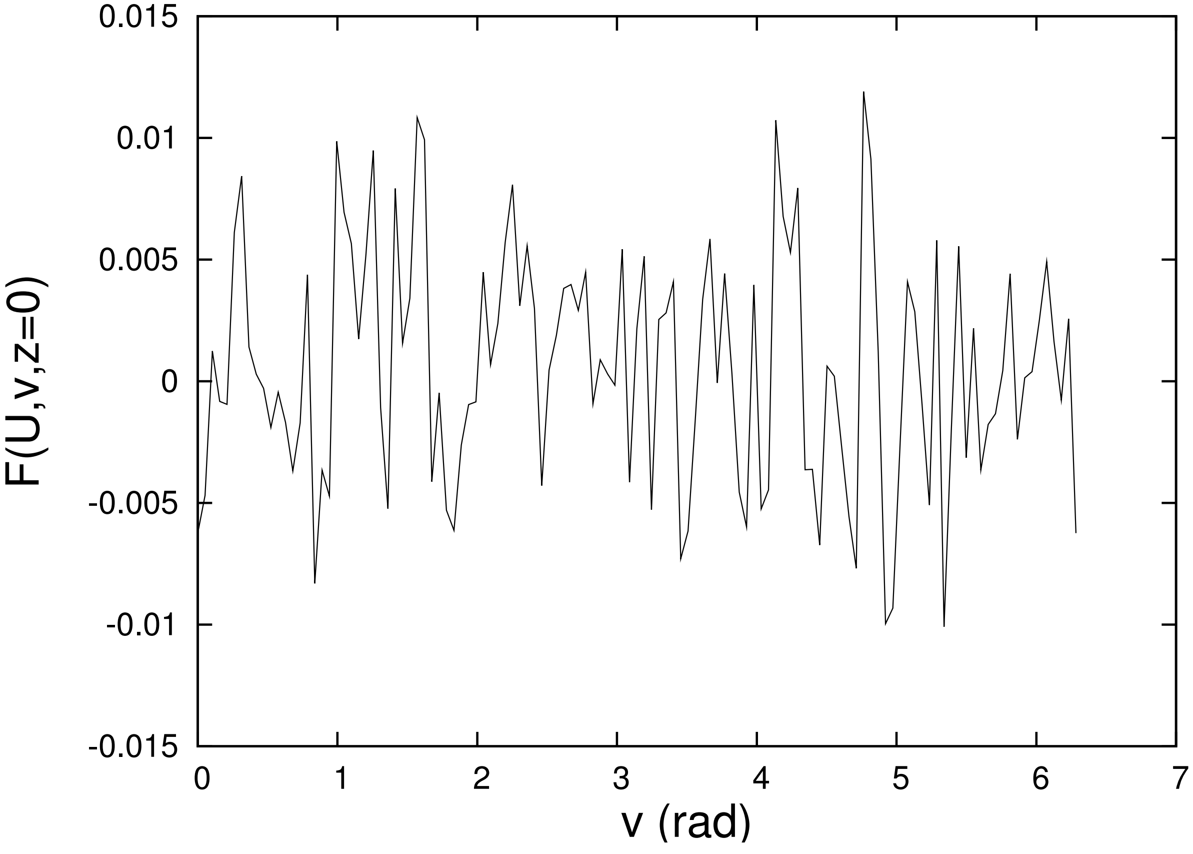


Figure 20.3.1: The function $F(U, v, z = 0)$ produced by a pure noise field.

According to (14.4.72) the second step is to perform a Fourier transform to produce $\tilde{F}(v, k)$. Figures 3.5 and 3.6 display the real parts of $\tilde{F}(v = \pi/2, k)$ for the two different seeds. The imaginary parts are comparable. Compare with Figure 16.2.3. In both cases $\tilde{F}(v = \pi/2, k)$ has support for large $|k|$ as is expected for noisy data.

The third step is to compute the Mathieu coefficients defined by (16.2.20) and (16.2.21). As described in Section 16.2.1, for the monopole doublet we are particularly interested in the coefficients $\tilde{\tilde{F}}_r^s(k)$ for odd r. Figures 3.7 and 3.8 display the real parts of the first few

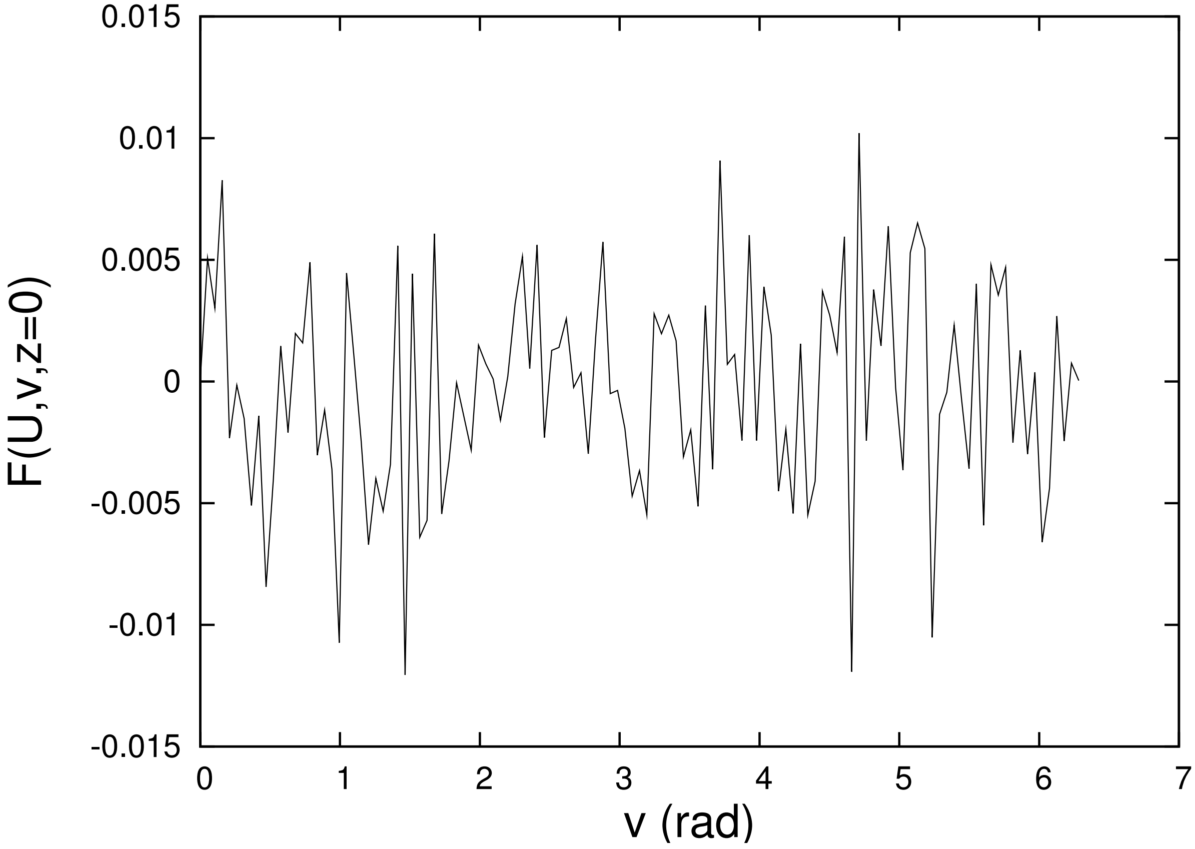


Figure 20.3.2: The function $F(U, v, z = 0)$ produced by a pure noise field arising from a second different random number seed.

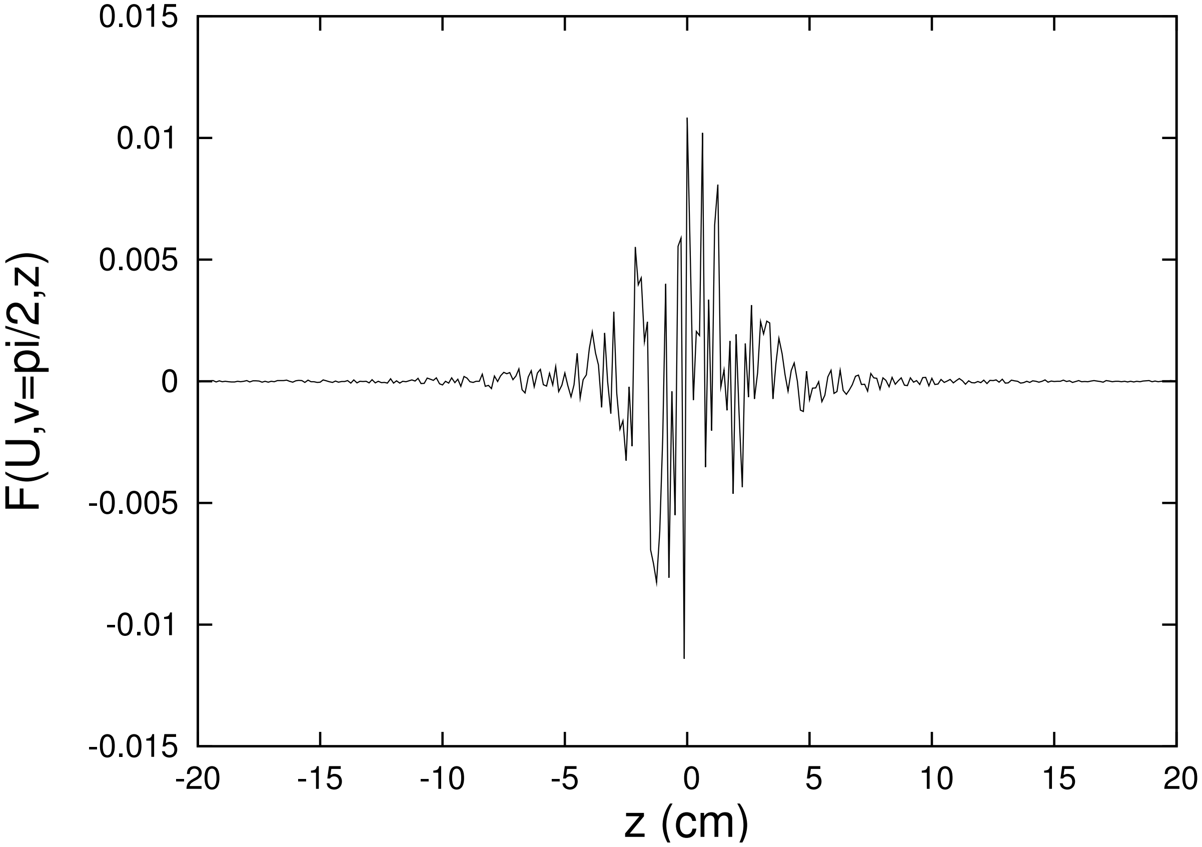


Figure 20.3.3: The function $F(U, v = \pi/2, z)$ produced by a pure noise field.

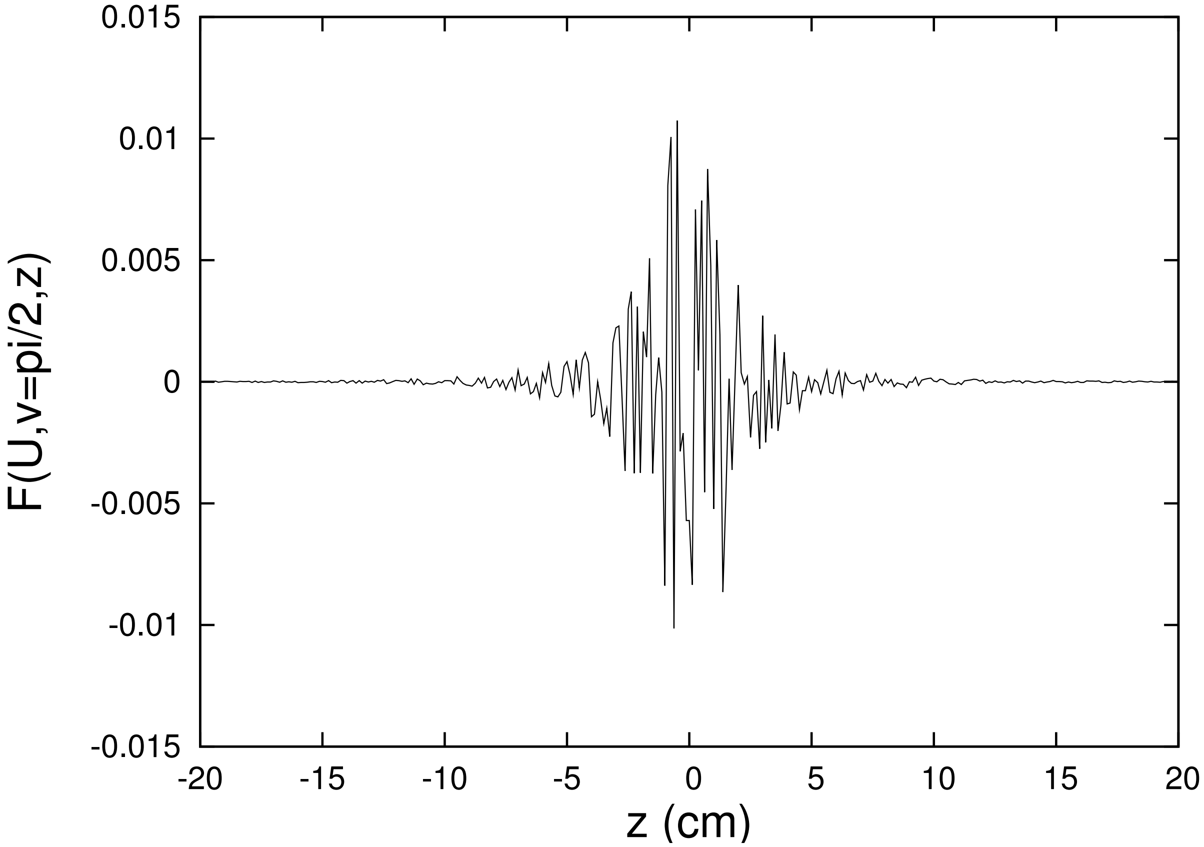


Figure 20.3.4: The function $F(U, v = \pi/2, z)$ produced by a pure noise field arising from a second different random number seed.

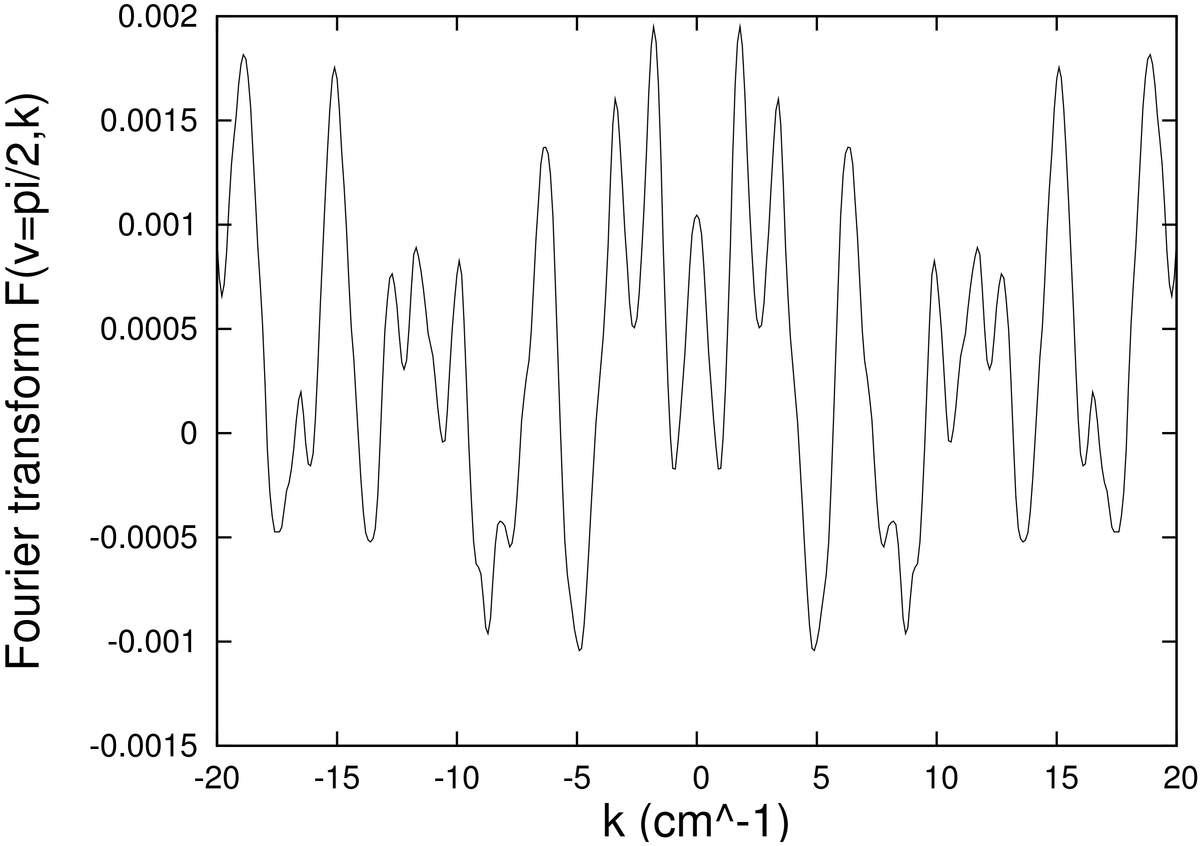


Figure 20.3.5: Real part of $\tilde{F}(v = \pi/2, k)$ produced by a pure noise field. The imaginary part is comparable.

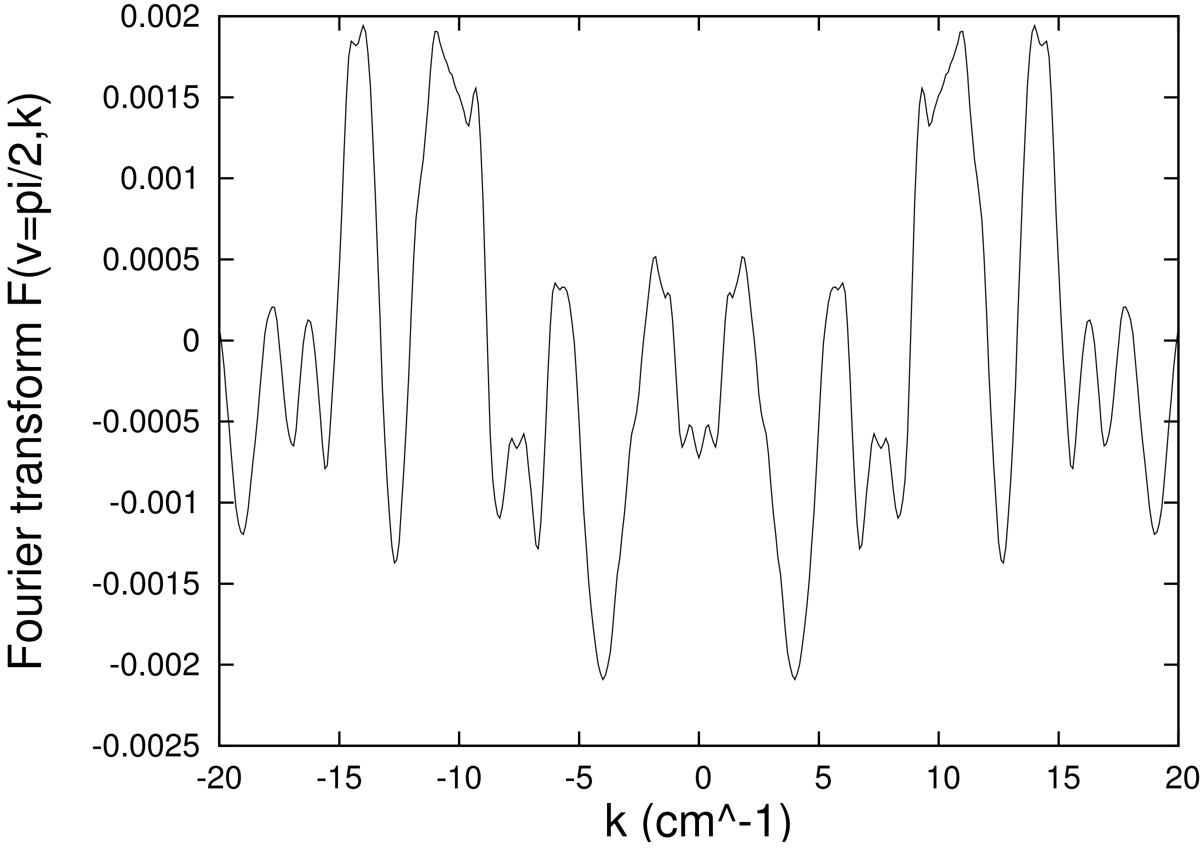


Figure 20.3.6: Real part of $\tilde{F}(v = \pi/2, k)$ produced by a pure noise field arising from a second different random number seed. The imaginary part is comparable.

of these, namely those for $r = 1, 3$, and 5, for the two different seeds. Compare with Figure 16.2.4. Note that they also have support for large $|k|$.

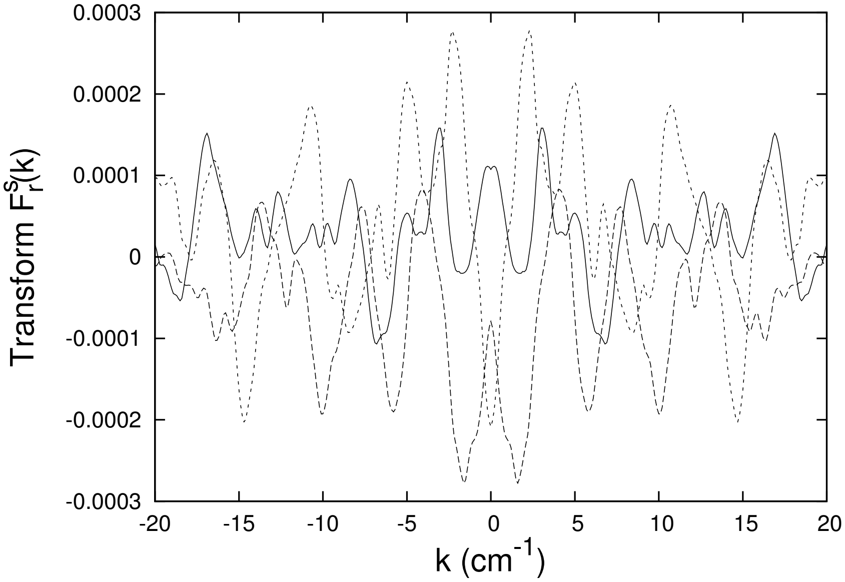


Figure 20.3.7: Real parts of the first few functions $\tilde{\tilde{F}}_r^s(k)$, those for $r = 1, 3$, and 5, produced by a pure noise field. The imaginary parts are comparable.

Suppose, for example, that we now again wish to compute the $C_{1,s}^{[0]}(z)$ produced by the pure noise field. Then, according to (14.8.86) and (14.8.84), we must compute the quantity $kG_{1,s}(k)$ by multiplying the $\tilde{\tilde{F}}_r^s(k)$ by the kernels $k\beta_1^r(k)/Se_r'(U, q)$ shown in Figure 16.2.7, and then summing over r. As described in Section 16.2.3, this sum is terminated at $r_{\max}(1) = 11$. Figures 3.9 and 3.10 display the real parts of the $kG_{1,s}(k)$ for the two different choices of random number seed. The imaginary parts are comparable. Note that, like the circular cylinder case, the kernels effectively filter out all the high frequency components.

This is a good place to compare the filtering provided by the use of an elliptic cylinder with that provided by the use of a circular cylinder. Figure 3.11 shows the circular cylinder $m = 1$ kernel $[1/I_1'(kR)]$ and the first few elliptical cylinder $m = 1$ kernels $[k\beta_1^r(k)/Se_r'(U, q)]$, and Figure 3.12 shows the same kernels all normalized to 1 at $k = 0$. From Figure 3.12 we see that the elliptic kernels for $r = 1$ and $r = 3$ fall off more rapidly with $|k|$ than the circular kernel, but that they fall off less rapidly for $r \geq 5$. The r value for which this transition occurs depends on the eccentricity of the ellipse: the larger the eccentricity the larger the transition r. Moreover, we see from Figure 3.11 that the elliptic kernels for small r dominate.

Finally, we need to carry out the integration in (14.4.86). Figures 3.13 and 3.14 show, as dashed lines, the $C_{1,s}^{[0]}(z)$ so obtained for each noise realization. Also shown, as solid lines, are the $C_{1,s}^{[0]}(z)$ obtained using a circular cylinder. (See Figures 2.9 and 2.10). Evidently use of the elliptic cylinder, in this case, has reduced the effect of noise by another factor of ~ 2.5 compared to the use of a circular cylinder. Comparison of these figures with Figure 16.1.8 shows that in this study a 1% noise in field data produces about .01% error in $C_{1,s}^{[0]}(z)$ when

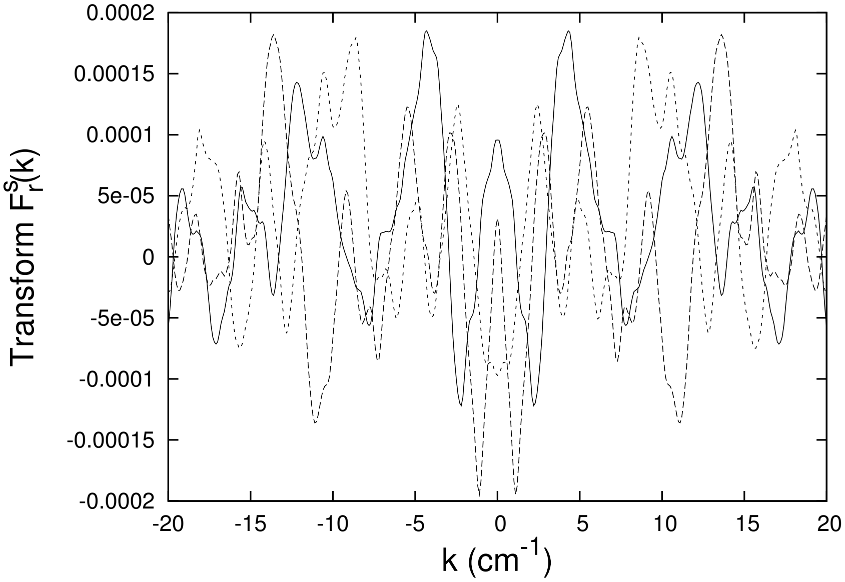


Figure 20.3.8: Real parts of the first few functions $\tilde{\tilde{F}}_r^s(k)$, those for $r = 1, 3$, and 5, produced by a pure noise field arising from a second different random number seed. The imaginary parts are comparable.

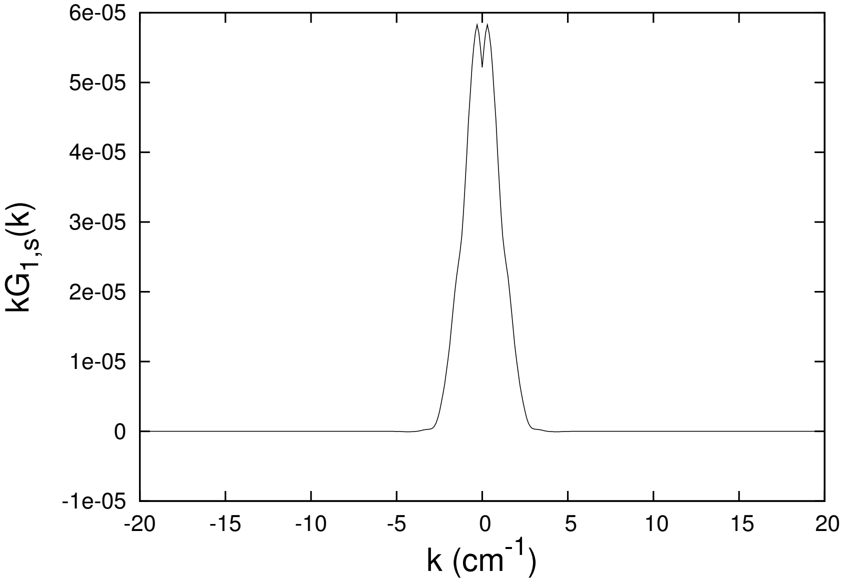


Figure 20.3.9: Real part of $kG_{1,s}(k)$ computed from $\tilde{\tilde{F}}_r^s(k)$ associated with the first seed. The imaginary part is comparable.

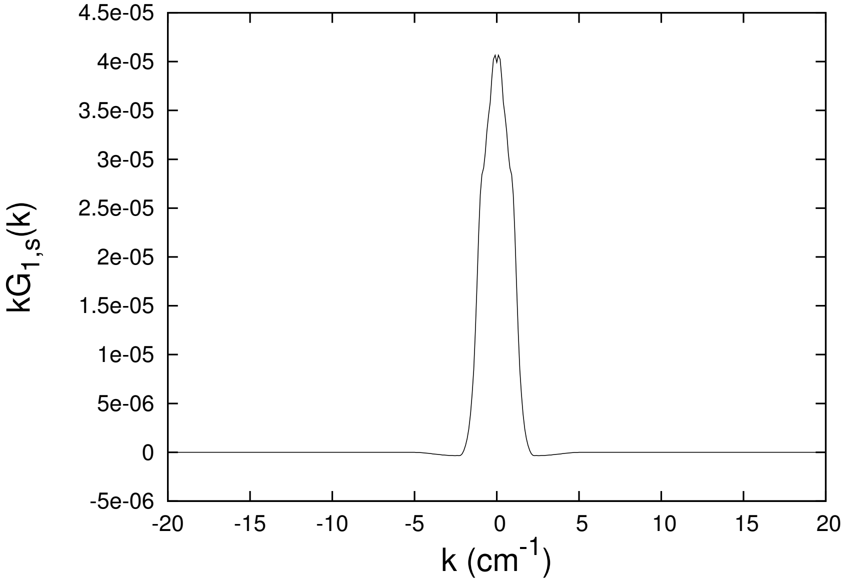


Figure 20.3.10: Real part of $kG_{1,s}(k)$ computed from $\tilde{\tilde{F}}_r^s(k)$ associated with the second seed. The imaginary part is comparable.

the elliptic cylinder is used.

What can be said about the other $C_{m,s}^{[n]}(z)$? Generally the use of an elliptic cylinder gives better results. But in some cases the circular and elliptic cylinder results are comparable, and sometimes the circular cylinder error is somewhat smaller. Results vary from seed to seed. For example, Figures 3.15 through 3.18 show the functions $C_{1,s}^{[6]}(z)$ and $C_{7,s}^{[0]}(z)$ computed using both elliptic and circular cylinders.

There are at least two remarks to be made. First, just as in the cases in Section 17.2 where the results of using circular cylinders with different radii were compared, statistical fluctuations can mask the effects of improved smoothing. Second, it is primarily the vertical magnetic field that contributes to the $C_{1,s}^{[n]}(z)$, $C_{3,s}^{[n]}(z)$, $C_{5,s}^{[n]}(z) \cdots$. However, what enters our calculation is the component of the magnetic field that is perpendicular to the surface. Although the elliptical cylinder surface has points that are farther from the axis than the circular cylinder, at these points the normal to the surface is primarily in the horizontal direction. Consequently, the points on the elliptical surface for which the field values actually contribute to the $C_{m,\alpha}^{[n]}(z)$ thus far examined are not very much farther from the axis than points on the circular surface.

We should also examine, for example, the quantities $C_{1,c}^{[n]}(z)$, $C_{3,c}^{[n]}(z)$, $C_{5,c}^{[n]}(z) \cdots$ for which the horizontal magnetic field makes substantial contributions. For these quantities we expect that noise results for circular and elliptic cylinders should be noticeably different. As a first exploratory step, let us again examine the relevant kernels. Figure 3.19 shows the circular cylinder $m = 1$ kernel $[1/I_1'(kR)]$ and the first few elliptical cylinder $m = 1$ kernels $[k\alpha_1^r(k)/Ce_r'(U,q)]$, and Figure 3.20 shows the same kernels all normalized to 1 at $k = 0$. From Figure 3.20 we see that the elliptic kernels for $r = 1$ and $r = 3$ fall off more rapidly with $|k|$ than the circular kernel, but that they fall off less rapidly for $r \geq 5$. The r value for which this transition occurs depends on the eccentricity of the ellipse: the larger the

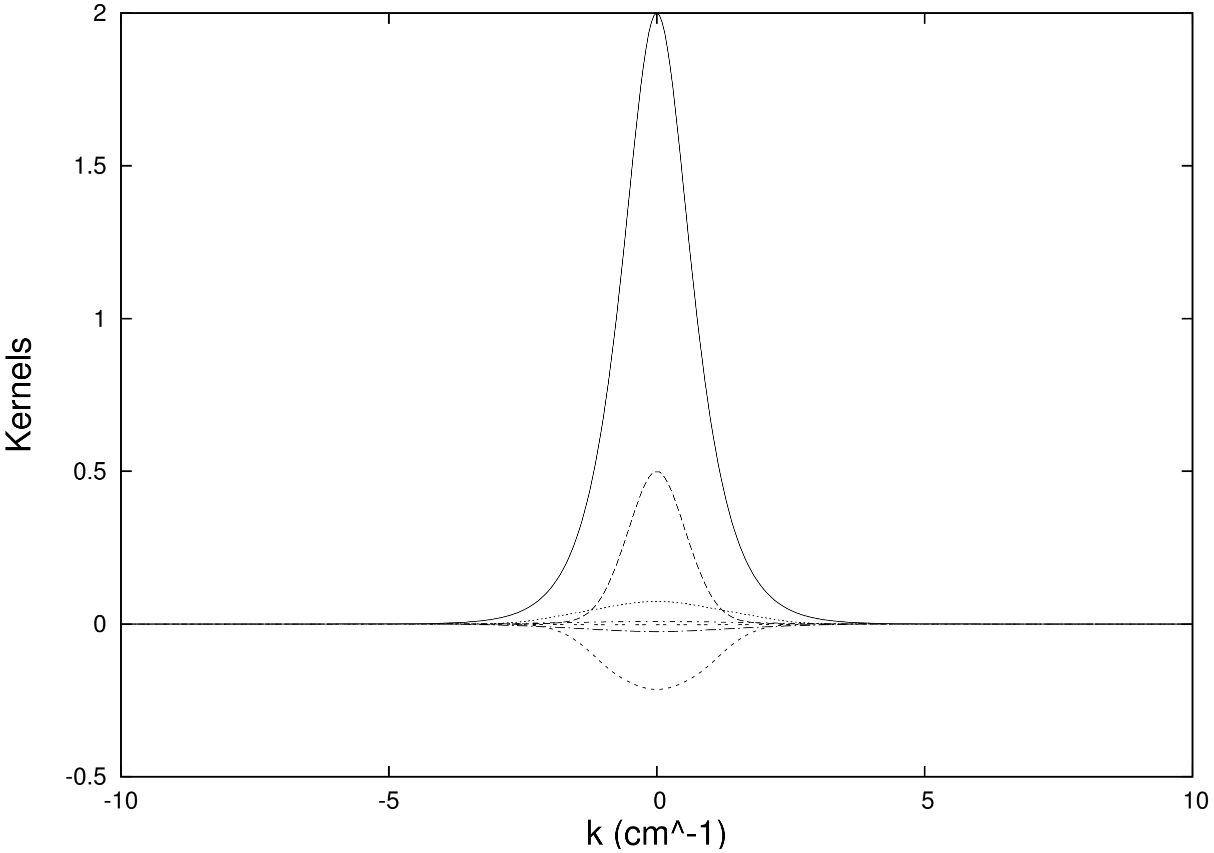


Figure 20.3.11: A comparison of the circular cylinder $m = 1$ kernel $[1/I_1'(kR)]$, shown as a solid line, and the first few relevant elliptical cylinder $m = 1$ kernels $[k\beta_1^r(k)/Se_r'(U,q)]$, namely those for $r = 1, 3, 5, 7, 9,$ and 11, shown as a dashed lines. The elliptic kernels alternate in sign, and their magnitude at $k = 0$ decreases with increasing r. See Figure 16.2.7.

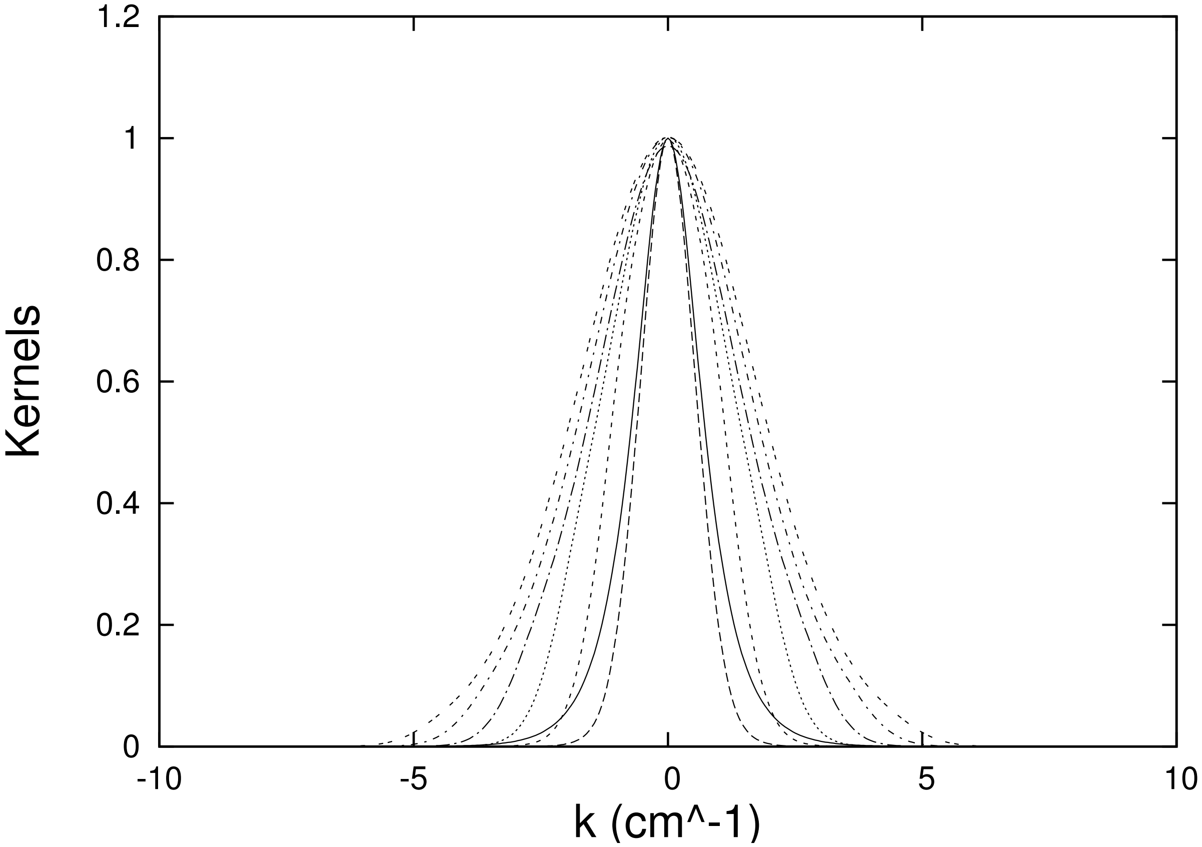


Figure 20.3.12: A comparison of the circular cylinder $m = 1$ kernel $[1/I_1'(kR)]$ and the first few elliptic cylinder $m = 1$ kernels $[k\beta_1^r(k)/Se_r'(U,q)]$, all normalized to 1 at $k = 0$.

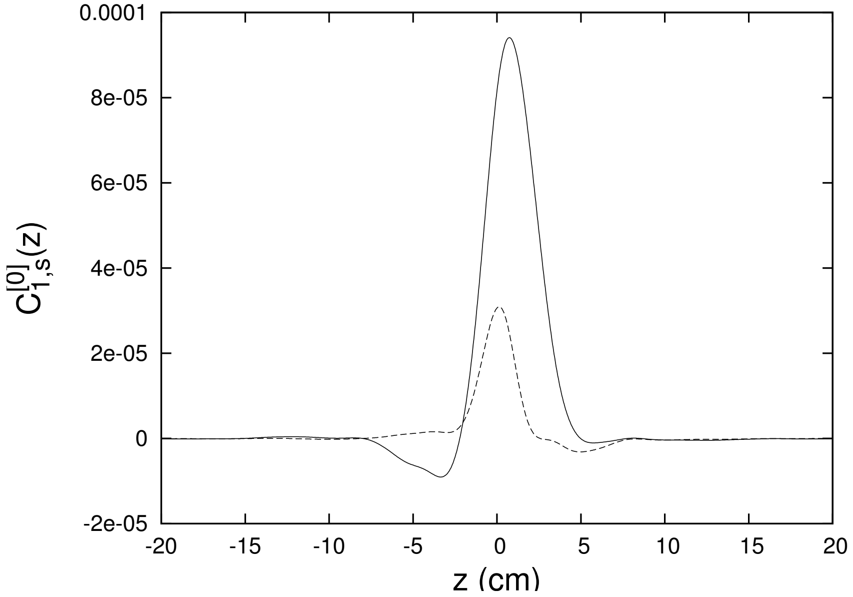


Figure 20.3.13: Dashed line: The function $C_{1,s}^{[0]}(z)$ produced by a pure noise field and using an elliptic cylinder. Solid line: The function $C_{1,s}^{[0]}(z)$ produced by a pure noise field and using a circular cylinder.

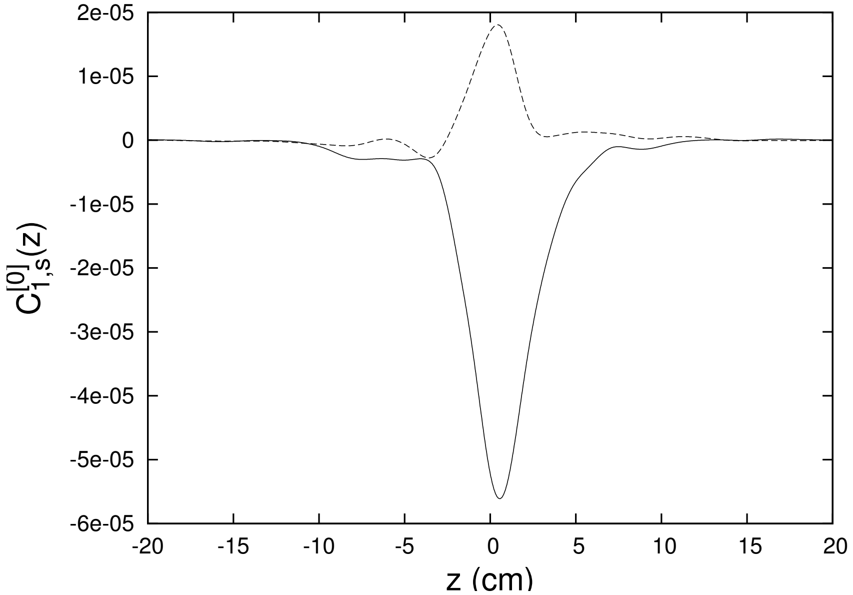


Figure 20.3.14: Results for the second random number seed. Dashed line: The function $C_{1,s}^{[0]}(z)$ produced by a pure noise field and using an elliptic cylinder. Solid line: The function $C_{1,s}^{[0]}(z)$ produced by a pure noise field and using a circular cylinder.

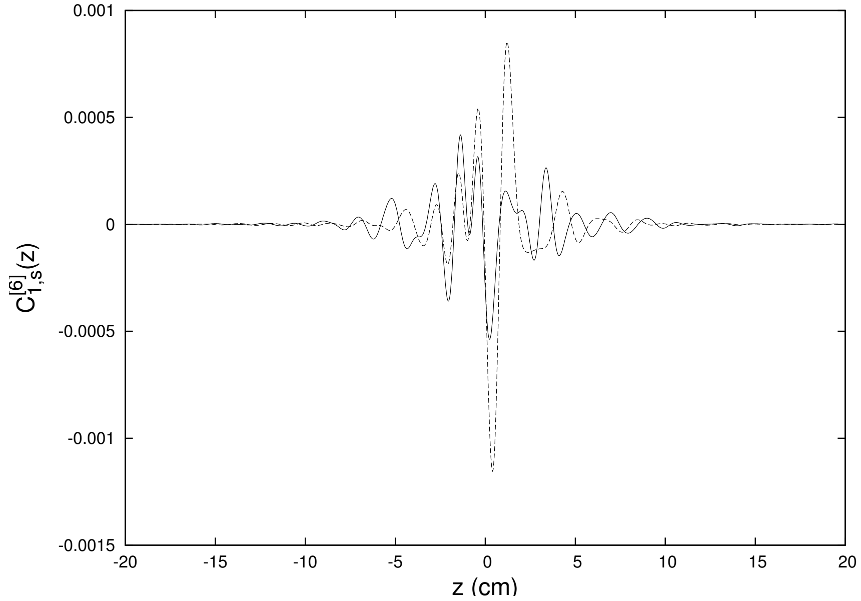


Figure 20.3.15: The function $C^{[6]}_{1,s}(z)$ produced by a pure noise field. Dashed line: Elliptic cylinder result. Solid line: Circular cylinder result.

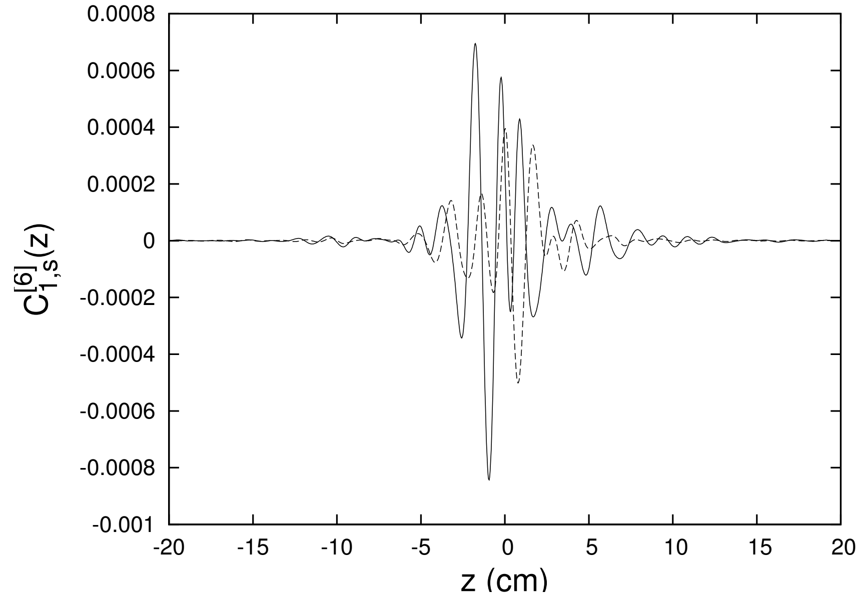


Figure 20.3.16: The function $C^{[6]}_{1,s}(z)$ produced by a pure noise field arising from a second different random number seed. Dashed line: Elliptic cylinder result. Solid line: Circular cylinder result.

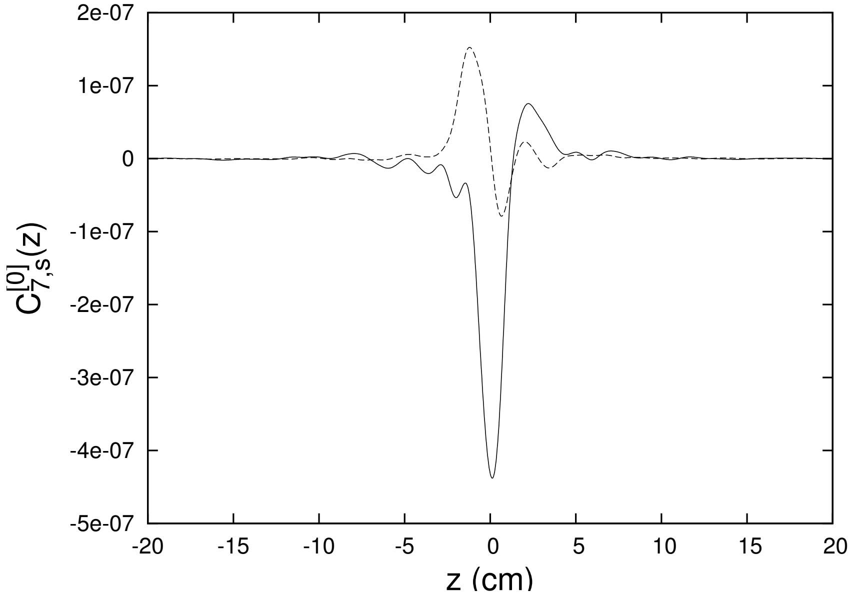


Figure 20.3.17: The function $C^{[0]}_{7,s}(z)$ produced by a pure noise field. Dashed line: Elliptic cylinder result. Solid line: Circular cylinder result.

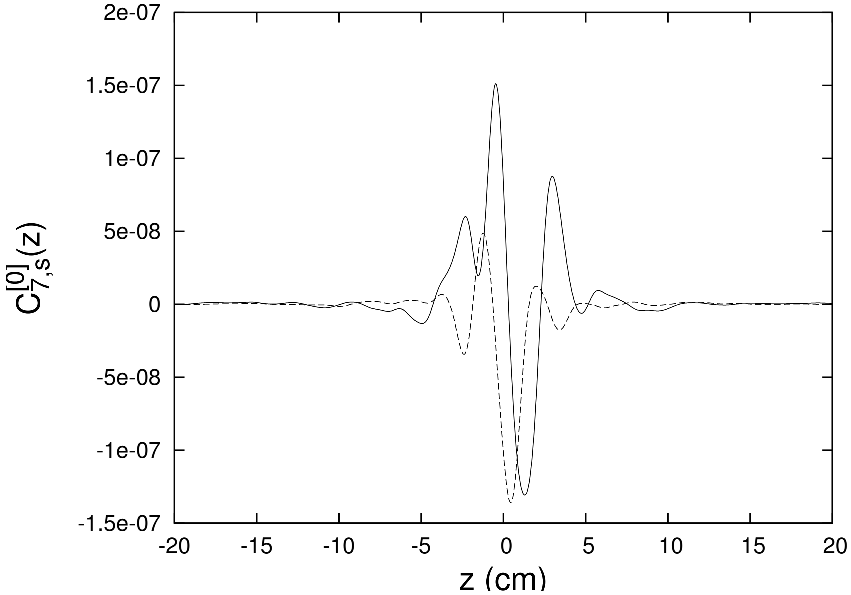


Figure 20.3.18: The function $C^{[0]}_{7,s}(z)$ produced by a pure noise field arising from a second different random number seed. Dashed line: Elliptic cylinder result. Solid line: Circular cylinder result.

eccentricity the larger the transition r. Moreover, we see from Figure 3.19 that the elliptic kernels for small r dominate.

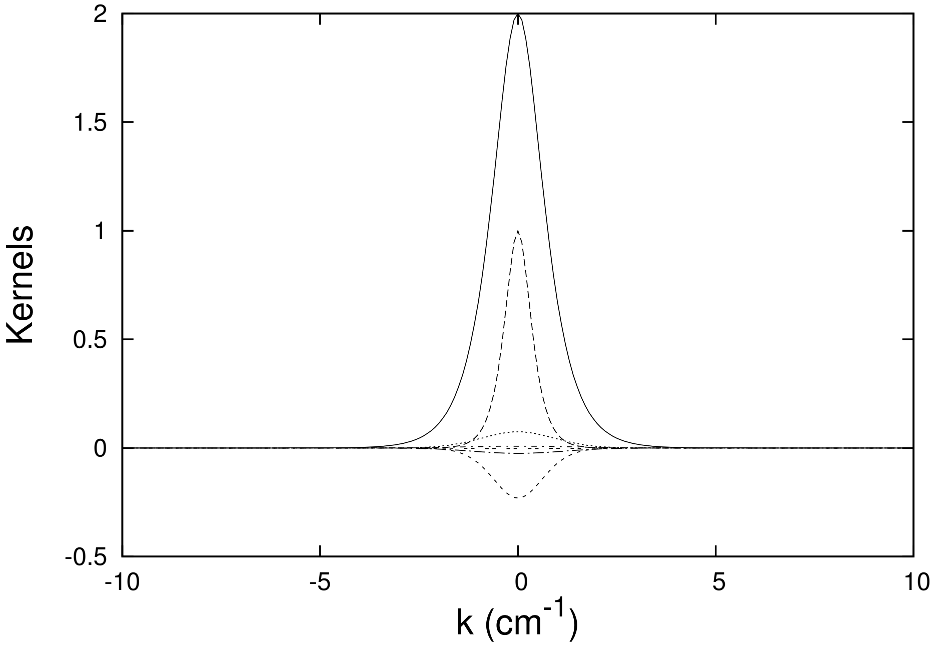


Figure 20.3.19: A comparison of the circular cylinder $m = 1$ kernel $[1/I_1'(kR)]$, shown as a solid line, and the first few relevant elliptical cylinder $m = 1$ kernels $[k\alpha_1^r(k)/Ce_r'(U,q)]$, namely those for $r = 1, 3, 5, 7, 9$, and 11, shown as a dashed lines. The elliptic kernels alternate in sign, and their magnitude at $k = 0$ decreases with increasing r.

Now we are prepared to examine the $C_{m,c}^{[n]}(z)$. Figures 3.21 and 3.22 show $C_{1,c}^{[0]}(z)$ for each noise realization. And Figures 3.23 through 3.26 show the associated quantities $C_{1,c}^{[6]}(z)$ and $C_{7,c}^{[0]}(z)$. Again there are statistical fluctuations, but the general trend is that the use of an elliptic cylinder yields on-axis gradients that have less sensitivity to errors in the grid data.

To study the problem more thoroughly, we should examine the results for a large number of seeds. We expect that when such results are examined, the effect of noise will average to zero (because the effect of noise can have either sign), but there will be a nonzero variance. What should be verified is that the variance is smaller for elliptic cylinders than for circular cylinders.

Figures 3.27 and 3.28 display circular and elliptical cylinder results for the $C_{1,c}^{[0]}(z)$ obtained for 12 seeds. They also show $< C_{1,c}^{[0]}(z) >$, the average of these results. Figures 3.29 through 3.32 do the same for $C_{1,c}^{[6]}(z)$ and $C_{7,c}^{[0]}(z)$. Evidently the averaged results are smaller than the individual results, thereby indicating that the average does indeed approach zero as the number of seeds is increased.

Figures 3.33 through 3.35 show the quantities $\{< [C_{1,c}^{[0]}(z)]^2 >\}^{1/2}$, $\{< [C_{1,c}^{[6]}(z)]^2 >\}^{1/2}$, and $\{< [C_{7,c}^{[0]}(z)]^2 >\}^{1/2}$, the root-mean-square values based on these 12 seeds. Results are shown for both the circular and elliptic cylinder. We see that, for the two cases of

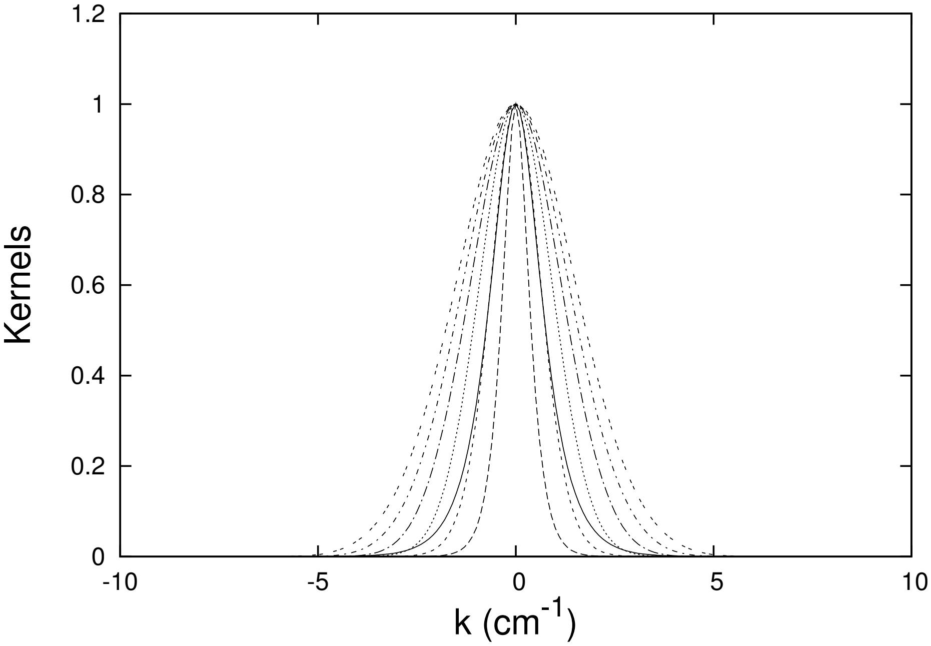


Figure 20.3.20: A comparison of the circular cylinder $m = 1$ kernel $[1/I_1'(kR)]$ and the first few elliptic cylinder $m = 1$ kernels $[k\alpha_1^r(k)/Ce_r'(U, q)]$, all normalized to 1 at $k = 0$.

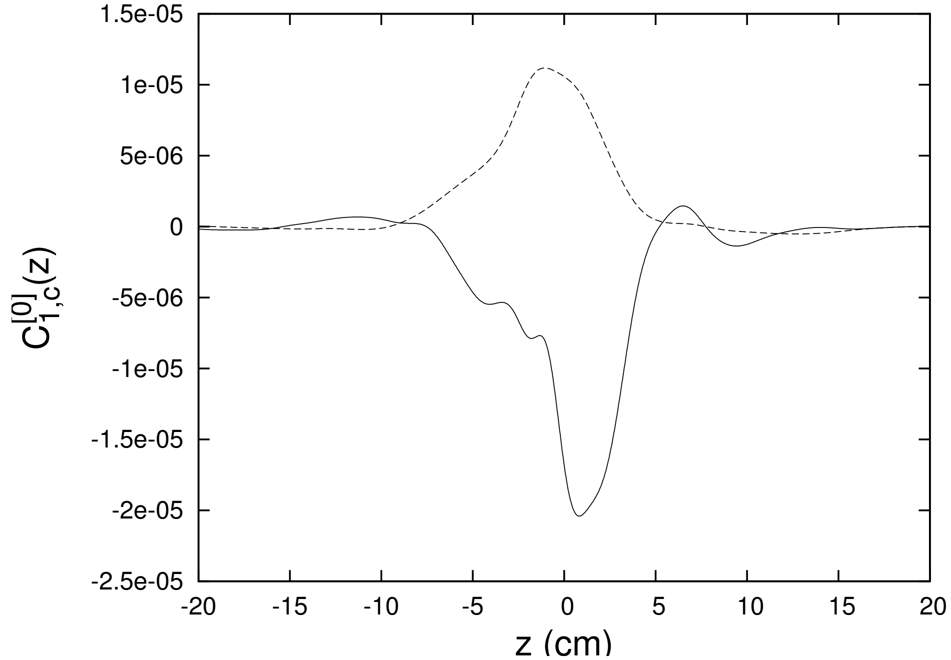


Figure 20.3.21: Dashed line: The function $C^{[0]}_{1,c}(z)$ produced by a pure noise field and using an elliptic cylinder. Solid line: The function $C^{[0]}_{1,c}(z)$ produced by a pure noise field and using a circular cylinder.

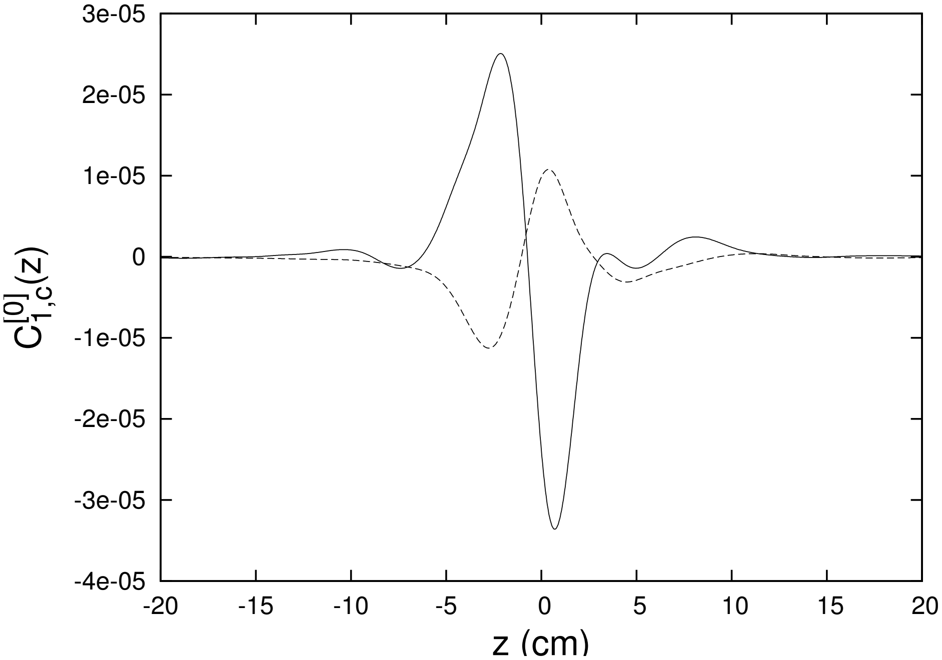


Figure 20.3.22: Results for the second random number seed. Dashed line: The function $C^{[0]}_{1,c}(z)$ produced by a pure noise field and using an elliptic cylinder. Solid line: The function $C^{[0]}_{1,c}(z)$ produced by a pure noise field and using a circular cylinder.

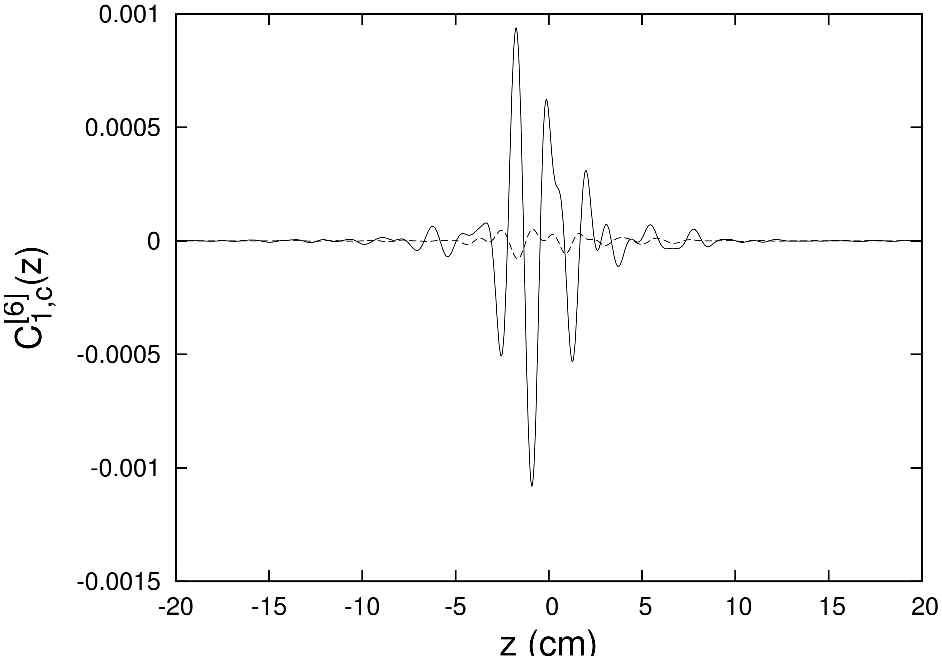


Figure 20.3.23: The function $C^{[6]}_{1,c}(z)$ produced by a pure noise field. Dashed line: Elliptic cylinder result. Solid line: Circular cylinder result.

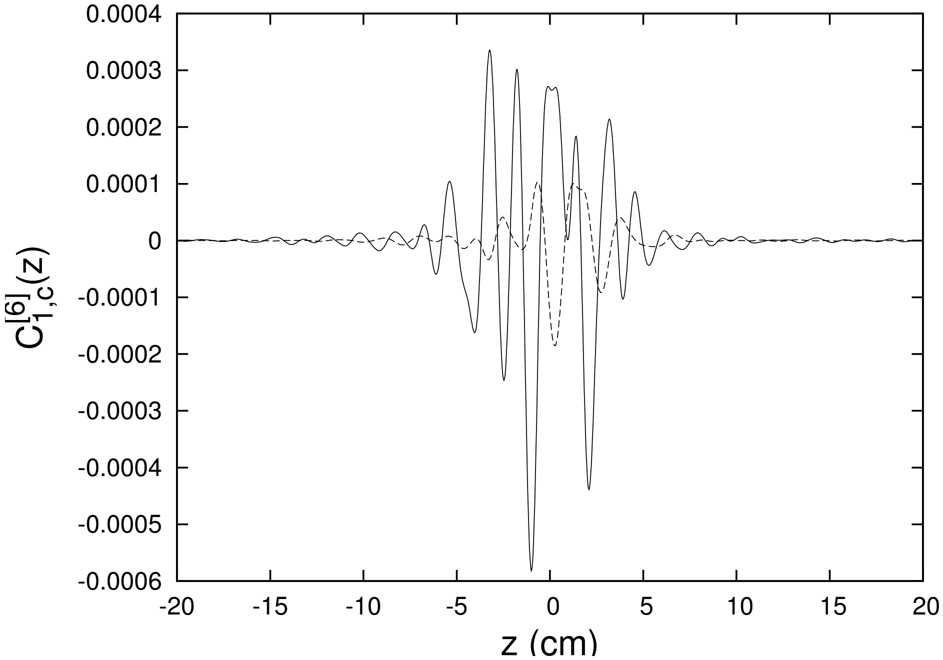


Figure 20.3.24: The function $C^{[6]}_{1,c}(z)$ produced by a pure noise field arising from a second different random number seed. Dashed line: Elliptic cylinder result. Solid line: Circular cylinder result.

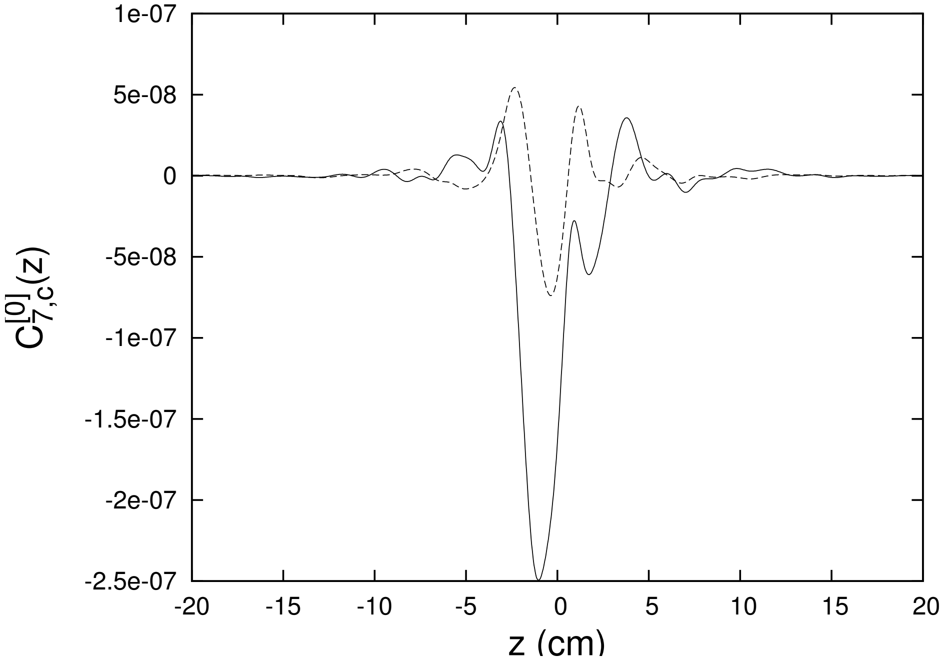


Figure 20.3.25: The function $C^{[0]}_{7,c}(z)$ produced by a pure noise field. Dashed line: Elliptic cylinder result. Solid line: Circular cylinder result.

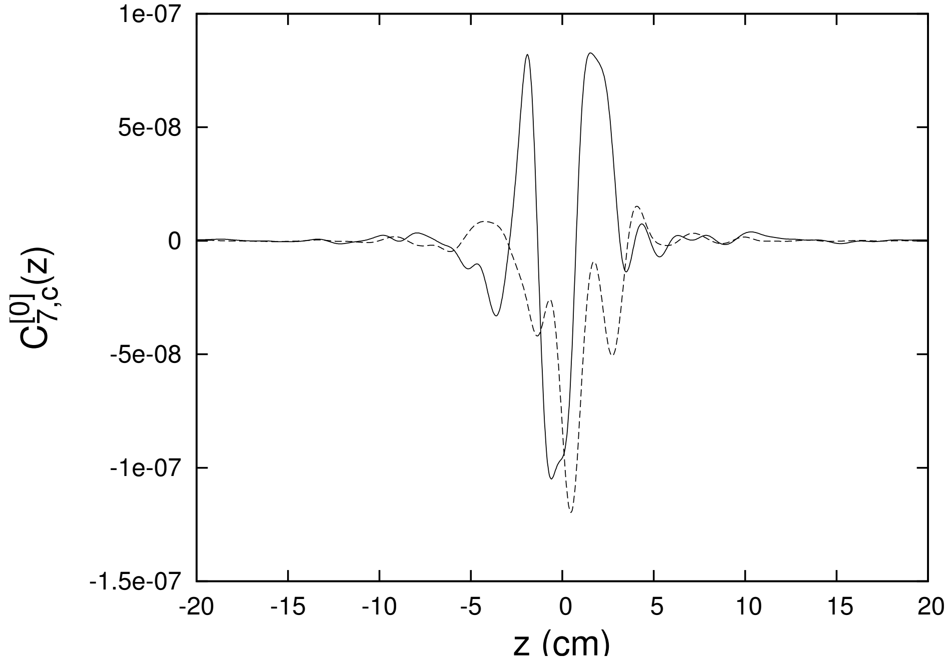


Figure 20.3.26: The function $C^{[0]}_{7,c}(z)$ produced by a pure noise field arising from a second different random number seed. Dashed line: Elliptic cylinder result. Solid line: Circular cylinder result.

$\{< [C^{[6]}_{1,c}(z)]^2 >\}^{1/2}$ and $\{< [C^{[0]}_{7,c}(z)]^2 >\}^{1/2}$, the root-mean-square values are indeed smaller for the elliptic cylinder compared to the circular cylinder.

The case for $\{< [C^{[0]}_{1,c}(z)]^2 >\}^{1/2}$ is inconclusive. It appears that the number of samples is still too small so that statistical fluctuations still overwhelm the expected effect. This hypothesis is validated by Figure 3.36. It shows the functions $C^{[0]}_{1,c}(z)$ produced by assigning a nonzero field value to only a *single* grid point. Consider the field value

$$(B_x, B_y, B_z) = (.01\,\mathrm{T}, 0, 0). \tag{20.3.1}$$

For the elliptic cylinder case we assign this field value to the grid point

$$(x, y, z) = (4\,\mathrm{cm}, 0, 0). \tag{20.3.2}$$

And for the circular cylinder case we assign this field value to the grid point

$$(x, y, z) = (2\,\mathrm{cm}, 0, 0). \tag{20.3.3}$$

All other grid points are assigned vanishing field values. As the figure shows, the elliptic cylinder result for $C^{[0]}_{1,c}(z)$ is indeed smaller than the circular cylinder result.

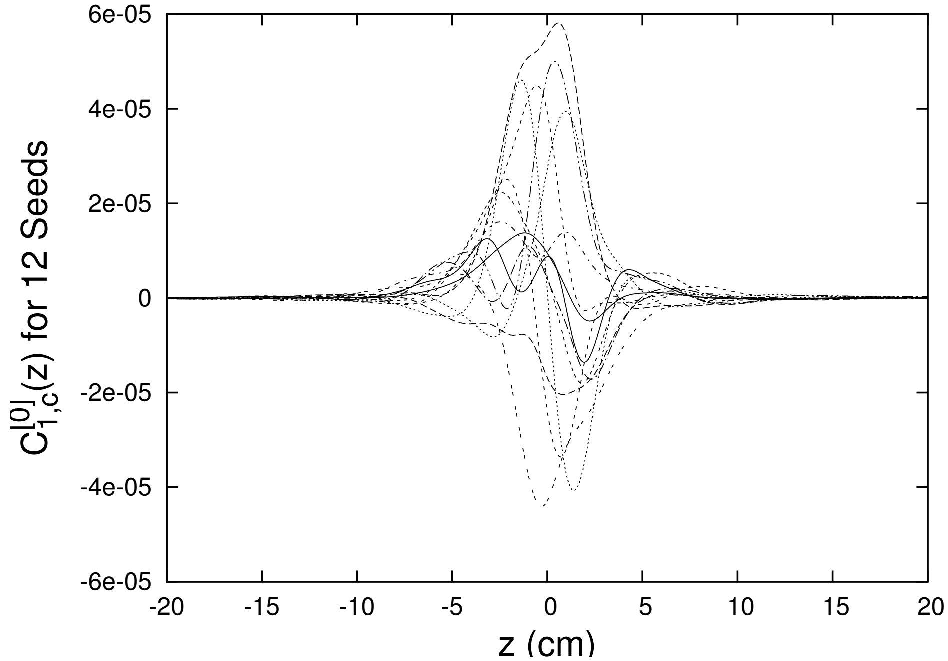


Figure 20.3.27: The functions $C^{[0]}_{1,c}(z)$ produced by pure noise fields generated by 12 seeds using data on a circular cylinder. Broken lines: Results from individual seeds. Solid line: Averaged results.

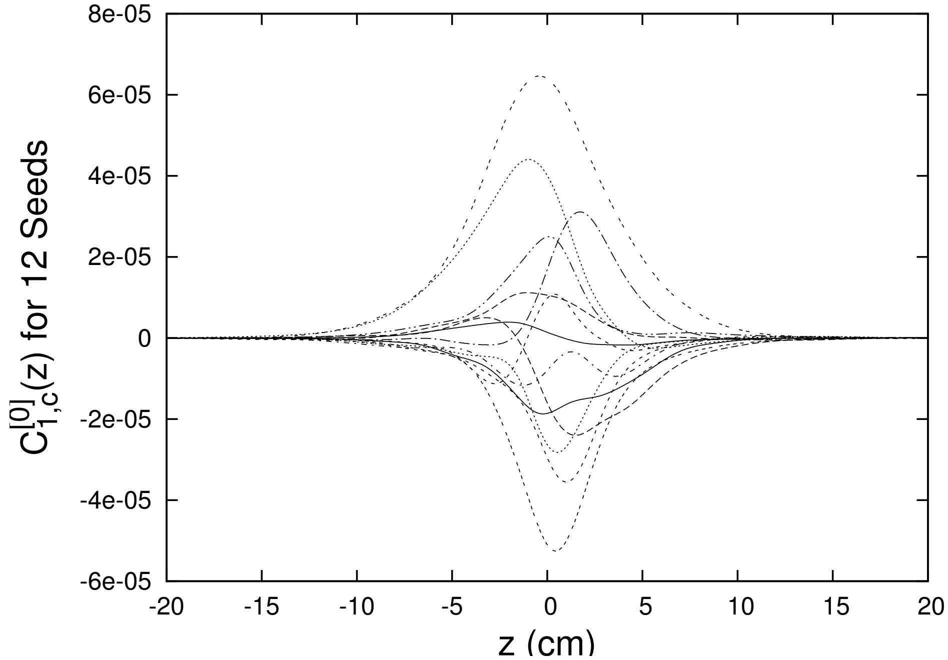


Figure 20.3.28: The functions $C^{[0]}_{1,c}(z)$ produced by pure noise fields generated by 12 seeds using data on an elliptical cylinder. Broken lines: Results from individual seeds. Solid line: Averaged results.

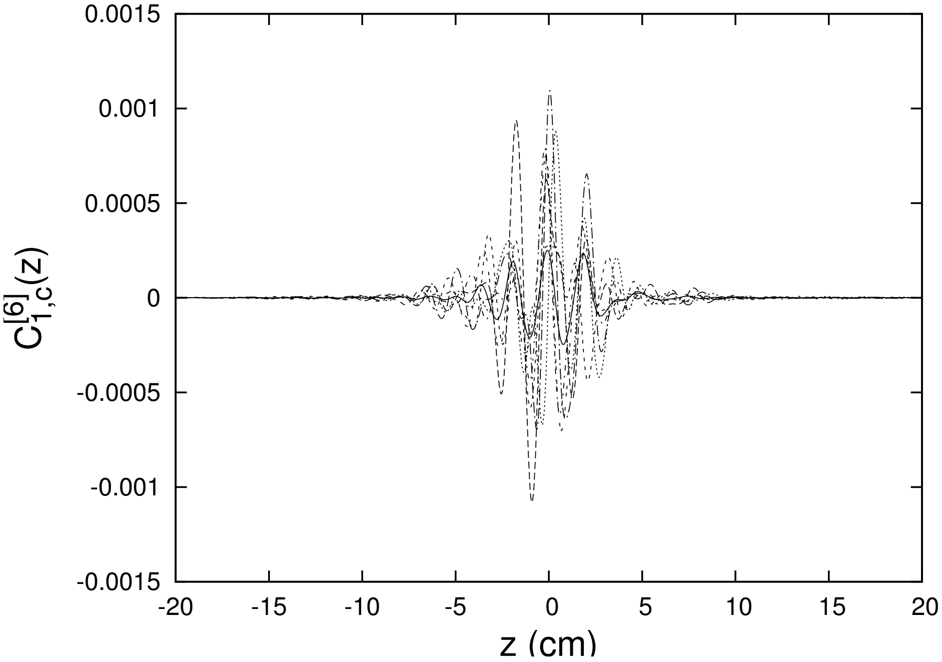


Figure 20.3.29: The functions $C^{[6]}_{1,c}(z)$ produced by pure noise fields generated by 12 seeds using data on a circular cylinder. Broken lines: Results from individual seeds. Solid line: Averaged results.

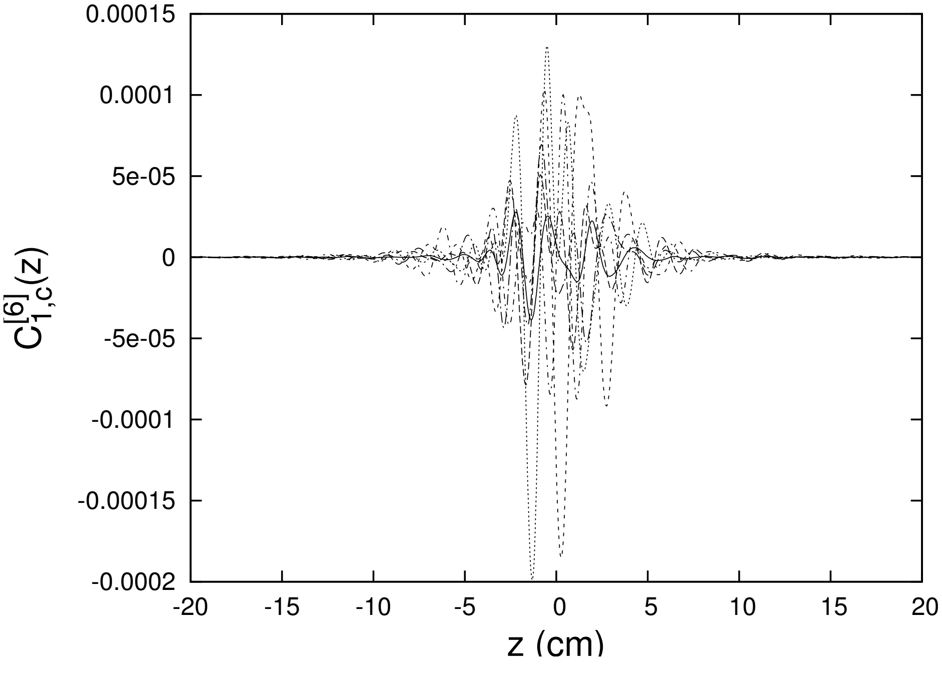


Figure 20.3.30: The functions $C^{[6]}_{1,c}(z)$ produced by pure noise fields generated by 6 seeds using data on an elliptical cylinder. Broken lines: Results from individual seeds. For clarity, in this graphic only results for 6 seeds are shown. Solid line: Averaged results. As in other related figures, results for 12 seeds were used in computing the average.

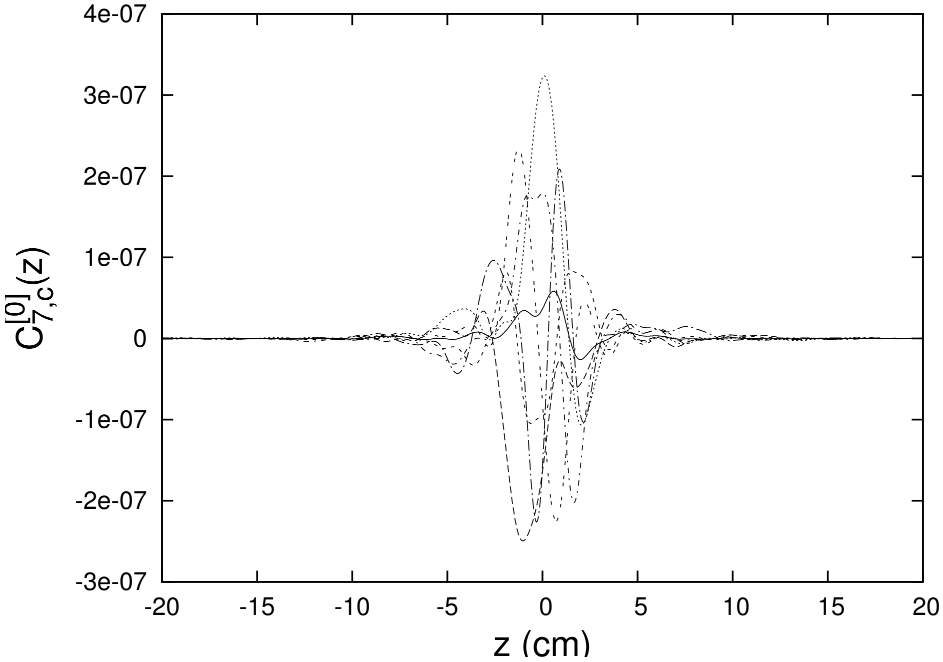


Figure 20.3.31: The functions $C^{[0]}_{7,c}(z)$ produced by pure noise fields generated by 12 seeds using data on a circular cylinder. Broken lines: Results from individual seeds. Solid line: Averaged results.

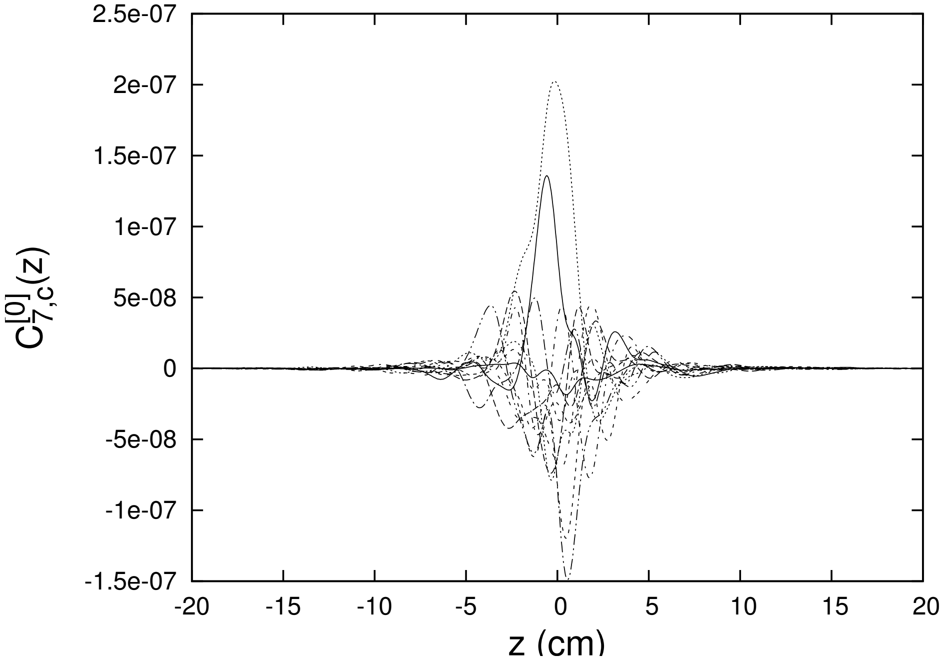


Figure 20.3.32: The functions $C^{[0]}_{7,c}(z)$ produced by pure noise fields generated by 12 seeds using data on an elliptical cylinder. Broken lines: Results from individual seeds. Solid line: Averaged results.

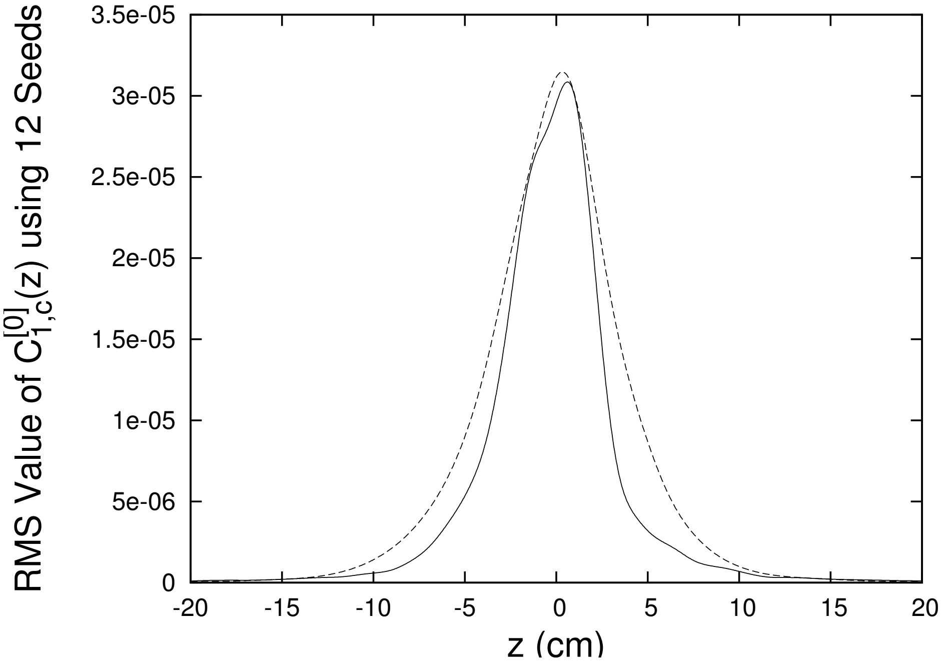


Figure 20.3.33: The function $\{<[C^{[0]}_{1,c}(z)]^2>\}^{1/2}$ produced by 12 pure noise fields. Dashed line: Result from using an elliptic cylinder. Solid line: Result from using a circular cylinder.

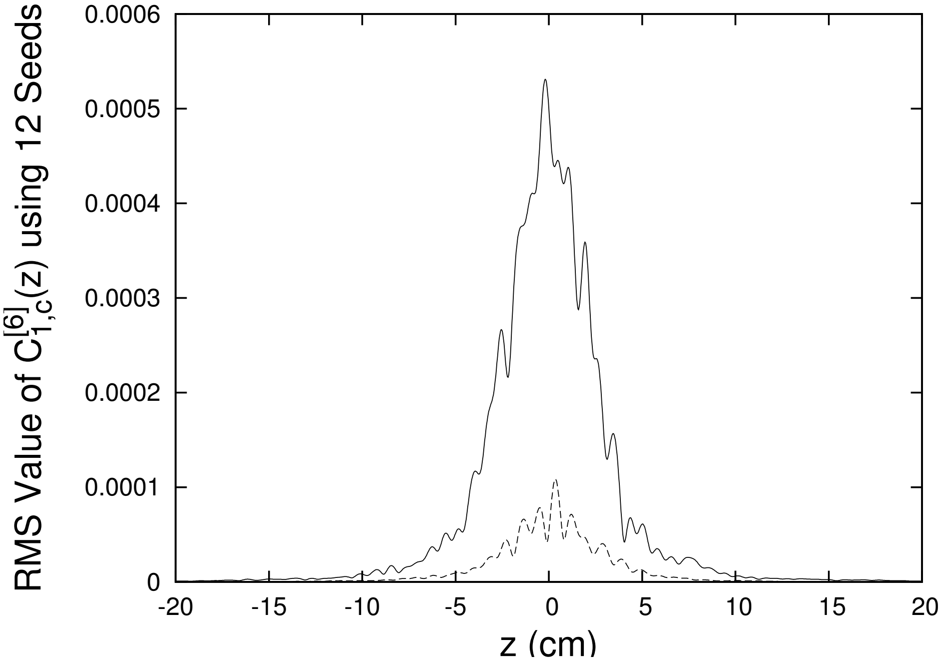


Figure 20.3.34: The function $\{<[C^{[6]}_{1,c}(z)]^2>\}^{1/2}$ produced by 12 pure noise fields. Dashed line: Result from using an elliptic cylinder. Solid line: Result from using a circular cylinder.

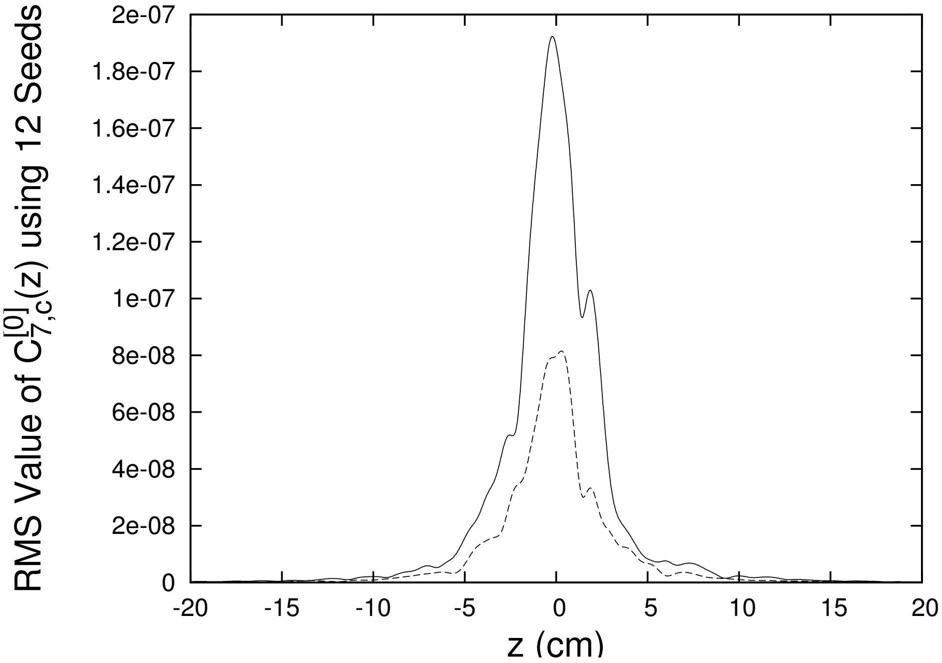


Figure 20.3.35: The function $\{< [C^{[0]}_{7,c}(z)]^2 >\}^{1/2}$ produced by 12 pure noise fields. Dashed line: Result from using an elliptic cylinder. Solid line: Result from using a circular cylinder.

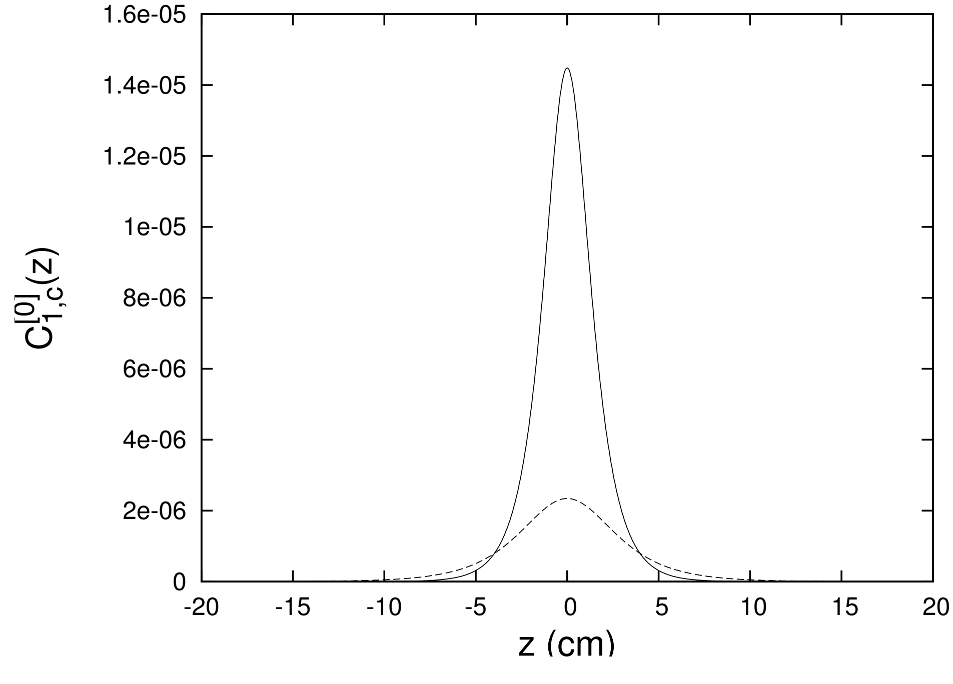


Figure 20.3.36: The functions $C^{[0]}_{1,c}(z)]$ produced by assigning a non-zero field value to only a single grid point. Dashed line: Result from using an elliptic cylinder. Solid line: Result from using a circular cylinder.

20.4 Rectangular Cylinders

Bibliography

General References

[1] M. Venturini and A. Dragt, "Accurate Computation of Transfer Maps from Magnetic Field Data", *Nuclear Instruments and Methods* **A427** (1991), p. 387.

[2] M. Venturini, Transfer Map for Printed-Circuit Magnetic Quadrupoles, Technical Note, Dept. of Physics, Univ. of Maryland (1995).

[3] C. Mitchell, "Calculation of Realistic Charged-Particle Transfer Maps", University of Maryland Physics Department Ph.D. Thesis (2007).

[4] C. Mitchell and A. Dragt, "Accurate transfer maps for realistic beam-line elements: Part I, straight elements", *Physical Review Special Topics - Accelerators and Beams* **13** 064001 (2010).

Chapter 21

Realistic Transfer Maps for General Straight Beam-Line Elements

Chapter 15 described cylindrical harmonic expansions for straight elements. This chapter utilizes these expansions and applies surface methods to several common magnetic beam-line elements.[1] It also summarizes various cases in which fields can be computed analytically. Particular attention is devoted to the way in which fringe fields fall off with increasing distance.

21.1 Solenoids

21.1.1 Preliminaries

A solenoid is a beam-line element whose field is described by a cylindrical harmonic expansion that contains (ideally) only an $m = 0$ term. We recall from Section 15.2.3 that in this case the magnetic scalar potential ψ has the expansion

$$\psi(x, y, z) = \psi_0(x, y, z) = C_0^{[0]}(z) - (1/4)(x^2 + y^2)C_0^{[2]}(z) + \cdots . \tag{21.1.1}$$

See (15.2.57). Correspondingly, the associated magnetic field has the expansion

$$B_x = \partial_x\psi_0 = -(1/2)xC_0^{[2]}(z) + \cdots , \tag{21.1.2}$$

$$B_y = \partial_y\psi_0 = -(1/2)yC_0^{[2]}(z) + \cdots , \tag{21.1.3}$$

$$B_z = \partial_z\psi_0 = C_0^{[1]}(z) - (1/4)(x^2 + y^2)C_0^{[3]}(z) + \cdots . \tag{21.1.4}$$

Finally, according to Section (15.4.1), there is a suitable associated vector potential $\hat{\boldsymbol{A}}^0$ given by the relation

$$\hat{A}_x^0 = -yU, \tag{21.1.5}$$

$$\hat{A}_y^0 = xU, \tag{21.1.6}$$

$$\hat{A}_z^0 = 0, \tag{21.1.7}$$

[1] Electrostatic beam-line elements can be treated in an analogous way.

where U is defined to be

$$U(\rho, z) = (1/2) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! (\ell+1)!} C_0^{[2\ell+1]}(z) \rho^{2\ell} \tag{21.1.8}$$

with

$$\rho^2 = x^2 + y^2. \tag{21.1.9}$$

From (1.2) through (1.9) we see that both $\boldsymbol{B}$ and $\hat{\boldsymbol{A}}^0$ are completely specified in terms of a single "master" function $C_0^{[1]}(z)$ and its derivatives. Moreover, according to (1.4), the on-axis field is given by the relation

$$B_z(0, 0, z) = C_0^{[1]}(z). \tag{21.1.10}$$

We will next see what can be said about the master function $C_0^{[1]}(z)$ in various cases.

21.1.2 Qualitatively Correct Iron-Dominated Solenoid Model

We next consider the case of an iron-dominated solenoid. Suppose an approximating signum function is defined by the rule

$$\mathrm{sgn}(z, a) = \tanh(z/a) \tag{21.1.11}$$

instead of the rule (1.21), and this approximating signum function is employed to define a soft-edge bump function with the use of (1.20). Then the relation (1.14) continues to hold, but with the soft-edge bump function defined in terms of (1.29) and (1.20).

In this case it can be verified that (1.15) through (1.18) remain true. See Exercise 1.2. However (1.26) and (1.27) are replaced by the asymptotic relations

$$\mathrm{sgn}(z, a) = 1 - 2\exp(-2z/a) + O[\exp(-4z/a)] \text{ as } z \to \infty, \tag{21.1.12}$$

$$\mathrm{bump}(z, a, L) = [\exp(2L/a) - 1] \exp(-2|z|/a) + O[\exp(-4|z|/a)] \text{ as } |z| \to \infty. \tag{21.1.13}$$

Consequently in this model (1.28) is replaced by the *exponential* fall off relation

$$C_0^{[1]}(z) = B(0, 0, z) = B[\exp(2L/a) - 1] \exp(-2|z|/a) + O[\exp(-4|z|/a)] \text{ as } |z| \to \infty. \tag{21.1.14}$$

The characteristic fall-off length is again governed by a.

This model does not correspond to any easily-described iron and current distribution. But it does have the property of exponential fall off, which is characteristic of iron-dominated solenoids when the coils are buried deep within the iron. See the next section. It therefore may be of some use in preliminary modeling studies of transfer maps for iron-dominated solenoids.

Exercises

21.1.1. Verify that $B_z(0, 0, z)$ as given by (1.11) describes the on-axis field of a simple air-core solenoid.

21.1.2. The purpose of this exercise is to verify the relations (1.15) through (1.18). Show that the approximating signum function (1.21) has the properties

$$\operatorname{sgn}(-z,a) = -\operatorname{sgn}(z,a), \tag{21.1.15}$$

$$\lim_{z\to\pm\infty} \operatorname{sgn}(z,a) = \pm 1. \tag{21.1.16}$$

Sketch $\operatorname{sgn}(z,a)$, $-\operatorname{sgn}(z-L,a)$, and $\operatorname{bump}(z,a,L)$ as given by (1.20) to verify the relations (1.15) through (1.17).

What remains is to prove the relation (1.18). Begin by writing

$$\int_{-\infty}^{\infty} \operatorname{bump}(z,a,L)dz = \lim_{w\to\infty} \int_{-w}^{w} \operatorname{bump}(z,a,L)dz. \tag{21.1.17}$$

Next verify from the representation (1.20) that

$$\int_{-w}^{w} \operatorname{bump}(z,a,L)dz = (1/2)\int_{-w}^{w} \operatorname{sgn}(z,a)dz - (1/2)\int_{-w}^{w} \operatorname{sgn}(z-L,a)dz. \tag{21.1.18}$$

Show that the first integral on the right side of (1.36) vanishes because of (1.33). Show that by making the change of variables $x = z - L$ the second integral on the right side of (1.36) becomes

$$\begin{aligned} -(1/2)\int_{-w}^{w} \operatorname{sgn}(z-L,a)dz &= -(1/2)\int_{-w-L}^{w-L} \operatorname{sgn}(x,a)dx \\ = -(1/2)\int_{-w-L}^{w+L} \operatorname{sgn}(x,a)dx &+ (1/2)\int_{w-L}^{w+L} \operatorname{sgn}(x,a)dx. \end{aligned} \tag{21.1.19}$$

Verify that the first integral in the second line of (1.37) vanishes, again because of (1.33). It follows that there is the result

$$\int_{-w}^{w} \operatorname{bump}(z,a,L)dz = (1/2)\int_{w-L}^{w+L} \operatorname{sgn}(x,a)dx. \tag{21.1.20}$$

Show from (1.34) that there is the result

$$\lim_{w\to\infty}(1/2)\int_{w-L}^{w+L} \operatorname{sgn}(x,a)dx = (1/2)2L = L. \tag{21.1.21}$$

Put all your intermediate results together to obtain the final result

$$\int_{-\infty}^{\infty} \operatorname{bump}(z,a,L)dz = L, \tag{21.1.22}$$

as desired. Note that the proof of this result has depended only on the representation (1.20) and properties (1.33) and (1.34), which are required properties of any approximating signum function.

21.1.3. Verify the limiting behavior (1.22). Verify the fall-off relations (1.26) through (1.28).

21.1.4. Verify for a long simple air-core solenoid that the on-axis field at either end ($z = 0$ or $z = L$) is $B/2$. Verify that the same is true for a long solenoid described by the tanh model (1.29), and for any bump function model that is constructed from approximating signum functions.

21.1.5. Verify the fall-off relations (1.30) through (1.32).

21.1.3 Improved Model for Iron-Dominated Solenoid

Chapter 16 described the computation of transfer maps for straight magnetic beam-line elements based on the *normal* component of the field on the surface of a cylinder. In this subsection we will use instead the *tangential* B_z component of the field on the surface of a cylinder. For simplicity, we will use the surface of a circular cylinder. This approach of employing the B_z component is particularly useful in the case of a solenoid.

From (15.2.13) we find the result

$$\begin{aligned} B_z(x,y,z) &= B_z(\rho,\phi,z) = \partial\psi(x,y,z)/\partial z \\ &= \sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} dk G_m(k)(ik)\exp(ikz)\exp(im\phi)I_m(k\rho). \end{aligned} \tag{21.1.23}$$

Now integrate both sides of (2.1) over $\phi \in [0, 2\pi]$ to find the relation

$$\tilde{B}_z(\rho,0,z) = \int_{-\infty}^{\infty} dk G_0(k)(ik)\exp(ikz)I_0(k\rho) \tag{21.1.24}$$

where

$$\tilde{B}_z(\rho,0,z) = [1/(2\pi)] \int_0^{2\pi} d\phi B_z(\rho,\phi,z). \tag{21.1.25}$$

Next, from the uniqueness of the Fourier transform, it follows that

$$G_0(k)(ik)I_0(kR) = \tilde{\tilde{B}}_z(R,0,k) \tag{21.1.26}$$

where

$$\tilde{\tilde{B}}_z(R,0,k) = [1/(2\pi)] \int_{-\infty}^{\infty} dz \exp(-ikz)\tilde{B}_z(R,0,z). \tag{21.1.27}$$

Inserting (2.4) into (2.2) gives the result

$$\tilde{B}_z(\rho,0,z) = \int_{-\infty}^{\infty} dk \exp(ikz)\tilde{\tilde{B}}_z(R,0,k)I_0(k\rho)/I_0(kR). \tag{21.1.28}$$

Finally, use of (15.2.71) or (15.4.36) gives the relation

$$C_0^{[1]}(z) = \int_{-\infty}^{\infty} dk \exp(ikz)\tilde{\tilde{B}}_z(R,0,k)/I_0(kR) \tag{21.1.29}$$

from which it follows that

$$C_0^{[n]}(z) = \int_{-\infty}^{\infty} dk \exp(ikz)\tilde{\tilde{B}}_z(R,0,k)[(ik)^{n-1}/I_0(kR)]. \tag{21.1.30}$$

We have found expressions for the $m = 0$ on-axis gradients in terms of the B_z component of the magnetic field on the surface of a circular cylinder. If desired, we could also find expressions for the $m \neq 0$ on-axis gradients in terms of B_z on the surface.[2]

As a simple application, the representation (2.7) can be used to find an *approximation* to the on-axis gradients (and hence the magnetic field) in the case that the field is produced by an iron-dominated magnetic solenoid of bore radius R with a small pole gap of length L centered at $z = 0$. Solenoids for use in electron microscopes are often of this type. See Figure 1.1. In this case we may make the approximation

$$\begin{aligned} &B_z(R,\phi,z) = B_{\text{gap}} \text{ for } z \in (-L/2, L/2), \\ &B_z(R,\phi,z) = 0 \text{ elsewhere.} \end{aligned} \tag{21.1.31}$$

That is, the tangential surface field exists only in the gap, and is constant there both in ϕ and in z.[3] With the assumption (2.9), the relations (2.3) and (2.5) become

$$\tilde{\tilde{B}}_z(R,0,k) = [B_{\text{gap}}/(2\pi)]\int_{-L/2}^{L/2} dz \exp(-ikz) = [B_{\text{gap}}/(\pi k)]\sin(kL/2). \tag{21.1.32}$$

Correspondingly, we find for the on-axis gradient the result

$$C_0^{[1]}(z) = [B_{\text{gap}}/(\pi)]\int_{-\infty}^{\infty} dk \exp(ikz)\sin(kL/2)/[kI_0(kR)]. \tag{21.1.33}$$

Examination of the integral representation (2.11) for $C_0^{[1]}(z)$ reveals that it depends on the dimensionless quantities z/R and L/R. Indeed, upon making the change of integration variable given by

$$\lambda = kR, \tag{21.1.34}$$

(2.11) takes the form

$$C_0^{[1]}(z) = B_{\text{gap}}(L/R)F(z/R, L/R) \tag{21.1.35}$$

where F is a *profile function* given by

$$\begin{aligned} F(z/R, L/R) &= (1/\pi)(R/L)\int_{-\infty}^{\infty} d\lambda \exp(i\lambda z/R)\sin[(\lambda/2)(L/R)]/[\lambda I_0(\lambda)] \\ &= [1/(2\pi)]\int_{-\infty}^{\infty} d\lambda \exp(i\lambda z/R)\frac{\sin[(\lambda/2)(L/R)]}{[(\lambda/2)(L/R)]}[1/I_0(\lambda)] \\ &= [1/(2\pi)]\int_{-\infty}^{\infty} d\lambda \exp(i\lambda z/R)\{\text{sinc}[(\lambda/2)(L/R)]\}[1/I_0(\lambda)]. \end{aligned} \tag{21.1.36}$$

[2] There might also be occasions in which one might want to use the azimuthal component B_ϕ on the surface.

[3] That the tangential magnetic field should be zero outside the gap follows from the idealized boundary condition for the interface between vacuum and a medium with infinite magnetic permeability. That the field should be constant in the gap is a further idealization.

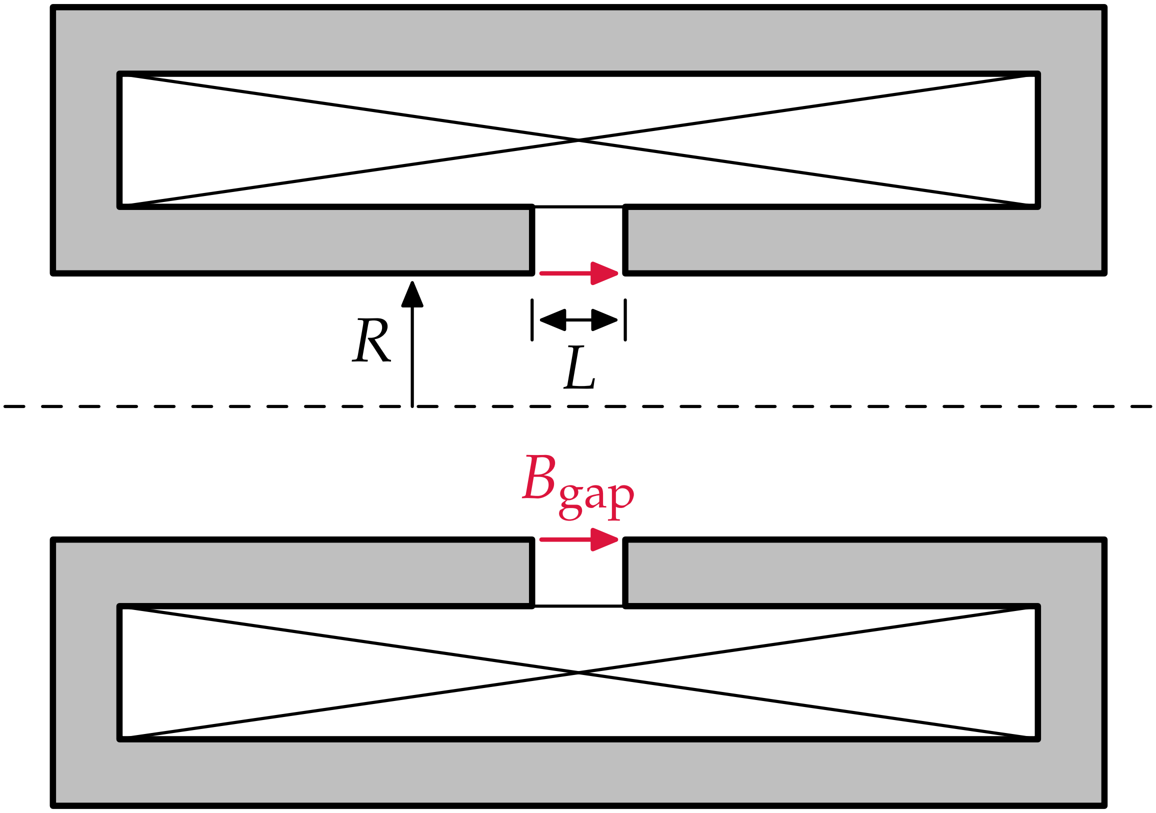


Figure 21.1.1: Schematic of an iron-dominated solenoid with an inter-pole gap L substantially smaller than the bore radius R.

Suppose that L is small compared to R so that L/R is small. Then, when the argument of the sine or sinc function in (2.14) differs significantly from zero, $[I_0(\lambda)]^{-1}$ will be essentially zero. Therefore, we can expand the sine or sinc function in a Taylor series and integrate term by term. We also note that the surface field model (2.9) is only reasonable in the limit of small L/R. Upon making the Taylor expansion just described, one finds that $F(z/R, L/R)$ differs from $F(z/R, 0)$ only by terms of order $(L/R)^2$ and

$$F(z/R, 0) = [1/(2\pi)] \int_{-\infty}^{\infty} d\lambda \exp(i\lambda z/R)/I_0(\lambda). \tag{21.1.37}$$

Therefore, for small L/R, it is useful to make the approximation

$$C_0^{[1]}(z) \simeq B_{\rm gap}(L/R)F(z/R, 0). \tag{21.1.38}$$

For example, Figure 1.2 displays $F(z/R, L/R)$ as a function of z/R for the two values $L/R = 0$ and $L/R = 1/2$. Evidently the two profiles nearly agree when $z = 0$, and are essentially identical away from $z = 0$.

Let us make the further and more drastic approximation of replacing $I_0(\lambda)$, the denominator in (2.15), by $\cosh(\lambda)$. In this approximation the integral (2.11) can be evaluated analytically to give the result

$$F(z/R, 0) \approx G(z/R) = [1/(2\pi)] \int_{-\infty}^{\infty} d\lambda \exp(i\lambda z/R)/\cosh(\lambda) = 1/\{2\cosh[\pi z/(2R)]\}. \tag{21.1.39}$$

Figure 1.2 also displays the approximate profile function $G(z/R)$.

There are three things that we can learn/observe from this approximation. First, although it is not a particularly good approximation, G becomes singular when $z = \pm iR$, which agrees with the discussion of analytic properties given in Subsection 19.1.2. [Note also that (1.12) is singular when $z = \pm ia$.] Second, G falls off *exponentially* as $\exp[-\pi|z|/(2R)]$ for large $|z|$. Third, as Figure 1.2 illustrates, $F(z/R, L/R)$ and correspondingly $C_0^{[1]}(z)$ essentially fall off for large $|z|$ in the same way as G. Thus, for example, at a distance of one bore diameter away from the center of the solenoid, when $|z| = 2R$, the on-axis field will have fallen from its maximum value by approximately a factor of $\exp(-\pi) \simeq .04$.

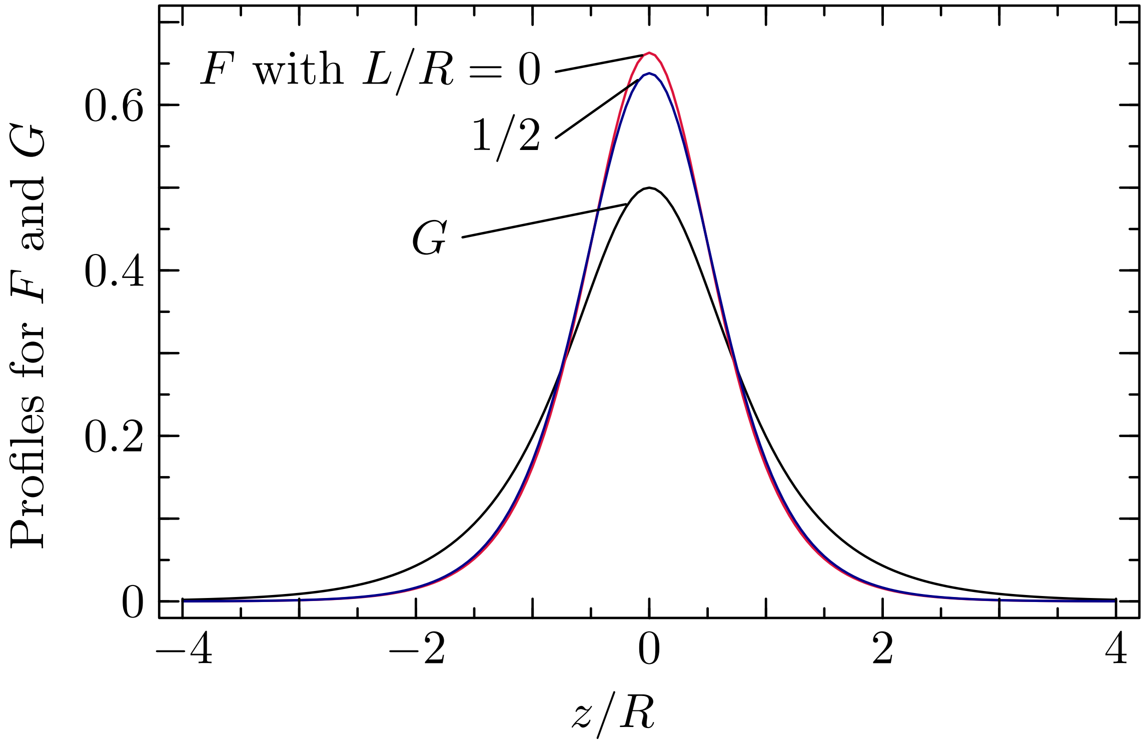


Figure 21.1.2: The profile function $F(z/R, L/R)$ as given by (2.14) in the cases $L/R = 0$ and $L/R = 1/2$, and the approximate profile function $G(z/R)$. The two curves that nearly agree are those for F, with the highest curve being that for F when $L/R = 0$. The third curve is that for G.

21.1.4 Quantitatively Correct Iron-dominated Solenoid

21.2 Realistic Wigglers/Undulators

21.2.1 An Iron-Dominated Superconducting Wiggler/Undulator

21.3 Quadrupoles

21.3.1 Validation of Circular Cylinder Surface Method

This subsection describes numerical tests performed for the case of a Lambertson type quadrupole. In this case the on-axis gradients and their derivatives can be determined analytically. The surface data $B_\rho(\rho = R, \phi, z)$ can also be found directly using the Biot-Savart law. We will first show that the gradients computed from the surface data following the method of Section 4 agree with the gradients determined analytically. Next we will add noise to the surface data, and again apply the method of Section 4 to this noisy data. We will find that this noise has no undue effect on the computed gradients. Finally, we will show that the noise also has no undue effect on the transfer map $\mathcal{M}$.

The method described in Section 4 has been implemented in the code MARYLIE 5.0 [**?**] as a user-defined routine. The routine reads from an external file the functions $a_m(R,z)$ and $b_m(R,z)$, see (4.1), evaluated on a discrete set of points z_i. It then generates the corresponding transfer map by using the built-in routine GENMAP to integrate the map equations (1.2). Since MARYLIE 5.0 is a 5^{th} order code, only the multipoles through $m=6$ need be considered.

The Fourier transforms (4.4) and (4.5) are calculated from the read-in values of $a_m(R,z)$ and $b_m(R,z)$ using Filon's method [3] for various values of k in the interval $[-k_{max}, k_{max}]$ where k_{max} is a suitable k cut-off for the integrals (4.2) and (4.3). For the cases described below, we have used the value $Rk_{max} = 20$. Filon's method requires interpolation of the functions $a_m(R,z)$ and $b_m(R,z)$; and for this purpose we use local parabolic fits.

The integration algorithm of GENMAP is based on a 11^{th} order multistep (Adams) method. Because the algorithm uses a fixed step size, one needs to provide values of the generalized gradients and their derivatives only at the predetermined locations in z required by GENMAP. The integrals (4.2) and (4.3) that provide the generalized on-axis gradients and their derivations are evaluated at the values of z needed by GENMAP, again using Filon's method. We emphasize that no interpolation of the generalized gradients is required by GENMAP.

We have tested the routine by treating the case of a Lambertson quadrupole [8, 9]. For this case only $b_2(R,z)$ and $b_6(R,z)$ are nonzero. Correspondingly only the functions $C^{[0]}_{2,s}$, $C^{[1]}_{2,s}$, $C^{[2]}_{2,s}$, $C^{[3]}_{2,s}$, $C^{[4]}_{2,s}$, and $C^{[0]}_{6,s}$ are required. The use of this case as an example has the virtue that the various C functions can be found exactly [10].

Also, the surface data $B_\rho(\rho = R, \phi, z)$ can be found directly using the Biot-Savart law, and this data can be integrated over ϕ to yield $b_2(R,z)$ and $b_6(R,z)$. In our test we evaluted $B_\rho(\rho = R, \phi, z)$ for 279 equally spaced z values within the interval

$$z \in [z_{min}, z_{max}] = [-7r, 7r] \tag{21.3.1}$$

according to the rule

$$z_i = z_{min} + \Delta(i-1) \text{ for } i = 1, 2, \cdots 279. \tag{21.3.2}$$

The cylinder on which we evaluated B_ρ had the radius

$$R = .75r. \tag{21.3.3}$$

Here $r = .128$ m is the radius of the quadrupole itself, and the length of the quadrupole is $2r$. Corresponding, $\Delta = 14r/278 = 6.44$ mm. The relatively large values of z_{min} and z_{max} were necessary because the large radius-to-length ratio of the quadrupole makes the fringe fields very extended.

For each z value the quantity $B_\rho(\rho = R, \phi, z)$ was evaluated for 256 equally spaced angles over the interval $[0, 2\pi]$, and these B_ρ values were used to compute the integrals

$$a_m(R,z) = \frac{1}{\pi}\int_0^{2\pi} d\phi \cos(m\phi) B_\rho(\rho = R, \phi, z), \tag{21.3.4}$$

$$b_m(R,z) = \frac{1}{\pi}\int_0^{2\pi} d\phi \sin(m\phi) B_\rho(\rho = R, \phi, z). \tag{21.3.5}$$

Because of the symmetries of a (normal) quadrupole only the functions b_2 and b_6 are non vanishing for $m \leq 6$. The net results of the steps just described are the values of these functions at the points (8.2). These functions are shown in Figures 8.1 and 8.2.

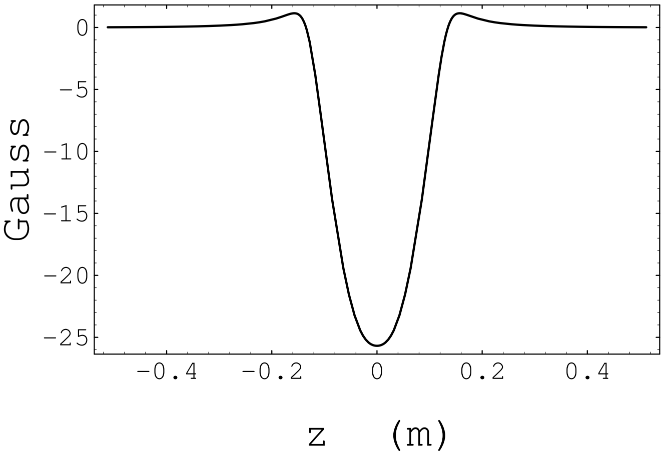


Figure 21.3.1: The angle integrated surface data $b_2(R,z)$. The magnet occupies the interval $z \in [-0.128\,\text{m}, 0.128\,\text{m}]$.

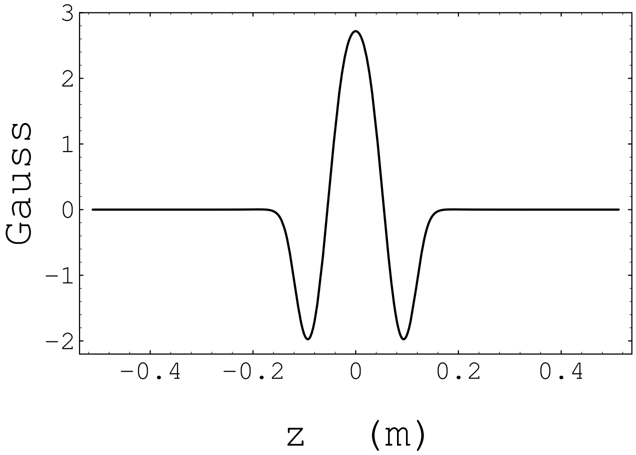


Figure 21.3.2: The angle integrated surface data $b_6(R,z)$.

Figure 8.3 shows $C_{2,s}^{[0]}$ determined both analytically and computed numerically from $b_2(R,z)$ using the method of Section 4 as described above. Evidently the agreement is excellent. Figure 8.4 shows analytic values of $C_{6,s}^{[0]}$ and values computed numerically from $b_6(R,z)$. Again the agreement is excellent. The most stringent test is a comparison of analytic values of $C_{2,s}^{[4]}$ with values computed numerically from $b_2(R,z)$. This comparison is given in Figure 8.5. Again the agreement is excellent.

As a final test of our routines, we compared the transfer maps for our Lambertson quadrupole obtained using either the analytically known on-axis gradients or on-axis gradients computed numerically from surface data. Table 8.1 shows that the (relative) difference

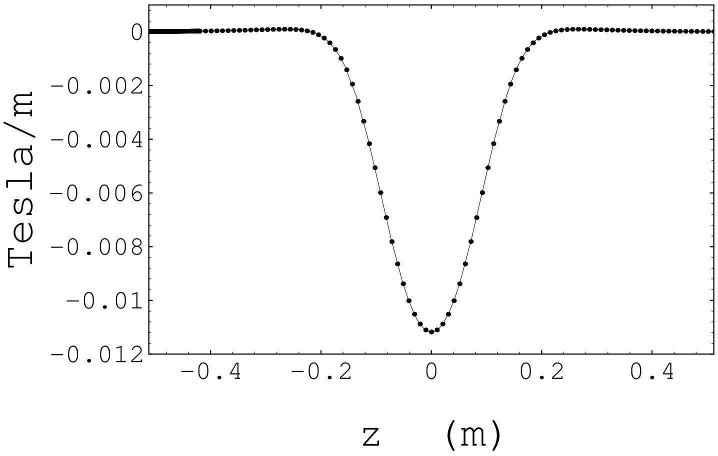


Figure 21.3.3: The function $C_{2,s}^{[0]}(z)$ as calculated numerically from surface data (dots) and analytically (solid line).

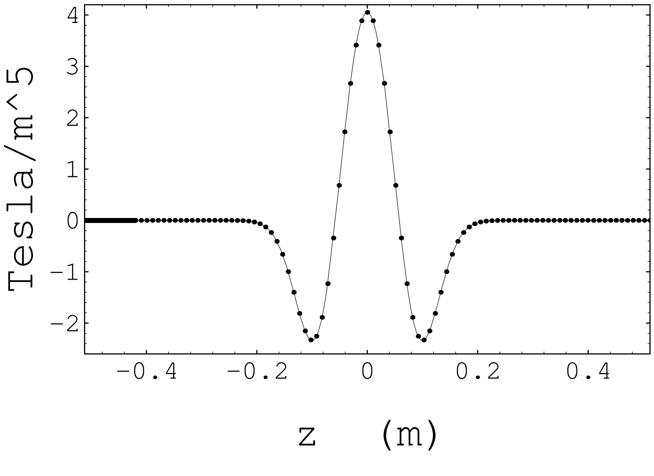


Figure 21.3.4: The function $C_{6,s}^{[0]}(z)$ as calculated numerically from surface data (dots) and analytically (solid line).

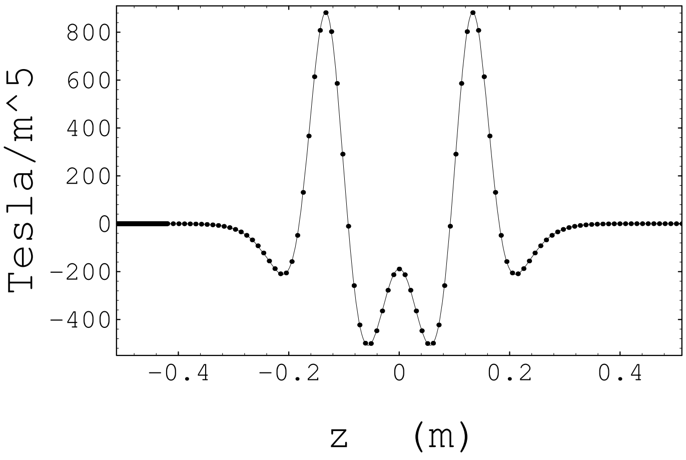


Figure 21.3.5: The function $C_{2,s}^{[4]}(z)$ as calculated numerically from surface data (dots) and analytically (solid line).

in the surface-data-based map, as compared to the exact map, is very small. Of course, apart from roundoff problems, we expect this difference will vanish as the number of sampling points in z and ϕ is made arbitrarily large.

Table 21.3.1: Relative difference between the surface-data-based map and the exact map.

map generators	relative difference
$\mathcal{R}_2$	$< 10^{-6}$
$f_3,\, f_4$	$< 10^{-5}$
$f_5,\, f_6$	$< 10^{-4}$

Now that the method of Section 4 has been verified to work, we will study the sensitivity of transfer map calculations to the presence of random errors (noise) in the surface data. As a simple model, consider the perturbed functions

$$b_2^{rnd}(R, z_i) = b_2(R, z_i)[1 + \epsilon_2(z_i)], \tag{21.3.6}$$

$$b_6^{rnd}(R, z_i) = b_6(R, z_i)[1 + \epsilon_6(z_i)], \tag{21.3.7}$$

where the $\epsilon_2(z_i)$, $\epsilon_6(z_i)$ are random variables uniformly distributed in the interval $[-\epsilon/2, \epsilon/2]$, and $b_2(R, z_i)$, $b_6(R, z_i)$ are the same as before. What effect do these errors have on the on-axis gradients computed from the (noisy) surface data?

Figure 8.6 shows $C_{2,s}^{[4]}$ for a particular realization of the error distribution (seed #2) and $\epsilon = 10^{-2}$. The solid line shows analytic results (the same as those of Figure 8.5) and the dots show results computed numerically from the noisy surface data. Close inspection of the figure shows that the points no longer fall exactly on the curve, as is to be expected in the case of noise. However, the size of the deviations from the curve is comparable to the size of the noise, and not unduly larger. To facilitate closer comparison, Figure 8.7 shows the difference between the analytic results and results computed numerically from the noisy surface data. Evidently the deviation is generally on the order of 1% or less, which is comparable with $\epsilon = 10^{-2}$.

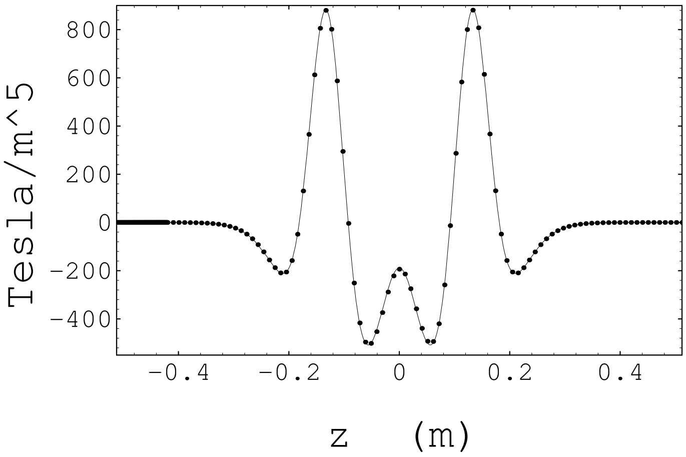


Figure 21.3.6: The function $C_{2,s}^{[4]}(z)$ as calculated numerically from noisy (seed #2) surface data (dots), and analytically (solid line).

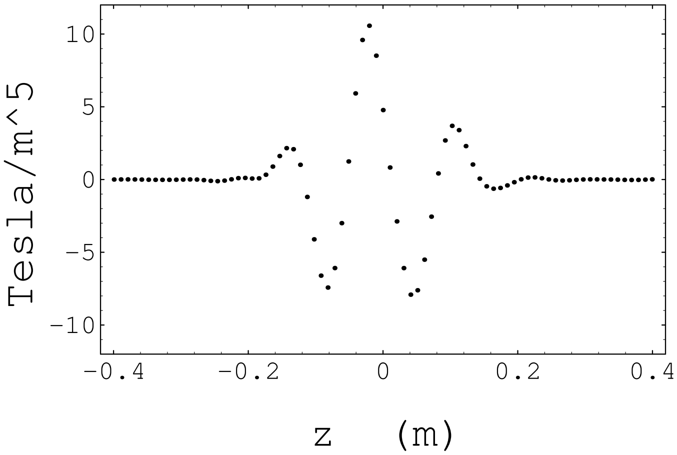


Figure 21.3.7: Difference between the solid line and dots of Figure 8.6.

Since, as we have seen, the computation of the on-axis gradients and their derivatives is not unduly affected by noise in the surface data, we expect the same will be true for the transfer map. This is indeed the case. Table 8.2 shows that the (relative) error in the noisy surface data based map (as compared to the exact map) is, at worst, on the order of the noise.

Table 21.3.2: Relative error of the noisy surface data based map compared to the exact map.

map generators	seed #1	seed#2	seed#3
$\mathcal{R}_2$	$< 3 \times 10^{-4}$	$< 8 \times 10^{-4}$	$< 5 \times 10^{-4}$
f_3, f_4	$< 10^{-3}$	$< 1.6 \times 10^{-3}$	$< 1.6 \times 10^{-3}$
f_5, f_6	$< 10^{-2}$	$< 1.5 \times 10^{-2}$	$< 1.3 \times 10^{-2}$

One might imagine that the introduction of noise would damage the differentiability (with respect to z) of the surface data $a_m(R, z)$ and $b_m(R, z)$. This is indeed true. Consider, for example, the function $b_2(R, z)$ of the previous section. Figure 9.1 shows its Fourier transform $\tilde{b}_2(R, k)$. Evidently the spectrum of b_2 cuts off beyond a wave number $k_{max} \simeq 150$ m^{-1}. For comparison, Figure 9.2 shows the Fourier transform of $\epsilon_2(z)$ for the case of seed #2. Note that the noise spectrum extends to $k'_{max} \simeq 600$ m^{-1}. This is to be expected since $\pi/\Delta \simeq 490$ m^{-1}. Finally, Figure 9.3 shows the Fourier transform of $b_2\epsilon_2$. Its spectrum extends to $k''_{max} \simeq 750$ m^{-1}. We see that $k''_{max} \simeq (k_{max} + k'_{max})$, as is also to be expected.

Although noise does damage the differentiability of surface data, it has considerably less effect on the on-axis gradients derived from the (noisy) surface data. Mathematically, this pleasant result arises from the spectral "cutoff" provided by the kernel $[k^{m+n-1}/I'_m(kR)]$ that occurs in (4.2) and (4.3). For example, Figure 9.4 shows the factor $[k^5/I'_2(kR)]$ that is relevant to the computation of $C^{[4]}_{2,s}$ for the quadrupole example of the last section. We see that it peaks at $k \simeq 50$ m^{-1}, and falls off rapidly beyond $k_{max} \simeq 200$ m^{-1}. Finally, Figure 9.5 shows the product of the two functions presented in Figures 9.3 and 9.4. From Figures 9.4 and 9.5 we see that the high wave-number part of the spectrum in Figure 9.3 is effectively filtered out. Correspondingly, as already seen in the previous section, noise in the surface data has no undue effect on the function $C^{[4]}_{2,s}$. We remark, as is obvious from our considerations, that noise has even less effect on the functions $C^{[n]}_{2,s}$ with $n < 4$. It also has no undue effect on $C^{[0]}_{6,s}$. We also note that the condition $Rk_{max} = 20$ used in Section 8 corresponds to $k_{max} \simeq 208$.

In summary, we have found that noise introduces high wave-number contributions to the spectrum of b^{rnd}_m where they are absent in the spectrum of b_m. These potentially damaging contributions are filtered by the kernel $[k^{m+n-1}/I'_m(kR)]$. Evidently this filtering becomes less effective with increasing n, and is improved by making R as large as possible. For the example studied, we found that smoothing was satisfactory for the selected value of R and the n values required for 5th-order calculations.

Finally we remark that, by making more detailed numerical studies, it should be possible to find the partial derivatives

$$\partial\mathcal{M}/\partial b_m(R, z_i) = \partial[\text{any selected coefficient of any generator for } \mathcal{M}]/\partial b_m(R, z_i). \quad (21.3.8)$$

That is, we can evaluate numerically how the map changes when the value of the surface data at any point is varied. In this way, we can get precise and complete information about the effect of possible noise.

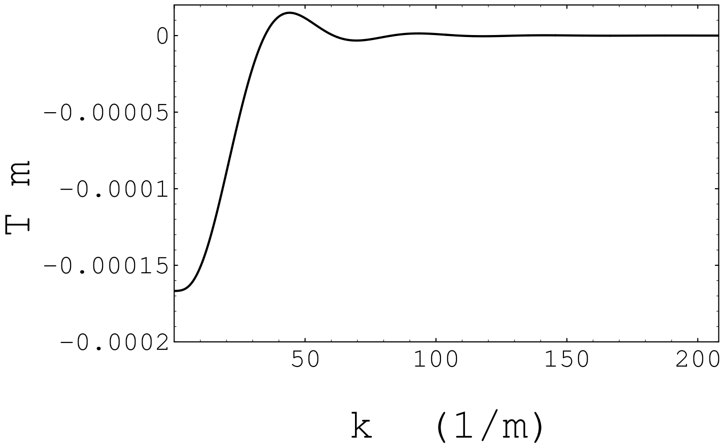


Figure 21.3.8: Real part of the function $\tilde{b}_2(R,k)$. The imaginary part vanishes.

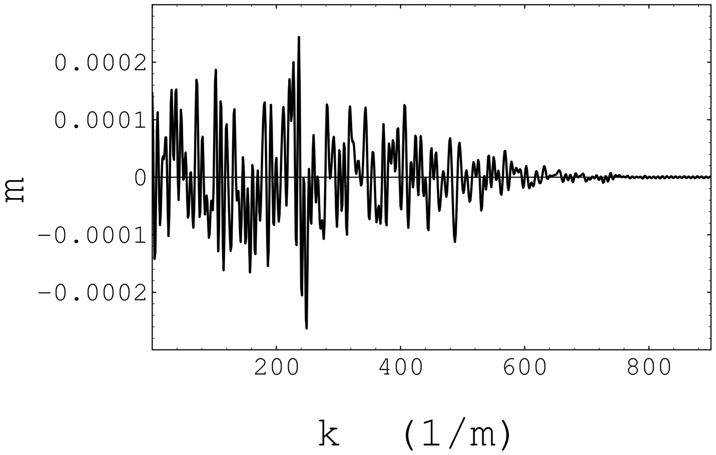


Figure 21.3.9: Real part of the function $\tilde{\epsilon}_2(k)$. The imaginary part has similar features.

21.3.2 Final Focus Quadrupoles

21.4 Closely Adjacent Quadrupoles and Sextupoles

21.5 Application to Radio-Frequency Cavities

Ackowledgement

We have benefitted greatly from many conversations with Peter Walstrom and the reading of some of his internal Technical Notes. We are also grateful to Thomas Mottershead, Filippo

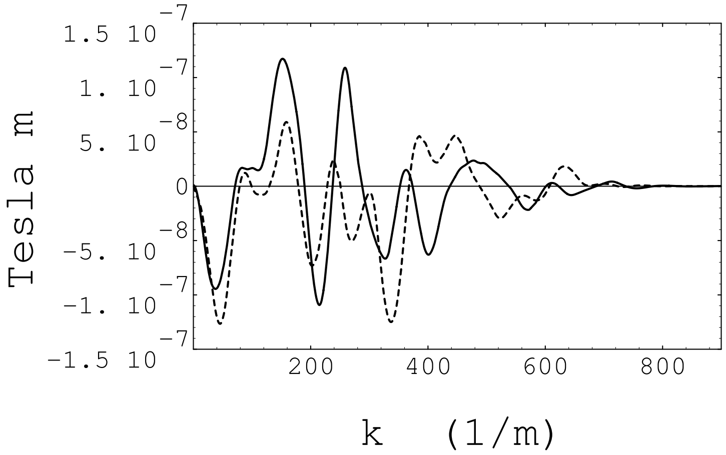


Figure 21.3.10: Real (solid line) and imaginary part (dashed line) of the Fourier transform of the function $b_2(R, z)\epsilon_2(z)$.

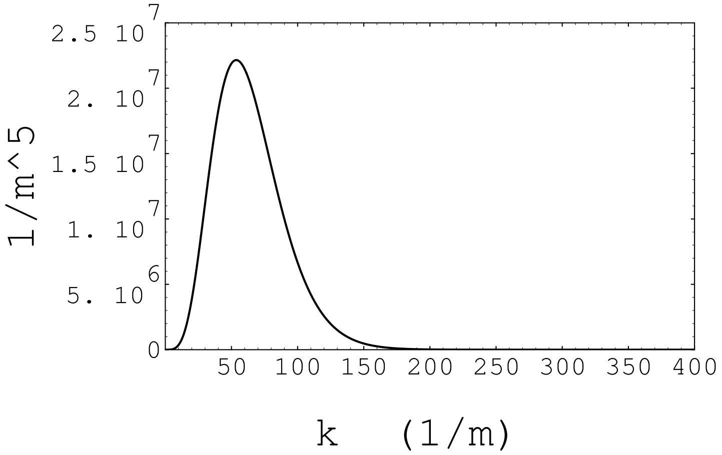


Figure 21.3.11: The factor $[k^5/I_2'(kR)]$ that appears in the calculation of $C_{2,s}^{[4]}$.

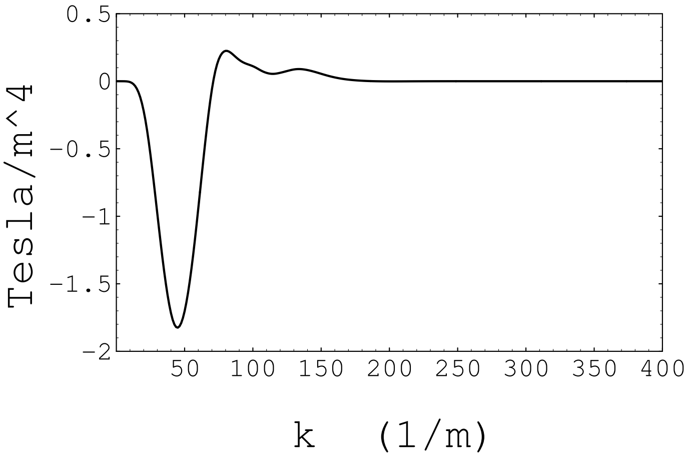


Figure 21.3.12: A plot of the real part of the product of the two functions of Figures 9.3 and 9.4. The imaginary part has similar features.

Neri, and Peter Walstrom for their many contributions to MARYLIE 5.0.

Bibliography

General References

[1] M. Venturini, "Lie Methods, Exact Map computation, and the Problem of Dispersion in space Charge Dominated Beams", University of Maryland College Park Physics Department Ph. D. thesis (1998).

[2] M. Venturini and A. Dragt, "Accurate Computation of Transfer Maps from Magnetic Field Data", *Nuclear Instruments and Methods* **A427**, p. 387 (1999).

[3] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover (1972). Also available on the Web by Googling "abramowitz and stegun 1972".

Solenoids

[4] A. El-Kareh and J. El-Kareh, *Electron Beams, Lenses, and Optics*, Vols. 1 and 2, Academic Press (1970).

[5] A. Dragt, "Numerical third-order transfer map for solenoid", *Nuclear Instruments and Methods in Physics Research* **A298**, p. 441 (1990).

[6] P. W. Hawkes and E. Kasper, *Principles of Electron Optics*, Vols. 1 through 3, Academic Press (1996).

Dipoles

[7] P. L. Walstrom, "Dipole-magnet field models based on a conformal map", *Physical Review Special Topics-Accelerators and Beams* **15**, 102401 (2012).

Wiggglers

Quadrupoles

[8] R.P. Avery, B.R. Lambertson, C.D. Pike, PAC 1971 Proceedings, p. 885.

[9] T.G. Godlove, S. Bernal, M. Reiser, Printed Circuit Quadrupole Design, PAC 1995 Proceedings, p. 2117

[10] M. Venturini, Transfer Map for Printed-Circuit Magnetic Quadrupoles, Technical Note, Dept. of Physics, Univ. of Maryland (1995).

[11] K. Halbach, “Physical and Optical Properties of Rare Earth Cobalt Magnets”, *Nuclear Instruments and Methods* **187**, pp. 109-117 (1981).

Chapter 22

Realistic Transfer Maps for General Curved Beam-Line Elements: Theory

22.1 Introduction

Surface methods based on the use of cylinders are appropriate for straight beam-line elements or for bent elements with small sagitta. However, cylinders cannot be employed for elements with large sagitta, such as dipoles, where no straight cylinder would fit within the aperture. For such cases more complicated surfaces are required. For example, Figure 1.1 shows a bent box with straight end legs. Its surface could be used to treat a dipole with large sagitta. In this case, the bent part of the box would lie within the body of the dipole, and the straight legs would enclose the fringe-field regions.

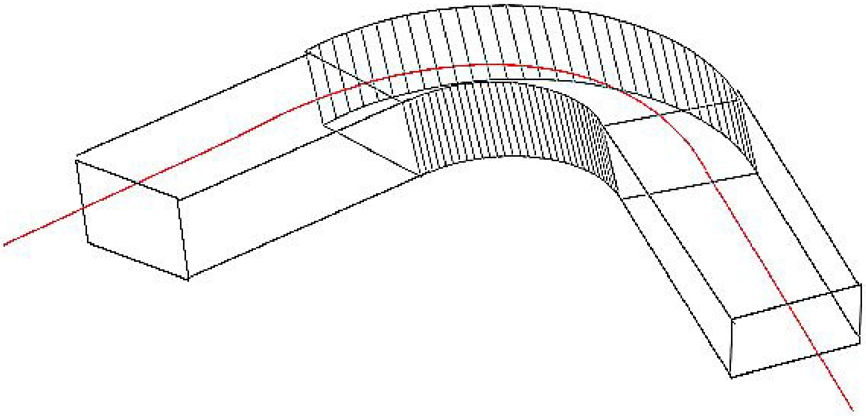


Figure 22.1.1: A bent box with straight end legs.

But now there is a complication: The *straight* cylinder methods succeeded because Laplace's equation is separable in circular, elliptical, and rectangular cylinder coordinates. Consequently, we were able to find a kernel that related the interior vector potential to the normal component of the surface magnetic field. However, there is no *bent* coordinate system with straight ends for which Laplace's equation is separable. The method of cylindrical multiples and on-axis gradients is only applicable to straight elements.

This problem can in principle be overcome if *both* the normal component of the magnetic field and the scalar potential for the magnetic field are known on the surface. (Note that a knowledge of the scalar potential on the surface is equivalent to a knowledge of the tangential component of the field on the surface.) Such data are in fact provided on a mesh by some 3-dimensional field solvers, and these data can be interpolated onto the surface.

Let V be some volume in three-dimensional space bounded by a surface S. Suppose that the magnetic field $\boldsymbol{B}(\boldsymbol{r})$ is source free when $\boldsymbol{r}$ is within V. That is, for $\boldsymbol{r}$ within V, $\boldsymbol{B}(\boldsymbol{r})$ satisfies the requirements

$$\nabla \cdot \boldsymbol{B}(\boldsymbol{r}) = 0, \tag{22.1.1}$$

$$\nabla \times \boldsymbol{B}(\boldsymbol{r}) = 0. \tag{22.1.2}$$

This will be the case for the magnetic field in an evacuated beam pipe. For a Hamiltonian treatment of trajectories, we need a vector potential $\boldsymbol{A}(\boldsymbol{r})$ such that

$$\boldsymbol{B}(\boldsymbol{r}) = \nabla \times \boldsymbol{A}(\boldsymbol{r}). \tag{22.1.3}$$

Let $\boldsymbol{n}'(\boldsymbol{r}')$ be the outward normal to S at the point $\boldsymbol{r}' \in S$. Then the normal component of $\boldsymbol{B}$ on S is given by the definition

$$B_n(\boldsymbol{r}') = \boldsymbol{n}'(\boldsymbol{r}') \cdot \boldsymbol{B}(\boldsymbol{r}'). \tag{22.1.4}$$

Also, let $\psi(\boldsymbol{r}')$ be the value of the magnetic scalar potential at the point $\boldsymbol{r}' \in S$. It satisfies the relation

$$\boldsymbol{B}(\boldsymbol{r}') = \nabla' \psi(\boldsymbol{r}'). \tag{22.1.5}$$

Then, with the aid of the vector potential for Dirac magnetic monopoles and Helmholtz's theorem, it can be shown that there are *kernels* $\boldsymbol{G}^n$ and $\boldsymbol{G}^t$ such that a suitable interior vector potential $\boldsymbol{A}(\boldsymbol{r})$ for $\boldsymbol{r}$ within V is given by the relation

$$\boldsymbol{A}(\boldsymbol{r}) = \boldsymbol{A}^n(\boldsymbol{r}) + \boldsymbol{A}^t(\boldsymbol{r}) \tag{22.1.6}$$

with

$$\boldsymbol{A}^n(\boldsymbol{r}) = \int_S dS' \; B_n(\boldsymbol{r}')\boldsymbol{G}^n(\boldsymbol{r}, \boldsymbol{r}') \tag{22.1.7}$$

and

$$\boldsymbol{A}^t(\boldsymbol{r}) = \int_S dS' \; \psi(\boldsymbol{r}')\boldsymbol{G}^t(\boldsymbol{r}, \boldsymbol{r}'). \tag{22.1.8}$$

(Here the superscripts n and t denote *normal* and *tangential*, respectively, and indicate the contributions to the vector potential made by the normal and tangential components of $\boldsymbol{B}$ on S.) Moreover, the constituents of $\boldsymbol{A}(\boldsymbol{r})$, and hence $\boldsymbol{A}(\boldsymbol{r})$ itself, satisfy the Coulomb gauge condition,

$$\nabla \cdot \boldsymbol{A}^n(\boldsymbol{r}) = \nabla \cdot \boldsymbol{A}^t(\boldsymbol{r}) = \nabla \cdot \boldsymbol{A}(\boldsymbol{r}) = 0. \tag{22.1.9}$$

A detailed exposition of this method, including expected accuracy and insensitivity to noise in the surface data, is the subject of this and the next few chapters. Thus, taken together, Chapters 15 through 21 and this chapter and Chapters 23 through 25 are intended to provide an extensive description of, and associated robust numerical algorithms for, the computation of transfer maps, including all fringe-field and higher-order multipole effects, for realistic beam-line elements having arbitrary geometry.[1]

Section 2 of this chapter describes the the mathematical tools required to treat general geometries. These tools are Dirac's magnetic monopole vector potential and Helmloltz's theorem. Sections 3 and 4 derive the relations (1.3) through (1.9), find the kernels $\boldsymbol{G}^n$ and $\boldsymbol{G}^t$, and describe their properties.

Before continuing on, we pause to advertise some of the virtues of what can be achieved with the use of general surface methods.

- The constituents $\boldsymbol{A}^n(\boldsymbol{r})$ and $\boldsymbol{A}^t(\boldsymbol{r})$ of $\boldsymbol{A}(\boldsymbol{r})$, and hence $\boldsymbol{A}(\boldsymbol{r})$ itself, are analytic functions of $\boldsymbol{r}$ for $\boldsymbol{r}$ within V, even when there are errors in the surface fields B_n and ψ, and no matter how poorly the integrals (1.7) and (1.8) are evaluated.

- The Maxwell equations for $\boldsymbol{B}(\boldsymbol{r})$, and the Coulomb gauge condition for $\boldsymbol{A}(\boldsymbol{r})$ and its constituents, are satisfied exactly even when there are errors in the surface fields B_n and ψ, and no matter how poorly the integrals (1.7) and (1.8) are evaluated.

- The kernels $\boldsymbol{G}^n$ and $\boldsymbol{G}^t$ are smoothing. Consequently, the $\boldsymbol{A}(\boldsymbol{r})$ given by (1.6) through (1.8) is relatively insensitive to noise in the surface fields B_n and ψ.

We hasten to add that the first two items above should not be taken to mean that there is no need to take care to evaluate integrals well. They just indicate that the worst disasters have been avoided. Subsequently we will learn that the kernels $\boldsymbol{G}^n$ and $\boldsymbol{G}^t$, and their $\boldsymbol{r}$ derivatives, can be strongly peaked in $\boldsymbol{r}'$ when $\boldsymbol{r}$ is near S. To obtain accurate results, this behavior of the kernels must be taken into account when integrating, with respect to $\boldsymbol{r}'$, over the surface S.

22.2 Mathematical Tools

22.2.1 Electric Dirac Strings

In this subsection we will motivate the subject of magnetic Dirac strings by treating the simpler electric case. Suppose $\boldsymbol{E}(\boldsymbol{r})$ is a vector field that obeys the equations

$$\nabla \times \boldsymbol{E} = 0, \tag{22.2.1}$$

$$\nabla \cdot \boldsymbol{E} = \rho. \tag{22.2.2}$$

From (2.1) we know there is a scalar potential ϕ such that

$$\boldsymbol{E} = -\nabla\phi, \tag{22.2.3}$$

[1] In this sentence we have used the term *multipole* loosely to refer, simply, to nonlinear terms arising from nonlinear magnetic field variations. As already emphasized earlier, the concept of cylindrical multipoles only applies to *straight* elements.

and from (2.2) it follows that

$$\nabla^2\phi = -\rho. \tag{22.2.4}$$

Introduce the notation

$$|\boldsymbol{r}-\boldsymbol{r}'| = ||\boldsymbol{r}-\boldsymbol{r}'|| = [(x-x')^2+(y-y')^2+(z-z')^2]^{1/2}. \tag{22.2.5}$$

Consider the function $1/|\boldsymbol{r}-\boldsymbol{r}'|$. It satisfies the relation

$$\nabla^2[1/|\boldsymbol{r}-\boldsymbol{r}'|] = -4\pi\delta_3(\boldsymbol{r}-\boldsymbol{r}') \tag{22.2.6}$$

where the indicated derivatives are to be taken with respect to the components of $\boldsymbol{r}$. Assuming that $\rho(\boldsymbol{r})$ falls off sufficiently rapidly at infinity, it follows that a solution to (2.4) is given by the relation

$$\phi(\boldsymbol{r}) = [1/(4\pi)]\int d^3\boldsymbol{r}'\rho(\boldsymbol{r}')/|\boldsymbol{r}-\boldsymbol{r}'|. \tag{22.2.7}$$

Moreover, (2.7) is the unique solution that vanishes at infinity.

For our discussion we will need some knowledge of low-order (spherical) multipole expansions, which we review briefly here. Suppose that the charge distribution ρ is nonzero only in some volume V surrounding the point $\boldsymbol{r}_d$. (Here the subscript d stands for *distribution*, and will later stand for *dipole*.) Then (2.7) becomes

$$\phi(\boldsymbol{r}) = [1/(4\pi)]\int_V d^3\boldsymbol{r}'\rho(\boldsymbol{r}')/|\boldsymbol{r}-\boldsymbol{r}'|. \tag{22.2.8}$$

Suppose also that $\boldsymbol{r}$ lies outside V so that the denominator in (2.8) never vanishes. Make the change of variables

$$\boldsymbol{r}' = \boldsymbol{r}_d + \boldsymbol{\xi} \tag{22.2.9}$$

so that (2.8) becomes

$$\phi(\boldsymbol{r}) = [1/(4\pi)]\int_{V_0} d^3\boldsymbol{\xi}\ \rho(\boldsymbol{r}_d+\boldsymbol{\xi})/|(\boldsymbol{r}-\boldsymbol{r}_d)-\boldsymbol{\xi}| \tag{22.2.10}$$

where V_0 is a volume surrounding the origin. Under the assumption that $\boldsymbol{r} \notin V$, the denominator factor in (2.10) can be expanded as a power series in the components of $\boldsymbol{\xi}$,

$$1/|(\boldsymbol{r}-\boldsymbol{r}_d)-\boldsymbol{\xi}| = [1/|\boldsymbol{r}-\boldsymbol{r}_d|][1+\boldsymbol{\xi}\cdot(\boldsymbol{r}-\boldsymbol{r}_d)/|\boldsymbol{r}-\boldsymbol{r}_d|^2 + O(\xi^2)]. \tag{22.2.11}$$

Put this expansion into the integral (2.10) to yield the result

$$\begin{aligned}\phi(\boldsymbol{r}) \;=\; & [1/(4\pi)][1/|\boldsymbol{r}-\boldsymbol{r}_d|]\int_{V_0} d^3\boldsymbol{\xi}\ \rho(\boldsymbol{r}_d+\boldsymbol{\xi}) \\ & +[1/(4\pi)][1/|\boldsymbol{r}-\boldsymbol{r}_d|^3](\boldsymbol{r}-\boldsymbol{r}_d)\cdot\int_{V_0} d^3\boldsymbol{\xi}\ \boldsymbol{\xi}\ \rho(\boldsymbol{r}_d+\boldsymbol{\xi}) + O(\xi^2).\end{aligned} \tag{22.2.12}$$

The integrals in (2.12) can be manipulated to bring them to the forms

$$\int_{V_0} d^3\boldsymbol{\xi}\ \rho(\boldsymbol{r}_d+\boldsymbol{\xi}) = \int_V d^3\boldsymbol{r}'\ \rho(\boldsymbol{r}') = Q, \tag{22.2.13}$$

$$\int_{V_0} d^3\boldsymbol{\xi}\ \boldsymbol{\xi}\ \rho(\boldsymbol{r}_d+\boldsymbol{\xi}) = \int_V d^3\boldsymbol{r}'\ (\boldsymbol{r}'-\boldsymbol{r}_d)\ \rho(\boldsymbol{r}') = \boldsymbol{p}_d. \tag{22.2.14}$$

Here Q, the total charge in V, is the monopole moment. And $\boldsymbol{p}_d$ is the dipole moment (with respect to the point $\boldsymbol{r}_d$) of the charge distribution in V. Thus, we find that

$$\phi(\boldsymbol{r}) = [Q/(4\pi)][1/|\boldsymbol{r}-\boldsymbol{r}_d|] + [1/(4\pi)][\boldsymbol{p}_d\cdot(\boldsymbol{r}-\boldsymbol{r}_d)]/|\boldsymbol{r}-\boldsymbol{r}_d|^3 + O(\xi^2). \tag{22.2.15}$$

That is, the potential arising from a charge distribution, at a point $\boldsymbol{r}$ outside the distribution, is a sum of monopole, dipole, and higher-order multipole contributions.

We recall that the prototypical example of a dipole consists of two opposite charges $\pm q$ separated by a distance 2ϵ in the limit that $\epsilon \to 0$ and $q \to \infty$ in such a way that the product $2q\epsilon$ remains constant. For example, suppose a charge $+q$ is placed at the location $\boldsymbol{r}_d + \boldsymbol{\epsilon}$ and a charge $-q$ is placed at the location $\boldsymbol{r}_d - \boldsymbol{\epsilon}$. Then we find that the potential due to this two-charge combination is given by the relation

$$\phi(\boldsymbol{r},\boldsymbol{r}_d) = [1/(4\pi)][q/|\boldsymbol{r}-(\boldsymbol{r}_d+\boldsymbol{\epsilon})| - q/|\boldsymbol{r}-(\boldsymbol{r}_d-\boldsymbol{\epsilon})|]. \tag{22.2.16}$$

Expansion of (2.16) in powers of $\boldsymbol{\epsilon}$ gives the result

$$\phi(\boldsymbol{r},\boldsymbol{r}_d) = [1/(4\pi)](2q\boldsymbol{\epsilon})\cdot(\boldsymbol{r}-\boldsymbol{r}_d)/|\boldsymbol{r}-\boldsymbol{r}_d|^3 + O(q\epsilon^2). \tag{22.2.17}$$

Now let $\boldsymbol{\epsilon} \to 0$ and $q \to \infty$ in such a way that

$$2q\boldsymbol{\epsilon} \to \boldsymbol{p}_d. \tag{22.2.18}$$

In this limit (2.17) becomes

$$\phi_d(\boldsymbol{r},\boldsymbol{r}_d) = [1/(4\pi)][\boldsymbol{p}_d\cdot(\boldsymbol{r}-\boldsymbol{r}_d)]/|\boldsymbol{r}-\boldsymbol{r}_d|^3, \tag{22.2.19}$$

in agreement with the second term in (2.15). We note, with the convention $q > 0$, that the dipole moment vector $\boldsymbol{p}_d$ points from the location of $-q$ to the location of $+q$.

We also note, for future use, that the field $\boldsymbol{E}_d(\boldsymbol{r},\boldsymbol{r}_d)$ at the point $\boldsymbol{r}$ arising from a dipole at the point $\boldsymbol{r}_d$ (with $\boldsymbol{r} \neq \boldsymbol{r}_d$) is given by the relation

$$\begin{aligned}\boldsymbol{E}_d(\boldsymbol{r},\boldsymbol{r}_d) &= -\nabla\phi_d(\boldsymbol{r},\boldsymbol{r}_d)\\ &= -[1/(4\pi)][\boldsymbol{p}_d/|\boldsymbol{r}-\boldsymbol{r}_d|^3] + [3/(4\pi)](\boldsymbol{r}-\boldsymbol{r}_d)[\boldsymbol{p}_d\cdot(\boldsymbol{r}-\boldsymbol{r}_d)]/|\boldsymbol{r}-\boldsymbol{r}_d|^5.\end{aligned} \tag{22.2.20}$$

We will now use the expression for the potential of a dipole, namely (2.19), to carry out an instructive construction and calculation. Suppose $\boldsymbol{r}_A$ and $\boldsymbol{r}_B$ are the locations of two points A and B. Imagine these two points to be joined by a line (path, *string*) L starting at $\boldsymbol{r}_A$ and ending at $\boldsymbol{r}_B$. See Figure 2.1. Divide the path into N segments, each of length Δs, and place a dipole of magnitude $g\Delta s$ at the center of each segment with the dipole moment vector pointing along the path at each point. Here g is some constant. Thus, the dipole moment $\Delta\boldsymbol{p}_d$ of each segment is given by the expression

$$\Delta\boldsymbol{p}_d = g\Delta s(\Delta\boldsymbol{r}/|\Delta\boldsymbol{r}|) = g\Delta\boldsymbol{r} \tag{22.2.21}$$

since $|\Delta \boldsymbol{r}| = \Delta s$. Let us compute the potential $\phi_s(\boldsymbol{r})$ produced by this *string* of dipoles. It will be the sum of the potentials of the individual dipoles. In the limit $\Delta s \to 0$ and $N \to \infty$ it is given by the integral

$$\begin{aligned}\phi_s(\boldsymbol{r}) &= [1/(4\pi)] \int_L d\boldsymbol{p}_d \cdot (\boldsymbol{r} - \boldsymbol{r}_d)/|\boldsymbol{r} - \boldsymbol{r}_d|^3 \\ &= [g/(4\pi)] \int_{\boldsymbol{r}_A}^{\boldsymbol{r}_B} d\boldsymbol{r}_d \cdot (\boldsymbol{r} - \boldsymbol{r}_d)/|\boldsymbol{r} - \boldsymbol{r}_d|^3. \qquad (22.2.22)\end{aligned}$$

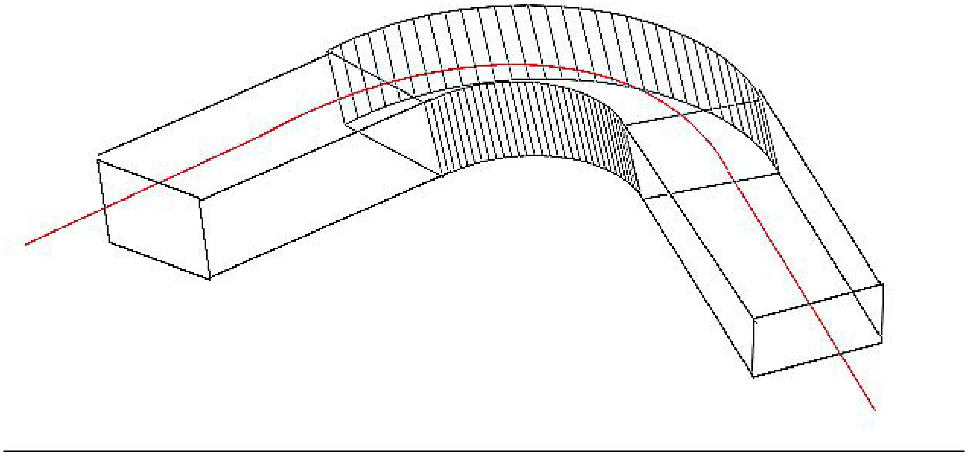


Figure 22.2.1: (Place Holder) A path L from the point A to the point B. Dipoles are laid out and aligned along the path to form a string.

Can the integral (2.22) be evaluated? Recall the identity

$$\nabla^d (1/|\boldsymbol{r} - \boldsymbol{r}_d|) = (\boldsymbol{r} - \boldsymbol{r}_d)/|\boldsymbol{r} - \boldsymbol{r}_d|^3 \qquad (22.2.23)$$

where ∇^d denotes differentiation with respect to the components of $\boldsymbol{r}_d$. This identity may be employed in (2.22) to yield the result

$$\begin{aligned}\phi_s(\boldsymbol{r}) &= [g/(4\pi)] \int_{\boldsymbol{r}_A}^{\boldsymbol{r}_B} d\boldsymbol{r}_d \cdot (\boldsymbol{r} - \boldsymbol{r}_d)/|\boldsymbol{r} - \boldsymbol{r}_d|^3 \\ &= [g/(4\pi)] \int_{\boldsymbol{r}_A}^{\boldsymbol{r}_B} d\boldsymbol{r}_d \cdot [\nabla^d (1/|\boldsymbol{r} - \boldsymbol{r}_d|)] \\ &= [g/(4\pi)]\{[(1/|\boldsymbol{r} - \boldsymbol{r}_B|)] - [(1/|\boldsymbol{r} - \boldsymbol{r}_A|)]\}. \qquad (22.2.24)\end{aligned}$$

We see that the potential $\phi_s(\boldsymbol{r})$ resulting from a string of dipoles is the same as the potential produced by a charge $-g$ located at $\boldsymbol{r}_A$ and a charge $+g$ located at $\boldsymbol{r}_B$. This mathematically derived result is also intuitive because we expect, for a string of dipoles arrayed head-to-tail, that adjacent head-tail pairs would cancel so all that would be left would be the initial negative tail and the final positive head.

Note that, as it stands, (2.22) is undefined for points $\boldsymbol{r} \in L$. However, since the integrand in (2.22) is a perfect differential, see (2.23), the path can be deformed at will to avoid any possible vanishings of the denominator in (2.22) without changing the value of the integral. Indeed, (2.24) shows that $\phi_s(\boldsymbol{r})$ depends only on the endpoints of the path, and is otherwise path independent.

22.2.2 Magnetic Dirac Strings

The General Case

In analogy to the work of the previous subsection, this subsection will describe calculations for the complementary case of a vector field $\boldsymbol{B}(\boldsymbol{r})$ that obeys the equations

$$\nabla \times \boldsymbol{B} = \boldsymbol{J}, \tag{22.2.25}$$

$$\nabla \cdot \boldsymbol{B} = 0. \tag{22.2.26}$$

Note, in order for (2.25) to make sense, we must require that

$$\nabla \cdot \boldsymbol{J} = \nabla \cdot (\nabla \times \boldsymbol{B}) = 0. \tag{22.2.27}$$

(Recall that the divergence of a curl vanishes.)

In the case of (2.25) and (2.26) it is often assumed that there is a vector potential $\boldsymbol{A}(\boldsymbol{r})$ such that

$$\boldsymbol{B} = \nabla \times \boldsymbol{A} \tag{22.2.28}$$

because (2.26) will then be satisfied automatically. Let us verify that this Ansatz is possible by construction. Substitution of (2.28) into (2.25) yields the hypothesis

$$\nabla \times (\nabla \times \boldsymbol{A}) = \boldsymbol{J}. \tag{22.2.29}$$

Recall the vector identity

$$\nabla \times (\nabla \times \boldsymbol{A}) = \nabla(\nabla \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A} \tag{22.2.30}$$

where here it is essential that Cartesian components be employed. Let us make the further Coulomb gauge assumption

$$\nabla \cdot \boldsymbol{A} = 0. \tag{22.2.31}$$

In this circumstance (2.29) and (2.30) become

$$\nabla^2 \boldsymbol{A} = -\boldsymbol{J}. \tag{22.2.32}$$

Thanks to (2.6), equation (2.32) has the immediate solution

$$\boldsymbol{A}(\boldsymbol{r}) = [1/(4\pi)] \int d^3\boldsymbol{r}'\ \boldsymbol{J}(\boldsymbol{r}')/|\boldsymbol{r} - \boldsymbol{r}'|. \tag{22.2.33}$$

Moreover, (2.33) is the unique solution that vanishes at infinity. But wait, we must also verify that (2.33) also satisfies (2.31). It does, as you will have the pleasure of showing in Exercise 2.4.

Next suppose that the current distribution $\boldsymbol{J}$ is nonzero only in some volume V surrounding the point $\boldsymbol{r}_d$. Then (2.33) becomes

$$\boldsymbol{A}(\boldsymbol{r}) = [1/(4\pi)] \int_V d^3\boldsymbol{r}' \, \boldsymbol{J}(\boldsymbol{r}')/|\boldsymbol{r} - \boldsymbol{r}'|. \tag{22.2.34}$$

Suppose also that $\boldsymbol{r}$ lies outside V so that the denominator in (2.34) never vanishes. Make the change of variables (2.9) so that (2.34) can be rewritten in the form

$$\boldsymbol{A}(\boldsymbol{r}) = [1/(4\pi)] \int_{V_0} d^3\boldsymbol{\xi} \, \boldsymbol{J}(\boldsymbol{r}_d + \boldsymbol{\xi})/|(\boldsymbol{r} - \boldsymbol{r}_d) - \boldsymbol{\xi}| \tag{22.2.35}$$

where V_0 is a volume surrounding the origin. As before, make the expansion (2.11) so that (2.35) can be written in the form

$$\begin{aligned} \boldsymbol{A}(\boldsymbol{r}) \;\; = \;\; & [1/(4\pi)][1/|\boldsymbol{r} - \boldsymbol{r}_d|] \int_{V_0} d^3\boldsymbol{\xi} \, \boldsymbol{J}(\boldsymbol{r}_d + \boldsymbol{\xi}) \\ & +[1/(4\pi)][1/|\boldsymbol{r} - \boldsymbol{r}_d|^3] \int_{V_0} d^3\boldsymbol{\xi} \, [(\boldsymbol{r} - \boldsymbol{r}_d)\cdot\boldsymbol{\xi}] \, \boldsymbol{J}(\boldsymbol{r}_d + \boldsymbol{\xi}) + O(\xi^2). \end{aligned} \tag{22.2.36}$$

The integrals in (2.36) can again be manipulated to bring them to more convenient forms. For the first integral we find that

$$\int_{V_0} d^3\boldsymbol{\xi} \, \boldsymbol{J}(\boldsymbol{r}_d + \boldsymbol{\xi}) = \int_V d^3\boldsymbol{r}' \, \boldsymbol{J}(\boldsymbol{r}') = 0. \tag{22.2.37}$$

Here use has been made of (2.27). See Exercise 2.5. The second integral can be brought to the form

$$\begin{aligned} \int_{V_0} d^3\boldsymbol{\xi} \, [(\boldsymbol{r} - \boldsymbol{r}_d)\cdot\boldsymbol{\xi}] \, \boldsymbol{J}(\boldsymbol{r}_d + \boldsymbol{\xi}) \;\; &= \;\; \int_V d^3\boldsymbol{r}' \, [(\boldsymbol{r} - \boldsymbol{r}_d)\cdot(\boldsymbol{r}' - \boldsymbol{r}_d)] \, \boldsymbol{J}(\boldsymbol{r}') \\ &= \;\; \boldsymbol{m}_d \times (\boldsymbol{r} - \boldsymbol{r}_d). \end{aligned} \tag{22.2.38}$$

Here use has again been made of (2.27), and $\boldsymbol{m}_d$ is the magnetic dipole moment defined by the integral

$$\boldsymbol{m}_d = (1/2) \int_V d^3\boldsymbol{r}' \, [(\boldsymbol{r}' - \boldsymbol{r}_d) \times \boldsymbol{J}(\boldsymbol{r}')]. \tag{22.2.39}$$

See Exercise 2.6. Thus, we find that

$$\boldsymbol{A}(\boldsymbol{r}) = \boldsymbol{A}_d(\boldsymbol{r}, \boldsymbol{r}_d) + O(\xi^2) \tag{22.2.40}$$

where

$$\boldsymbol{A}_d(\boldsymbol{r}, \boldsymbol{r}_d) = [1/(4\pi)][\boldsymbol{m}_d \times (\boldsymbol{r} - \boldsymbol{r}_d)]/|\boldsymbol{r} - \boldsymbol{r}_d|^3. \tag{22.2.41}$$

We see that the vector potential arising from a current distribution, at a point $\boldsymbol{r}$ outside the distribution, is a sum of dipole and higher order multipole contributions. Unlike the electric case, there is no monopole contribution. We also remark that $\boldsymbol{A}(\boldsymbol{r}, \boldsymbol{r}_d)$ satisfies the Coulomb gauge condition (2.31),

$$\nabla \cdot \boldsymbol{A}(\boldsymbol{r}, \boldsymbol{r}_d) = 0. \tag{22.2.42}$$

See Exercise 2.7.

We recall that the prototypical example of a magnetic dipole consists of a small circular and planar ring of radius R, surrounding an area A and carrying a current I, in the limit that $A \to 0$ and $I \to \infty$ in such a way that the product AI remains constant. For example, suppose the ring is placed in the x, y plane and centered around the origin. Suppose also that the current I circulates in the counterclockwise direction when viewed from above (looking down from positive z toward the origin). Then we find that (2.39) takes the form

$$\boldsymbol{m}_d = (1/2) \int_V d^3\boldsymbol{r}' \; [\boldsymbol{r}' \times \boldsymbol{J}(\boldsymbol{r}')] = AI\boldsymbol{e}_z. \tag{22.2.43}$$

For the field $\boldsymbol{B}_d(\boldsymbol{r}, \boldsymbol{r}_d)$ at the point $\boldsymbol{r}$ arising from a magnetic dipole at the point $\boldsymbol{r}_d$ (with $\boldsymbol{r} \neq \boldsymbol{r}_d$) we find the result

$$\begin{aligned} \boldsymbol{B}_d(\boldsymbol{r}, \boldsymbol{r}_d) &= \nabla \times \boldsymbol{A}_d(\boldsymbol{r}, \boldsymbol{r}_d) \\ &= -[1/(4\pi)][\boldsymbol{m}_d/|\boldsymbol{r} - \boldsymbol{r}_d|^3] + [3/(4\pi)](\boldsymbol{r} - \boldsymbol{r}_d)[\boldsymbol{m}_d \cdot (\boldsymbol{r} - \boldsymbol{r}_d)]/|\boldsymbol{r} - \boldsymbol{r}_d|^5. \end{aligned} \tag{22.2.44}$$

Note that the right sides of (2.20) and (2.44) agree if $\boldsymbol{p}_d = \boldsymbol{m}_d$. Thus, we have the key mathematical relation

$$-\nabla\phi_d(\boldsymbol{r}, \boldsymbol{r}_d) = \nabla \times \boldsymbol{A}_d(\boldsymbol{r}, \boldsymbol{r}_d) \text{ when } \boldsymbol{p}_d = \boldsymbol{m}_d \text{ and } \boldsymbol{r} \neq \boldsymbol{r}_d. \tag{22.2.45}$$

In analogy to what was done in the previous subsection for a string of electric dipoles, let us compute the vector potential $\boldsymbol{A}_s(\boldsymbol{r})$ arising from a string of magnetic dipoles. Again we will initially divide the path into N equal segments, and the magnetic dipole moment of each segment will be given by the relation

$$\Delta\boldsymbol{m}_d = g\Delta s(\Delta\boldsymbol{r}/|\Delta\boldsymbol{r}|) = g\Delta\boldsymbol{r}. \tag{22.2.46}$$

In the limit $\Delta s \to 0$ and $N \to \infty$ the vector potential due to the string is given by the integral

$$\begin{aligned} \boldsymbol{A}_s(\boldsymbol{r}) &= [1/(4\pi)] \int_L d\boldsymbol{m}_d \times (\boldsymbol{r} - \boldsymbol{r}_d)/|\boldsymbol{r} - \boldsymbol{r}_d|^3 \\ &= [g/(4\pi)] \int_{\boldsymbol{r}_A}^{\boldsymbol{r}_B} d\boldsymbol{r}_d \times (\boldsymbol{r} - \boldsymbol{r}_d)/|\boldsymbol{r} - \boldsymbol{r}_d|^3. \end{aligned} \tag{22.2.47}$$

Recall (2.41). Note that, as it stands, (2.47) is undefined for points $\boldsymbol{r} \in L$. As before, the path can be deformed to avoid any possible vanishings of the denominator. However, unlike the electric case and as will soon be seen, so doing changes the value of $\boldsymbol{A}_s(\boldsymbol{r})$. We also note that the current distribution associated with a string of magnetic dipoles (all aligned along the string) is that of an infinitesimally thin solenoid bent into the shape of the string.

What is the nature of the magnetic field $\boldsymbol{B}_s(\boldsymbol{r})$ given by

$$\boldsymbol{B}_s(\boldsymbol{r}) = \nabla \times \boldsymbol{A}_s(\boldsymbol{r})? \tag{22.2.48}$$

We claim, for $\boldsymbol{r} \notin L$, that

$$\nabla \times \boldsymbol{A}_s(\boldsymbol{r}) = -\nabla \phi_s(\boldsymbol{r}). \tag{22.2.49}$$

We will prove this assertion shortly. Assuming it is true, the right side of (2.48) can be evaluated easily using (2.49). In view of (2.24), there is the relation

$$-\nabla \phi_s(\boldsymbol{r}) = [g/(4\pi)][(\boldsymbol{r}-\boldsymbol{r}_B)/|\boldsymbol{r}-\boldsymbol{r}_B|^3] - [g/(4\pi)][(\boldsymbol{r}-\boldsymbol{r}_A)/|\boldsymbol{r}-\boldsymbol{r}_A|^3]. \tag{22.2.50}$$

It follows that

$$\boldsymbol{B}_s(\boldsymbol{r}) = [g/(4\pi)][(\boldsymbol{r}-\boldsymbol{r}_B)/|\boldsymbol{r}-\boldsymbol{r}_B|^3] - [g/(4\pi)][(\boldsymbol{r}-\boldsymbol{r}_A)/|\boldsymbol{r}-\boldsymbol{r}_A|^3]. \tag{22.2.51}$$

We see that the field $\boldsymbol{B}_s(\boldsymbol{r})$ is that produced by two magnetic *monopoles*, one located at $\boldsymbol{r}_B$ with strength g, and a second located at $\boldsymbol{r}_A$ with strength $-g$.

At this juncture two comments are in order. First, the $\boldsymbol{B}_s(\boldsymbol{r})$ given by (2.51) evidently *is not* divergence free at the points $\boldsymbol{r}_A$ and $\boldsymbol{r}_B$. But the $\boldsymbol{B}_s(\boldsymbol{r})$ given by (2.48) is a curl, and we again recall the theorem that a curl *is* divergence free. The resolution to this apparent paradox is that $\boldsymbol{A}_s(\boldsymbol{r})$ is singular for $\boldsymbol{r} \in L$, and every neighborhood of the points $\boldsymbol{r}_A$ and $\boldsymbol{r}_B$ contains such singular points, and therefore the conditions for the theorem are not met. Correspondingly, (2.51) holds only for points $\boldsymbol{r} \notin L$.

The second comment is equally subtle. Suppose two different strings s and s' (but with the same endpoints) are used to compute $\boldsymbol{B}_s(\boldsymbol{r})$ and $\boldsymbol{B}_{s'}(\boldsymbol{r})$. Then, according to (2.51), these fields should agree except possibly at the points for which $\boldsymbol{r} \in L$ and/or $\boldsymbol{r} \in L'$. Thus, we have the relation

$$\nabla \times [\boldsymbol{A}_s(\boldsymbol{r}) - \boldsymbol{A}_{s'}(\boldsymbol{r})] = 0 \text{ for } \boldsymbol{r} \notin L \text{ and } \boldsymbol{r} \notin L'. \tag{22.2.52}$$

Let Σ be some surface spanning the two strings s and s'. See Figure 2.2. Three-dimensional Euclidean space with the surface Σ excluded is still simply connected. It follows that there is a function $\psi_{ss'}(\boldsymbol{r})$ such that

$$\boldsymbol{A}_s(\boldsymbol{r}) - \boldsymbol{A}_{s'}(\boldsymbol{r}) = \nabla \psi_{ss'}(\boldsymbol{r}) \text{ for } \boldsymbol{r} \notin \Sigma. \tag{22.2.53}$$

That is, the vector potentials associated with two different strings (but with the same endpoints) are related by a gauge transformation. From (2.53) we see that $\psi_{ss'}(\boldsymbol{r})$ will be singular for both $\boldsymbol{r} \in L$ and $\boldsymbol{r} \in L'$. It can be shown that $\psi_{ss'}(\boldsymbol{r})$ is also harmonic,

$$\nabla^2 \psi_{ss'}(\boldsymbol{r}) = 0 \text{ for } \boldsymbol{r} \notin L \text{ and } \boldsymbol{r} \notin L'. \tag{22.2.54}$$

See Exercise 2.13.

Finally, suppose we let $\boldsymbol{r}_B \to \infty$. In this limit, the first term on the right side of (2.51) vanishes, and we have the result

$$\boldsymbol{B}_s(\boldsymbol{r}) = -[g/(4\pi)][(\boldsymbol{r}-\boldsymbol{r}_A)/|\boldsymbol{r}-\boldsymbol{r}_A|^3], \tag{22.2.55}$$

which is the field of a monopole located at $\boldsymbol{r}_A$ and having strength $-g$. Correspondingly, the upper limit in the integral (2.47) is infinite, and the string s, which we will call a *half-infinite Dirac string*, extends from $\boldsymbol{r}_A$ to infinity. And the field (2.55) may be viewed as that of a *Dirac* magnetic monopole.

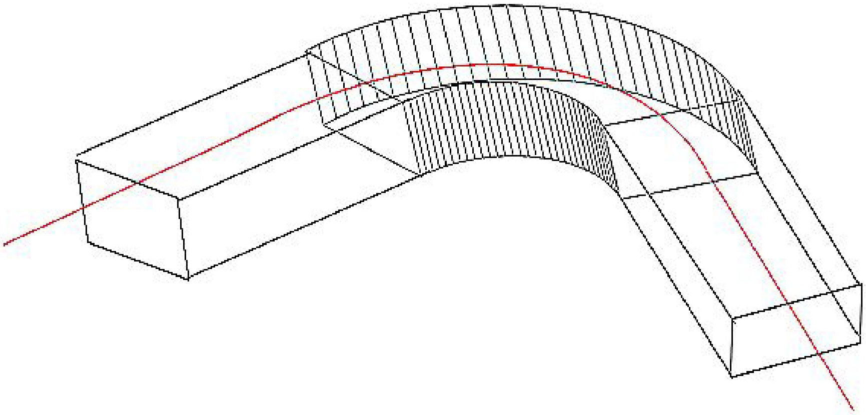


Figure 22.2.2: (Place holder.) A surface Σ spanning the two strings s and s'.

Straight Half-Infinite Strings

For future use, there is a special class of half-infinite strings that is particularly convenient. Let $\boldsymbol{m}$ be some unit vector. Consider the straight string (path) from $\boldsymbol{r}_A$ to infinity parameterized as

$$\boldsymbol{r}_d(\lambda) = \boldsymbol{r}_A + \lambda\boldsymbol{m} \text{ with } \lambda \in [0, \infty]. \tag{22.2.56}$$

See Figure 2.3. Then, on this path, $\boldsymbol{m}_d$ is in the direction of $\boldsymbol{m}$, and we also have the relation

$$d\boldsymbol{r}_d(\lambda) = \boldsymbol{m}d\lambda. \tag{22.2.57}$$

For this class of strings the integral (2.47) can be evaluated analytically. We begin by rewriting (2.47) in the form

$$\boldsymbol{A}_s(\boldsymbol{r}; \boldsymbol{r}_A, \boldsymbol{m}) = [g/(4\pi)] \int_0^\infty d\lambda\, \boldsymbol{m} \times [\boldsymbol{r} - \boldsymbol{r}_d(\lambda)]/|\boldsymbol{r} - \boldsymbol{r}_d(\lambda)|^3. \tag{22.2.58}$$

From (2.56) we see that

$$\boldsymbol{r} - \boldsymbol{r}_d(\lambda) = \boldsymbol{r} - \boldsymbol{r}_A - \lambda\boldsymbol{m} \tag{22.2.59}$$

and therefore

$$\boldsymbol{m} \times [\boldsymbol{r} - \boldsymbol{r}_d(\lambda)] = \boldsymbol{m} \times (\boldsymbol{r} - \boldsymbol{r}_A). \tag{22.2.60}$$

Consequently, the integral (2.58) simplifies to the form

$$\boldsymbol{A}_s(\boldsymbol{r}; \boldsymbol{r}_A, \boldsymbol{m}) = [g/(4\pi)][\boldsymbol{m} \times (\boldsymbol{r} - \boldsymbol{r}_A)] \int_0^\infty d\lambda/|\boldsymbol{r} - \boldsymbol{r}_d(\lambda)|^3. \tag{22.2.61}$$

As shown in Exercise 2.14, the integral appearing in (2.61) can be evaluated to yield the result

$$\begin{aligned}\int_0^\infty d\lambda/|\boldsymbol{r} - \boldsymbol{r}_d(\lambda)|^3 &= \int_0^\infty d\lambda/|\boldsymbol{r} - \boldsymbol{r}_A - \lambda\boldsymbol{m}|^3 \\ &= 1/\{|\boldsymbol{r} - \boldsymbol{r}_A|[|\boldsymbol{r} - \boldsymbol{r}_A| - \boldsymbol{m} \cdot (\boldsymbol{r} - \boldsymbol{r}_A)]\}. \end{aligned} \tag{22.2.62}$$

Therefore, $\boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m})$ takes the final explicit form

$$\boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m}) = [g/(4\pi)][\boldsymbol{m}\times(\boldsymbol{r}-\boldsymbol{r}_A)]/\{|\boldsymbol{r}-\boldsymbol{r}_A|[|\boldsymbol{r}-\boldsymbol{r}_A|-\boldsymbol{m}\cdot(\boldsymbol{r}-\boldsymbol{r}_A)]\}. \tag{22.2.63}$$

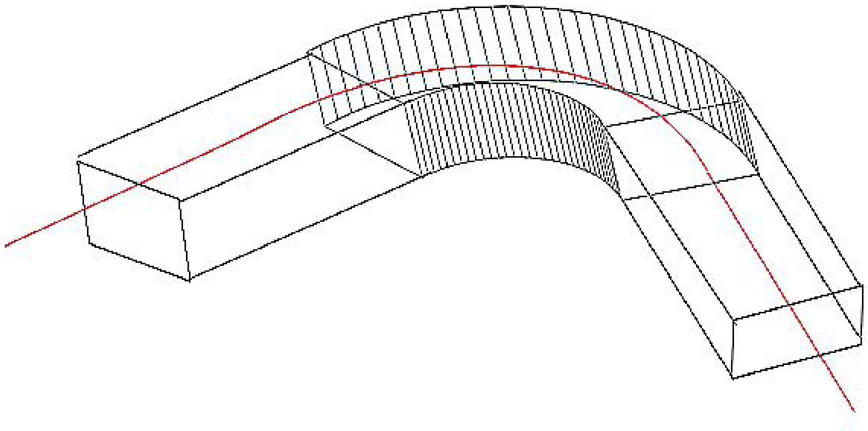


Figure 22.2.3: (Place holder.) A straight half-infinite string extending from $\boldsymbol{r}_A$ to infinity in the direction $\boldsymbol{m}$.

Remaining Verifications

It remains to be verified that (2.49) holds. Suppose that (2.19) is written in the form

$$\phi_d(\boldsymbol{r},\boldsymbol{r}_d;|\boldsymbol{p}_d|,\boldsymbol{n}_d) = [1/(4\pi)][|\boldsymbol{p}_d|\boldsymbol{n}_d\cdot(\boldsymbol{r}-\boldsymbol{r}_d)]/|\boldsymbol{r}-\boldsymbol{r}_d|^3 \tag{22.2.64}$$

where $\boldsymbol{n}_d$ is the unit vector in the direction of $\boldsymbol{p}_d$. Then (2.22) takes the form

$$\begin{aligned}\phi_s(\boldsymbol{r}) &= [1/(4\pi)]\int_L d\boldsymbol{p}_d\cdot(\boldsymbol{r}-\boldsymbol{r}_d)/|\boldsymbol{r}-\boldsymbol{r}_d|^3 \\ &= [g/(4\pi)]\int_{\boldsymbol{r}_A}^{\boldsymbol{r}_B} d\boldsymbol{r}_d\cdot(\boldsymbol{r}-\boldsymbol{r}_d)/|\boldsymbol{r}-\boldsymbol{r}_d|^3 \\ &= \int_L \phi_d(\boldsymbol{r},\boldsymbol{r}_d;gds,d\boldsymbol{r}_d/|d\boldsymbol{r}_d|),\end{aligned} \tag{22.2.65}$$

and therefore

$$-\nabla\phi_s(\boldsymbol{r}) = \int_L -\nabla\phi_d(\boldsymbol{r},\boldsymbol{r}_d;gds,d\boldsymbol{r}_d/|d\boldsymbol{r}_d|). \tag{22.2.66}$$

Suppose also that (2.41) is written in the form

$$\boldsymbol{A}_d(\boldsymbol{r},\boldsymbol{r}_d;|\boldsymbol{m}_d|,\boldsymbol{n}_d) = [1/(4\pi)][|\boldsymbol{m}_d|\boldsymbol{n}_d\times(\boldsymbol{r}-\boldsymbol{r}_d)]/|\boldsymbol{r}-\boldsymbol{r}_d|^3. \tag{22.2.67}$$

Then (2.47) takes the form

$$\begin{aligned}\boldsymbol{A}_s(\boldsymbol{r}) &= [1/(4\pi)]\int_L d\boldsymbol{m}_d \times (\boldsymbol{r}-\boldsymbol{r}_d)/|\boldsymbol{r}-\boldsymbol{r}_d|^3 \\ &= [g/(4\pi)]\int_{\boldsymbol{r}_A}^{\boldsymbol{r}_B} d\boldsymbol{r}_d \times (\boldsymbol{r}-\boldsymbol{r}_d)/|\boldsymbol{r}-\boldsymbol{r}_d|^3 \\ &= \int_L \boldsymbol{A}_d(\boldsymbol{r},\boldsymbol{r}_d; gds, d\boldsymbol{r}_d/|d\boldsymbol{r}_d|), \end{aligned} \tag{22.2.68}$$

and therefore

$$\nabla \times \boldsymbol{A}_s(\boldsymbol{r}) = \int_L \nabla \times \boldsymbol{A}_d(\boldsymbol{r},\boldsymbol{r}_d; gds, d\boldsymbol{r}_d/|d\boldsymbol{r}_d|). \tag{22.2.69}$$

Now compare the integrands on the right sides of (2.66) and (2.69). We see that they have identical arguments. Consequently, by (2.45), they are equal. It follows that the left sides of (2.66) and (2.69) are equal, and therefore (2.49) is correct.

There are still two final matters. First, (2.68) shows that $\boldsymbol{A}_s(\boldsymbol{r})$ is a superposition (integration over $\boldsymbol{r}_d$) of the $\boldsymbol{A}_d(\boldsymbol{r},\boldsymbol{r}_d)$ and, for each $\boldsymbol{A}_d(\boldsymbol{r},\boldsymbol{r}_d)$, we know that the relation (2.42) holds. It follows that $\boldsymbol{A}_s(\boldsymbol{r})$ also satisfies the Coulomb gauge condition,

$$\nabla \cdot \boldsymbol{A}_s(\boldsymbol{r}) = 0. \tag{22.2.70}$$

In particular, there is the relation

$$\nabla \cdot \boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m}) = 0. \tag{22.2.71}$$

Second, since $\boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m})$, is a magnetic monopole vector potential, there is the relation

$$\nabla \times \boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m}) = -[g/(4\pi)][(\boldsymbol{r}-\boldsymbol{r}_A)/|\boldsymbol{r}-\boldsymbol{r}_A|^3] = [g/(4\pi)]\nabla(1/|\boldsymbol{r}-\boldsymbol{r}_A|). \tag{22.2.72}$$

It follows that

$$\nabla \times [\nabla \times \boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m})] = 0. \tag{22.2.73}$$

The relations (2.71) and (2.73) will be of subsequent use.

Fully Infinite (Two) String Monopole Vector Potential

The previous discussion treated the half-infinite string vector potential for a magnetic monopole. In particular, (2.63) gives the vector potential $\boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m})$ for a magnetic monopole of strength $-g$, see (2.55), located at $\boldsymbol{r}_A$ with a straight string extending from $\boldsymbol{r}_A$ to ∞ in the direction $\boldsymbol{m}$. This vector potential is singular on the line

$$\boldsymbol{r} = \boldsymbol{r}_A + \lambda\boldsymbol{m} \text{ with } \lambda \in [0,\infty]. \tag{22.2.74}$$

For completeness we will now describe what we will call the *fully infinite* string monopole vector potential.[2] Suppose we form the average of $\boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m})$ and $\boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,-\boldsymbol{m})$ by writing

$$\boldsymbol{A}_{2s}(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m}) = (1/2)[\boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m}) + \boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,-\boldsymbol{m})]. \tag{22.2.75}$$

[2] The fully infinite string monopole vector potential is sometimes called the *Schwinger potential*.

This vector potential, which we will also call a *two*-string monopole vector potential, will (by superposition) also produce the monopole field (2.55), and will be singular along the full line

$$\boldsymbol{r} = \boldsymbol{r}_A + \lambda \boldsymbol{m} \text{ with } \lambda \in [-\infty, \infty]. \tag{22.2.76}$$

See Figure 2.4.

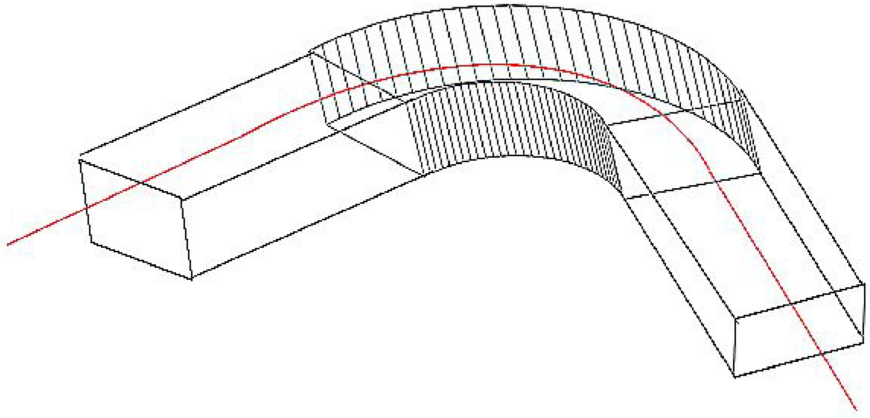


Figure 22.2.4: (Place holder.) A straight full infinite string extending from $\boldsymbol{r}_A$ to infinity in the directions ±$\boldsymbol{m}$.

By superposition, and the use of (2.71) and (2.73), the fully infinite string monopole vector potential satisfies the analogous relations

$$\nabla \cdot \boldsymbol{A}_{2s}(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m}) = 0 \tag{22.2.77}$$

and

$$\nabla \times [\nabla \times \boldsymbol{A}_{2s}(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m})] = 0, \tag{22.2.78}$$

provided $\boldsymbol{r}$ is not on the line (2.76). Finally, from (2.63) and the definition (2.75), we find that $\boldsymbol{A}_{2s}(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m})$ has the explicit form

$$\begin{aligned}
\boldsymbol{A}_{2s}(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m}) &= [g/(8\pi)][\boldsymbol{m}\times(\boldsymbol{r}-\boldsymbol{r}_A)]/\{|\boldsymbol{r}-\boldsymbol{r}_A|[|\boldsymbol{r}-\boldsymbol{r}_A| - \boldsymbol{m}\cdot(\boldsymbol{r}-\boldsymbol{r}_A)]\} \\
&\quad -[g/(8\pi)][\boldsymbol{m}\times(\boldsymbol{r}-\boldsymbol{r}_A)]/\{|\boldsymbol{r}-\boldsymbol{r}_A|[|\boldsymbol{r}-\boldsymbol{r}_A| + \boldsymbol{m}\cdot(\boldsymbol{r}-\boldsymbol{r}_A)]\} \\
&= \frac{[g/(4\pi)][\boldsymbol{m}\times(\boldsymbol{r}-\boldsymbol{r}_A)][\boldsymbol{m}\cdot(\boldsymbol{r}-\boldsymbol{r}_A)]}{|\boldsymbol{r}-\boldsymbol{r}_A|\{|\boldsymbol{r}-\boldsymbol{r}_A|^2 - [\boldsymbol{m}\cdot(\boldsymbol{r}-\boldsymbol{r}_A)]^2\}} \\
&= \frac{[g/(4\pi)][\boldsymbol{m}\times(\boldsymbol{r}-\boldsymbol{r}_A)][\boldsymbol{m}\cdot(\boldsymbol{r}-\boldsymbol{r}_A)]}{|\boldsymbol{r}-\boldsymbol{r}_A||\boldsymbol{m}\times(\boldsymbol{r}-\boldsymbol{r}_A)|^2}.
\end{aligned} \tag{22.2.79}$$

22.2.3 Helmholtz Decomposition

Suppose V is some simply connected volume in 3-dimensional space bounded by a surface S, and suppose $\boldsymbol{F}(\boldsymbol{r})$ is some 3-dimensional vector field defined in V. Then, according to a theorem of *Helmholtz*, there are scalar and vector potentials $\phi(\boldsymbol{r})$ and $\boldsymbol{A}(\boldsymbol{r})$ such that

$$\boldsymbol{F}(\boldsymbol{r}) = -\nabla\phi(\boldsymbol{r}) + \nabla \times \boldsymbol{A}(\boldsymbol{r}) \text{ for } \boldsymbol{r} \in \mathrm{V}. \tag{22.2.80}$$

Moreover, $\boldsymbol{A}(\boldsymbol{r})$ will have the property

$$\nabla \cdot \boldsymbol{A}(\boldsymbol{r}) = 0 \text{ for } \boldsymbol{r} \in \mathrm{V}. \tag{22.2.81}$$

Finally, let Let $G(\boldsymbol{r}, \boldsymbol{r}')$ be the function

$$G(\boldsymbol{r}, \boldsymbol{r}') = 1/|\boldsymbol{r} - \boldsymbol{r}'|. \tag{22.2.82}$$

Then, the scalar and vector potentials are given in terms of $\boldsymbol{F}(\boldsymbol{r})$, with $\boldsymbol{r} \in V$, by the relations

$$\phi(\boldsymbol{r}) = -[1/(4\pi)] \int_S dS'\, \boldsymbol{n}' \cdot \boldsymbol{F}(\boldsymbol{r}')G(\boldsymbol{r}, \boldsymbol{r}') + [1/(4\pi)] \int_V d^3\boldsymbol{r}'\, G(\boldsymbol{r}, \boldsymbol{r}')\nabla' \cdot \boldsymbol{F}(\boldsymbol{r}'), \tag{22.2.83}$$

$$\boldsymbol{A}(\boldsymbol{r}) = -[1/(4\pi)] \int_S dS'\, [\boldsymbol{n}' \times G(\boldsymbol{r}, \boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')] + [1/(4\pi)] \int_V d^3\boldsymbol{r}'\, G(\boldsymbol{r}, \boldsymbol{r}')\nabla' \times \boldsymbol{F}(\boldsymbol{r}'). \tag{22.2.84}$$

Here $\boldsymbol{n}'$ is the outward normal to S at the point $\boldsymbol{r}'$. We emphasize, as is evident from (2.80), (2.83), and (2.84), that for $\boldsymbol{r} \in V$ the vector field $\boldsymbol{F}(\boldsymbol{r})$ is completely specified in terms of the divergence and curl of $\boldsymbol{F}$ within V and the values of $\boldsymbol{F}$ on the bounding surface S. No information is required outside of V.

We will derive this result in stages. Before doing so, some remarks are in order. There two cases of special interest. If $\boldsymbol{F}(\boldsymbol{r})$ is globally defined and falls off at infinity at least as fast as $1/|\boldsymbol{r}|^2$, then we may take the surface S to infinity and find that the surface integrals vanish. This result shows that, with suitable boundary conditions (fall off) imposed at infinity, $\boldsymbol{F}(\boldsymbol{r})$ is completely specified in terms of its divergence and curl. That the operations of divergence and curl are necessary and sufficient to determine $\boldsymbol{F}(\boldsymbol{r})$ is a consequence of two things: the fact that we are working in *three* dimensions and certain properties of the Euclidean group in three dimensions. See Exercise 2.21.

The second case, of special interest for our purposes, is that for which $\boldsymbol{F}(\boldsymbol{r})$ is divergence and curl free (source free) in V,

$$\nabla \cdot \boldsymbol{F}(\boldsymbol{r}) = 0 \text{ for } \boldsymbol{r} \in \mathrm{V} \tag{22.2.85}$$

and

$$\nabla \times \boldsymbol{F}(\boldsymbol{r}) = 0 \text{ for } \boldsymbol{r} \in \mathrm{V}. \tag{22.2.86}$$

In this case, only the surface terms appear in (2.83) and (2.84), and we obtain the results

$$\phi(\boldsymbol{r}) = -[1/(4\pi)] \int_S dS'\, \boldsymbol{n}' \cdot \boldsymbol{F}(\boldsymbol{r}')G(\boldsymbol{r}, \boldsymbol{r}'), \tag{22.2.87}$$

$$\boldsymbol{A}(\boldsymbol{r}) = -[1/(4\pi)] \int_S dS' \, [\boldsymbol{n}' \times \boldsymbol{F}(\boldsymbol{r}')] G(\boldsymbol{r}, \boldsymbol{r}'). \tag{22.2.88}$$

We will eventually apply these results to the case of a magnetic field $\boldsymbol{B}(\boldsymbol{r})$ that is assumed to be source free within V, as in (1.1) and (1.2). We take the opportunity at this point to note that $G(\boldsymbol{r}, \boldsymbol{r}')$ as given by (2.82), and for fixed $\boldsymbol{r}'$, is an *analytic* function of the components of $\boldsymbol{r}$ for $\boldsymbol{r} \neq \boldsymbol{r}'$. It follows from the representations (2.87) and (2.88), under very mild assumptions on the surface behavior of $\boldsymbol{F}(\boldsymbol{r})$, boundedness and continuity will do, that $\phi(\boldsymbol{r})$ and $\boldsymbol{A}(\boldsymbol{r})$ are analytic functions of the components of $\boldsymbol{r}$ for $\boldsymbol{r}$ within V. Correspondingly, from (2.80), $\boldsymbol{F}(\boldsymbol{r})$ must then also be analytic for $\boldsymbol{r}$ within V.

We begin the proof of Helmholtz's theorem by noting that $G(\boldsymbol{r}, \boldsymbol{r}')$ has the properties

$$\nabla G(\boldsymbol{r}, \boldsymbol{r}') = -\nabla' G(\boldsymbol{r}, \boldsymbol{r}') = -(\boldsymbol{r} - \boldsymbol{r}')/|\boldsymbol{r} - \boldsymbol{r}'|^3, \tag{22.2.89}$$

$$\nabla^2 G(\boldsymbol{r}, \boldsymbol{r}') = (\nabla')^2 G(\boldsymbol{r}, \boldsymbol{r}') = -4\pi\delta_3(\boldsymbol{r} - \boldsymbol{r}'), \tag{22.2.90}$$

where ∇' denotes differentiation with respect to the components of $\boldsymbol{r}'$. As a result of (2.90) there is, for $\boldsymbol{r} \in V$, the identity

$$\begin{aligned}
\boldsymbol{F}(\boldsymbol{r}) &= \int_V d^3\boldsymbol{r}' \, \delta_3(\boldsymbol{r} - \boldsymbol{r}') \boldsymbol{F}(\boldsymbol{r}') \\
&= -[1/(4\pi)] \int_V d^3\boldsymbol{r}' \, \boldsymbol{F}(\boldsymbol{r}') \nabla^2 G(\boldsymbol{r}, \boldsymbol{r}') \\
&= -[1/(4\pi)] \nabla^2 \int_V d^3\boldsymbol{r}' \, \boldsymbol{F}(\boldsymbol{r}') G(\boldsymbol{r}, \boldsymbol{r}') \\
&= -\nabla^2 \boldsymbol{H}(\boldsymbol{r})
\end{aligned} \tag{22.2.91}$$

where

$$\boldsymbol{H}(\boldsymbol{r}) = [1/(4\pi)] \int_V d^3\boldsymbol{r}' \, \boldsymbol{F}(\boldsymbol{r}') G(\boldsymbol{r}, \boldsymbol{r}'). \tag{22.2.92}$$

Invoke again the vector identity

$$-\nabla^2 \boldsymbol{H}(\boldsymbol{r}) = \nabla \times [\nabla \times \boldsymbol{H}(\boldsymbol{r})] - \nabla[\nabla \cdot \boldsymbol{H}(\boldsymbol{r})]. \tag{22.2.93}$$

It follows that

$$\boldsymbol{F}(\boldsymbol{r}) = \nabla \times [\nabla \times \boldsymbol{H}(\boldsymbol{r})] - \nabla[\nabla \cdot \boldsymbol{H}(\boldsymbol{r})], \tag{22.2.94}$$

and therefore (2.80) holds with the definitions

$$\phi(\boldsymbol{r}) = \nabla \cdot \boldsymbol{H}(\boldsymbol{r}), \tag{22.2.95}$$

$$\boldsymbol{A}(\boldsymbol{r}) = \nabla \times \boldsymbol{H}(\boldsymbol{r}). \tag{22.2.96}$$

It remains to work out computationally useful expressions for $\phi(\boldsymbol{r})$ and $\boldsymbol{A}(\boldsymbol{r})$. Doing so requires a flurry of vector manipulations. Begin with $\phi(\boldsymbol{r})$. According to (2.92) and (2.95) it can be written as

$$\phi(\boldsymbol{r}) = [1/(4\pi)] \int_V d^3\boldsymbol{r}' \, \nabla \cdot [\boldsymbol{F}(\boldsymbol{r}') G(\boldsymbol{r}, \boldsymbol{r}')]. \tag{22.2.97}$$

Manipulate the integrand in (2.97) to find the result

$$\begin{aligned}\nabla\cdot[\boldsymbol{F}(\boldsymbol{r}')G(\boldsymbol{r},\boldsymbol{r}')] &= \boldsymbol{F}(\boldsymbol{r}')\cdot\nabla G(\boldsymbol{r},\boldsymbol{r}')] = -\boldsymbol{F}(\boldsymbol{r}')\cdot\nabla' G(\boldsymbol{r},\boldsymbol{r}')]\\ &= -\nabla'\cdot[\boldsymbol{F}(\boldsymbol{r}')G(\boldsymbol{r},\boldsymbol{r}')] + G(\boldsymbol{r},\boldsymbol{r}')\nabla'\cdot\boldsymbol{F}(\boldsymbol{r}'). \qquad (22.2.98)\end{aligned}$$

Employ this result in (2.97) to rewrite it in the form

$$\phi(\boldsymbol{r}) = [1/(4\pi)]\int_V d^3\boldsymbol{r}'\ \{-\nabla'\cdot[\boldsymbol{F}(\boldsymbol{r}')G(\boldsymbol{r},\boldsymbol{r}')] + G(\boldsymbol{r},\boldsymbol{r}')\nabla'\cdot\boldsymbol{F}(\boldsymbol{r}')\}. \qquad (22.2.99)$$

Finally, use the divergence theorem to transform the first term on the right side of (2.99) to yield the result

$$\phi(\boldsymbol{r}) = -[1/(4\pi)]\int_S dS'\ \boldsymbol{n}'\cdot\boldsymbol{F}(\boldsymbol{r}')G(\boldsymbol{r},\boldsymbol{r}') + [1/(4\pi)]\int_V d^3\boldsymbol{r}'\ G(\boldsymbol{r},\boldsymbol{r}')\nabla'\cdot\boldsymbol{F}(\boldsymbol{r}'), \quad (22.2.100)$$

in agreement with (2.83).

The case of $\boldsymbol{A}(\boldsymbol{r})$ requires somewhat more effort. Combining (2.92) and (2.96) gives the result

$$\boldsymbol{A}(\boldsymbol{r}) = [1/(4\pi)]\int_V d^3\boldsymbol{r}'\ \nabla\times[\boldsymbol{F}(\boldsymbol{r}')G(\boldsymbol{r},\boldsymbol{r}')]. \qquad (22.2.101)$$

Manipulate the integrand in (2.101) to find the result

$$\begin{aligned}\nabla\times[\boldsymbol{F}(\boldsymbol{r}')G(\boldsymbol{r},\boldsymbol{r}')] &= [\nabla G(\boldsymbol{r},\boldsymbol{r}')]\times\boldsymbol{F}(\boldsymbol{r}') = -\boldsymbol{F}(\boldsymbol{r}')\times\nabla G(\boldsymbol{r},\boldsymbol{r}')\\ &= \boldsymbol{F}(\boldsymbol{r}')\times\nabla' G(\boldsymbol{r},\boldsymbol{r}'). \qquad (22.2.102)\end{aligned}$$

There is also the vector identity

$$\nabla'\times[\boldsymbol{F}(\boldsymbol{r}')G(\boldsymbol{r},\boldsymbol{r}')] = G(\boldsymbol{r},\boldsymbol{r}')\nabla'\times\boldsymbol{F}(\boldsymbol{r}') - \boldsymbol{F}(\boldsymbol{r}')\times\nabla' G(\boldsymbol{r},\boldsymbol{r}'). \qquad (22.2.103)$$

Combining (2.102) and (2.103) gives the result

$$\nabla\times[\boldsymbol{F}(\boldsymbol{r}')G(\boldsymbol{r},\boldsymbol{r}')] = G(\boldsymbol{r},\boldsymbol{r}')\nabla'\times\boldsymbol{F}(\boldsymbol{r}') - \nabla'\times[\boldsymbol{F}(\boldsymbol{r}')G(\boldsymbol{r},\boldsymbol{r}')]. \qquad (22.2.104)$$

Employ this result in (2.101) to rewrite it in the form

$$\boldsymbol{A}(\boldsymbol{r}) = [1/(4\pi)]\int_V d^3\boldsymbol{r}'\ G(\boldsymbol{r},\boldsymbol{r}')\nabla'\times\boldsymbol{F}(\boldsymbol{r}') - [1/(4\pi)]\int_V d^3\boldsymbol{r}'\ \nabla'\times[\boldsymbol{F}(\boldsymbol{r}')G(\boldsymbol{r},\boldsymbol{r}')]. \qquad (22.2.105)$$

Now work on the second integral appearing on the right side of (2.105). Let $\boldsymbol{c}$ be any constant vector. By the divergence theorem there is the relation

$$\int_V d^3\boldsymbol{r}'\ \nabla'\cdot[\boldsymbol{c}\times G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')] = \int_S dS'\ \boldsymbol{n}'\cdot[\boldsymbol{c}\times G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')]. \qquad (22.2.106)$$

There is also the vector identity

$$\begin{aligned}\boldsymbol{n}'\cdot[\boldsymbol{c}\times G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')] &= -\boldsymbol{n}'\cdot[G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')\times\boldsymbol{c}]\\ &= -[\boldsymbol{n}'\times G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')]\cdot\boldsymbol{c}\\ &= -\boldsymbol{c}\cdot[\boldsymbol{n}'\times G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')]. \qquad (22.2.107)\end{aligned}$$

Consequently, (2.106) can be rewritten in the form

$$\int_V d^3\boldsymbol{r}'\ \nabla' \cdot [\boldsymbol{c} \times G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')] = -\boldsymbol{c} \cdot \int_S dS'\ [\boldsymbol{n}' \times G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')]. \tag{22.2.108}$$

Next manipulate the integrand on the left side of (2.108) to find the result

$$\begin{aligned} \nabla' \cdot [\boldsymbol{c} \times G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')] &= -\nabla' \cdot [G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}') \times \boldsymbol{c}] \\ &= -\{\nabla' \times [G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')]\} \cdot \boldsymbol{c} \\ &= -\boldsymbol{c} \cdot \{\nabla' \times [G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')]\}. \end{aligned} \tag{22.2.109}$$

Therefore (2.108) can be rewritten as

$$-\boldsymbol{c} \cdot \int_V d^3\boldsymbol{r}'\ \nabla' \times [G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')] = -\boldsymbol{c} \cdot \int_S dS'\ [\boldsymbol{n}' \times G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')], \tag{22.2.110}$$

from which it follows, because $\boldsymbol{c}$ is arbitrary, that

$$\int_V d^3\boldsymbol{r}'\ \nabla' \times [G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')] = \int_S dS'\ [\boldsymbol{n}' \times G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')]. \tag{22.2.111}$$

The last step is to employ (2.111) in (2.105) to obtain the final result

$$\boldsymbol{A}(\boldsymbol{r}) = [1/(4\pi)] \int_V d^3\boldsymbol{r}'\ G(\boldsymbol{r},\boldsymbol{r}')\nabla' \times \boldsymbol{F}(\boldsymbol{r}') - [1/(4\pi)] \int_S dS'\ [\boldsymbol{n}' \times G(\boldsymbol{r},\boldsymbol{r}')\boldsymbol{F}(\boldsymbol{r}')], \tag{22.2.112}$$

in agreement with (2.84).

It still remains to be shown that, for the definitions made, $\nabla \cdot \boldsymbol{A}(\boldsymbol{r}) = 0$. Look at (2.96). Since the divergence of a curl vanishes, when suitable smoothness conditions are met by the functions involved, it follows that under these conditions $\boldsymbol{A}(\boldsymbol{r})$ as given by (2.96), and therefore also by (2.112), is indeed divergence free. From (2.92) we see that the analytic properties of $\boldsymbol{H}(\boldsymbol{r})$ are determined by those of $\boldsymbol{F}(\boldsymbol{r})$. In general $\boldsymbol{H}(\boldsymbol{r})$ will be smoother than $\boldsymbol{F}(\boldsymbol{r})$. See Appendix F. Therefore, under mild conditions on $\boldsymbol{F}(\boldsymbol{r})$, the vector potential $\boldsymbol{A}(\boldsymbol{r})$ will be divergence free.

Exercises

22.2.1. Verify the expansions (2.11) and (2.17).

22.2.2. Verify (2.20).

22.2.3. Verify the identity (2.23) and its use to evaluate the integral (2.24).

22.2.4. Verify that $\boldsymbol{A}(\boldsymbol{r})$ as given by (2.33) satisfies (2.31).

22.2.5. The purpose of this exercise is to verify (2.37) using (2.27).

22.2.6. The purpose of this exercise is to verify (2.38) using the definition (2.39).

22.2.7. The purpose of this exercise is to verify (2.42) using the definition (2.41).

22.2.8. Verify (2.43).

22.2.9. Verify (2.44).

22.2.10. Show that the integral (2.47) can be written in the form

$$[g/(4\pi)]\int_{\boldsymbol{r}_A}^{\boldsymbol{r}_B} d\boldsymbol{r}_d \times (\boldsymbol{r}-\boldsymbol{r}_d)/|\boldsymbol{r}-\boldsymbol{r}_d|^3 = -[g/(4\pi)]\int_{\boldsymbol{r}_A}^{\boldsymbol{r}_B} d\boldsymbol{r}_d \times \nabla[1/|\boldsymbol{r}-\boldsymbol{r}_d|]. \quad (22.2.113)$$

22.2.11. Verify (2.50).

22.2.12. Nature of thin solenoid and nature of field at the end of a thin solenoid.

22.2.13. The purpose of this exercise is to verify (2.54).

22.2.14. The purpose of this exercise is to verify (2.62).

22.2.15. Evaluate $\boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m})$ as given by (2.63) for the case

$$\boldsymbol{r}_A = 0 \quad (22.2.114)$$

and

$$\boldsymbol{m} = \boldsymbol{e}_z. \quad (22.2.115)$$

Show, using spherical coordinates, that in this case $\boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m})$ has only a ϕ component A^s_ϕ given by

$$A^s_\phi(\boldsymbol{r};0,\boldsymbol{e}_z) = [g/(4\pi)](1+\cos\theta)/[r\sin\theta] = [g/(4\pi)](1/r)\cot(\theta/2) \quad (22.2.116)$$

Verify that A^s_ϕ is singular on the positive z axis, but not on the negative z axis. Show, by explicit calculation, that

$$\nabla \times \boldsymbol{A}_s(\boldsymbol{r};0,\boldsymbol{e}_z) = -[g/(4\pi)][\boldsymbol{r}/|\boldsymbol{r}|^3], \quad (22.2.117)$$

as expected.

Repeat the above calculations for the case $\boldsymbol{m} = -\boldsymbol{e}_z$. Show that again $\boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m})$ has only a ϕ component A^s_ϕ now given by

$$\begin{aligned} A^s_\phi(\boldsymbol{r};0,-\boldsymbol{e}_z) &= -[g/(4\pi)](1-\cos\theta)/[r\sin\theta] = -[g/(4\pi)](1/r)\sin\theta/(1+\cos\theta) \\ &= -[g/(4\pi)](1/r)\tan(\theta/2). \quad (22.2.118) \end{aligned}$$

Verify that this A^s_ϕ is singular on the negative z axis, but not on the posititve z axis. Show, by explicit calculation, that

$$\nabla \times \boldsymbol{A}_s(\boldsymbol{r};0,-\boldsymbol{e}_z) = -[g/(4\pi)][\boldsymbol{r}/|\boldsymbol{r}|^3], \quad (22.2.119)$$

again as expected.

Verify that $\boldsymbol{A}_s(\boldsymbol{r};0,-\boldsymbol{e}_z)$ and $\boldsymbol{A}_s(\boldsymbol{r};0,\boldsymbol{e}_z)$ are related by a gauge transformation,

$$\boldsymbol{A}_s(\boldsymbol{r};0,-\boldsymbol{e}_z)=\boldsymbol{A}_s(\boldsymbol{r};0,\boldsymbol{e}_z)+\nabla\chi \tag{22.2.120}$$

with

$$\chi=-[g/(2\pi)]\phi. \tag{22.2.121}$$

Form the fully infinite string vector potential $\boldsymbol{A}_{2s}(\boldsymbol{r};0,\boldsymbol{e}_z)$ using (2.75). Show that $\boldsymbol{A}_{2s}(\boldsymbol{r};0,\boldsymbol{e}_z)$ also has only a ϕ component given by

$$\begin{aligned}A_\phi^{2s}(\boldsymbol{r};0,\boldsymbol{e}_z) &= (1/2)A_\phi^{s}(\boldsymbol{r};0,\boldsymbol{e}_z)+(1/2)A_\phi^{s}(\boldsymbol{r};0,-\boldsymbol{e}_z)\\ &= [g/(4\pi)](1/r)(\cot\theta).\end{aligned} \tag{22.2.122}$$

Verify that $A_\phi^{2s}(\boldsymbol{r};0,\boldsymbol{e}_z)$ is singular everywhere on the z axis. Verify by explicit calculation that

$$\nabla\times\boldsymbol{A}_{2s}(\boldsymbol{r};0,\boldsymbol{e}_z)=-[g/(4\pi)][\boldsymbol{r}/|\boldsymbol{r}|^3], \tag{22.2.123}$$

also as expected.

According to Subsection 2.2, a magnetic Dirac string can be viewed as an infinitesimally thin solenoid. In the case that the string is straight, one can assign a definite vector to the string that points in the direction of current flow. Show that for a string directed along the z axis, as is the case for this exercise, the current is in the + (or perhaps −) $\boldsymbol{e}_\phi$ direction, which is the same direction as the associated vector potential $\boldsymbol{A}$.

22.2.16. Exercise on the singularity structure of the vector potential for a straight half-infinite Dirac string.

22.2.17. Let $\boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m})$ and $\boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m}')$ be equal strength monopole vector potentials produced by straight-line strings both originating at $\boldsymbol{r}_A$ but extending to infinity in the directions $\boldsymbol{m}$ and $\boldsymbol{m}'$. See (2.63). Show that both produce the same magnetic field (2.55) at points off the strings. Show that these vector potentials are related by a gauge transformation.

22.2.18. Verify (2.72) and (2.73).

22.2.19. Show from (2.79) that

$$|\boldsymbol{A}_{2s}(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m})|=\frac{|[g/(4\pi)][\boldsymbol{m}\cdot(\boldsymbol{r}-\boldsymbol{r}_A)]|}{|\boldsymbol{r}-\boldsymbol{r}_A||\boldsymbol{m}\times(\boldsymbol{r}-\boldsymbol{r}_A)|}. \tag{22.2.124}$$

Verify that $\boldsymbol{A}_{2s}(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m})$ is singular on, and only on, the line (2.76).

22.2.20. Suppose a vector field $\boldsymbol{F}(\boldsymbol{r})$ is specified in some volume V. Surround this volume by a *thin shell* Σ. Extend $\boldsymbol{F}(\boldsymbol{r})$ to all of space by requiring that it vanish outside Σ and go to zero smoothly within Σ. That is, on the boundary of V, which is the inner surface of Σ, $\boldsymbol{F}$ may have finite values; but within Σ it goes smoothly to zero so that it vanishes on the outer surface of Σ and beyond. It is a standard result in analysis that this can be done in such a way that $\boldsymbol{F}(\boldsymbol{r})$ will have as many derivatives as desired in Σ. Find formulas for $\phi(\boldsymbol{r})$ and $\boldsymbol{A}(\boldsymbol{r})$ in this case. Now let the shell shrink to zero thickness while keeping V unchanged so that Σ becomes the surface S. Show that the relations (2.80), (2.83), and (2.84) continue to give $\boldsymbol{F}(\boldsymbol{r})$ for $\boldsymbol{r}\in V$, and give $\boldsymbol{F}(\boldsymbol{r})=0$ for $\boldsymbol{r}\notin V$.

22.2.21. Suppose a vector field $\boldsymbol{F}(\boldsymbol{r})$ is globally defined and falls off at infinity at least as fast as $1/|\boldsymbol{r}|^2$. Show that, when the surface S in (2.83) and (2.84) is taken to infinity, the surface integrals then vanish. Consequently, (2.83) and (2.84) then take the form

$$\phi(\boldsymbol{r}) = [1/(4\pi)] \int d^3\boldsymbol{r}'\ G(\boldsymbol{r}, \boldsymbol{r}')\nabla' \cdot \boldsymbol{F}(\boldsymbol{r}'), \tag{22.2.125}$$

$$\boldsymbol{A}(\boldsymbol{r}) = [1/(4\pi)] \int d^3\boldsymbol{r}'\ G(\boldsymbol{r}, \boldsymbol{r}')\nabla' \times \boldsymbol{F}(\boldsymbol{r}'). \tag{22.2.126}$$

Thus, in view of (2.80), such a vector field is completely specified by a knowledge of its divergence and curl.

Why should this be the case? Suppose that $\boldsymbol{F}(\boldsymbol{r})$ has the Fourier representation

$$\boldsymbol{F}(\boldsymbol{r}) = \int d^3\boldsymbol{k}\ \exp(i\boldsymbol{k} \cdot \boldsymbol{r})\ \tilde{\boldsymbol{F}}(\boldsymbol{k}). \tag{22.2.127}$$

Such a representation is possible in any number of dimensions, and its existence is a consequence of the completeness of the unitary representations of the translation part of the Euclidean group. Show that there are the relations

$$\boldsymbol{\nabla} \cdot \boldsymbol{F}(\boldsymbol{r}) = i \int d^3\boldsymbol{k}\ \exp(i\boldsymbol{k} \cdot \boldsymbol{r})\ \boldsymbol{k} \cdot \tilde{\boldsymbol{F}}(\boldsymbol{k}), \tag{22.2.128}$$

$$\boldsymbol{\nabla} \times \boldsymbol{F}(\boldsymbol{r}) = i \int d^3\boldsymbol{k}\ \exp(i\boldsymbol{k} \cdot \boldsymbol{r})\ \boldsymbol{k} \times \tilde{\boldsymbol{F}}(\boldsymbol{k}). \tag{22.2.129}$$

Consequently, if the functions $\boldsymbol{\nabla} \cdot \boldsymbol{F}(\boldsymbol{r})$ and $\boldsymbol{\nabla} \times \boldsymbol{F}(\boldsymbol{r})$ are assumed known, then, by the Fourier inversion theorem, the functions $\boldsymbol{k} \cdot \tilde{\boldsymbol{F}}(\boldsymbol{k})$ and $\boldsymbol{k} \times \tilde{\boldsymbol{F}}(\boldsymbol{k})$ are also known. Recall the vector identity

$$\boldsymbol{a} \times (\boldsymbol{b} \times \boldsymbol{c}) = \boldsymbol{b}\ (\boldsymbol{a} \cdot \boldsymbol{c}) - \boldsymbol{c}\ (\boldsymbol{a} \cdot \boldsymbol{b}). \tag{22.2.130}$$

Use this identity to show that

$$\boldsymbol{k} \times (\boldsymbol{k} \times \tilde{\boldsymbol{F}}) = \boldsymbol{k}\ (\boldsymbol{k} \cdot \tilde{\boldsymbol{F}}) - \tilde{\boldsymbol{F}}\ (\boldsymbol{k} \cdot \boldsymbol{k}), \tag{22.2.131}$$

and therefore

$$\tilde{\boldsymbol{F}} = [1/(\boldsymbol{k} \cdot \boldsymbol{k})][\boldsymbol{k}(\boldsymbol{k} \cdot \tilde{\boldsymbol{F}})] - [1/(\boldsymbol{k} \cdot \boldsymbol{k})][\boldsymbol{k} \times (\boldsymbol{k} \times \tilde{\boldsymbol{F}})]. \tag{22.2.132}$$

Thus, the function $\tilde{\boldsymbol{F}}(\boldsymbol{k})$ is known if the functions $\boldsymbol{k} \cdot \tilde{\boldsymbol{F}}(\boldsymbol{k})$ and $\boldsymbol{k} \times \tilde{\boldsymbol{F}}(\boldsymbol{k})$ are known. Correspondingly, the function $\boldsymbol{F}(\boldsymbol{r})$ is determined if the functions $\boldsymbol{\nabla} \cdot \boldsymbol{F}(\boldsymbol{r})$ and $\boldsymbol{\nabla} \times \boldsymbol{F}(\boldsymbol{r})$ are assumed known. Finally, we note that the identity (2.130) may be viewed as a Lie algebraic relation for the cross-product Lie algebra. See Section 3.7.4. From Exercise 3.7.31 we know that the cross-product Lie algebra is equivalent to $so(3)$, and therefore (2.130) is also a property of $so(3)$. Finally, $so(3)$ is a subalgebra of the Lie algebra of the three-dimensional Euclidean group. Thus, the fact that a vector field in three dimensions is specified, if its divergence and curl are known, is a consequence of the properties of the three-dimensional Euclidean group.

22.3 Construction of Kernels G^n and G^t

22.3.1 Background

Let us apply the results of the previous section to the case of a magnetic field $\boldsymbol{B}(\boldsymbol{r})$ in a volume V under the assumption that there are no sources in V. See (1.1) and (1.2). As stated earlier, this would be the case of interest for charged particles propagating through an evacuated beam pipe. In this circumstance we may use (2.80), (2.87), and (2.88) to write

$$\boldsymbol{B}(\boldsymbol{r}) = -\nabla\phi^n(\boldsymbol{r}) + \nabla \times \boldsymbol{A}^t(\boldsymbol{r}) \text{ for } \boldsymbol{r} \in V \tag{22.3.1}$$

with

$$\phi^n(\boldsymbol{r}) = -[1/(4\pi)]\int_S dS'\, \boldsymbol{n}' \cdot \boldsymbol{B}(\boldsymbol{r}')G(\boldsymbol{r},\boldsymbol{r}'), \tag{22.3.2}$$

$$\boldsymbol{A}^t(\boldsymbol{r}) = -[1/(4\pi)]\int_S dS'\, [\boldsymbol{n}' \times \boldsymbol{B}(\boldsymbol{r}')]G(\boldsymbol{r},\boldsymbol{r}'). \tag{22.3.3}$$

Here, as before, the superscripts n and t denote *normal* and *tangential* since the quantities so denoted involve normal and tangential components of $\boldsymbol{B}$.

The relations (3.1) through (3.3) could be employed if one wished to integrate Newton's equations of motion, and also find Taylor maps based on these equations, for all that would then be required is the magnetic field $\boldsymbol{B}(\boldsymbol{r})$. See, for example, the equations of motion (1.6.68) and (1.6.69), or (1.6.135) through (1.6.138) and (1.6.145) through (1.6.147). However, if one wishes instead to employ a Hamiltonian formulation in order to reap the benefits of symplectic symmetry, then it is necessary to have the magnetic field specified *entirely* in terms of a vector potential rather than in terms of both a scalar and vector potential as in (3.1). What we need is a vector potential $\boldsymbol{A}^n(\boldsymbol{r})$ such that

$$\nabla \times \boldsymbol{A}^n(\boldsymbol{r}) = -\nabla\phi^n(\boldsymbol{r}). \tag{22.3.4}$$

Then, with the definition

$$\boldsymbol{A}(\boldsymbol{r}) = \boldsymbol{A}^n(\boldsymbol{r}) + \boldsymbol{A}^t(\boldsymbol{r}), \tag{22.3.5}$$

there would be the result

$$\boldsymbol{B}(\boldsymbol{r}) = \nabla \times \boldsymbol{A}(\boldsymbol{r}). \tag{22.3.6}$$

The construction of an $\boldsymbol{A}^n(\boldsymbol{r})$ that satisfies (3.4) can be accomplished with the aid of the Dirac monopole vector potential. Inspection of $\phi^n(\boldsymbol{r})$, as given by (3.2), shows that it appears to arise from a distribution of magnetic monopoles described by a magnetic charge surface density spread over the surface S. Therefore, it should be possible to find an equivalent vector potential based on the vector potential for a magnetic monopole.

22.3.2 Construction of G^n Using Half-Infinite String Monopoles

Let us make this idea precise. To do so, for simplicity, will use half-infinite string Dirac monopoles. (Fully infinite string Dirac monopoles can also be used. See Exercise 3.2.) Define $\boldsymbol{B}^n$ by the rule

$$\boldsymbol{B}^n = -\nabla\phi^n \tag{22.3.7}$$

so that the $\boldsymbol{A}^n$ that we seek satisfies

$$\nabla \times \boldsymbol{A}^n = \boldsymbol{B}^n. \tag{22.3.8}$$

Combining (3.2) and (3.7) gives the result

$$\boldsymbol{B}^n(\boldsymbol{r}) = [1/(4\pi)] \int_S dS' \; \boldsymbol{n}' \cdot \boldsymbol{B}(\boldsymbol{r}')\nabla G(\boldsymbol{r}, \boldsymbol{r}'). \tag{22.3.9}$$

From (2.89) we know that

$$\nabla G(\boldsymbol{r}, \boldsymbol{r}') = -(\boldsymbol{r} - \boldsymbol{r}')/|\boldsymbol{r} - \boldsymbol{r}'|^3. \tag{22.3.10}$$

But, from (2.72), we also have the relation

$$(4\pi/g)\nabla \times \boldsymbol{A}_s(\boldsymbol{r}; \boldsymbol{r}', \boldsymbol{m}') = -[(\boldsymbol{r} - \boldsymbol{r}')/|\boldsymbol{r} - \boldsymbol{r}'|^3]. \tag{22.3.11}$$

Define a quantity $\boldsymbol{K}(\boldsymbol{r}; \boldsymbol{r}', \boldsymbol{m}')$ by the rule

$$\begin{aligned} \boldsymbol{K}(\boldsymbol{r}; \boldsymbol{r}', \boldsymbol{m}') &= (4\pi/g)\boldsymbol{A}_s(\boldsymbol{r}; \boldsymbol{r}', \boldsymbol{m}') \\ &= [\boldsymbol{m}' \times (\boldsymbol{r} - \boldsymbol{r}')]/\{|\boldsymbol{r} - \boldsymbol{r}'|[|\boldsymbol{r} - \boldsymbol{r}'| - \boldsymbol{m}' \cdot (\boldsymbol{r} - \boldsymbol{r}')]\}. \end{aligned} \tag{22.3.12}$$

See (2.63). In view of (3.10) through (3.12), we have established the key relation

$$\nabla G(\boldsymbol{r}, \boldsymbol{r}') = \nabla \times \boldsymbol{K}(\boldsymbol{r}; \boldsymbol{r}', \boldsymbol{m}'). \tag{22.3.13}$$

See Exercise 2.15 for a specific instance of this relation.

We are almost done. Insertion of (3.13) into (3.9) gives the result

$$\begin{aligned} \boldsymbol{B}^n &= [1/(4\pi)] \int_S dS' \; \boldsymbol{n}' \cdot \boldsymbol{B}(\boldsymbol{r}')\nabla \times \boldsymbol{K}(\boldsymbol{r}; \boldsymbol{r}', \boldsymbol{m}') \\ &= [1/(4\pi)]\nabla \times \int_S dS' \; \boldsymbol{n}' \cdot \boldsymbol{B}(\boldsymbol{r}')\boldsymbol{K}(\boldsymbol{r}; \boldsymbol{r}', \boldsymbol{m}'). \end{aligned} \tag{22.3.14}$$

Comparison of (3.8) and (3.14) shows that we may make the definition

$$\boldsymbol{A}^n(\boldsymbol{r}) = \boldsymbol{A}^{n1s}(\boldsymbol{r}) \tag{22.3.15}$$

with

$$\boldsymbol{A}^{n1s}(\boldsymbol{r}) = [1/(4\pi)] \int_S dS' \; \boldsymbol{n}' \cdot \boldsymbol{B}(\boldsymbol{r}')\boldsymbol{K}(\boldsymbol{r}; \boldsymbol{r}', \boldsymbol{m}'). \tag{22.3.16}$$

Here we have used the superscript $n1s$ to indicate that the vector potential for *one* half-infinite Dirac *string* has been employed. Finally, we make the definitions

$$B_n(\boldsymbol{r}') = \boldsymbol{n}' \cdot \boldsymbol{B}(\boldsymbol{r}') \tag{22.3.17}$$

and

$$\begin{aligned} \boldsymbol{G}^{n1s}(\boldsymbol{r}; \boldsymbol{r}', \boldsymbol{m}') &= [1/(4\pi)]\boldsymbol{K}(\boldsymbol{r}; \boldsymbol{r}', \boldsymbol{m}') \\ &= \{\boldsymbol{m}'(\boldsymbol{r}') \times (\boldsymbol{r} - \boldsymbol{r}')\}/\{4\pi|\boldsymbol{r} - \boldsymbol{r}'|[|\boldsymbol{r} - \boldsymbol{r}'| - \boldsymbol{m}'(\boldsymbol{r}') \cdot (\boldsymbol{r} - \boldsymbol{r}')]\}. \end{aligned} \tag{22.3.18}$$

Here $\boldsymbol{n}'(\boldsymbol{r}')$ is the outward normal to S at the point $\boldsymbol{r}'$. With these definitions we have the result

$$\boldsymbol{A}^{n1s}(\boldsymbol{r}) = \int_S dS'\ B_n(\boldsymbol{r}')\boldsymbol{G}^{n1s}(\boldsymbol{r};\boldsymbol{r}',\boldsymbol{m}'). \tag{22.3.19}$$

Together (3.15) and (3.17) through (3.19) provide a realization of the relation (1.7).

In evaluating the integral (3.19) it necessary to specify $\boldsymbol{m}'(\boldsymbol{r}')$, the direction of the straight half-infinite Dirac string, as $\boldsymbol{r}'$ varies over S. There is considerable freedom in doing so, and different choices simply result in different gauges for $\boldsymbol{A}^{n1s}(\boldsymbol{r})$. There is only one major consideration. No string should intersect the volume V because it is desirable that $\boldsymbol{A}^{n1s}(\boldsymbol{r})$ be analytic for $\boldsymbol{r} \in V$. For many geometries a convenient choice is to require that $\boldsymbol{m}'(\boldsymbol{r}')$ be normal to and point outward from S,

$$\boldsymbol{m}'(\boldsymbol{r}') = \boldsymbol{n}'(\boldsymbol{r}'). \tag{22.3.20}$$

Other choices may also be convenient and useful.

22.3.3 Discussion

Let $\boldsymbol{A}^n(\boldsymbol{r})$ denote the $\boldsymbol{A}^{n1s}(\boldsymbol{r})$ given by (3.19) and let $\boldsymbol{G}^n(\boldsymbol{r},\boldsymbol{r}')$ denote the $\boldsymbol{G}^{n1s}(\boldsymbol{r};\boldsymbol{r}',\boldsymbol{m}')$ given by (3.18). At this point we can take pleasure in observing that $\boldsymbol{A}^n(\boldsymbol{r})$ and $\boldsymbol{G}^n(\boldsymbol{r},\boldsymbol{r}')$ have several desirable properties: First, as long as the Dirac strings for $\boldsymbol{r}' \in S$ do not intersect V, the functions $\boldsymbol{G}^n(\boldsymbol{r},\boldsymbol{r}')$, for every $\boldsymbol{r}' \in S$, are analytic in $\boldsymbol{r}$ for all $\boldsymbol{r} \in V$. It follows from (3.19), under mild conditions on $B_n(\boldsymbol{r}')$ for $\boldsymbol{r}' \in S$, that $\boldsymbol{A}^n(\boldsymbol{r})$ is analytic in V. Second, since the kernel $\boldsymbol{G}^n(\boldsymbol{r},\boldsymbol{r}')$ is essentially the vector potential for a Dirac magnetic monopole, see (3.12) and (3.18), it has, for $\boldsymbol{r} \in V$, the properties

$$\nabla \cdot [\boldsymbol{G}^n(\boldsymbol{r},\boldsymbol{r}')] = 0, \tag{22.3.21}$$

$$\nabla \times [\nabla \times \boldsymbol{G}^n(\boldsymbol{r},\boldsymbol{r}')] = 0. \tag{22.3.22}$$

See (2.71) and (2.73). It follows from (3.19), again under mild conditions on $B_n(\boldsymbol{r}')$, that $\boldsymbol{A}^n(\boldsymbol{r})$ has these same properties,

$$\nabla \cdot [\boldsymbol{A}^n(\boldsymbol{r})] = 0, \tag{22.3.23}$$

$$\nabla \times [\nabla \times \boldsymbol{A}^n(\boldsymbol{r})] = 0. \tag{22.3.24}$$

In practical applications, the surface values $B_n(\boldsymbol{r}')$ will only be known approximately, and the integrals (3.19) may be evaluated numerically with limited precision. It is comforting to know that, nevertheless, the resulting $\boldsymbol{A}^n(\boldsymbol{r})$ will be analytic in V and will satisfy the relations (3.23) and (3.24) exactly no matter what errors are present in the surface values $B_n(\boldsymbol{r}')$ and no matter how poorly the integrals (3.19) are evaluated. All that matters is that the kernel $\boldsymbol{G}^n$ be evaluated to high precision.

22.3.4 Construction of $\mathbf{G}^t$

What can be said about the properties of $\boldsymbol{A}^t(\boldsymbol{r})$ as given by (3.3)? Just as is the case for $\boldsymbol{A}^n(\boldsymbol{r})$, we would like $\boldsymbol{A}^t(\boldsymbol{r})$ to be analytic in V and to satisfy properties analogous to (3.23) and (3.24). That is, we desire the relations

$$\nabla \cdot [\boldsymbol{A}^t(\boldsymbol{r})] = 0, \tag{22.3.25}$$

$$\nabla \times [\nabla \times \boldsymbol{A}^t(\boldsymbol{r})] = 0, \tag{22.3.26}$$

and we would like to have them hold no matter how poorly the integral (3.3) is evaluated. As the expression (3.3) for $\boldsymbol{A}^t(\boldsymbol{r})$ stands, this is not the case. However, we can transform (3.3) into a form that meets all our hopes.

Since, by assumption, $\boldsymbol{B}(\boldsymbol{r}')$ is curl free for $\boldsymbol{r}' \in V$, there exists a scalar potential $\psi(\boldsymbol{r}')$ such that

$$\boldsymbol{B}(\boldsymbol{r}') = +\nabla'\psi(\boldsymbol{r}'). \tag{22.3.27}$$

[Note, by convention, we have used a minus sign in (2.3) and a plus sign in (3.27). See also (15.2.1) and (15.2.6).] Consequently, (3.3) can be rewritten in the form

$$\boldsymbol{A}^t(\boldsymbol{r}) = -[1/(4\pi)] \int_S dS' \, [\boldsymbol{n}' \times \nabla'\psi(\boldsymbol{r}')]G(\boldsymbol{r}, \boldsymbol{r}'). \tag{22.3.28}$$

[Also note, as observed earlier, that a knowledge of the tangential component of $\nabla'\psi(\boldsymbol{r}')$, which is what is involved in (3.28) and is equivalent to a knowledge of $\psi(\boldsymbol{r}')$ on S, is in turn equivalent to a knowledge of the tangential component of $\boldsymbol{B}(\boldsymbol{r}')$ on S under the assumption that $\boldsymbol{B}(\boldsymbol{r}')$ is curl free.] Next observe that there is the identity

$$[\nabla'\psi(\boldsymbol{r}')]G(\boldsymbol{r}, \boldsymbol{r}') = \nabla'[\psi(\boldsymbol{r}')G(\boldsymbol{r}, \boldsymbol{r}')] - \psi(\boldsymbol{r}')\nabla'G(\boldsymbol{r}, \boldsymbol{r}'). \tag{22.3.29}$$

Therefore (3.28) can also be written in the form

$$\begin{aligned} \boldsymbol{A}^t(\boldsymbol{r}) \;\; &= \;\; -[1/(4\pi)] \int_S dS' \, \{\boldsymbol{n}' \times \nabla'[\psi(\boldsymbol{r}')G(\boldsymbol{r}, \boldsymbol{r}')]\} \\ &\quad\; +[1/(4\pi)] \int_S dS' \, \{\boldsymbol{n}' \times [\psi(\boldsymbol{r}')\nabla'G(\boldsymbol{r}, \boldsymbol{r}')]\}. \end{aligned} \tag{22.3.30}$$

It can be shown that the first integral on the right side of (3.30) vanishes,

$$-[1/(4\pi)] \int_S dS' \, \{\boldsymbol{n}' \times \nabla'[\psi(\boldsymbol{r}')G(\boldsymbol{r}, \boldsymbol{r}')]\} = 0. \tag{22.3.31}$$

See Exercise 3.1. Moreover, the second integral can be rewritten in the form

$$[1/(4\pi)] \int_S dS' \, \{\boldsymbol{n}' \times [\psi(\boldsymbol{r}')\nabla'G(\boldsymbol{r}, \boldsymbol{r}')]\} = [1/(4\pi)] \int_S dS' \, \psi(\boldsymbol{r}')[\boldsymbol{n}' \times \nabla'G(\boldsymbol{r}, \boldsymbol{r}')]. \tag{22.3.32}$$

Consequently $\boldsymbol{A}^t(\boldsymbol{r})$ can also be written in the form

$$\boldsymbol{A}^t(\boldsymbol{r}) = [1/(4\pi)] \int_S dS' \, \psi(\boldsymbol{r}')[\boldsymbol{n}' \times \nabla'G(\boldsymbol{r}, \boldsymbol{r}')]. \tag{22.3.33}$$

Finally, let $\boldsymbol{G}^t(\boldsymbol{r},\boldsymbol{r'})$ be the kernel

$$\boldsymbol{G}^t(\boldsymbol{r},\boldsymbol{r'}) = [1/(4\pi)][\boldsymbol{n'}(\boldsymbol{r'}) \times \nabla' G(\boldsymbol{r},\boldsymbol{r'})]. \tag{22.3.34}$$

With this definition, $\boldsymbol{A}^t(\boldsymbol{r})$ takes the final form

$$\boldsymbol{A}^t(\boldsymbol{r}) = \int_S dS' \; \psi(\boldsymbol{r'})\boldsymbol{G}^t(\boldsymbol{r},\boldsymbol{r'}). \tag{22.3.35}$$

And working out (3.34) explicitly gives the result

$$\boldsymbol{G}^t(\boldsymbol{r},\boldsymbol{r'}) = [\boldsymbol{n'}(\boldsymbol{r'}) \times (\boldsymbol{r} - \boldsymbol{r'})]/[4\pi|\boldsymbol{r} - \boldsymbol{r'}|^3]. \tag{22.3.36}$$

We have derived the relation (1.8) with $\boldsymbol{G}^t$ given by (3.36).

At this point we should verify that we have achieved our desired goals. First, it is evident from (3.36) that $\boldsymbol{G}^t(\boldsymbol{r},\boldsymbol{r'})$ is analytic in the components of $\boldsymbol{r}$ for $\boldsymbol{r} \in V$ and $\boldsymbol{r'} \in S$. Therefore, from the representation (3.35), we see that, under mild conditions on $\psi(\boldsymbol{r'})$, $\boldsymbol{A}^t(\boldsymbol{r})$ will be analytic in V.

Next, let us compute $\nabla \cdot \boldsymbol{G}^t(\boldsymbol{r},\boldsymbol{r'})$ and $\nabla \times [\nabla \times \boldsymbol{G}^t(\boldsymbol{r},\boldsymbol{r'})]$. We will see that they both vanish for $\boldsymbol{r} \in V$. Recall the vector identity

$$\nabla \cdot (\boldsymbol{C} \times \boldsymbol{D}) = \boldsymbol{D} \cdot (\nabla \times \boldsymbol{C}) - \boldsymbol{C} \cdot (\nabla \times \boldsymbol{D}). \tag{22.3.37}$$

From this identity, (2.89), (3.34), and the fact that the curl of a gradient vanishes, it follows that

$$\begin{aligned}\nabla \cdot \boldsymbol{G}^t(\boldsymbol{r},\boldsymbol{r'}) &= -[1/(4\pi)]\boldsymbol{n'}(\boldsymbol{r'}) \cdot \{\nabla \times [\nabla' G(\boldsymbol{r},\boldsymbol{r'})]\} \\ &= [1/(4\pi)]\boldsymbol{n'}(\boldsymbol{r'}) \cdot \{\nabla \times [\nabla G(\boldsymbol{r},\boldsymbol{r'})]\} = 0. \end{aligned} \tag{22.3.38}$$

Also, it is evident from (2.90) and (3.34) that

$$\begin{aligned}\nabla^2 \boldsymbol{G}^t(\boldsymbol{r},\boldsymbol{r'}) &= [1/(4\pi)][\nabla^2][\boldsymbol{n'}(\boldsymbol{r'}) \times \nabla' G(\boldsymbol{r},\boldsymbol{r'})] \\ &= [1/(4\pi)]\{\boldsymbol{n'}(\boldsymbol{r'}) \times \nabla' [\nabla^2 G(\boldsymbol{r},\boldsymbol{r'})]\} \\ &= 0 \text{ for } \boldsymbol{r} \text{ within } V \text{ and } \boldsymbol{r'} \in S. \end{aligned} \tag{22.3.39}$$

Finally, again invoke the vector identity

$$\nabla \times (\nabla \times \boldsymbol{C}) = \nabla(\nabla \cdot \boldsymbol{C}) - \nabla^2 \boldsymbol{C}. \tag{22.3.40}$$

When applied to $\boldsymbol{G}^t(\boldsymbol{r},\boldsymbol{r'})$, in view of (3.38) and (3.39), it yields the relation

$$\nabla \times [\nabla \times \boldsymbol{G}^t(\boldsymbol{r},\boldsymbol{r'})] = 0 \text{ for } \boldsymbol{r} \text{ within } V \text{ and } \boldsymbol{r'} \in S. \tag{22.3.41}$$

We have seen that the kernel $\boldsymbol{G}^t(\boldsymbol{r},\boldsymbol{r'})$ satisfies the relations (3.38) and (3.41), and note that these relations are analogous to the relations (3.21) and (3.22) for $\boldsymbol{G}^n(\boldsymbol{r},\boldsymbol{r'})$. It follows, by the same reasoning used in the case of $\boldsymbol{G}^n(\boldsymbol{r},\boldsymbol{r'})$ and $\boldsymbol{A}^n(\boldsymbol{r})$, that $\boldsymbol{A}^t(\boldsymbol{r})$ satisfies the relations (3.25) and (3.26), and these relations hold exactly even in the presence of errors in the surface values $\psi(\boldsymbol{r'})$ and no matter how poorly the integrals (3.35) are evaluated. Similar to to the case of $\boldsymbol{G}^n$, all that matters is that the kernel $\boldsymbol{G}^t$ be evaluated to high precision.

22.3.5 Final Discussion

Let us put together what we have learned about analyticity and "exactness". Look at (3.5) and (3.6). Since $\boldsymbol{A}^n(\boldsymbol{r})$ and $\boldsymbol{A}^t(\boldsymbol{r})$ are both analytic in V, $\boldsymbol{A}(\boldsymbol{r})$ will be analytic in V. And since (3.23) through (3.26) hold, analogous results will hold for $\boldsymbol{A}(\boldsymbol{r})$,

$$\nabla \cdot [\boldsymbol{A}(\boldsymbol{r})] = 0, \tag{22.3.42}$$

$$\nabla \times [\nabla \times \boldsymbol{A}(\boldsymbol{r})] = 0. \tag{22.3.43}$$

Moreover, analyticity and the relations (3.42) and (3.43) will still hold exactly even in the presence of errors in the surface values B_n and ψ, and no matter how poorly the relevant integrals are evaluated. Finally, in view of (3.6), the Maxwell equation

$$\nabla \cdot \boldsymbol{B} = 0 \tag{22.3.44}$$

will be satisfied exactly. And, in view of (3.6) and (3.43), the second Maxwell equation

$$\nabla \times \boldsymbol{B} = 0 \tag{22.3.45}$$

will also be satisfied exactly.

Exercises

22.3.1. The purpose of this exercise is to verify the relation (3.31).

22.3.2. Subsection 3.2 described the construction of the kernel we called $\boldsymbol{G}^{\mathrm{n1s}}$ using the vector potential for a half-infinite string Dirac monopole. Another possibility is to use the vector potential for fully infinite string (two string) Dirac monopole to construct an analogous kernel we will call $\boldsymbol{G}^{\mathrm{n2s}}$. The purpose of this exercise is of explore that possibility.

The vector potential $\boldsymbol{A}_{2s}(\boldsymbol{r};\boldsymbol{r}_A,\boldsymbol{m})$ given by (2.75) also produces the monopole field (2.55) so that there is the relation

$$(4\pi/g)\nabla \times \boldsymbol{A}_{2s}(\boldsymbol{r};\boldsymbol{r}',\boldsymbol{m}') = -[(\boldsymbol{r}-\boldsymbol{r}')/|\boldsymbol{r}-\boldsymbol{r}'|^3]. \tag{22.3.46}$$

Now define a quantity $\boldsymbol{K}(\boldsymbol{r};\boldsymbol{r}',\boldsymbol{m}')$ by the rule

$$\begin{aligned}\boldsymbol{K}(\boldsymbol{r};\boldsymbol{r}',\boldsymbol{m}') &= (4\pi/g)\boldsymbol{A}_{2s}(\boldsymbol{r};\boldsymbol{r}',\boldsymbol{m}') \\ &= \frac{[\boldsymbol{m}\times(\boldsymbol{r}-\boldsymbol{r}_A)][\boldsymbol{m}\cdot(\boldsymbol{r}-\boldsymbol{r}_A)]}{|\boldsymbol{r}-\boldsymbol{r}_A||\boldsymbol{m}\times(\boldsymbol{r}-\boldsymbol{r}_A)|^2}.\end{aligned} \tag{22.3.47}$$

See (2.79). In view of (3.10), (3.21), and (3.22), we have also established the key relation

$$\nabla G(\boldsymbol{r},\boldsymbol{r}') = \nabla \times \boldsymbol{K}(\boldsymbol{r};\boldsymbol{r}',\boldsymbol{m}') \tag{22.3.48}$$

with $\boldsymbol{K}$ now given by (3.22). Also see Exercise 2.15 for a specific instance of this relation.

Next, insertion of (3.23) into (3.9) gives the result

$$\begin{aligned}\boldsymbol{B}^n &= [1/(4\pi)]\int_S dS'\ \boldsymbol{n}'\cdot\boldsymbol{B}(\boldsymbol{r}')\nabla\times\boldsymbol{K}(\boldsymbol{r};\boldsymbol{r}',\boldsymbol{m}')\\ &= [1/(4\pi)]\nabla\times\int_S dS'\ \boldsymbol{n}'\cdot\boldsymbol{B}(\boldsymbol{r}')\boldsymbol{K}(\boldsymbol{r};\boldsymbol{r}',\boldsymbol{m}').\end{aligned} \tag{22.3.49}$$

Comparison of (3.8) and (3.24) shows that we may also make the definition

$$\boldsymbol{A}^n(\boldsymbol{r}) = \boldsymbol{A}^{n2s}(\boldsymbol{r}) \tag{22.3.50}$$

with

$$\boldsymbol{A}^{n2s}(\boldsymbol{r}) = [1/(4\pi)]\int_S dS'\ \boldsymbol{n}'\cdot\boldsymbol{B}(\boldsymbol{r}')\boldsymbol{K}(\boldsymbol{r};\boldsymbol{r}',\boldsymbol{m}'). \tag{22.3.51}$$

Here we have used the superscript $n2s$ to indicate that the vector potential for a fully infinite (*2*-sided) Dirac *string* has been employed. Finally we may write (3.26) in the form

$$\boldsymbol{A}^{n2s}(\boldsymbol{r}) = \int_S dS'\ B_n(\boldsymbol{r}')\boldsymbol{G}^{n2s}(\boldsymbol{r};\boldsymbol{r}',\boldsymbol{m}') \tag{22.3.52}$$

where (3.17) is again employed and $\boldsymbol{G}^{n2s}(\boldsymbol{r};\boldsymbol{r}',\boldsymbol{m}')$ is the kernel

$$\begin{aligned}\boldsymbol{G}^{n2s}(\boldsymbol{r};\boldsymbol{r}',\boldsymbol{m}') &= [1/(4\pi)]\boldsymbol{K}(\boldsymbol{r};\boldsymbol{r}',\boldsymbol{m}')\\ &= \frac{[\boldsymbol{m}'\times(\boldsymbol{r}-\boldsymbol{r}_A)][\boldsymbol{m}'\cdot(\boldsymbol{r}-\boldsymbol{r}_A)]}{4\pi|\boldsymbol{r}-\boldsymbol{r}_A||\boldsymbol{m}'\times(\boldsymbol{r}-\boldsymbol{r}_A)|^2}.\end{aligned} \tag{22.3.53}$$

Together (3.25), (3.27), and (3.28) provide another realization of the relation (1.7). Show that this fully infinite Dirac string kernel obeys relations analogous to (3.21) and (3,22) and therefore the relations analogous to (3.23) and (3.24) are also satisfied.

In evaluating the integral (3.28) it again necessary to specify $\boldsymbol{m}'(\boldsymbol{r}')$, now the direction of the straight fully infinite Dirac string, as $\boldsymbol{r}'$ varies over S. As before, there is considerable freedom in doing so, and different choices simply result in different gauges for $\boldsymbol{A}^{n1s}(\boldsymbol{r})$. The major considerations are again that no string intersect the volume V and that the vector potential fall off rapidly in fringe-field regions. We also note that one may use $\boldsymbol{A}^{n1s}(\boldsymbol{r})$ for some parts of S and $\boldsymbol{A}^{n2s}(\boldsymbol{r})$ for other parts.

22.3.3. At the beginning of this section it was mentioned that (3.1) through (3.3) could be used to integrate Newton's equations of motion in terms of $\boldsymbol{B}(\boldsymbol{r})$. However the $\boldsymbol{B}(\boldsymbol{r})$ obtained using (3.1) is not guaranteed to satisfy the Maxwell equations if there are errors in surface values and/or the integrals are not evaluated accurately. Verify that, in this regard, there is no difficulty in the use of (3.2) by showing that it is guaranteed to satisfy

$$\nabla^2\phi^n(\boldsymbol{r}) = 0, \tag{22.3.54}$$

and therefore (3.43) is satisfied. Show that if (3.3) is replaced by (3.32), then (3.44) is also guaranteed.

22.3.4. Suppose $\boldsymbol{B}(\boldsymbol{r})$ is source free in a volume V bounded by a surface S, as in (1.1) and (1.2), and suppose $B_n(\boldsymbol{r}')$ and $\psi(\boldsymbol{r}')$ are known on S. The aim of this exercise is to compute $\boldsymbol{B}(\boldsymbol{r})$ in terms of $B_n(\boldsymbol{r}')$ and $\psi(\boldsymbol{r}')$ using the representation given by (1.3), (1.6) through (1.8), (1.10), and (1.11). Verify that

$$\nabla \times \boldsymbol{G}^n(\mathbf{r},\mathbf{r}') = [1/(4\pi)]\nabla G(\mathbf{r},\mathbf{r}') \tag{22.3.55}$$

from which it follows that

$$\begin{aligned}\boldsymbol{B}^n(\boldsymbol{r}) &= \nabla \times \boldsymbol{A}^n(\boldsymbol{r}) = \int_S dS'\ B_n(\boldsymbol{r}')\nabla \times \boldsymbol{G}^n(\boldsymbol{r},\boldsymbol{r}') \\ &= [1/(4\pi)]\int_S dS'\ B_n(\boldsymbol{r}')\nabla G(\mathbf{r},\mathbf{r}') \\ &= -[1/(4\pi)]\int_S dS'\ B_n(\boldsymbol{r}')(\mathbf{r}-\mathbf{r}')/|\mathbf{r}-\mathbf{r}'|^3,\end{aligned} \tag{22.3.56}$$

in accord with (3.9). Recall the vector identity

$$\nabla \times (\boldsymbol{C}\times\boldsymbol{D}) = (\boldsymbol{D}\cdot\nabla)\boldsymbol{C} + \boldsymbol{C}(\nabla\cdot\boldsymbol{D}) - (\boldsymbol{C}\cdot\nabla)\boldsymbol{D} - \boldsymbol{D}(\nabla\cdot\boldsymbol{C}). \tag{22.3.57}$$

Using (3.31) and (3.48), show that

$$\nabla \times \boldsymbol{G}^t(\boldsymbol{r},\boldsymbol{r}') = -[1/(4\pi)]\boldsymbol{n}'(\boldsymbol{r}')/|\boldsymbol{r}-\boldsymbol{r}'|^3 + [3/(4\pi)][\boldsymbol{n}'(\boldsymbol{r}')\cdot(\boldsymbol{r}-\boldsymbol{r}')](\boldsymbol{r}-\boldsymbol{r}')/|\boldsymbol{r}-\boldsymbol{r}'|^5, \tag{22.3.58}$$

from which it follows that

$$\begin{aligned}\boldsymbol{B}^t(\boldsymbol{r}) &= \nabla \times \boldsymbol{A}^t(\boldsymbol{r}) = \int_S dS'\ \psi(\boldsymbol{r}')\nabla \times \boldsymbol{G}^t(\boldsymbol{r},\boldsymbol{r}') \\ &= -[1/(4\pi)]\int_S dS'\ \psi(\boldsymbol{r}')\boldsymbol{n}'(\mathbf{r}')/|\mathbf{r}-\mathbf{r}'|^3 \\ &\quad + [3/(4\pi)]\int_S dS'\ \psi(\boldsymbol{r}')[\boldsymbol{n}'(\boldsymbol{r}')\cdot(\boldsymbol{r}-\boldsymbol{r}')](\boldsymbol{r}-\boldsymbol{r}')/|\boldsymbol{r}-\boldsymbol{r}'|^5.\end{aligned} \tag{22.3.59}$$

Observe that, if we wish, we may define kernels $\boldsymbol{K}^n(\boldsymbol{r},\boldsymbol{r}')$ and $\boldsymbol{K}^t(\boldsymbol{r},\boldsymbol{r}')$ by the rules

$$\begin{aligned}\boldsymbol{K}^n(\boldsymbol{r},\boldsymbol{r}') &= \nabla \times \boldsymbol{G}^n(\mathbf{r},\mathbf{r}') = [1/(4\pi)]\nabla G(\mathbf{r},\mathbf{r}') \\ &= -[1/(4\pi)](\mathbf{r}-\mathbf{r}')/|\mathbf{r}-\mathbf{r}'|^3\end{aligned} \tag{22.3.60}$$

and

$$\begin{aligned}\boldsymbol{K}^t(\boldsymbol{r},\boldsymbol{r}') &= \nabla \times \boldsymbol{G}^t(\boldsymbol{r},\boldsymbol{r}') \\ &= -[1/(4\pi)]\boldsymbol{n}'(\boldsymbol{r}')/|\boldsymbol{r}-\boldsymbol{r}'|^3 + [3/(4\pi)][\boldsymbol{n}'(\boldsymbol{r}')\cdot(\boldsymbol{r}-\boldsymbol{r}')](\boldsymbol{r}-\boldsymbol{r}')/|\boldsymbol{r}-\boldsymbol{r}'|^5.\end{aligned} \tag{22.3.61}$$

With the aid of these definitions, (3.47) and (3.50) take the form

$$\boldsymbol{B}^n(\boldsymbol{r}) = \int_S dS'\ B_n(\boldsymbol{r}')\boldsymbol{K}^n(\boldsymbol{r},\boldsymbol{r}') \tag{22.3.62}$$

and

$$\boldsymbol{B}^t(\boldsymbol{r}) = \int_S dS' \; \psi(\boldsymbol{r}')\boldsymbol{K}^t(\boldsymbol{r}, \boldsymbol{r}'). \tag{22.3.63}$$

Finally, write

$$\boldsymbol{B}(\boldsymbol{r}) = \boldsymbol{B}^n(\boldsymbol{r}) + \boldsymbol{B}^t(\boldsymbol{r}). \tag{22.3.64}$$

Show that, for fixed $\boldsymbol{r}'$, $\boldsymbol{K}^n(\boldsymbol{r}, \boldsymbol{r}')$ falls off as $1/r^2$ for large $r = |\boldsymbol{r}|$ and $\boldsymbol{K}^t(\boldsymbol{r}, \boldsymbol{r}')$ falls off as $1/r^3$.

22.3.5. Show that the Cartesian components of $\boldsymbol{A}(\boldsymbol{r})$, as given by (3.5), (3.17), and (3.32), are harmonic functions,

$$\nabla^2 \boldsymbol{A}(\boldsymbol{r}) = 0. \tag{22.3.65}$$

22.3.6. According to (3.12) and (3.18), $\boldsymbol{G}^{n1s}$ and $\boldsymbol{A}_s$ are proportional. Consequently, Exercise 2.15 provides a description of the direction of the vector $\boldsymbol{G}^n$ and, and hence the associated $\boldsymbol{A}^n$ produced using (1.7). The purpose of this exercise is to determine the direction of $\boldsymbol{G}^t$, and hence the associated $\boldsymbol{A}^t$ produced using (1.8). Consider, for purposes of calculation, a small patch of surface $\Delta S'$ located at the point

$$\boldsymbol{r}' = d\boldsymbol{e}_y \text{ with } d > 0 \tag{22.3.66}$$

and whose normal is given by the relation

$$\boldsymbol{n}'(\boldsymbol{r}') = \boldsymbol{e}_y. \tag{22.3.67}$$

Then, from (3.44), verify that there is the result

$$\boldsymbol{G}^t(\boldsymbol{r}, d\boldsymbol{e}_y) = [\boldsymbol{e}_y \times (\boldsymbol{r} - d\boldsymbol{e}_y)]/[4\pi|\boldsymbol{r} - d\boldsymbol{e}_y|^3]. \tag{22.3.68}$$

Also, verify the relation

$$\boldsymbol{e}_y \times (\boldsymbol{r} - d\boldsymbol{e}_y) = -x\boldsymbol{e}_z + z\boldsymbol{e}_x. \tag{22.3.69}$$

Show, therefore, that in this case, $\boldsymbol{G}^t$ is given by the relation

$$\boldsymbol{G}^t = [1/(4\pi)](-x\boldsymbol{e}_z + z\boldsymbol{e}_x)/[x^2 + (y-d)^2 + z^2]^{3/2}. \tag{22.3.70}$$

Recall that in cylindrical coordinates there is the relation

$$\boldsymbol{r} = x\boldsymbol{e}_x + y\boldsymbol{e}_y + z\boldsymbol{e}_z = \rho\cos\phi\boldsymbol{e}_x + \rho\sin\phi\boldsymbol{e}_y + z\boldsymbol{e}_z. \tag{22.3.71}$$

See (13.2.3) and (13.2.4). Consequently there is the relation

$$\partial\boldsymbol{r}/\partial\phi = -\rho\sin\phi\boldsymbol{e}_x + \rho\cos\phi\boldsymbol{e}_y = -y\boldsymbol{e}_x + x\boldsymbol{e}_y. \tag{22.3.72}$$

We also know that

$$\boldsymbol{e}_\phi = (\partial\boldsymbol{r}/\partial\phi)/|\partial\boldsymbol{r}/\partial\phi| = (-y\boldsymbol{e}_x + x\boldsymbol{e}_y)/\rho. \tag{22.3.73}$$

In the case of cylindrical coordinates the vector $\boldsymbol{e}_\phi$ circles around the z axis. Verify, by geometric analogy, that the vector $\boldsymbol{G}^t$ given by (3.70) circles about the y axis. And, according to (3.67), this axis is the $\boldsymbol{n}'(\boldsymbol{r}')$ axis.

22.4 Expansion of Kernels

22.4.1 Our Goal

22.4.2 Binomial Theorem

Since Newton's discovery we have known the binomial expansion

$$(1+x)^\alpha = \sum_{k=0}^{\infty} \left(\begin{array}{c} \alpha \\ k \end{array} \right) x^k. \tag{22.4.1}$$

Moreover, the binomial coefficients obey the recursion relations

$$\left(\begin{array}{c} \alpha \\ 0 \end{array} \right) = 1, \tag{22.4.2}$$

$$\left(\begin{array}{c} \alpha \\ k+1 \end{array} \right) = [(\alpha - k)/(k+1)] \left(\begin{array}{c} \alpha \\ k \end{array} \right), \tag{22.4.3}$$

and therefore can easily be computed sequentially.

22.4.3 Expansion of $\mathbf{G}^t(\mathbf{r}, \mathbf{r}')$

22.4.4 Expansion of $\mathbf{G}^n(\mathbf{r}, \mathbf{r}')$

Bibliography

Electromagnetism and Magnetic Monopoles

[1] J.D. Jackson, *Classical Electrodynamics*, John Wiley (1999).

[2] J. Reitz and F. Milford, *Foundations of Electromagnetic Theory*, Second Edition, Addison-Wesley (1967).

[3] R. Plonsey and R. Collin, *Principles and Applications of Electromagnetic Fields*, McGraw Hill (1961).

[4] S. Russenschuck, *Field Computation for Accelerator Magnets: Analytical and Numerical Methods for Electromagnetic Design and Optimization*, Wiley (2010).

[5] Ya. Shnir, *Magnetic Monopoles*, Springer (2005).

[6] P. Dirac,"Quantized singularities in the electromagnetic field", *Proceedings of the Royal Society of London* **A133**, pp. 60-72 (1931).

[7] The idea of constructing kernels using Helmholtz's theorem and Dirac's magnetic monopole vector potential is due to Peter Walstrom.

General References

[8] C. Mitchell, "Calculation of Realistic Charged-Particle Transfer Maps", University of Maryland Physics Department Ph.D. Thesis (2007).

Chapter 23

Realistic Transfer Maps for General Curved Beam-Line Elements: Exact Monopole Doublet Results

How do the surface methods for curved elements, described by relations (22.1.3) through (22.1.8) of Section 22.1 with the $\boldsymbol{G}^n$ and $\boldsymbol{G}^t$ found in Section 22.3, work in practice? The purpose of this chapter and the next two is to explore this question.

This chapter finds exact results for the case of a monopole doublet. Chapter 24 finds bent box monopole doublet results. Comparison of the results of these two chapters provides a benchmark for the accuracy of surface methods for curved beam-line elements. Chapter 25 applies surface methods to the case of a realistic storage-ring dipole.

23.1 Magnetic Monopole Doublet Vector Potential

Consider the monopole doublet magnetic field described by Equations (15.8.1) through (15.8.6) and Figures 15.8.1 through 15.8.5 of Section 15.8. In order to set up the Hamiltonian that will describe particle motion in this field, we need a vector potential $\boldsymbol{A}(\boldsymbol{r})$ such that

$$\nabla \times \boldsymbol{A}(\boldsymbol{r}) = \nabla\psi(\boldsymbol{r}) \tag{23.1.1}$$

with ψ given by (15.8.3). For this purpose we will employ the string vector potential given by (22.2.63). The desired vector potential will describe two Dirac magnetic monopoles of opposite sign. The upper, with strength $4\pi g$, will be situated at $\boldsymbol{r}^+ = a\boldsymbol{e}_y$, and will be taken to have a half-infinite string extending from $\boldsymbol{r}^+$ to infinity along the positive y axis. The lower, with strength $-4\pi g$, will be situated at $\boldsymbol{r}^- = -a\boldsymbol{e}_y$, and will be taken to have a half-infinite string extending from $\boldsymbol{r}^-$ to infinity along the negative y axis. See (22.2.63) and Figure 1.1. Thus, $\boldsymbol{A}(\boldsymbol{r})$ will be given by the relation

$$\boldsymbol{A}(\boldsymbol{r}) = \boldsymbol{A}^+(\boldsymbol{r}) + \boldsymbol{A}^-(\boldsymbol{r}) \tag{23.1.2}$$

with

$$\begin{aligned} \boldsymbol{A}^{+}(\boldsymbol{r}) &= -\boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}^{+}\rightarrow +\infty\boldsymbol{e}_y) \\ &= -g[\boldsymbol{e}_y\times(\boldsymbol{r}-a\boldsymbol{e}_y)]/\{|\boldsymbol{r}-a\boldsymbol{e}_y|[|\boldsymbol{r}-a\boldsymbol{e}_y|-\boldsymbol{e}_y\cdot(\boldsymbol{r}-a\boldsymbol{e}_y)]\} \\ &= -g(\boldsymbol{e}_y\times\boldsymbol{r})/\{|\boldsymbol{r}-a\boldsymbol{e}_y|[|\boldsymbol{r}-a\boldsymbol{e}_y|-y+a]\}, \end{aligned} \tag{23.1.3}$$

and

$$\begin{aligned} \boldsymbol{A}^{-}(\boldsymbol{r}) &= -(-1)\boldsymbol{A}_s(\boldsymbol{r};\boldsymbol{r}^{-}\rightarrow -\infty\boldsymbol{e}_y) \\ &= -(-g)[-\boldsymbol{e}_y\times(\boldsymbol{r}+a\boldsymbol{e}_y)]/\{|\boldsymbol{r}+a\boldsymbol{e}_y|[|\boldsymbol{r}+a\boldsymbol{e}_y|+\boldsymbol{e}_y\cdot(\boldsymbol{r}+a\boldsymbol{e}_y)]\} \\ &= -g(\boldsymbol{e}_y\times\boldsymbol{r})/\{|\boldsymbol{r}+a\boldsymbol{e}_y|[|\boldsymbol{r}+a\boldsymbol{e}_y|+y+a]\}. \end{aligned} \tag{23.1.4}$$

Here we have used the notation $\boldsymbol{r}^{+}\rightarrow +\infty\boldsymbol{e}_y$ to denote a string extending from $\boldsymbol{r}^{+}$ to infinity along the positive y axis, and have used the notation $\boldsymbol{r}^{-}\rightarrow -\infty\boldsymbol{e}_y$ to denote a string extending from $\boldsymbol{r}^{-}$ to infinity along the negative y axis. Also, as in Section 15.9.1, we have taken the monopoles to have strengths $\pm 4\pi g$ so as to avoid the appearance of 4π factors in subsequent formulas such as (1.3) and (1.4).

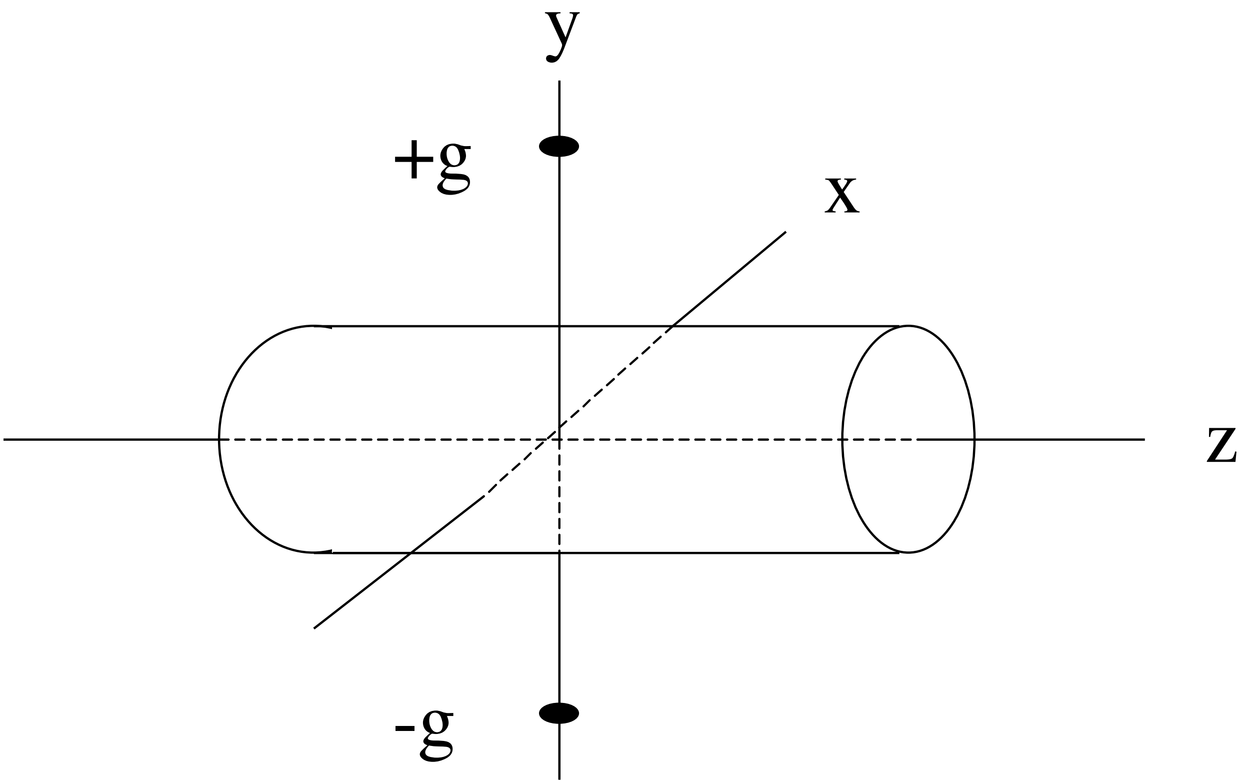


Figure 23.1.1: (Place holder) A monopole doublet consisting of two magnetic monopoles of equal and opposite sign placed on the y axis and centered on the origin. Also shown are half-infinite Dirac strings extending from the $+g$ monopole along the positive y axis and from the $-g$ monopole along the negative y axis.

Note that

$$\boldsymbol{e}_y\times\boldsymbol{r} = -x\boldsymbol{e}_z + z\boldsymbol{e}_x. \tag{23.1.5}$$

Therefore, in terms of components, the relation (1.2) takes the explicit form

$$\begin{aligned} A_x(x,y,z) &= -\frac{gz}{[x^2+(y-a)^2+z^2]^{1/2}\{[x^2+(y-a)^2+z^2]^{1/2}-y+a\}} \\ &\quad -\frac{gz}{[x^2+(y+a)^2+z^2]^{1/2}\{[x^2+(y+a)^2+z^2]^{1/2}+y+a\}}, \end{aligned} \tag{23.1.6}$$

$$A_y(x,y,z)=0, \tag{23.1.7}$$

$$\begin{aligned} A_z(x,y,z) &= +\frac{gx}{[x^2+(y-a)^2+z^2]^{1/2}\{[x^2+(y-a)^2+z^2]^{1/2}-y+a\}} \\ &\quad +\frac{gx}{[x^2+(y+a)^2+z^2]^{1/2}\{[x^2+(y+a)^2+z^2]^{1/2}+y+a\}}. \end{aligned} \tag{23.1.8}$$

Examination of (1.6) reveals that $A_x(x,y,z)$ is even in x and y, and odd in z; and examination of (1.8) shows that $A_z(x,y,z)$ is odd in x and even in y and z.

From (22.1.3) and (1.6) through (1.8), and with some algebraic effort, it can be checked that

$$\begin{aligned} B_x &= \partial_y A_z - \partial_z A_y = \partial_y A_z = \\ &= gx[x^2+(y-a)^2+z^2]^{-3/2} - gx[x^2+(y+a)^2+z^2]^{-3/2}, \end{aligned} \tag{23.1.9}$$

$$\begin{aligned} B_y &= \partial_z A_x - \partial_x A_z \\ &= g(y-a)\{[x^2+(y-a)^2+z^2]^{-3/2} - g(y+a)[x^2+(y+a)^2+z^2]^{-3/2}, \end{aligned} \tag{23.1.10}$$

$$\begin{aligned} B_z &= \partial_x A_y - \partial_y A_x = -\partial_y A_x = \\ &= gz[x^2+(y-a)^2+z^2]^{-3/2} - gz[x^2+(y+a)^2+z^2]^{-3/2}, \end{aligned} \tag{23.1.11}$$

in agreement with (15.9.4) through (15.9.6).

Figure 1.2 displays the quantity $A_x(x,y,z)$ as a function of z along the line $x=y=0$. Here, for convenience in plotting and as done before, we have used the values

$$a = 2.5 \text{ cm} = .025 \text{ m} \tag{23.1.12}$$

and

$$g = 1 \text{ Tesla (cm)}^2 = 1\times 10^{-4} \text{ Tesla m}^2. \tag{23.1.13}$$

Evidently A_x along this line falls off very slowly with increasing $|z|$. Indeed, for large $|z|$, we see from (1.6) that $A_x(x,y,z)$ has the asymptotic behavior

$$A_x(x,y,z) \simeq -2g/z. \tag{23.1.14}$$

Figure 1.3 displays A_z as a function of z along the line given by the conditions $x = -1/2$ cm and $y = 0$. It falls off somewhat more rapidly. From (1.8) we see that, for large $|z|$, it has the asymptotic behavior

$$A_z(x, y, z) \simeq 2gx/z^2. \tag{23.1.15}$$

Neither A_x nor A_z falls off as rapidly as $B_y(0, 0, z)$, which falls off as $1/|z|^3$ for large $|z|$. See Section 15.8.1 and Figure 15.8.3. We also note that if a cylindrical harmonic expansion is employed as in Section 16.3, which involves the use of on-axis gradients, then all components of the associated vector potential fall off as $1/|z|^3$ for large $|z|$. What we are observing is that the asymptotic behavior of the vector potential depends on the choice of gauge. Why not, then, employ a cylindrical harmonic expansion for which the asymptotic behavior of the associated vector potential is optimal? The reason is that we wish to treat cases for which the design orbit is significantly bent so that on-axis expansions are not applicable.

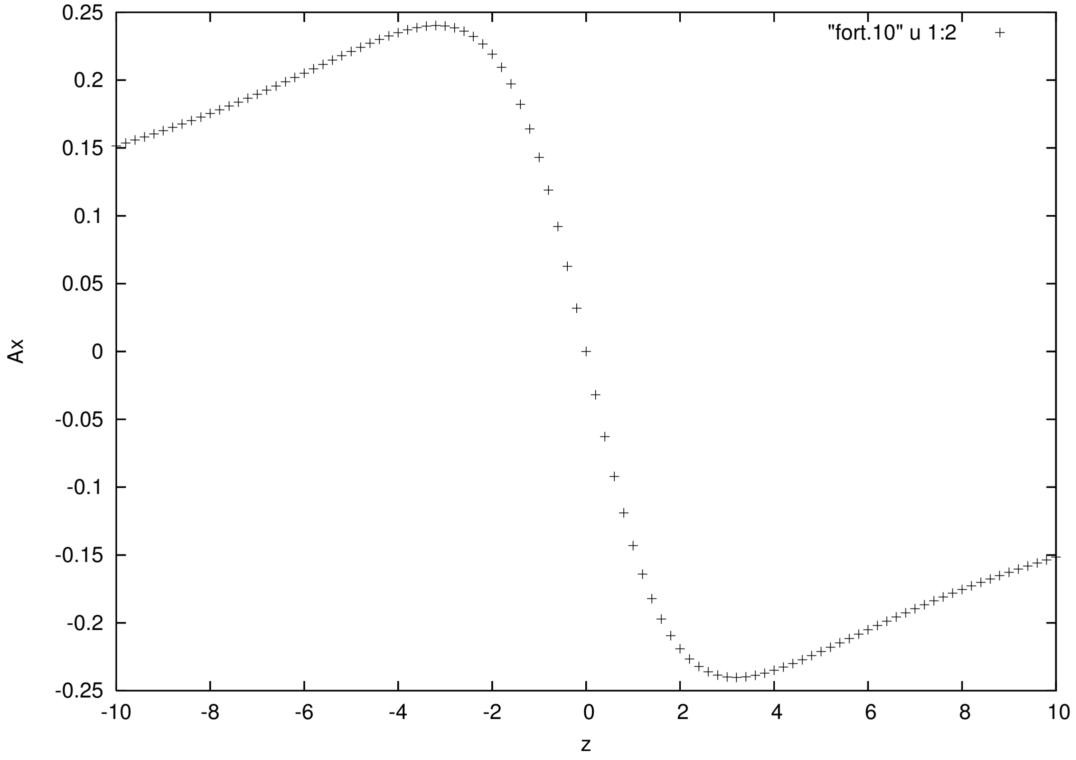


Figure 23.1.2: Behavior of A_x on the line $(0, 0, z)$. The quantity z is in cm.

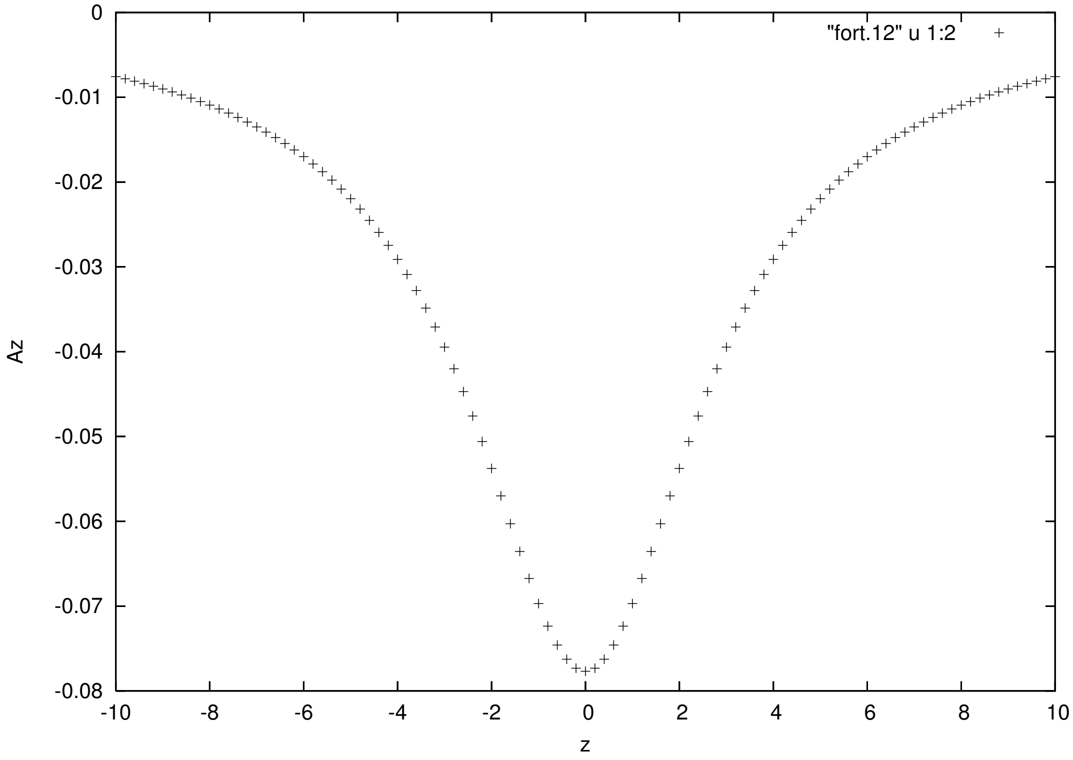


Figure 23.1.3: Behavior of A_z on the line $(-1/2, 0, z)$. The quantity z is in cm.

Exercises

23.1.1. Using (1.6) through (1.8), verify (1.9) through (1.11).

23.2 Selection of Hamiltonian and Scaled Variables

To compute orbits (and maps) it is convenient to use z as the independent variable. In this case, and for the vector potential given by (1.6) through (1.8), the Hamiltonian becomes

$$K = -[p_t^2/c^2 - m^2c^2 - (p_x - qA_x)^2 - p_y^2]^{1/2} - qA_z. \tag{23.2.1}$$

See (1.6.16). Let β and γ be the usual relativistic factors defined by

$$\beta = v/c, \tag{23.2.2}$$

$$\gamma = (1 - \beta^2)^{-1/2} \tag{23.2.3}$$

where v is the particle velocity. Then the magnitude of the mechanical momentum is given by the relation

$$p = \gamma m v = \gamma\beta m c \tag{23.2.4}$$

and the quantity p_t has the value

$$p_t = -(m^2c^4 + p^2c^2)^{1/2} = -\gamma m c^2. \tag{23.2.5}$$

Since K is independent of t, the quantities p_t and p will be constants of motion. Finally, let p^0 be the momentum for the design orbit.

At this point it is useful to introduce dimensionless/scaled variables by the rules

$$\hat{x} = x/\ell, \tag{23.2.6}$$

$$\hat{y} = y/\ell, \tag{23.2.7}$$

$$\tau = ct/\ell, \tag{23.2.8}$$

$$\hat{p}_x = p_x/p^0, \tag{23.2.9}$$

$$\hat{p}_y = p_y/p^0, \tag{23.2.10}$$

$$p_\tau = p_t/(p^0 c). \tag{23.2.11}$$

Here ℓ is a convenient scale length, and is not to be confused with the path length introduced in Exercise 1.7.8.

The dimensionless variables satisfy the Poisson bracket rules

$$[\hat{x}, \hat{p}_x] = [\hat{y}, \hat{p}_y] = [\tau, p_\tau] = 1/(\ell p^0). \tag{23.2.12}$$

From now on we will redefine their Poisson brackets so that conjugate variables again have unity Poisson brackets. This is permissible providing the Hamiltonian K is replaced by a properly scaled new Hamiltonian H given by the relation

$$\begin{aligned} H &= -[1/(\ell p^0)]\{[(p^0 c)^2 p_\tau^2/c^2 - m^2c^2 - (p^0\hat{p}_x - qA_x)^2 - (p^0)^2\hat{p}_y^2]^{1/2} + qA_z\} \\ &= -(1/\ell)\{p_\tau^2 - (mc/p^0)^2 - (\hat{p}_x - \mathcal{A}_x)^2 - \hat{p}_y^2]^{1/2} + \mathcal{A}_z\} \end{aligned} \tag{23.2.13}$$

where

$$\mathcal{A}_x(\hat{x}, \hat{y}, z) = (q/p^0)A_x(\ell\hat{x}, \ell\hat{y}, z), \tag{23.2.14}$$

$$\mathcal{A}_z(\hat{x}, \hat{y}, z) = (q/p^0)A_z(\ell\hat{x}, \ell\hat{y}, z). \tag{23.2.15}$$

(See Appendix D.)

23.3 Design Orbit and Fields

How should we choose a design orbit? We would like it to lie in the $y = 0$ plane, to pass through the origin, and to be symmetric about $z = 0$. How do we know that it is possible for there to be an orbit that lies in the $y = 0$ plane? Let us evaluate (1.9) through (1.11) to find $\boldsymbol{B}$ when $y = 0$. So doing gives the results

$$B_x(x, 0, z) = 0, \tag{23.3.1}$$

$$B_y(x, 0, z) = -2ga\{[x^2 + a^2 + z^2]^{-3/2}, \tag{23.3.2}$$

$$B_z(x, 0, z) = 0. \tag{23.3.3}$$

We see that if a particle is initially in the $y = 0$ plane and moving with a velocity in this plane, then the Lorentz force is also in this plane: thus there is no force acting to accelerate the particle out of this plane and it must remain in this plane. Next observe from (1.6) through (1.8) that $\boldsymbol{A}(\boldsymbol{r})$ vanishes at the origin,

$$\boldsymbol{A}(0, 0, 0) = 0. \tag{23.3.4}$$

Therefore the canonical and mechanical momenta agree at the origin. See (1.5.30). Consequently, and by symmetry, one way to achieve the desired design orbit is to select, for $z = 0$, the initial conditions

$$\hat{x} = \hat{y} = \tau = 0, \tag{23.3.5}$$

$$\hat{p}_x = \hat{p}_y = 0, \tag{23.3.6}$$

and then integrate both backward and forward in z to obtain the complete orbit. Note that for a orbit lying in the $y = 0$ plane the relations

$$\hat{y} = \hat{p}_y = 0 \tag{23.3.7}$$

hold for all z.

What remains is to select the values of p_τ and p^0. From (2.4) we see that for the design orbit there is the relation

$$p^0 = \gamma^0 \beta^0 mc. \tag{23.3.8}$$

From (2.5) we see that the energy on this orbit will be given by

$$p_t^0 = -\gamma^0 mc^2. \tag{23.3.9}$$

Therefore, on this orbit p_τ has the value

$$p_\tau = p_\tau^0 = p_t^0/(p^0 c) = -\gamma^0 mc^2/(\gamma^0 \beta^0 mcc) = -1/\beta^0. \tag{23.3.10}$$

And, with regard to the ingredients in (2.13), we see that

$$(p_\tau^0)^2 - (mc/p^0)^2 = (1/\beta^0)^2 - [1/(\gamma^0\beta^0)^2] = 1. \tag{23.3.11}$$

Therefore, on the design orbit, H becomes

$$H = -(1/\ell)\{[1 - (\hat{p}_x - \mathcal{A}_x)^2 - \hat{p}_y^2]^{1/2} + \mathcal{A}_z\}. \tag{23.3.12}$$

Finally, we should select (by trial and error) the quantity p^0, which now appears only in (2.14) and (2.15), in such a way that, for the specified values of a and g, the design orbit has some desired bend angle ϕ_{bend}. For purposes of illustration, we will require that ϕ_{bend} for an electron be approximately 30°.

Let us work out the spatial equations of motion associated with H as given by (3.12). For convenience we will take the scale length to have the value

$$\ell = 1 \text{ cm}. \tag{23.3.13}$$

We find the results

$$\hat{x}' = \partial H/\partial \hat{p}_x = (\hat{p}_x - \mathcal{A}_x)/[1 - (\hat{p}_x - \mathcal{A}_x)^2 - \hat{p}_y^2]^{1/2}, \tag{23.3.14}$$

$$\hat{y}' = \partial H/\partial \hat{p}_y = \hat{p}_y/[1 - (\hat{p}_x - \mathcal{A}_x)^2 - \hat{p}_y^2]^{1/2}, \tag{23.3.15}$$

$$\hat{p}_x' = -\partial H/\partial \hat{x} = (\partial \mathcal{A}_x/\partial \hat{x})(\hat{p}_x - \mathcal{A}_x)/[1 - (\hat{p}_x - \mathcal{A}_x)^2 - \hat{p}_y^2]^{1/2} + (\partial \mathcal{A}_z/\partial \hat{x}), \tag{23.3.16}$$

$$\hat{p}_y' = -\partial H/\partial \hat{y} = (\partial \mathcal{A}_x/\partial \hat{y})(\hat{p}_x - \mathcal{A}_x)/[1 - (\hat{p}_x - \mathcal{A}_x)^2 - \hat{p}_y^2]^{1/2} + (\partial \mathcal{A}_z/\partial \hat{y}). \tag{23.3.17}$$

Here a prime denotes d/dz.

We have already remarked that for this vector potential A_x and A_z are even in y, and therefore we may write

$$\mathcal{A}_x(\hat{x}, -\hat{y}, z) = \mathcal{A}_z(\hat{x}, \hat{y}, z), \tag{23.3.18}$$

$$\mathcal{A}_z(\hat{x}, -\hat{y}, z) = \mathcal{A}_z(\hat{x}, \hat{y}, z). \tag{23.3.19}$$

See (1.6) through (1.8). It follows that

$$[\partial \mathcal{A}_x(\hat{x}, \hat{y}, z)/\partial \hat{y}]|_{\hat{y}=0} = [\partial \mathcal{A}_z(\hat{x}, \hat{y}, z)/\partial \hat{y}]|_{\hat{y}=0} = 0. \tag{23.3.20}$$

Upon combining the information provided by (3.20) with the $(\hat{y}, \hat{p}_y)$ equations of motion (3.15) and (3.17) we see that there are orbits, one of which which will be the design orbit, that satisfy the conditions (3.7) for all z. Moreover, on these orbits, the $(\hat{x}, \hat{p}_x)$ equations of motion take the form

$$\hat{x}' = (\hat{p}_x - \mathcal{A}_x)/[1 - (\hat{p}_x - \mathcal{A}_x)^2]^{1/2}, \tag{23.3.21}$$

$$\hat{p}_x' = (\partial \mathcal{A}_x/\partial \hat{x})(\hat{p}_x - \mathcal{A}_x)/[1 - (\hat{p}_x - \mathcal{A}_x)^2]^{1/2} + (\partial \mathcal{A}_z/\partial \hat{x}), \tag{23.3.22}$$

and it is only this pair we need integrate. For the record we note that, on the design orbit so that (3.7) holds, there are the relations

$$\mathcal{A}_x|_{\hat{y}=0} = \mathcal{A}_x(\hat{x}, 0, z) = -\frac{(2gq/p^0)z}{(\hat{x}^2 + a^2 + z^2)^{1/2}[(\hat{x}^2 + a^2 + z^2)^{1/2} + a]}, \tag{23.3.23}$$

$$\mathcal{A}_z|_{\hat{y}=0} = \mathcal{A}_z(\hat{x}, 0, z) = +\frac{(2gq/p^0)\hat{x}}{(\hat{x}^2 + a^2 + z^2)^{1/2}[(\hat{x}^2 + a^2 + z^2)^{1/2} + a]}, \tag{23.3.24}$$

$$(\partial \mathcal{A}_x/\partial \hat{x})|_{\hat{y}=0} = +\frac{(2gq/p^0)(\hat{x}z)[2(\hat{x}^2 + a^2 + z^2)^{1/2} + a]}{(\hat{x}^2 + a^2 + z^2)^{3/2}[(\hat{x}^2 + a^2 + z^2)^{1/2} + a]^2}, \tag{23.3.25}$$

$$\begin{aligned}(\partial \mathcal{A}_z/\partial \hat{x})|_{\hat{y}=0} \;=\; & +\frac{(2gq/p^0)}{(\hat{x}^2 + a^2 + z^2)^{1/2}[(\hat{x}^2 + a^2 + z^2)^{1/2} + a]} \\ & -\frac{(2gq/p^0)(\hat{x}^2)[2(\hat{x}^2 + a^2 + z^2)^{1/2} + a]}{(\hat{x}^2 + a^2 + z^2)^{3/2}[(\hat{x}^2 + a^2 + z^2)^{1/2} + a]^2}.\end{aligned} \tag{23.3.26}$$

See (1.6) through (1.8), (2.14), and (2.15). Finally, imposing the initial conditions (3.5) and (3.6) and a suitable value for p^0 yield the design orbit.

Figures 3.1 and 3.2 display, in canonical coordinates, the design orbit that results from integrating the equations of motion (3.21) and (3.22), with the initial conditions (3.5) and (3.6), when the design momentum p^0 is selected to satisfy the relation

$$qg/p^0 = -.3291331 \text{ cm}. \tag{23.3.27}$$

(Recall that $q < 0$ for an electron.) Figure 3.1 shows the spatial part of the design orbit. Figure 3.2 displays the canonical momentum $\hat{p}_x$ on this orbit. Note that the canonical momentum depends on the choice of gauge.

To provide further insight, Figure 3.3 displays the *mechanical* scaled momentum $\hat{p}_x^{\text{mech}}$ related to the canonical momentum by the rule

$$\hat{p}_x^{\text{mech}} = \hat{p}_x - \mathcal{A}_x. \tag{23.3.28}$$

Note that the mechanical momentum does not depend on the choice of gauge. Finally observe that, with the aid of (3.28), the relation (3.21) can be rewritten in the form

$$\hat{x}' = \hat{p}_x^{\text{mech}}/[1 - (\hat{p}_x^{\text{mech}})^2]^{1/2}. \tag{23.3.29}$$

Figure 3.4 displays $\hat{x}'(z)$. We also reiterate that $\hat{y} = 0$ and $\hat{p}_y = 0$ on a design orbit.

On this design orbit there are, for $z = \mp 20$ cm, the end values

$$\hat{x}(\mp 20) = -4.75976218485406, \tag{23.3.30}$$

$$\hat{p}_x(\mp 20) = \pm.?, \tag{23.3.31}$$

$$\hat{p}_x^{\text{mech}}(\mp 20) = \pm.2588190579162489, \tag{23.3.32}$$

$$\hat{x}'(\mp 20) = \pm.2679492066493081. \tag{23.3.33}$$

Correspondingly, we find that over the interval $z \in [-20, 20]$ the bend angle has the value

$$\phi_{\text{bend}} = 30.000001520142693^\circ. \tag{23.3.34}$$

See Exercise 3.2.

For the design orbit the magnetic rigidity has the value

$$\begin{aligned} p^0/|q| &= g/(.3291331 \text{ cm}) = 1 \text{ Tesla (cm)}^2/(.3291331 \text{ cm}) \\ &= 3.0382845116458963 \text{ Tesla cm} \\ &= 3.0382845116458963 \times 10^{-2} \text{ Tesla m}. \end{aligned} \tag{23.3.35}$$

See (1.6.116). Correspondingly, we find the values

$$p^0 = 9.108547817 \text{ MeV/c}, \tag{23.3.36}$$

$$p_t^0 = -9.122870347 \text{ MeV}, \tag{23.3.37}$$

$$\begin{aligned} &\text{kinetic energy} = -p_t^0 - m_e c^2 \\ &= (\gamma^0 - 1) m_e c^2 \\ &= 8.611871287313742 \text{ MeV}, \end{aligned} \tag{23.3.38}$$

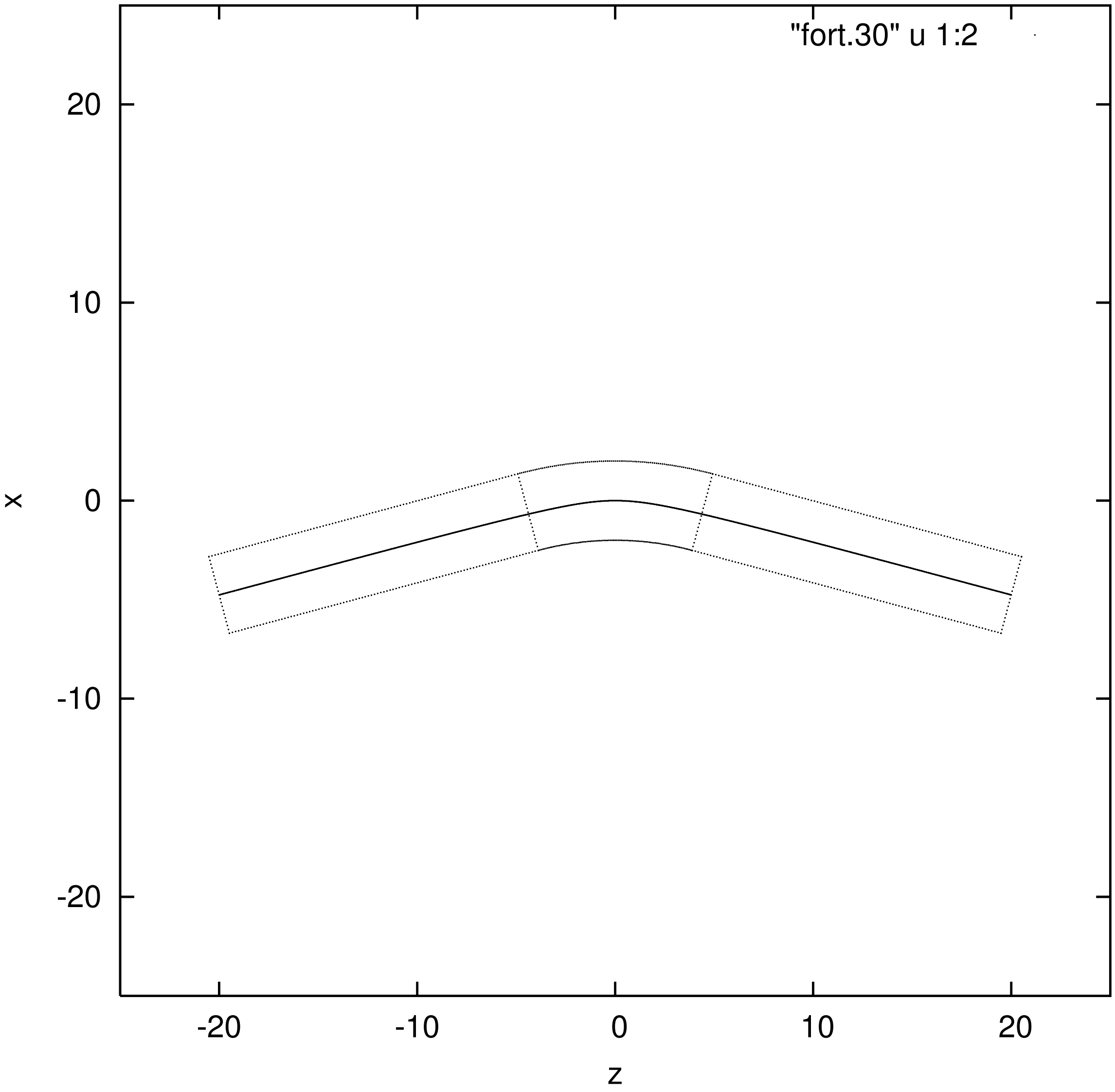


Figure 23.3.1: Design orbit $x(z) = \hat{x}(z)$. Also shown is a surrounding bent box with straight end legs. It will be employed in Chapter 24. The center curve is the design orbit. The outer curves are the boundary of the surrounding bent box with with straight end legs. For ease of visualization, the seams between the bent box and the straight end legs are also shown. The quantities x and z are in cm.

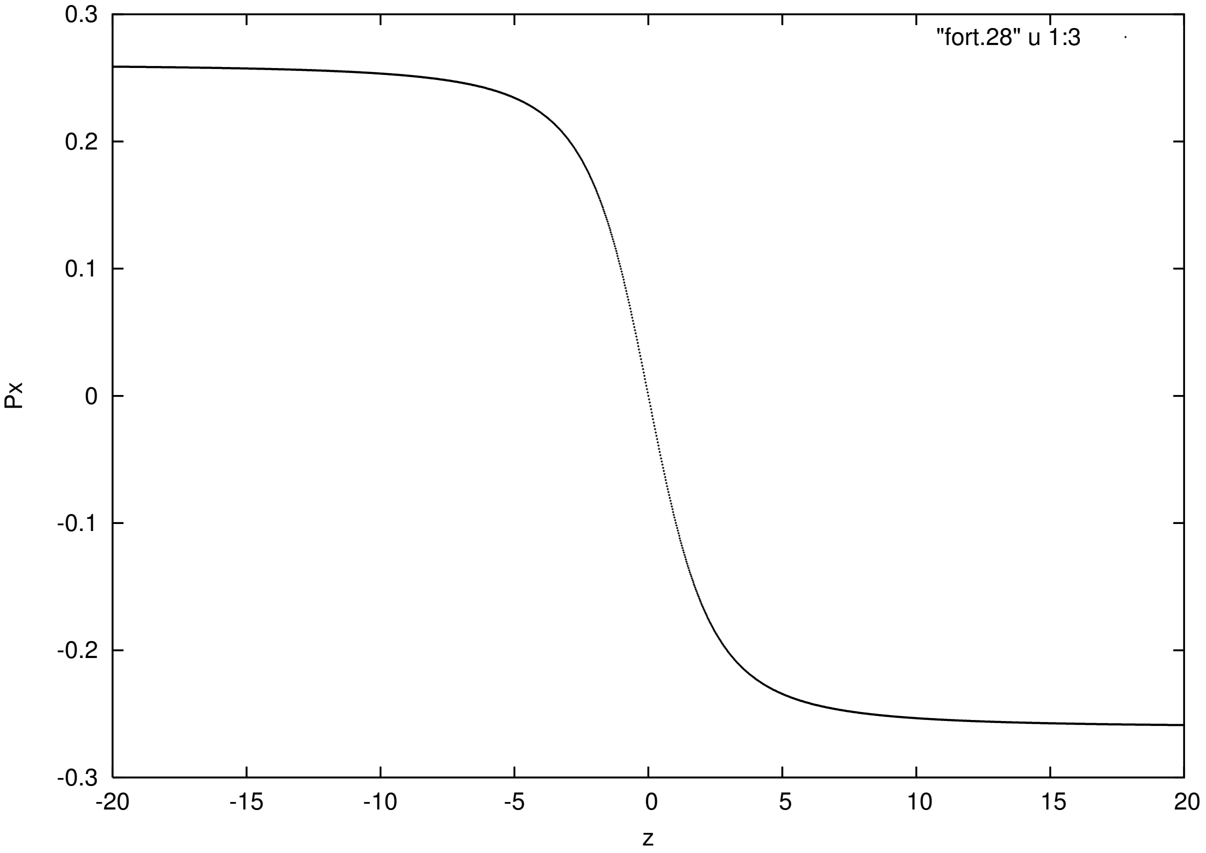


Figure 23.3.2: (Place holder) The canonical momentum $\hat{p}_x(z)$ on the design orbit. The quantity z is in cm.

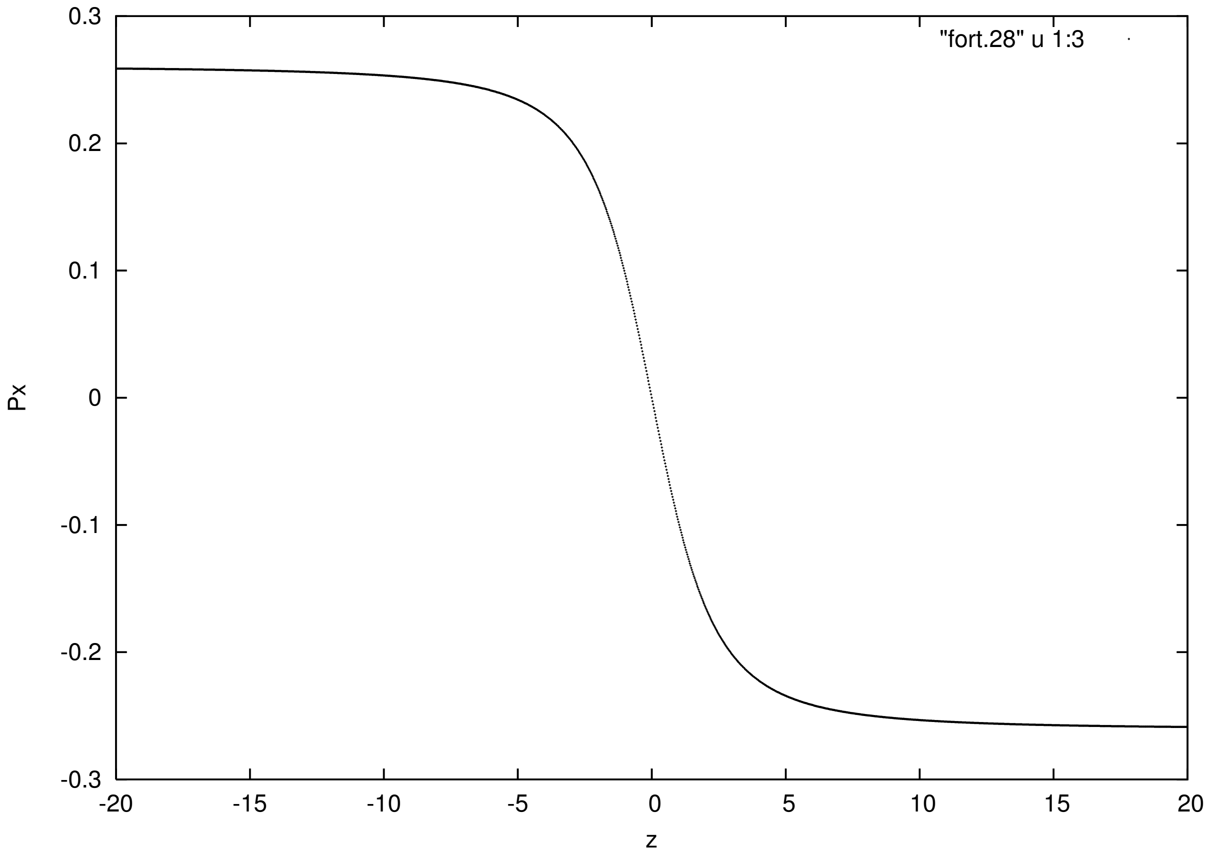


Figure 23.3.3: The scaled mechanical momentum $\hat{p}_x^{\text{mech}}(z)$ on the design orbit. The quantity z is in cm.

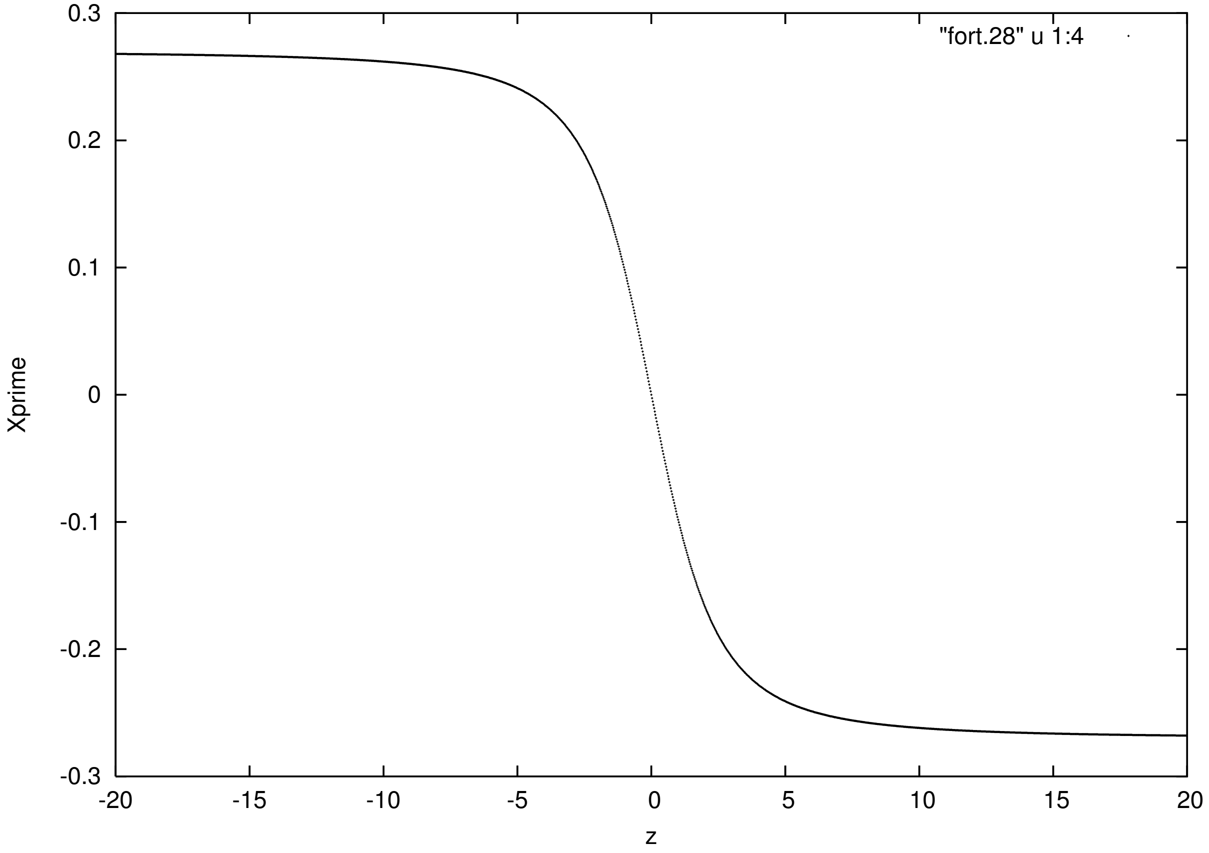


Figure 23.3.4: The quantity $\hat{x}'(z)$ on the design orbit. The quantity z is in cm.

$$\beta^0 = .9984300412295174, \tag{23.3.39}$$

$$\gamma^0 = 17.853008080511426. \tag{23.3.40}$$

Here we have used the values

$$m_e c^2 = .51099906 \text{ MeV}, \tag{23.3.41}$$

$$-q = e = 1.60217733 \times 10^{-19} \text{ coulomb}, \tag{23.3.42}$$

$$c = 2.99792458 \times 10^8 \text{ m/s}. \tag{23.3.43}$$

It is also useful to have graphics of the quantities B_y, $\mathcal{A}_x$, and $\mathcal{A}_z$ along the design orbit. They are displayed in Figures 3.5 through 3.7. Note that B_y falls off quite rapidly with increasing $|z|$, i.e. $\sim 1/|z|^3$, as expected for a monopole doublet. However, $\mathcal{A}_x$ and $\mathcal{A}_z$ fall off less rapidly on the design orbit. From (1.14) we expect for $\mathcal{A}_x$ a fall off $\sim 1/|z|$. And, from Figure 3.1 we see that on the design orbit $|x|$ grows linearly with $|z|$ for large $|z|$. Therefore, if (1.15) provides any indication, we expect that $\mathcal{A}_z$ will also fall off only as $1/|z|$ for large $|z|$.

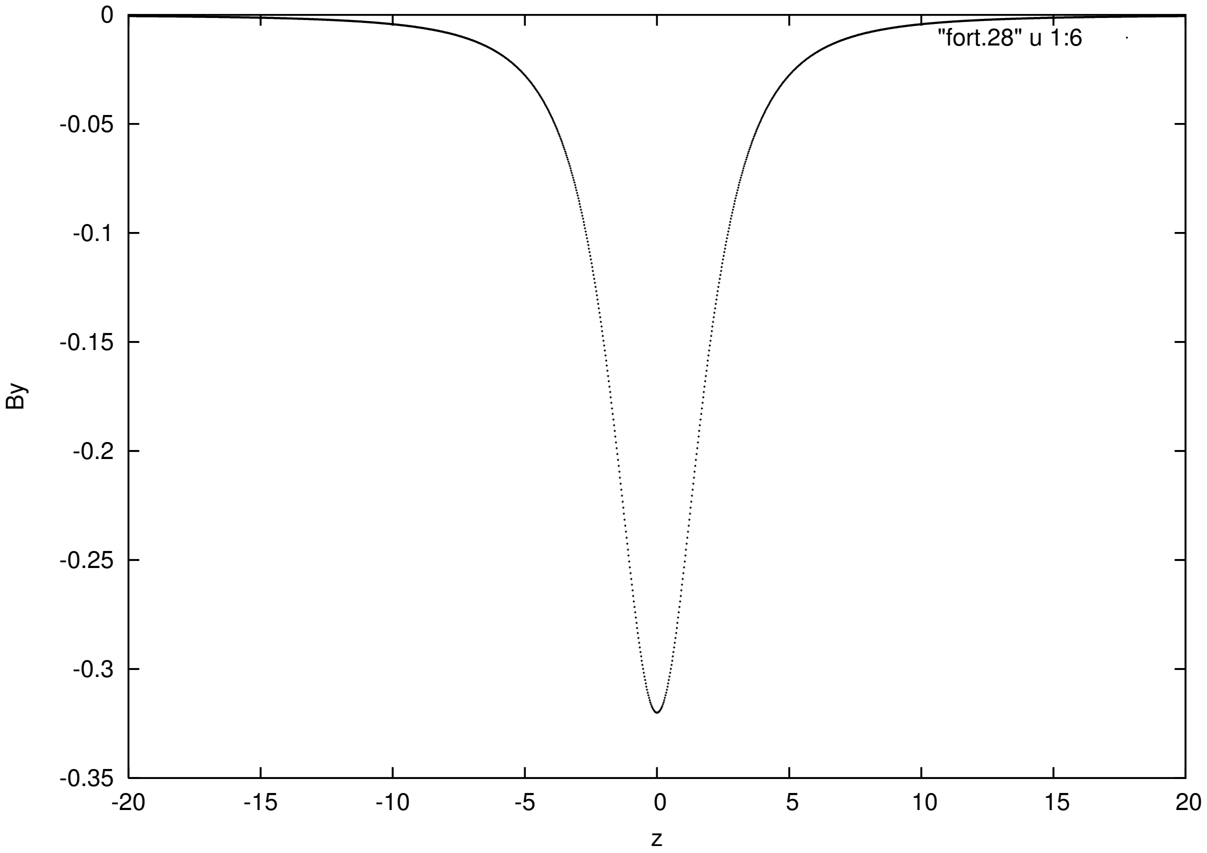


Figure 23.3.5: The quantity B_y on the design orbit. The quantity z is in cm.

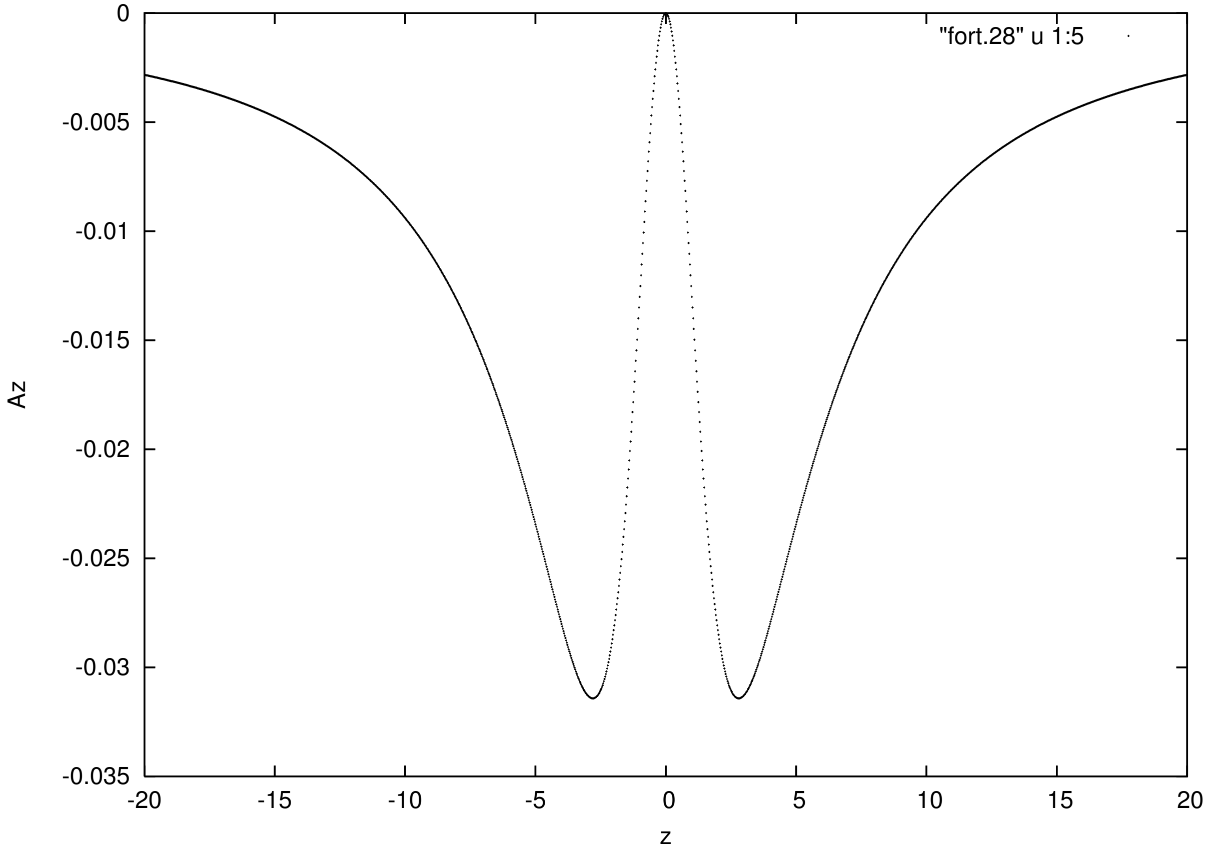


Figure 23.3.6: (Place holder) The quantity $\mathcal{A}_x$ on the design orbit. The quantity z is in cm.

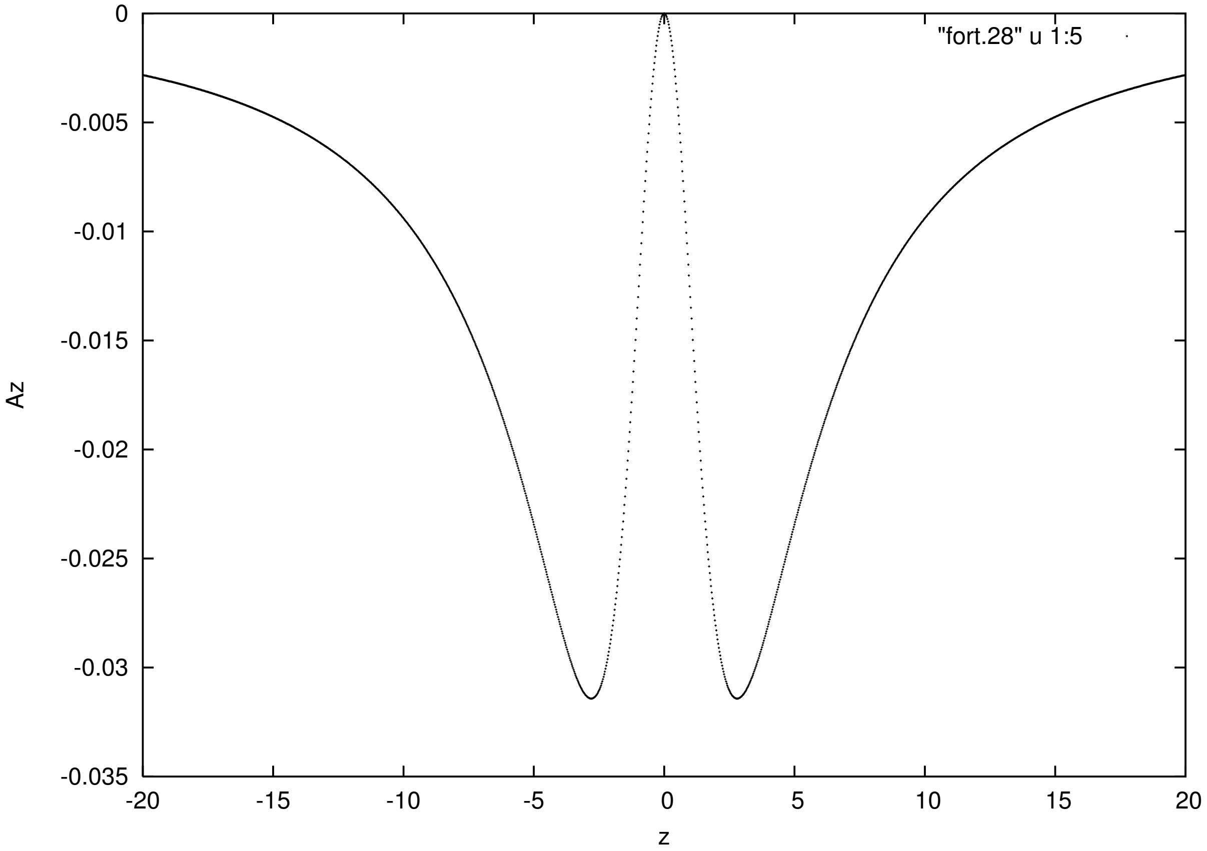


Figure 23.3.7: (Place Holder?) The quantity $\mathcal{A}_z$ on the design orbit. The quantity z is in cm.

Precise field values at some key points on the design orbit are given by the relations

$$B_y(-4.7597\cdots, 0, \mp 20) = -5.6290\cdots \times 10^{-4}, \tag{23.3.44}$$

$$B_y(0,0,0) = -.32, \tag{23.3.45}$$

$$B_y(-4.7597\cdots, 0, \mp 20)/B_y(0,0,0) \simeq 1.8 \times 10^{-3}, \tag{23.3.46}$$

$$\mathcal{A}_x(-4.7597\cdots, 0, \mp 20) = \pm 2.78\cdots \times 10^{-2}, \tag{23.3.47}$$

$$\mathcal{A}_z(-4.7597\cdots, 0, \mp 20) = -6.6\cdots \times 10^{-3}. \tag{23.3.48}$$

From (3.45) we see that B_y on the design orbit has fallen by a factor of $\simeq 1.8 \times 10^{-3}$ at the end points. By contrast, comparison of (3.31) and (3.36) shows that the vector potential for the half-infinite string choice of gauge still makes a significant contribution to the canonical momentum $\hat{p}_x$ at the end points. Compare also (3.30) and (3.31). Therefore it is important to use some other gauge for end-field termination.

Exercises

23.3.1. In Section 3 the design orbit was found by integrating the canonical pair (3.21) and (3.22). This exercise describes an alternate approach. It has the feature of illustrating that in mechanical variables the design orbit is manifestly gauge independent, as we know it should be.

Recall the relations (3.28) and (3.29). Suppose (3.28) is differentiated with respect to z and along the design orbit. Verify that doing so gives the result

$$(\hat{p}_x^{\rm mech})' = \hat{p}_x' - (\partial \mathcal{A}_x/\partial \hat{x})\hat{x}' - (\partial \mathcal{A}_x/\partial z). \tag{23.3.49}$$

Next employ (3.21) to rewrite the second term on the right of (3.49) in the form

$$-(\partial \mathcal{A}_x/\partial \hat{x})\hat{x}' = -(\partial \mathcal{A}_x/\partial \hat{x})(\hat{p}_x - \mathcal{A}_x)/[1-(\hat{p}_x - \mathcal{A}_x)^2]^{1/2}. \tag{23.3.50}$$

Observe that the right side of (3.50) agrees with the first term on the right side of (3.22) save for a sign. Show, therefore, that use of (3.22) in (3.49) yields, following a glorious cancellation, the simple result

$$(\hat{p}_x^{\rm mech})' = (\partial \mathcal{A}_z/\partial \hat{x}) - (\partial \mathcal{A}_x/\partial z). \tag{23.3.51}$$

Also, from (1.10), (2.14), and (2.15), we have the result

$$[(\partial \mathcal{A}_z/\partial \hat{x}) - (\partial \mathcal{A}_x/\partial z)]|_{\hat{y}=0} = -(q/p^0)B_y(x,y,z)|_{y=0} \tag{23.3.52}$$

and, again from (1.10), we see that

$$B_y(x,y,z)|_{y=0} = -2ga/[x^2+a^2+z^2]^{3/2}. \tag{23.3.53}$$

It follows that (3.51) can be written in the final form

$$(\hat{p}_x^{\rm mech})' = -(q/p^0)B_y(x,y,z)|_{y=0} = (q/p^0)2ga/[\hat{x}^2+a^2+z^2]^{3/2}. \tag{23.3.54}$$

Together (3.29) and (3.54) form a convenient coupled set for numerical integration. Once the pair $\{\hat{x}(z), \hat{p}_x^{\rm mech}(z)\}$ has been found, the canonical momentum $\hat{p}_x(z)$ is given by the relation (3.28) rewritten in the form

$$\hat{p}_x = \hat{p}_x^{\rm mech} + \mathcal{A}_x|_{\hat{y}=0} \tag{23.3.55}$$

with $\mathcal{A}_x|_{\hat{y}=0}$ given by (3.23). Verify that $\mathcal{A}_x$ and $\mathcal{A}_z$ on the design orbit are given by the relations (3.23) and (3.24). Verify (3.25) and (3.26).

23.3.2. Verify (3.34) based on (3.33).

23.4 Terminating End Fields

23.4.1 Minimum Vector Potential for End Fields

The first few terms in the expansion [about the point $(X_0, 0, Y_0)$] of the minimum vector potential for a magnetic monopole doublet were found in Section 15.10. We recall the results

$$\boldsymbol{A}^{\min 1}(\boldsymbol{r}; X_0, Z_0) = [ga/(X_0^2 + Z_0^2 + a^2)^{3/2}](-z\boldsymbol{e}_x + x\boldsymbol{e}_z), \tag{23.4.1}$$

$$\begin{gathered}\boldsymbol{A}^{\min 2}(\boldsymbol{r}; X_0, Z_0) = [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ [(Z_0y^2 - Z_0z^2 - X_0xz)\boldsymbol{e}_x + (X_0yz - Z_0xy)\boldsymbol{e}_y + (X_0x^2 + Z_0xz - X_0y^2)\boldsymbol{e}_z].\end{gathered} \tag{23.4.2}$$

See (15.10.7) and (15.10.8). The still higher-order terms (the terms for $n > 2$) can be found in an analogous way.

23.4.2 Associated Termination Error

Suppose we wish to initiate or terminate the magnetic field of a magnetic monopole doublet at the point $(X_0, 0, Y_0)$. Then we need to find the minimum vector potential expansion about this point in terms of variables appropriate to the relevant reference planes. As an example, supposed the field is initiated at the point $(X_0 = -4.7597\cdots, 0, Z_0 = -20)$ corresponding to the beginning of the left leg of the bent box in Figure 3.1. Then the relevant reference plane would be the incoming face of the left leg.

To be more precise, let $\boldsymbol{e}_\xi$ and $\boldsymbol{e}_\eta$ be unit vectors in this reference plane, and let $\boldsymbol{e}_\zeta$ be a unit vector perpendicular to this plane. These requirements can be met by making the definitions

$$\boldsymbol{e}_\xi = \cos\theta\boldsymbol{e}_x - \sin\theta\boldsymbol{e}_z, \tag{23.4.3}$$

$$\boldsymbol{e}_\eta = \boldsymbol{e}_y, \tag{23.4.4}$$

$$\boldsymbol{e}_\zeta = \sin\theta\boldsymbol{e}_x + \cos\theta e_z. \tag{23.4.5}$$

Here θ is the angle between the reference plane and the plane $Y_0 = 0$. See Figure 4.1. For the problem at hand, θ is given by the relation

$$\theta = (1/2)\phi_{\rm bend} \simeq 15°. \tag{23.4.6}$$

As the notation is intended to convey, the vectors $\boldsymbol{e}_\xi$, $\boldsymbol{e}_\eta$, $\boldsymbol{e}_\zeta$ comprise a right-handed orthonormal triad. Consequently, there are relations of the form

$$\boldsymbol{e}_\zeta \times \boldsymbol{e}_\xi = -\sin^2\theta(\boldsymbol{e}_x \times e_z) + \cos^2\theta(\boldsymbol{e}_z \times e_x) = \boldsymbol{e}_y = \boldsymbol{e}_\eta, \text{etc.} \tag{23.4.7}$$

Next we observe that, associated with the unit vectors $\boldsymbol{e}_\xi$, $\boldsymbol{e}_\eta$, $\boldsymbol{e}_\zeta$, we may define local expansion coordinates ξ, η, ζ by writing

$$\xi = x\cos\theta - z\sin\theta, \tag{23.4.8}$$

$$\eta = y, \tag{23.4.9}$$

$$\zeta = x\sin\theta + z\cos\theta. \tag{23.4.10}$$

Finally, the definitions (4.3) through (4.5) and (4.8) through (4.10) may be inverted to yield the relations

$$\boldsymbol{e}_x = \cos\theta\boldsymbol{e}_\xi + \sin\theta\boldsymbol{e}_\zeta, \tag{23.4.11}$$

$$\boldsymbol{e}_y = \boldsymbol{e}_\eta, \tag{23.4.12}$$

$$\boldsymbol{e}_z = -\sin\theta\boldsymbol{e}_\xi + \cos\theta\boldsymbol{e}_\zeta; \tag{23.4.13}$$

$$x = \xi\cos\theta + \zeta\sin\theta, \tag{23.4.14}$$

$$y = \eta, \tag{23.4.15}$$

$$z = -\xi\sin\theta + \zeta\cos\theta. \tag{23.4.16}$$

We also record that, as expected, there are the relations

$$\boldsymbol{r} = x\boldsymbol{e}_x + y\boldsymbol{e}_y + z\boldsymbol{e}_z = \xi\boldsymbol{e}_\xi + \eta\boldsymbol{e}_\eta + \zeta\boldsymbol{e}_\zeta. \tag{23.4.17}$$

With all these relations at hand, let us express the minimum vector potential for a magnetic monopole doublet in terms of the variables ξ, η, ζ and their associated unit vectors. From (4.1) and using (4.11) through (4.17) we find the result

$$\boldsymbol{A}^{\min 1}(\xi,\eta,\zeta;X_0,Z_0) = \boldsymbol{A}^{\min 1}(\boldsymbol{r};X_0,Z_0) = [ga/(X_0^2+Z_0^2+a^2)^{3/2}](-\zeta\boldsymbol{e}_\xi + \xi\boldsymbol{e}_\zeta). \tag{23.4.18}$$

And, from (4.2) and again using (4.11) through (4.17), we find the result

$$\begin{gathered}\boldsymbol{A}^{\min 2}(\xi,\eta,\zeta;X_0,Z_0) = \boldsymbol{A}^{\min 2}(\boldsymbol{r};X_0,Z_0) = -2ga/(X_0^2+Z_0^2+a^2)^{5/2}] \times \\ [(Z_0y^2 - Z_0z^2 - X_0xz)\boldsymbol{e}_\xi + (X_0yz - Z_0xy)\boldsymbol{e}_\eta + (X_0x^2 + Z_0xz - X_0y^2)\boldsymbol{e}_\zeta].\end{gathered} \tag{23.4.19}$$

Following the discussion of Section 16.1, what interests us with regard to the discontinuities in the transverse mechanical momenta associated with the field termination approximation are the ξ and η components of $\boldsymbol{A}^{\min}$ evaluated at $\zeta = 0$. From (4.18) and (4.19) we see that the lowest order contributions to these discontinuities are given by the relations

$$A_\xi^{\min 1}(\xi,\eta,0;X_0,Z_0) = A_\eta^{\min 1}(\xi,\eta,0;X_0,Z_0) = 0, \tag{23.4.20}$$

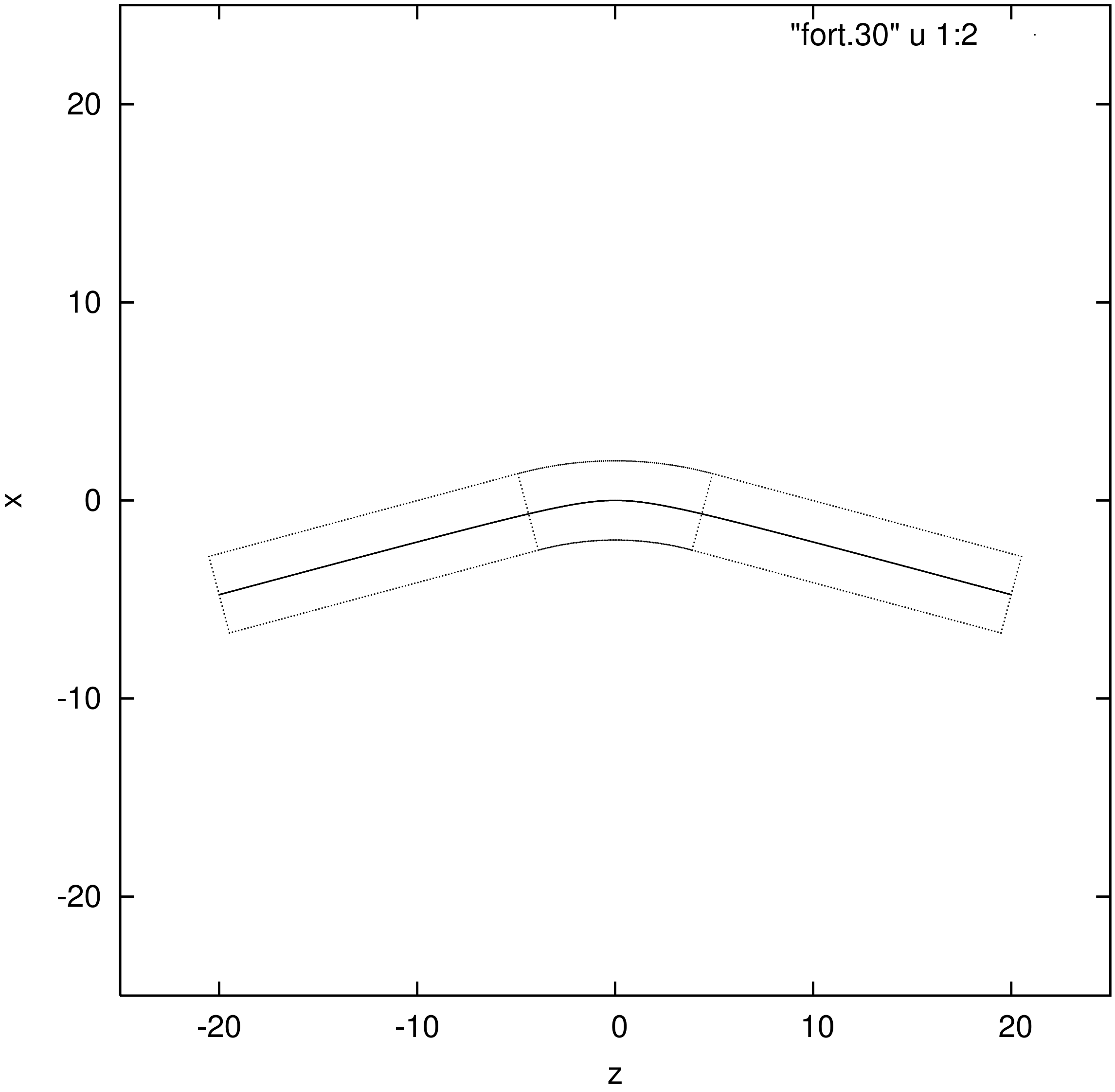


Figure 23.4.1: (Place Holder) The orthonormal triad $\boldsymbol{e}_\xi$, $\boldsymbol{e}_\eta$, $\boldsymbol{e}_\zeta$ and associated local deviation variables ξ, η, ζ for the entry of the left leg of the bent box with legs.

$$A_\xi^{\min 2}(\xi,\eta,0;X_0,Z_0) = [-2ga/(X_0^2+Z_0^2+a^2)^{5/2}](X_0\sin\theta+Z_0\cos\theta)\eta^2, \quad (23.4.21)$$

$$A_\eta^{\min 2}(\xi,\eta,0;X_0,Z_0) = [2ga/(X_0^2+Z_0^2+a^2)^{5/2}](X_0\sin\theta+Z_0\cos\theta)\xi\eta. \quad (23.4.22)$$

We conclude that $\boldsymbol{A}^{\min 1}$ makes no contributions to the discontinuities and that $\boldsymbol{A}^{\min 2}$ makes contributions that are quadratic in the deviation variables ξ and η.

Let us examine the (upon entry) discontinuities associated with $\boldsymbol{A}^{\min}$. As a measure of these, define the *dimensionless* quantities δ_ξ and δ_η by the relations

$$\delta_{\xi,\eta} = (1/p^0)\Delta_{\xi,\eta}^{\text{mech}}. \quad (23.4.23)$$

See (16.1.30) and (16.1.31). Upon employing (16.1.30) and (16.1.31) in (4.23), and with the use of (4.21) and (4.22), we find the results

$$\delta_\xi^2(\xi,\eta,0;X_0,Z_0) = (q/p^0)[-2ga/(X_0^2+Z_0^2+a^2)^{5/2}](X_0\sin\theta+Z_0\cos\theta)\eta^2, \quad (23.4.24)$$

$$\delta_\eta^2(\xi,\eta,0;X_0,Z_0) = (q/p^0)[2ga/(X_0^2+Z_0^2+a^2)^{5/2}](X_0\sin\theta+Z_0\cos\theta)\xi\eta. \quad (23.4.25)$$

Let us evaluate these discontinuities when the transverse deviations from the design orbit have the substantial values $\xi=\eta=1$ cm. So doing, and recalling (3.23) and (4.26), we find the results

$$\delta_\xi^2(\xi=1,\eta=1,0;X_0=-4.7597\cdots,Z_0=-20) = -8.8\cdots\times10^{-6}, \quad (23.4.26)$$

$$\delta_\eta^2(\xi=1,\eta=1,0;X_0=-4.7597\cdots,Z_0=-20) = 8.8\cdots\times10^{-6}. \quad (23.4.27)$$

These numbers are pleasantly small, and we conclude that there is relatively little discontinuity error associated with terminating the field of the monopole doublet to the left of $Z_0=-20$ providing the minimal vector potential is employed. The same is true for termination to the right of $Z_0=+20$.[1]

23.4.3 Taylor Expansion of String Vector Potential

$$\boldsymbol{A}^{\text{ex}}(\boldsymbol{r};\boldsymbol{R}_0) = \boldsymbol{A}^{\text{ex}}(x,y,z;X_0,Z_0) = \boldsymbol{A}(\boldsymbol{R}_0+\boldsymbol{r}) = \sum_{n=0}^{\infty}\boldsymbol{A}^{\text{ex}\,n}(x,y,z;X_0,Z_0) \quad (23.4.28)$$

where $\boldsymbol{A}^{\text{ex}\,n}(x,y,z;X_0,Z_0)$ is a homogeneous polynomial vector field of degree n in the components if $\boldsymbol{r}$.

$$A_x(\boldsymbol{R_0}+\boldsymbol{r}) = \\ -\frac{g(Z_0+z)}{[(X_0+x)^2+(y-a)^2+(Z_0+z)^2]^{1/2}\{[(X_0+x)^2+(y-a)^2+(Z_0+z)^2]^{1/2}-y+a\}} \\ -\frac{g(Z_0+z)}{[(X_0+x)^2+(y+a)^2+(Z_0+z)^2]^{1/2}\{[(X_0+x)^2+(y+a)^2+(Z_0+z)^2]^{1/2}+y+a\}}, \quad (23.4.29)$$

[1] However, one should not be overly sanguine. It turns out that the design orbit continues to bend by as much as a degree as one continues to the left of $Z_0=-20$ and the right of $Z_0=+20$. That is, true *asymptopia* has not been reached even when $Z_0=\pm20$ and (3.41) holds. The magnetic monopole doublet field, and the field of any iron-free dipole, are problematic to treat because of their slow fringe-field fall off.

$$A_y(\boldsymbol{R_0}+\boldsymbol{r}) = 0, \tag{23.4.30}$$

$$A_z(\boldsymbol{R_0}+\boldsymbol{r}) = \\ +\frac{g(X_0+x)}{[(X_0+x)^2+(y-a)^2+(Z_0+z)^2]^{1/2}\{[(X_0+x)^2+(y-a)^2+(Z_0+z)^2]^{1/2}-y+a\}} \\ +\frac{g(X_0+x)}{[(X_0+x)^2+(y+a)^2+(Z_0+z)^2]^{1/2}\{[(X_0+x)^2+(y+a)^2+(Z_0+z)^2]^{1/2}+y+a\}}. \tag{23.4.31}$$

$$A_x^{\text{ex}\,0}(x,y,z;X_0,Z_0) = -\frac{2gZ_0}{[X_0^2+a^2+Z_0^2]^{1/2}\{[X_0^2+a^2+Z_0^2]^{1/2}+a\}}, \tag{23.4.32}$$

$$A_y^{\text{ex}\,0}(x,y,z;X_0,Z_0) = 0, \tag{23.4.33}$$

$$A_z^{\text{ex}\,0}(x,y,z;X_0,Z_0) = \frac{2gX_0}{[X_0^2+a^2+Z_0^2]^{1/2}\{[X_0^2+a^2+Z_0^2]^{1/2}+a\}}. \tag{23.4.34}$$

23.4.4 Finding the Associated Gauge Function

23.5 Gauge Transformation Map

23.6 Pole Face Rotation

23.7 Computation of Transfer Map

Exercises

23.7.1. Using (1.7) through (1.9), verify (1.10) through (1.12).

23.7.2. Review the last paragraph of Exercise 22.2.15. Use the geometric insight provided in that paragraph to conclude that the direction of the vector potential found in Subsection 1.2 follows from the orientations of the Dirac strings assigned to the monopoles making up the monopole doublet.

23.8 Scraps

****

$$\boldsymbol{B}^0(\boldsymbol{r};X_0,Z_0) = -[2ga/(X_0^2+Z_0^2+a^2)^{3/2}]\boldsymbol{e}_\eta. \tag{23.8.1}$$

$$\begin{aligned} \boldsymbol{A}^{\min 1}(\xi,\eta,\zeta;X_0,Z_0) &= -(1/2)\boldsymbol{r}\times\boldsymbol{B}^0(\boldsymbol{r};X_0,Z_0) \\ &= [ga/(X_0^2+Z_0^2+a^2)^{3/2}](-\zeta\boldsymbol{e}_\xi+\xi\boldsymbol{e}_\zeta), \end{aligned} \tag{23.8.2}$$

$$\boldsymbol{A}^{\min 1}(\xi,\eta,0;X_0,Z_0) = [ga/(X_0^2+Z_0^2+a^2)^{3/2}](\xi\boldsymbol{e}_\zeta), \tag{23.8.3}$$

$$A_\xi^{\min 1}(\xi,\eta,0;X_0,Z_0) = A_\eta^{\min 1}(\xi,\eta,0;X_0,Z_0) = 0. \tag{23.8.4}$$

$$A_\xi^{\min 2}(\xi,\eta,0;X_0,Z_0) = [-2ga/(X_0^2+Z_0^2+a^2)^{5/2}](X_0\sin\theta+Z_0\cos\theta)\eta^2, \tag{23.8.5}$$

$$A_\eta^{\min 2}(\xi,\eta,0;X_0,Z_0) = [2ga/(X_0^2+Z_0^2+a^2)^{5/2}](X_0\sin\theta+Z_0\cos\theta)\xi\eta. \tag{23.8.6}$$

$$\begin{aligned} \boldsymbol{B}^1(\boldsymbol{r};X_0,Z_0) &= [6ga/(X_0^2+Z_0^2+a^2)^{5/2}]\times \\ &\quad [(X_0x+Z_0z)\boldsymbol{e}_y+y(X_0\boldsymbol{e}_x+Z_0\boldsymbol{e}_z)]. \end{aligned} \tag{23.8.7}$$

$$(X_0x+Z_0z)\boldsymbol{e}_y = [x_0(\xi\cos\theta+\zeta\sin\theta)+Z_0(-\xi\sin\theta+\zeta\cos\theta)]\boldsymbol{e}_\eta, \tag{23.8.8}$$

$$\begin{aligned} y(X_0\boldsymbol{e}_x+Z_0\boldsymbol{e}_z) &= \eta[X_0(\cos\theta\boldsymbol{e}_\xi+\sin\theta\boldsymbol{e}_\zeta)+Z_0(-\sin\theta\boldsymbol{e}_\xi+\cos\theta\boldsymbol{e}_\zeta)] \\ &= \eta[(X_0\cos\theta-Z_0\sin\theta)\boldsymbol{e}_\xi+(X_0\sin\theta+Z_0\cos\theta)\boldsymbol{e}_\zeta]. \end{aligned} \tag{23.8.9}$$

$$\begin{aligned} \boldsymbol{B}^1(\xi,\eta,\zeta;X_0,Z_0) &= [6ga/(X_0^2+Z_0^2+a^2)^{5/2}]\times \\ &\quad \{[X_0(\xi\cos\theta+\zeta\sin\theta)+Z_0(-\xi\sin\theta+\zeta\cos\theta)]\boldsymbol{e}_\eta \\ &\quad +\eta[(X_0\cos\theta-Z_0\sin\theta)\boldsymbol{e}_\xi+(X_0\sin\theta+Z_0\cos\theta)\boldsymbol{e}_\zeta]\}. \end{aligned} \tag{23.8.10}$$

$$\begin{aligned} \boldsymbol{B}^1(\xi,\eta,0;X_0,Z_0) &= [6ga/(X_0^2+Z_0^2+a^2)^{5/2}]\times \\ &\quad \{[X_0(\xi\cos\theta)+Z_0(-\xi\sin\theta)]\boldsymbol{e}_\eta \\ &\quad +\eta[(X_0\cos\theta-Z_0\sin\theta)\boldsymbol{e}_\xi+(X_0\sin\theta+Z_0\cos\theta)\boldsymbol{e}_\zeta]\}. \end{aligned} \tag{23.8.11}$$

$$\boldsymbol{r}(\xi,\eta,\zeta) = \xi\boldsymbol{e}_\xi+\eta\boldsymbol{e}_\eta+\zeta\boldsymbol{e}_\zeta. \tag{23.8.12}$$

$$\boldsymbol{r}(\xi,\eta,0)=\xi\boldsymbol{e}_\xi+\eta\boldsymbol{e}_\eta. \tag{23.8.13}$$

$$\begin{aligned}\boldsymbol{e}_\eta\times\boldsymbol{B}^1(\xi,\eta,0;X_0,Z_0) &= [6ga/(X_0^2+Z_0^2+a^2)^{5/2}]\times\\ &\quad \{\eta[(-X_0\cos\theta+Z_0\sin\theta)\boldsymbol{e}_\zeta+(X_0\sin\theta+Z_0\cos\theta)\boldsymbol{e}_\xi]\}.\end{aligned} \tag{23.8.14}$$

$$\begin{aligned}[\eta\boldsymbol{e}_\eta\times\boldsymbol{B}^1(\xi,\eta,0;X_0,Z_0)]_\xi &= [6ga/(X_0^2+Z_0^2+a^2)^{5/2}]\times\\ &\quad \{\eta^2[(X_0\sin\theta+Z_0\cos\theta)]\}.\end{aligned} \tag{23.8.15}$$

$$\begin{aligned}\boldsymbol{e}_\xi\times\boldsymbol{B}^1(\xi,\eta,0;X_0,Z_0) &= [6ga/(X_0^2+Z_0^2+a^2)^{5/2}]\times\\ &\quad \{[X_0(\xi\cos\theta)+Z_0(-\xi\sin\theta)]\boldsymbol{e}_\zeta\\ &\quad -\eta[(X_0\sin\theta+Z_0\cos\theta)\boldsymbol{e}_\eta]\}.\end{aligned} \tag{23.8.16}$$

$$\begin{aligned}[\xi\boldsymbol{e}_\xi\times\boldsymbol{B}^1(\xi,\eta,0;X_0,Z_0)]_\eta &= [6ga/(X_0^2+Z_0^2+a^2)^{5/2}]\times\\ &\quad \{-\xi\eta[(X_0\sin\theta+Z_0\cos\theta)]\}.\end{aligned} \tag{23.8.17}$$

$$\boldsymbol{A}^{\min 2}(\xi,\eta,0;X_0,Z_0) = -(1/3)\boldsymbol{r}(\xi,\eta,0)\times\boldsymbol{B}^1(\xi,\eta,0;X_0,Z_0). \tag{23.8.18}$$

$$A_\xi^{\min 2}(\xi,\eta,0;X_0,Z_0)=[-2ga/(X_0^2+Z_0^2+a^2)^{5/2}](X_0\sin\theta+Z_0\cos\theta)\eta^2, \tag{23.8.19}$$

$$A_\eta^{\min 2}(\xi,\eta,0;X_0,Z_0)=[2ga/(X_0^2+Z_0^2+a^2)^{5/2}](X_0\sin\theta+Z_0\cos\theta)\xi\eta. \tag{23.8.20}$$

Chapter 24

Realistic Transfer Maps for General Curved Beam-Line Elements: Bent Box Monopole Doublet Results

24.1 Choice of Surrounding Bent Box

Also shown in Figure 5.1 is the top view of a suitable bent box that surrounds this orbit. The top and bottom of the box are superposed in the figure, and lie in the planes $y = \pm 2$ cm. The circular arcs that comprise the bent portion of the box have the common center

$$(x_c, z_c) = (-17 \text{ cm}, 0) \tag{24.1.1}$$

and have radii

$$r_{\text{out}} = 19 \text{ cm}, \tag{24.1.2}$$

$$r_{\text{in}} = 15 \text{ cm}. \tag{24.1.3}$$

Both subtend an angle of 30°, and are extended by straight lines thereby forming the straight ends of the box.

How was this bent box determined? Again by trial and error. Note that the construction of the bent box is not critical. All that is required is that the bent box well surround the design orbit. Consider all circular arcs that pass through the origin and are symmetric about the x axis. Such arcs will have their centers on the x axis. Also require that each arc subtend an angle of 30°. With these restrictions the only remaining quantity to be selected is the radius of an arc. Finally, require that the optimal arc, when extended by straight lines at both ends, well fit the design orbit. Figure 6.1 below shows that, for the problem at hand, a good fit occurs when the arc radius has the value

$$r_{\text{fit}} = 17 \text{ cm}. \tag{24.1.4}$$

By construction, the center of this arc is given by (5.55).

Now determine the outer and inner boundaries of the box by requiring that they also be circular arcs with straight-line extensions. Further require that both arcs have a common

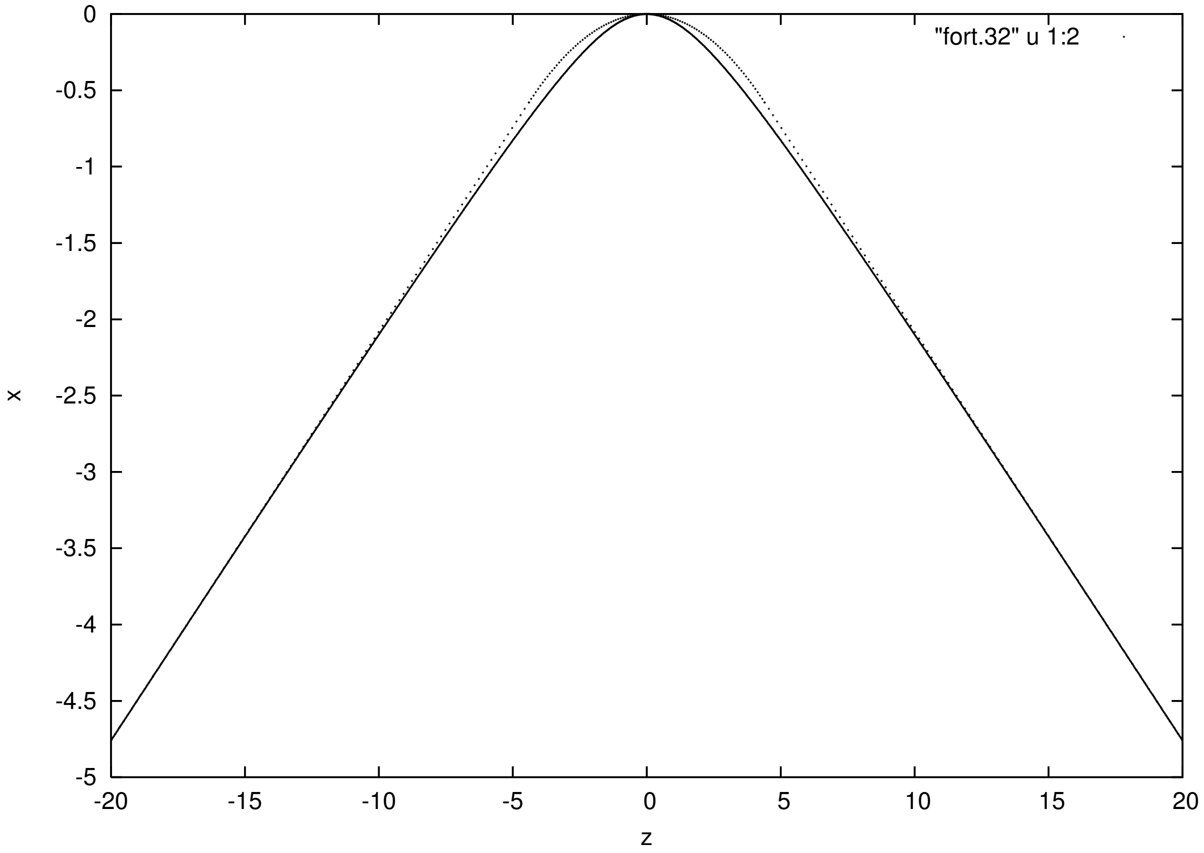


Figure 24.1.1: Design orbit $x(z)$ and best approximating circular arc with straight-line extensions. The solid line is the design orbit, and the dotted line is the best approximating circular arc with straight-line extensions. The quantities x and z are in cm.

center given by (5.55), and that both arcs subtend an angle of 30°. The last step is to specify the radius of each arc. This is conveniently done by the prescription

$$r_{\text{out}} = r_{\text{fit}} + w, \tag{24.1.5}$$

$$r_{\text{in}} = r_{\text{fit}} - w, \tag{24.1.6}$$

where $2w$ is the width of the box. For our illustration we have chosen the value

$$w = 2 \text{ cm}. \tag{24.1.7}$$

24.2 Comparison of Fields

How well do surface methods work for general geometries? In this subsection we will take magnetic field and scalar potential values for a monopole doublet, interpolate them onto the bent box surface found in the previous subsection, and then use surface methods to compute the interior field at various sample points. This computed interior field will then be compared with the actual monopole doublet field at these sample points, thereby providing a test of the accuracy of the method. Note that the bent box we have chosen has a square cross section with side 4 cm, and therefore is comparable in cross section to the 4 cm diameter cylinder used in Section 19.1.

24.2.1 Preliminaries

Since it is our intent to compute the interior field from surface values of B_n and ψ, it would be good to have some feel for how these quantities behave on the surface of the bent box with legs. Figure 5.7 displays B_n on the upper face, $y = 2$ cm, of the bent box with legs directly above the design orbit. Figure 5.8 does the same for ψ. Up to signs, similar results hold for the bottom face, $y = -2$ cm. The observation to be made is that B_n falls off fairly rapidly with increasing $|z|$, like $\sim 1/|z|^3$, and ψ falls off somewhat less rapidly, like $\sim 1/|z|^2$.

Something should also be said about the behavior of B_n and ψ on the sides of the box with legs. It is easily checked that, for a monopole doublet, B_x, B_z, and ψ are odd functions of y, and therefore must vanish in the midplane $y = 0$. From this fact, and from considerations of field-line geometry for the case of a monopole doublet field, we conclude that the values of B_n and ψ on the sides of the box with legs will be comparable to, and usually smaller than, their values on the top and bottom faces.

Taking into account the behavior of B_n and ψ on the entire surface of the box with legs, we conclude that it is only necessary to integrate over some bounded portion of the surface in order to compute the interior vector potential $\boldsymbol{A}$ accurately.

With this background in mind, we are prepared to make some numerical tests. We begin by imbedding the bent box with legs of Figure 5.1 (also see Figure 1.1) within a three-dimensional rectangular mesh,

$$x \in [x_{min}, x_{max}], \tag{24.2.1}$$

$$y \in [y_{min}, y_{max}], \tag{24.2.2}$$

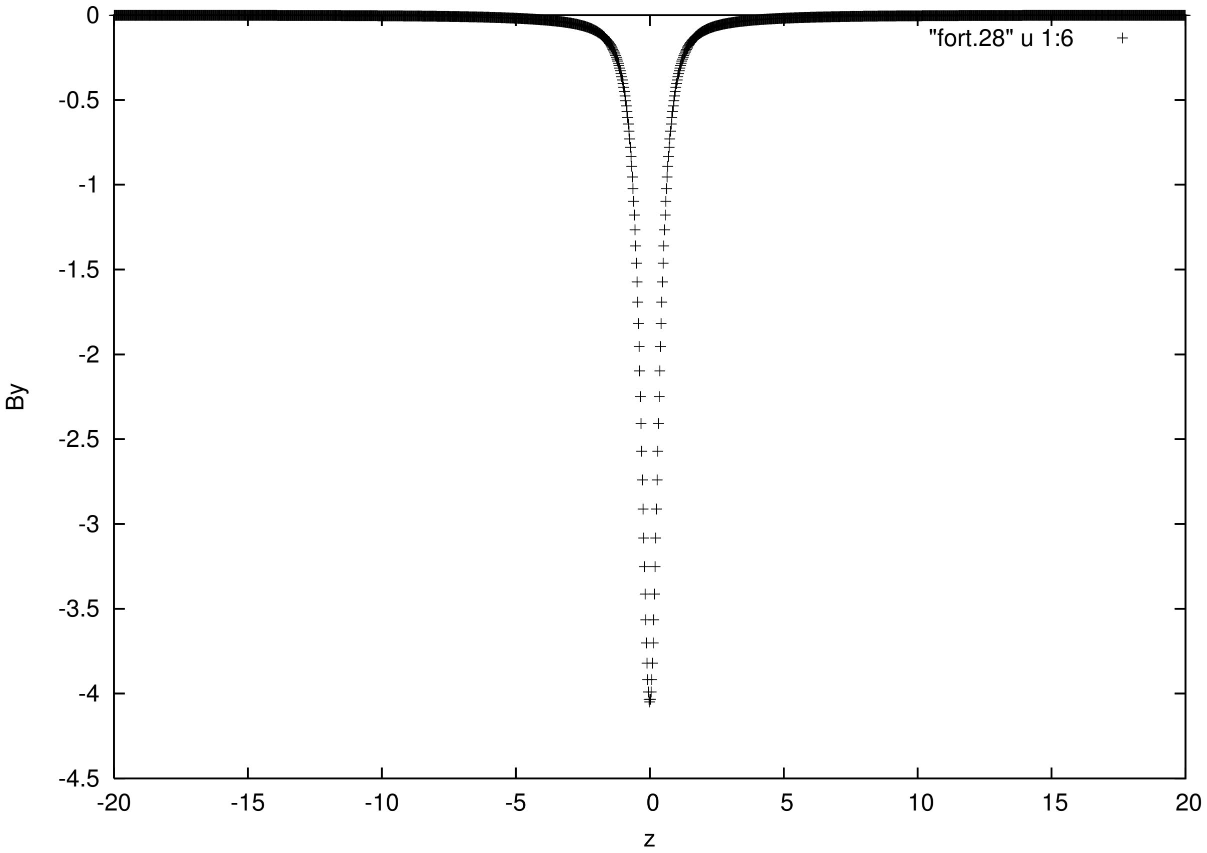


Figure 24.2.1: The quantity $B_n = B_y$ on the upper face, $y = 2$ cm, and directly above the design orbit. The quantity z is in cm.

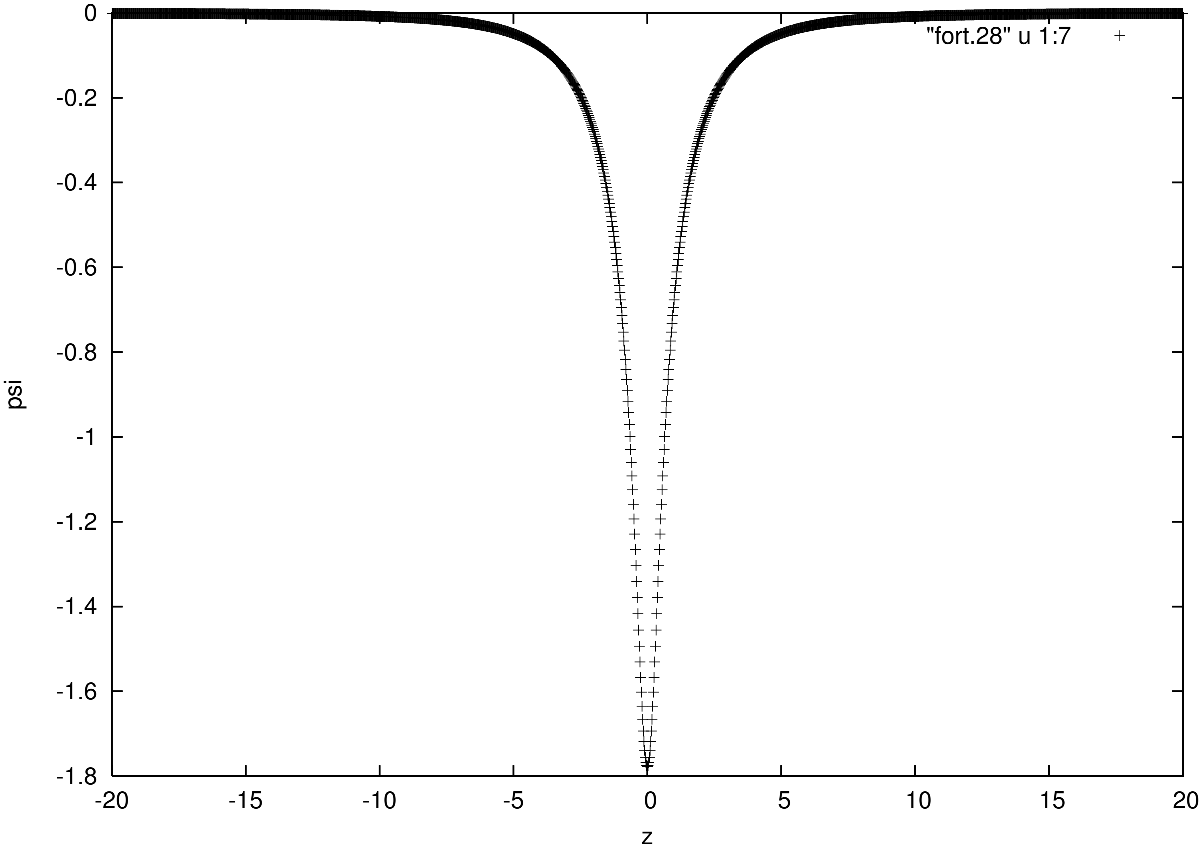


Figure 24.2.2: The quantity ψ on the upper face, $y = 2$ cm, and directly above the design orbit. The quantity z is in cm.

$$z \in [z_{min}, z_{max}], \tag{24.2.3}$$

with mesh-point spacings h_x, h_y, and h_z, respectively. By looking at Figure 5.1 we see that convenient end-point values are given by the relations

$$x_{min} = -7,\ x_{max} = 3, \tag{24.2.4}$$

$$y_{min} = -3,\ y_{max} = 3, \tag{24.2.5}$$

$$z_{min} = -25,\ z_{max} = 25. \tag{24.2.6}$$

See also the corner coordinates given in (5.76) through (5.79). For the mesh-point spacings we take the values

$$h_x = h_y = h_z = .05. \tag{24.2.7}$$

At each of these grid points we compute and store the quantities $\boldsymbol{B}$ and ψ for the monopole doublet field. It is data if this kind that could be expected for the output of some electromagnetic solver.

24.2.2 Evaluation of Surface Integrals

Our task now is to use this data to evaluate surface integrals of the kind (1.7) and (1.8). In this case the surface S, a bent box with legs, consists of a bent sector with straight end legs. When viewed from above, and as described earlier, the sector has inner and outer radii given by (5.59) and (5.60), and subtends an angle of 30°. It has corners at the locations

$$z^{\rm s}_{\ell\ell} = -3.882285676537811,\ x^{\rm s}_{\ell\ell} = -2.511112605663975, \tag{24.2.8}$$

$$z^{\rm s}_{{\rm u}\ell} = -4.917561856947894,\ x^{\rm s}_{{\rm u}\ell} = 1.352590699492298, \tag{24.2.9}$$

$$z^{\rm s}_{\rm ur} = 4.917561856947894,\ x^{\rm s}_{\rm ur} = 1.352590699492298, \tag{24.2.10}$$

$$z^{\rm s}_{\ell{\rm r}} = 3.882285676537811,\ x^{\rm s}_{\ell{\rm r}} = -2.511112605663975. \tag{24.2.11}$$

The left straight end leg, again when viewed form above, has leftmost corners at the locations

$$z^{\ell}_{\ell\ell} = -19.482361909794935,\ x^{\ell}_{\ell\ell} = -6.69114043422916, \tag{24.2.12}$$

$$z^{\ell}_{{\rm u}\ell} = -20.51763809020501,\ x^{\ell}_{{\rm u}\ell} = -2.8274371290729103. \tag{24.2.13}$$

The right straight end leg has rightmost corners at the locations

$$z^{\rm r}_{\rm ur} = 20.51763809020501,\ x^{\rm r}_{\rm ur} = -2.8274371290729103, \tag{24.2.14}$$

$$z^{\rm r}_{\ell{\rm r}} = 19.482361909794935,\ x^{\rm r}_{\ell{\rm r}} = -6.69114043422916. \tag{24.2.15}$$

Each straight end leg has a length of 16.150387336872548 cm.

We will decompose S into 12 pieces. The first 8 will be the top, bottom, inner, and outer faces of the two straight legs. The remaining 4 will be the top, bottom, inner, and outer faces of the bent sector. See Figures 1.1 and 5.1. The first 8 surfaces, those for the straight legs, are all rectangular, and can be conveniently integrated over using rectangular coordinates. Integrals over the remaining 4 surfaces, those for the bent sector, are most easily evaluated using polar/cylindrical coordinates. Our task is to parameterize these 12 pieces and find expressions for dS' for each. In particular, we will convert the integration over each of these pieces into a related integration over a unit square.

Consider the straight legs. We will present results for the left leg. Results for the right leg are analogous.

Note that the right end of the left leg abuts the left end of the bent sector. Again see Figure 5.1. Consequently, the top face of the left leg has corners at $(z^{\ell}_{\ell\ell}, x^{\ell}_{\ell\ell})$, $(z^{\ell}_{\mathrm{u}\ell}, x^{\ell}_{\mathrm{u}\ell})$, $(z^{\mathrm{s}}_{\mathrm{u}\ell}, x^{\mathrm{s}}_{\mathrm{u}\ell})$, and $(z^{\mathrm{s}}_{\ell\ell}, x^{\mathrm{s}}_{\ell\ell})$. It can be described in terms of parameters u and v by writing

$$z^{\ell\mathrm{t}}(u,v) = z^{\ell}_{\ell\ell} + (z^{\mathrm{s}}_{\ell\ell} - z^{\ell}_{\ell\ell})u + (z^{\ell}_{\mathrm{u}\ell} - z^{\ell}_{\ell\ell})v, \tag{24.2.16}$$

$$x^{\ell\mathrm{t}}(u,v) = x^{\ell}_{\ell\ell} + (x^{\mathrm{s}}_{\ell\ell} - x^{\ell}_{\ell\ell})u + (x^{\ell}_{\mathrm{u}\ell} - x^{\ell}_{\ell\ell})v, \tag{24.2.17}$$

$$y^{\ell\mathrm{t}} = 2, \tag{24.2.18}$$

with

$$u, v \in [0,1]. \tag{24.2.19}$$

In this case one finds for the surface element the relation

$$dS' = dz\,dx = [(z^{\mathrm{s}}_{\ell\ell} - z^{\ell}_{\ell\ell})(x^{\ell}_{\mathrm{u}\ell} - x^{\ell}_{\ell\ell}) - (z^{\ell}_{\mathrm{u}\ell} - z^{\ell}_{\ell\ell})(x^{\mathrm{s}}_{\ell\ell} - x^{\ell}_{\ell\ell})]du\,dv. \tag{24.2.20}$$

Similar results hold for the bottom face of the left leg.

The inner face of the left leg can be described in terms of parameters u and v by writing

$$z^{\ell\mathrm{i}}(u,v) = z^{\ell}_{\ell\ell} + (z^{\mathrm{s}}_{\ell\ell} - z^{\ell}_{\ell\ell})u, \tag{24.2.21}$$

$$x^{\ell\mathrm{i}}(u,v) = x^{\ell}_{\ell\ell} + (x^{\mathrm{s}}_{\ell\ell} - x^{\ell}_{\ell\ell})u, \tag{24.2.22}$$

$$y^{\ell\mathrm{i}}(u,v) = -2 + 4v, \tag{24.2.23}$$

again with

$$u, v \in [0,1]. \tag{24.2.24}$$

In this case one finds for the surface element the relation

$$dS' = [4/\cos(\pi/12)](z^{\mathrm{s}}_{\ell\ell} - z^{\ell}_{\ell\ell})du\,dv. \tag{24.2.25}$$

Similar results hold for the outer face of the left leg.

Consider the bent sector. The top face of the bent sector can be described in terms of cylindrical coordinates ρ, ϕ, and y by writing

$$z^{\mathrm{st}} = \rho\sin\phi, \tag{24.2.26}$$

$$x^{\mathrm{st}} = \rho\cos\phi - 17, \tag{24.2.27}$$

$$y^{\mathrm{st}} = 2. \tag{24.2.28}$$

Introduce parameters u and v by writing

$$\rho(u,v) = 15 + 4v, \tag{24.2.29}$$

$$\phi(u,v) = -\pi/12 + (\pi/6)u, \tag{24.2.30}$$

again with the understanding (5.83). Combining (5.90) through (5.94) gives the results

$$z^{\rm st}(u,v) = (15+4v)\sin(-\pi/12+\pi u/6), \tag{24.2.31}$$

$$x^{\rm st}(u,v) = (15+4v)\cos(-\pi/12+\pi u/6) - 17, \tag{24.2.32}$$

$$y^{\rm st}(u,v) = 2. \tag{24.2.33}$$

In this case one finds for the surface element of the top face of the bent sector the relation

$$dS' = \rho d\rho\, d\phi = (2\pi/3)(15+4v)du\, dv. \tag{24.2.34}$$

Similar results hold for the bottom face of the bent sector.

The inner face of the bent sector can be described in terms of relations analogous to (5.90) and (5.91) with $\rho = 15$,

$$z^{\rm si}(u,v) = 15\sin(-\pi/12+\pi u/6), \tag{24.2.35}$$

$$x^{\rm si}(u,v) = 15\cos(-\pi/12+\pi u/6) - 17. \tag{24.2.36}$$

Here we have again used (5.94). We also write

$$y^{\rm si}(u,v) = -2+4v. \tag{24.2.37}$$

In this case we find for the surface element the relation

$$dS' = \rho d\phi\, dy = 10\pi du\, dv. \tag{24.2.38}$$

Similar results hold for the outer face of the bent sector.

The result of the work so far is that the integrations over the 12 pieces of S have been converted into integrations over 12 unit squares of the form (5.83). For each piece, changes in u produce longitudinal displacements, and changes in v produce transverse displacements. We next select points within each unit square to be used in evaluating the various surface integrals numerically.

For the straight legs this is achieved as follows: Each unit square corresponding to a leg surface is decomposed into $100 \times 160 = 16,000$ small rectangles by the prescription

$$h_u = 1/100,\ h_v = 1/160. \tag{24.2.39}$$

Thus, there are 100 subdivisions in the longitudinal direction and 160 subdivisions in the transverse directions.

For the surfaces of the bent sector each corresponding unit square is decomposed into $160 \times 160 = 25,600$ small squares by the prescription

$$h_u = 1/160,\ h_v = 1/160. \tag{24.2.40}$$

Thus, for these surfaces there are 160 subdivisions for both the longitudinal and transverse directions.[1]

[1] More subdivisions are used for the bent sector surfaces because the fields are expected to vary more rapidly over these surfaces. See Figures 5.7 and 5.8.

The integral over each small rectangle or small square is approximated using a 7-point cubature formula. (For a discussion of cubature formulas, see Appendix T.) The values of the integrands at the cubature points are obtained from the values of $\boldsymbol{B}$ and ψ at the grid points using 3-dimensional cubic spline interpolation.[2] Finally, all the small rectangle and small square results are summed to obtain the required integrals over S.

24.2.3 Resulting Vector Potential

Figures 5.9 and 5.10 show the components A_x^{sd} and A_z^{sd} of the vector potential computed along the design trajectory based on bent-box *surface data*.[3] The A_y^{sd} component vanishes in the midplane, and therefore is not shown. For the contribution from the B_n the kernel $\boldsymbol{G}^{n2s}$ was used for the top and bottom faces, and the kernel $\boldsymbol{G}^{n1s}$ was used for the side faces. In all cases the strings were taken to lie on lines parallel to the x axis. For the contribution from ψ the kernel $\boldsymbol{G}^t$ was used.

Recall that the vector potential used in the previous subsection and given by (2.102) through (2.104) has no x component and only a z component on the design orbit. By contrast, A_x^{sd} as shown in Figure 5.9, although small, is not zero on the design orbit. Note also that A_z as displayed in Figure 5.4 and the A_z^{sd} displayed in Figure 5.10, while similar, are not the same. This apparent discrepancy arises from the fact that the vector potential given by (2.102) through (2.104) and the vector potential computed from surface data differ by a gauge transformation. We also remark that examination of the numerical results reveals that the nonzero contribution to A_x^{sd} arises from surface ψ values. See (3.43).

24.2.4 Comparison of Fields

If the interior vector potential $\boldsymbol{A}^{sd}$ has been computed successfully using surface methods, so that it differs from the vector potential given by (2.102) through (2.104) at most only by a gauge transformation, in the interior of the box it should also give rise to the monopole-doublet $\boldsymbol{B}$ field (2.105) through (2.107). Let $\boldsymbol{B}^{sd}$ be the magnetic field given by

$$\boldsymbol{B}^{sd} = \nabla \times \boldsymbol{A}^{sd} \tag{24.2.41}$$

and let $\boldsymbol{B}^e$ be the exact $\boldsymbol{B}$ field. Then, for example, use of (5.105) to compute B_y^{sd} on the design orbit should produce a graphic similar to Figure 5.5.[4] This is indeed the case. A plot of B_y^{sd} on the design orbit is indistinguishable to the eye from Figure 5.5.

To give a better indication of the error involved, define a relative error $\boldsymbol{\Delta}$ by the relation

$$\boldsymbol{\Delta} = (\boldsymbol{B}^{sd} - \boldsymbol{B}^e)/B_y^{maxmag}. \tag{24.2.42}$$

[2] Note that, like the case of cylindrical surfaces, the data at most of the data points on the grid are unused. For each cubature point on S there is an associated point in x,y,z space, and only data at the grid points near these points are actually used in interpolation.

[3] That is, as just described, grid data were manufactured and interpolated onto the 12 pieces of the surface S at the points required for the repeated use of a 7-point cubature formula. The results from these surface values were processed, by repeated application of this cubature formula, to find A_x^{sd} and A_z^{sd} along the design trajectory.

[4] Note that the use of (5.105) requires a knowledge of spatial derivatives of the components of $\boldsymbol{A}$. These derivatives are obtained by differentiating the kernels $\boldsymbol{G}$ under the integral sign prior to carrying out the required surface integrations. See Subsections 4.3 and 4.4.

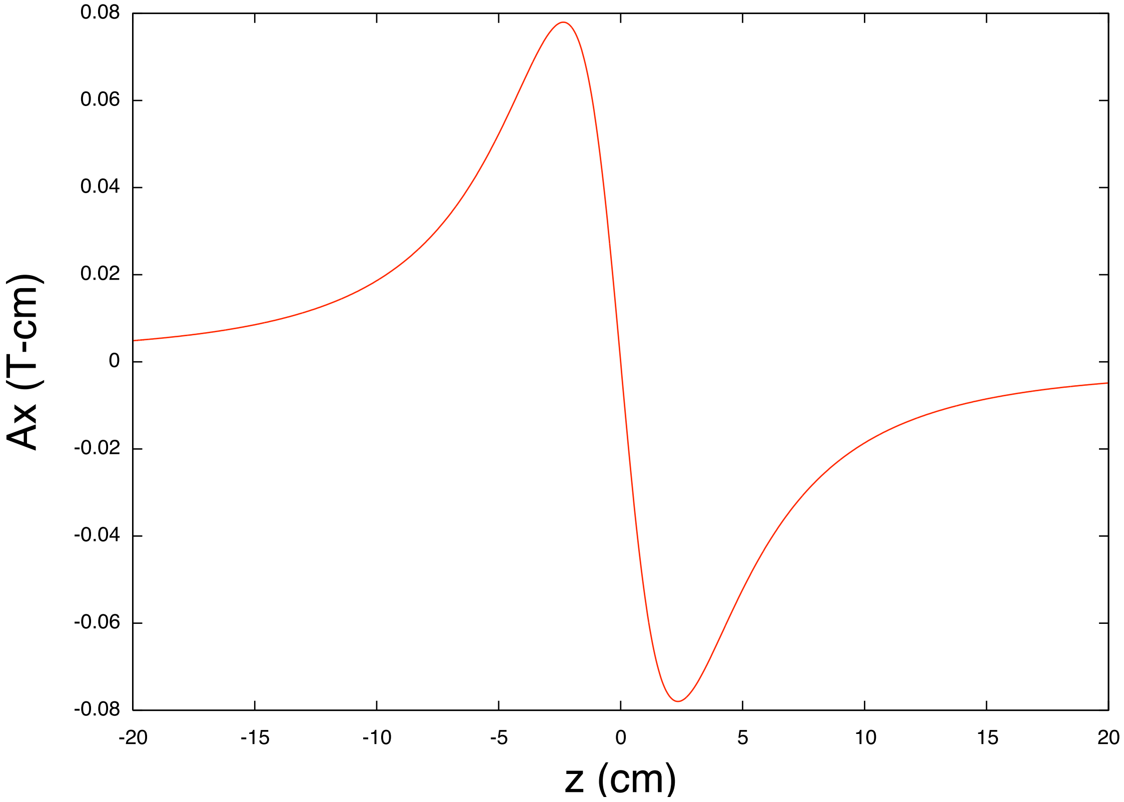


Figure 24.2.3: The quantity A_x^{sd} on the design orbit. The quantity z is in cm.

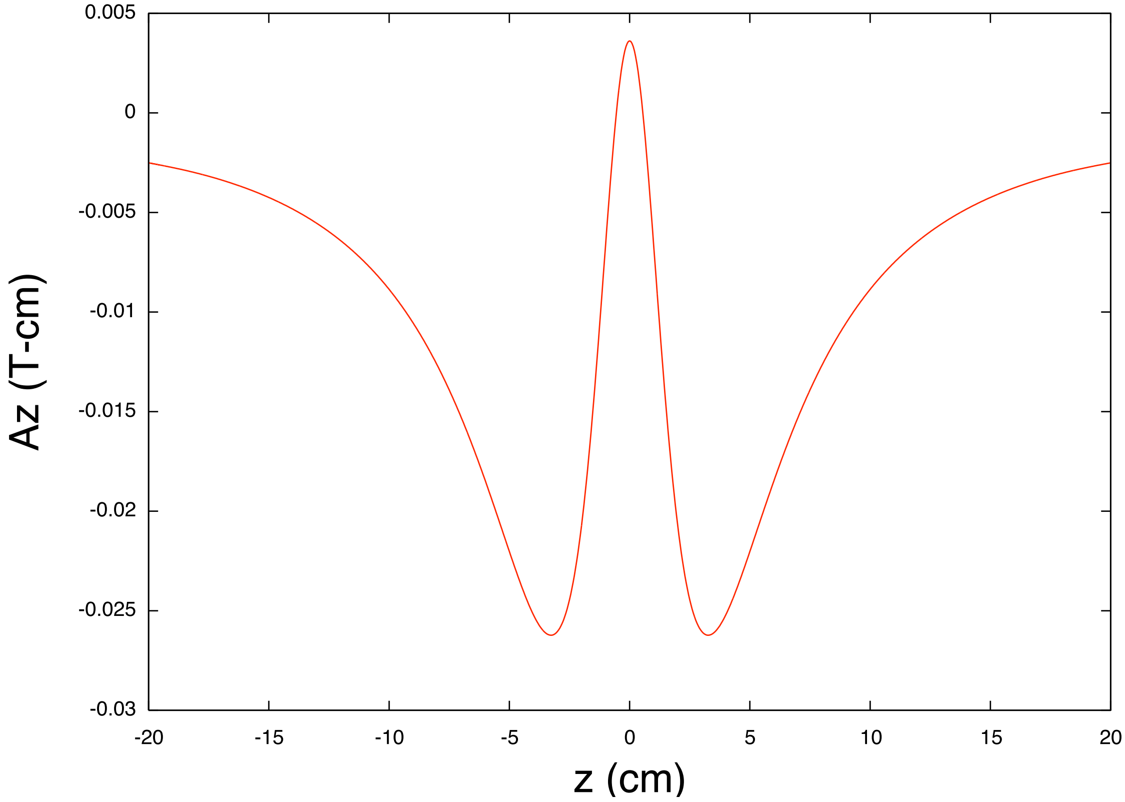


Figure 24.2.4: The quantity A_z^{sd} on the design orbit. The quantity z is in cm.

Here B_y^{maxmag} is the *maximum* value of the *magnitude* of B_y^e on the design orbit,

$$B_y^{maxmag} = |B_y(x = 0, y = 0, z = 0)| = 2g/a^2 = .32 \text{ Tesla}. \tag{24.2.43}$$

See (15.8.5) and Figure 15.8.3. Figure 5.11 displays the value of Δ_y on the design orbit as a function of z. We see that Δ_y is very small over most of the interval $z \in [-20, 20]$, but rises very rapidly to a value of $\simeq 5 \times 10^{-4}$ at the endpoints.

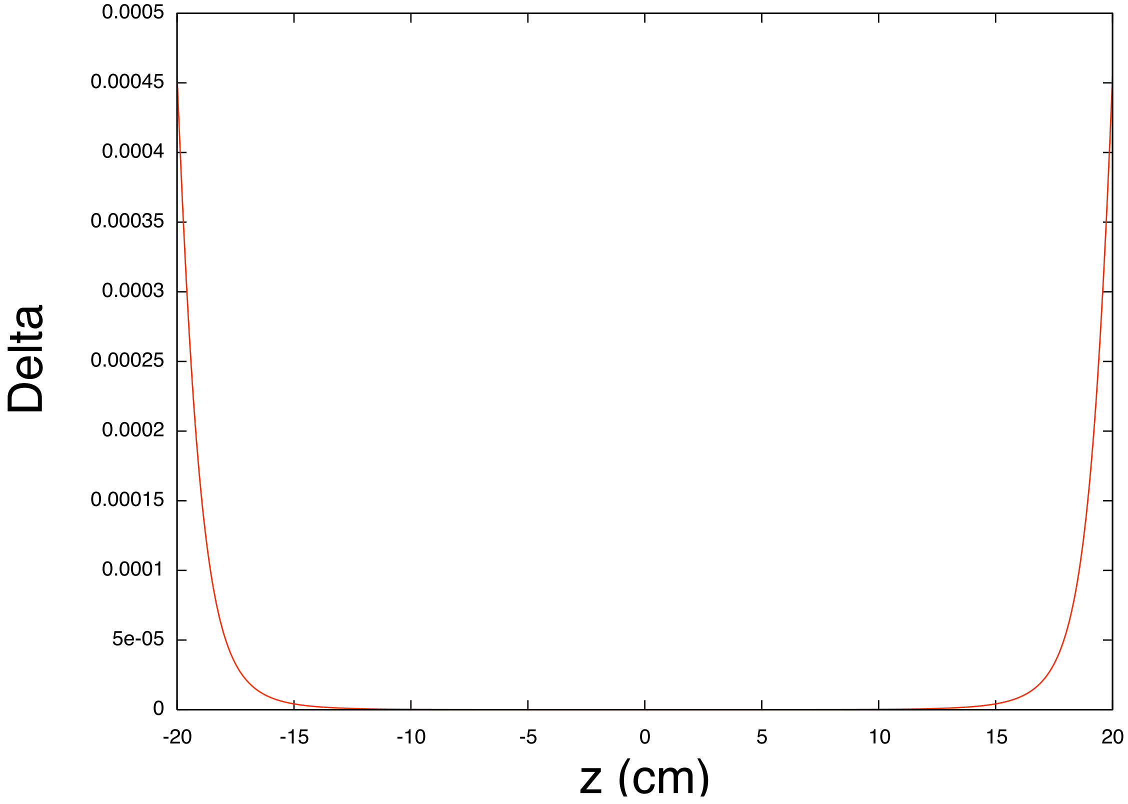


Figure 24.2.5: The relative error Δ_y on the design orbit. The quantity z is in cm.

While the very small error for most of the interval error is very satisfying, the rapid increase of the error at the endpoints might seem alarming. It is not. The relative error Δ_y remains bounded and eventually goes to zero as $|z|$ goes to infinity. Moreover, both B_y^{sd} and B_y^e are small for $|z| \geq 20$, and go to zero as $|z|$ goes to infinity.

To elaborate on these assertions, we begin by noting that both B_y^{sd} and B_y^e are negative. See Figure 5.5. But B_y^{sd} is slightly less negative than B_y^e because, by terminating the straight legs of the box at $z = \pm 20$, the surface fields that serve as a “source” for the interior field are effectively set to zero beyond $z \in [-20, 20]$. Correspondingly, Δ_y is positive. Observe that, because B_y^{sd} is negative, Δ_y always obeys the crude bound

$$\Delta_y < -B_y^e/B_y^{maxmag}. \tag{24.2.44}$$

At the end points the coordinates x, y, z have the values

$$x \simeq -4.76, y = 0, z = \pm 20. \tag{24.2.45}$$

See (5.45). Using these values in (15.8.5) we find that B_y^e has the value

$$B_y^e(x \simeq -4.76, y = 0, z = \pm 20) \simeq -5.63 \times 10^{-4} \text{ Tesla.} \tag{24.2.46}$$

Correspondingly, at worst and for $|z| \geq 20$, Δ_y can never exceed

$$5.63 \times 10^{-4}/.32 \simeq 1.8 \times 10^{-3}. \tag{24.2.47}$$

And since B_y^e for large $|z|$ falls off as $|z|^{-3}$, Δ_y must eventually go to zero for large $|z|$ as $|z|^{-3}$.

Figure 5.11 displays the relative error in the y component of $\boldsymbol{B}^{sd}$ on the design orbit. We are also interested in examining the relative error in all the components of $\boldsymbol{B}^{sd}$ in the vicinity of the design orbit. For this purpose it is convenient to introduce a deviation variable ξ by writing

$$x = x^d + \xi \tag{24.2.48}$$

where x^d is the design orbit shown in Figure 5.6. Figure 5.12 shows Δ, the magnitude of $\boldsymbol{\Delta}$, over the domain $\xi \in [-1, 1], z \in [0, 20]$ in the plane $y = 0$. Figure 5.13 shows Δ over the same domain in the plane $y = 1$. For ease of visualization, values are shown only for $y \geq 0$ and $z \geq 0$ since Δ is even in these variables.

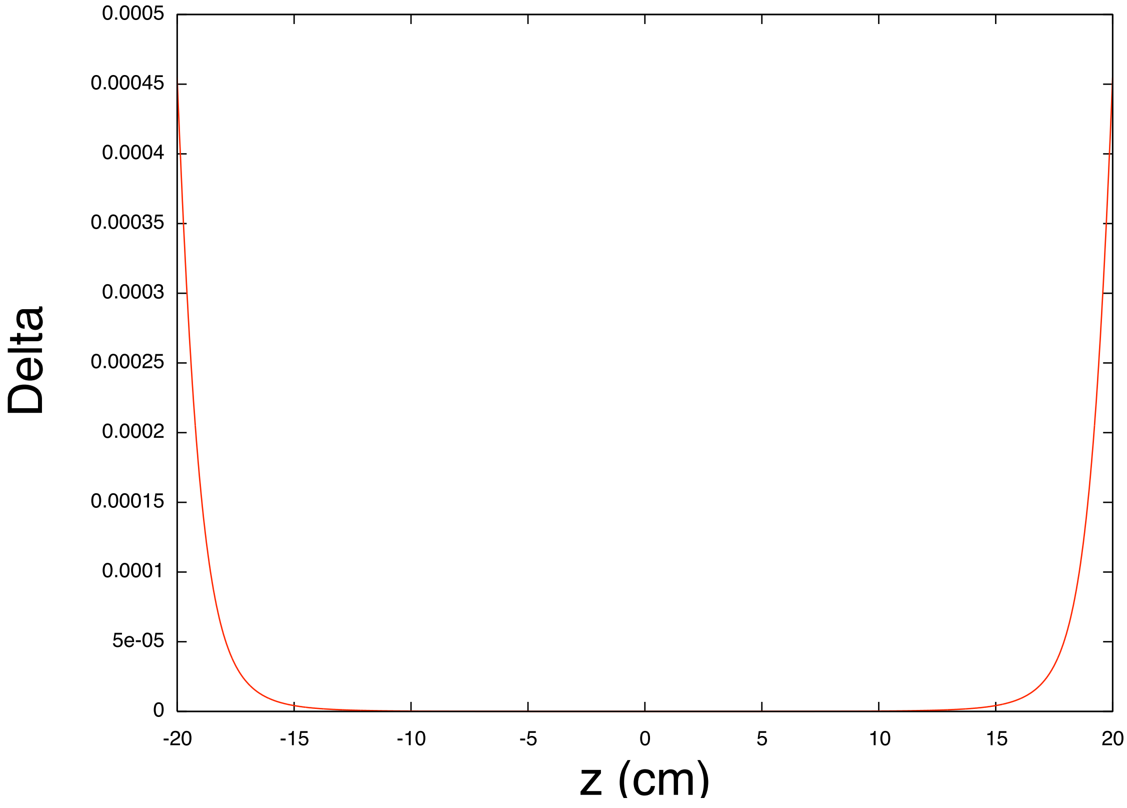


Figure 24.2.6: Place holder. The quantity $\Delta = |\boldsymbol{\Delta}|$ as a function of ξ and z in the vicinity of the design orbit and in the plane $y = 0$. The quantities ξ, y, and z are in cm.

Upon examining these figures we see that $\cdots$.

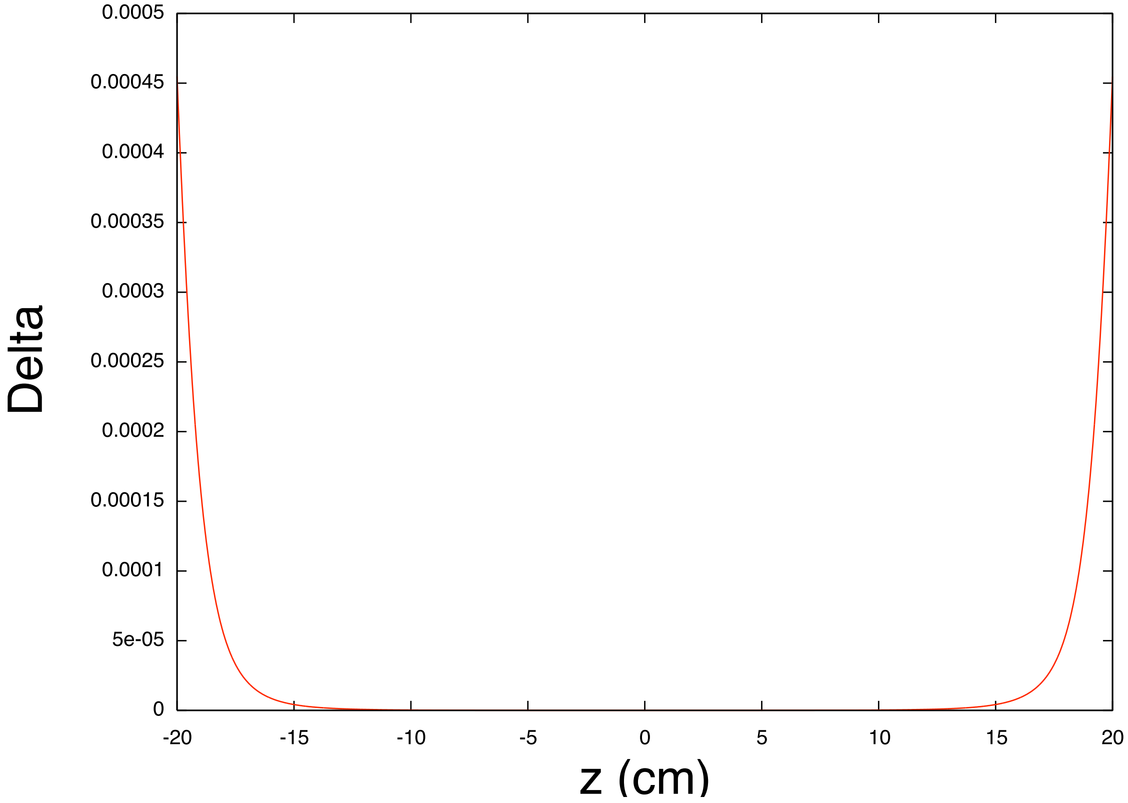


Figure 24.2.7: Place holder. The quantity $\Delta = |\boldsymbol{\Delta}|$ as a function of ξ and z in the vicinity of the design orbit and in the plane $y = 1$. The quantities ξ, y, and z are in cm

24.3 Comparison of Design Orbits

How accurate are design orbits computed using surface methods? In this subsection we will use surface methods to compute a design orbit for the case of a magnetic monopole doublet. Comparison of this design orbit with the design orbit selected in Subsection 5.1 will provide a further indication of the accuracy of surface methods.

24.4 Terminating End Fields

Let us compute the magnetic field $\boldsymbol{B}$ associated with the first two terms in (9.3). We find the result

$$\begin{aligned}\boldsymbol{B}(\boldsymbol{r};X_0,Z_0) &= -[2ga/(X_0^2+Z_0^2+a^2)^{3/2}]\boldsymbol{e}_y\\ &\quad +[6ga/(X_0^2+Z_0^2+a^2)^{5/2}](X_0x+Z_0z)\boldsymbol{e}_y\\ &\quad +[6ga/(X_0^2+Z_0^2+a^2)^{5/2}][y(X_0\boldsymbol{e}_x+Z_0\boldsymbol{e}_z)]. \qquad (24.4.1)\end{aligned}$$

Next let us find the minimum vector potential $\boldsymbol{A}^{\rm min}$ associated with the first two terms in (9.3). Begin by decomposing $\boldsymbol{B}$ into homogeneous polynomials by rewriting (9.4) in the form (2.109) with

$$\boldsymbol{B}^0(\boldsymbol{r};X_0,Z_0) = -[2ga/(X_0^2+Z_0^2+a^2)^{3/2}]\boldsymbol{e}_y \qquad (24.4.2)$$

and

$$\boldsymbol{B}^1(\boldsymbol{r};X_0,Z_0) = [6ga/(X_0^2+Z_0^2+a^2)^{5/2}][(X_0x+Z_0z)\boldsymbol{e}_y+y(X_0\boldsymbol{e}_x+Z_0\boldsymbol{e}_z)]. \qquad (24.4.3)$$

The minimum vector potential associated with this magnetic field is given by the relations (2.109) through (2.111). Working out the indicated cross products yields the results

$$\boldsymbol{A}^{\min 1}(\boldsymbol{r};X_0,Z_0) = [ga/(X_0^2+Z_0^2+a^2)^{3/2}](-z\boldsymbol{e}_x+x\boldsymbol{e}_z), \qquad (24.4.4)$$

$$\begin{gathered}\boldsymbol{A}^{\min 2}(\boldsymbol{r};X_0,Z_0) = [-2ga/(X_0^2+Z_0^2+a^2)^{5/2}]\times\\ [(Z_0y^2-Z_0z^2-X_0xz)\boldsymbol{e}_x+(X_0yz-Z_0xy)\boldsymbol{e}_y+(X_0x^2+Z_0xz-X_0y^2)\boldsymbol{e}_z].\end{gathered} \qquad (24.4.5)$$

Simple calculation verifies that there are the relations

$$\nabla\times\boldsymbol{A}^{\min 1}(\boldsymbol{r};X_0,Z_0) = \boldsymbol{B}^0(\boldsymbol{r};X_0,Z_0), \qquad (24.4.6)$$

$$\nabla\times\boldsymbol{A}^{\min 2}(\boldsymbol{r};X_0,Z_0) = \boldsymbol{B}^1(\boldsymbol{r};X_0,Z_0), \qquad (24.4.7)$$

as desired. We note that $\boldsymbol{A}^{\min 1}$ falls off as $(1/|X_0|)^3$ or $(1/|Z_0|)^3$ for large $|X_0|$ or $|Z_0|$, and $\boldsymbol{A}^{\min 2}$ falls off as $(1/|X_0|)^4$ or $(1/|Z_0|)^4$. In general, successive $\boldsymbol{A}^{\min n}$ fall off with ever increasing powers of $(1/|X_0|)$ or $(1/|Z_0|)$.

$$\boldsymbol{e}_\xi = \cos\theta\boldsymbol{e}_x - \sin\theta\boldsymbol{e}_z, \qquad (24.4.8)$$

$$\boldsymbol{e}_\eta = \boldsymbol{e}_y, \tag{24.4.9}$$

$$\boldsymbol{e}_\zeta = \sin\theta \boldsymbol{e}_x + \cos\theta e_z. \tag{24.4.10}$$

$$\boldsymbol{e}_\zeta \times \boldsymbol{e}_\xi = -\sin^2\theta(\boldsymbol{e}_x \times e_z) + \cos^2\theta(\boldsymbol{e}_z \times e_x) = \boldsymbol{e}_y = \boldsymbol{e}_\eta; \tag{24.4.11}$$

$$\boldsymbol{e}_x = \cos\theta \boldsymbol{e}_\xi + \sin\theta \boldsymbol{e}_\zeta, \tag{24.4.12}$$

$$\boldsymbol{e}_y = \boldsymbol{e}_\eta \tag{24.4.13}$$

$$\boldsymbol{e}_z = -\sin\theta \boldsymbol{e}_\xi + \cos\theta \boldsymbol{e}_\zeta. \tag{24.4.14}$$

$$\boldsymbol{r} = x\boldsymbol{e}_x + y\boldsymbol{e}_y + z\boldsymbol{e}_z = \xi\boldsymbol{e}_\xi + \eta\boldsymbol{e}_\eta + \zeta\boldsymbol{e}_\zeta. \tag{24.4.15}$$

$$\begin{aligned} x &= \boldsymbol{r} \cdot e_x = (\xi\boldsymbol{e}_\xi + \eta\boldsymbol{e}_\eta + \zeta\boldsymbol{e}_\zeta) \cdot \boldsymbol{e}_x \\ &= \xi\boldsymbol{e}_\xi \cdot \boldsymbol{e}_x + \eta\boldsymbol{e}_\eta \cdot e_x + \zeta\boldsymbol{e}_\zeta \cdot e_x \\ &= \xi\cos\theta + \zeta\sin\theta. \end{aligned} \tag{24.4.16}$$

$$\begin{aligned} y &= \boldsymbol{r} \cdot e_y = (\xi\boldsymbol{e}_\xi + \eta\boldsymbol{e}_\eta + \zeta\boldsymbol{e}_\zeta) \cdot \boldsymbol{e}_y \\ &= \xi\boldsymbol{e}_\xi \cdot \boldsymbol{e}_y + \eta\boldsymbol{e}_\eta \cdot e_y + \zeta\boldsymbol{e}_\zeta \cdot e_y \\ &= \eta. \end{aligned} \tag{24.4.17}$$

$$\begin{aligned} z &= \boldsymbol{r} \cdot e_z = (\xi\boldsymbol{e}_\xi + \eta\boldsymbol{e}_\eta + \zeta\boldsymbol{e}_\zeta) \cdot \boldsymbol{e}_z \\ &= \xi\boldsymbol{e}_\xi \cdot \boldsymbol{e}_z + \eta\boldsymbol{e}_\eta \cdot e_z + \zeta\boldsymbol{e}_\zeta \cdot e_z \\ &= -\xi\sin\theta + \zeta\cos\theta. \end{aligned} \tag{24.4.18}$$

$$\boldsymbol{B}^0(\boldsymbol{r}; X_0, Z_0) \;=\; -[2ga/(X_0^2 + Z_0^2 + a^2)^{3/2}]\boldsymbol{e}_\eta. \tag{24.4.19}$$

$$\begin{aligned} \boldsymbol{A}^{\min 1}(\xi,\eta,\zeta; X_0, Z_0) &= -(1/2)\boldsymbol{r} \times \boldsymbol{B}^0(\boldsymbol{r}; X_0, Z_0) \\ &= [ga/(X_0^2 + Z_0^2 + a^2)^{3/2}](-\zeta\boldsymbol{e}_\xi + \xi\boldsymbol{e}_\zeta), \end{aligned} \tag{24.4.20}$$

$$\boldsymbol{A}^{\min 1}(\xi,\eta,0; X_0, Z_0) \;=\; [ga/(X_0^2 + Z_0^2 + a^2)^{3/2}](\xi\boldsymbol{e}_\zeta), \tag{24.4.21}$$

$$A_\xi^{\min 1}(\xi,\eta,0; X_0, Z_0) = A_\eta^{\min 1}(\xi,\eta,0; X_0, Z_0) = 0. \tag{24.4.22}$$

$$\begin{aligned} \boldsymbol{B}^1(\boldsymbol{r}; X_0, Z_0) &= [6ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ &\quad [(X_0x + Z_0z)\boldsymbol{e}_y + y(X_0\boldsymbol{e}_x + Z_0\boldsymbol{e}_z)]. \end{aligned} \tag{24.4.23}$$

$$(X_0x + Z_0z)\boldsymbol{e}_y = [x_0(\xi\cos\theta + \zeta\sin\theta) + Z_0(-\xi\sin\theta + \zeta\cos\theta)]\boldsymbol{e}_\eta, \tag{24.4.24}$$

$$\begin{aligned} y(X_0\boldsymbol{e}_x + Z_0\boldsymbol{e}_z) &= \eta[X_0(\cos\theta\boldsymbol{e}_\xi + \sin\theta\boldsymbol{e}_\zeta) + Z_0(-\sin\theta\boldsymbol{e}_\xi + \cos\theta\boldsymbol{e}_\zeta)] \\ &= \eta[(X_0\cos\theta - Z_0\sin\theta)\boldsymbol{e}_\xi + (X_0\sin\theta + Z_0\cos\theta)\boldsymbol{e}_\zeta]. \end{aligned} \tag{24.4.25}$$

$$\begin{aligned} \boldsymbol{B}^1(\xi,\eta,\zeta;X_0,Z_0) &= [6ga/(X_0^2+Z_0^2+a^2)^{5/2}] \times \\ &\quad \{[X_0(\xi\cos\theta + \zeta\sin\theta) + Z_0(-\xi\sin\theta + \zeta\cos\theta)]\boldsymbol{e}_\eta \\ &\qquad + \eta[(X_0\cos\theta - Z_0\sin\theta)\boldsymbol{e}_\xi + (X_0\sin\theta + Z_0\cos\theta)\boldsymbol{e}_\zeta]\}. \end{aligned} \tag{24.4.26}$$

$$\begin{aligned} \boldsymbol{B}^1(\xi,\eta,0;X_0,Z_0) &= [6ga/(X_0^2+Z_0^2+a^2)^{5/2}] \times \\ &\quad \{[X_0(\xi\cos\theta) + Z_0(-\xi\sin\theta)]\boldsymbol{e}_\eta \\ &\qquad + \eta[(X_0\cos\theta - Z_0\sin\theta)\boldsymbol{e}_\xi + (X_0\sin\theta + Z_0\cos\theta)\boldsymbol{e}_\zeta]\}. \end{aligned} \tag{24.4.27}$$

$$\boldsymbol{r}(\xi,\eta,\zeta) = \xi\boldsymbol{e}_\xi + \eta\boldsymbol{e}_\eta + \zeta\boldsymbol{e}_\zeta. \tag{24.4.28}$$

$$\boldsymbol{r}(\xi,\eta,0) = \xi\boldsymbol{e}_\xi + \eta\boldsymbol{e}_\eta. \tag{24.4.29}$$

$$\begin{aligned} \boldsymbol{e}_\eta \times \boldsymbol{B}^1(\xi,\eta,0;X_0,Z_0) &= [6ga/(X_0^2+Z_0^2+a^2)^{5/2}] \times \\ &\quad \{\eta[(-X_0\cos\theta + Z_0\sin\theta)\boldsymbol{e}_\zeta + (X_0\sin\theta + Z_0\cos\theta)\boldsymbol{e}_\xi]\}. \end{aligned} \tag{24.4.30}$$

$$\begin{aligned} [\eta\boldsymbol{e}_\eta \times \boldsymbol{B}^1(\xi,\eta,0;X_0,Z_0)]_\xi &= [6ga/(X_0^2+Z_0^2+a^2)^{5/2}] \times \\ &\quad \{\eta^2[(X_0\sin\theta + Z_0\cos\theta)]\}. \end{aligned} \tag{24.4.31}$$

$$\begin{aligned} \boldsymbol{e}_\xi \times \boldsymbol{B}^1(\xi,\eta,0;X_0,Z_0) &= [6ga/(X_0^2+Z_0^2+a^2)^{5/2}] \times \\ &\quad \{[X_0(\xi\cos\theta) + Z_0(-\xi\sin\theta)]\boldsymbol{e}_\zeta \\ &\qquad - \eta[(X_0\sin\theta + Z_0\cos\theta)\boldsymbol{e}_\eta]\}. \end{aligned} \tag{24.4.32}$$

$$\begin{aligned} [\xi\boldsymbol{e}_\xi \times \boldsymbol{B}^1(\xi,\eta,0;X_0,Z_0)]_\eta &= [6ga/(X_0^2+Z_0^2+a^2)^{5/2}] \times \\ &\quad \{-\xi\eta[(X_0\sin\theta + Z_0\cos\theta)]\}. \end{aligned} \tag{24.4.33}$$

$$\boldsymbol{A}^{\min 2}(\xi,\eta,0;X_0,Z_0) \;=\; -(1/3)\boldsymbol{r}(\xi,\eta,0)\times\boldsymbol{B}^1(\xi,\eta,0;X_0,Z_0). \tag{24.4.34}$$

$$A_\xi^{\min 2}(\xi,\eta,0;X_0,Z_0) = [-2ga/(X_0^2+Z_0^2+a^2)^{5/2}](X_0\sin\theta+Z_0\cos\theta)\eta^2, \tag{24.4.35}$$

$$A_\eta^{\min 2}(\xi,\eta,0;X_0,Z_0) = [2ga/(X_0^2+Z_0^2+a^2)^{5/2}](X_0\sin\theta+Z_0\cos\theta)\xi\eta. \tag{24.4.36}$$

24.5 Gauge Transformation Map

24.6 Pole Face Rotation

24.7 Comparison of Maps

How accurate are maps computed using surface methods? In this subsection we will use surface methods to compute the transfer map about the design orbit found in Subsection 5.3. We will also compute the exact transfer map for the case of a magnetic monopole doublet. Comparison of these maps will provide a final indication of the accuracy of surface methods.

24.8 Smoothing and Insensitivity to Errors

Exercises

24.8.1. Show that any orbit having the initial conditions $Y=0$ and $P_y=0$ when $z=0$ must lie in the $y=0$ plane.

24.8.2. Show that, in the case of a 30° bend produced by a magnetic monopole doublet, one expects the asymptotic behavior

$$\lim_{z\to\mp\infty} X'(z) = \pm\tan(15^\circ) = \pm.267949\cdots. \tag{24.8.1}$$

Actually, in the numerical computations for Section 20.5.1, p^0 was chosen so that

$$X'(z=\mp 20) = \pm\tan(15^\circ) = \pm.267949\cdots. \tag{24.8.2}$$

See (5.47) and (5.48). From Figure 5.3 we observe that "asymptopia" has essentially been achieved when $|z|\geq 20$ so that the requirements (5.113) and (5.114) are nearly equivalent.

24.8.3. Consider $A_z(x,y,z)$ as given by (2.104). Under the assumption that $|x|$ increases linearly with $|z|$, as it does for large $|z|$ on the design orbit shown in Figure 5.1, find the midplane, $y=0$, asymptotic behavior of $A_z(x,y,z)$ for large $|z|$. Do the same for $B_y(x,y,z)$. Verify that the results you obtain are consistent with Figures 5.4 and 5.5.

24.8.4. Verify the parameterizations (5.80) through (5.83), (5.85) through (5.88), (5.95) through (5.97), and (5.99) through (5.101). Verify the surface elements (5.84), (5.89), (5.98), and (5.102).

Chapter 25

Realistic Transfer Maps for General Curved Beam-Line Elements: Application to a Storage-Ring Dipole

Bibliography

General References

[1] C. Mitchell, “Calculation of Realistic Charged-Particle Transfer Maps”, University of Maryland Physics Department Ph.D. Thesis (2007).

Chapter 26

Error Effects and the Euclidean Group

Chapter 27

Representations of $sp(2n)$ and Related Matters

Historically there are several mathematical groups that have been studied in detail because of their relevance to our understanding of the physical world. A detailed knowledge of the 3-dimensional rotation group is of great use in many areas including rigid body dynamics, condensed matter physics, chemistry, atomic physics, nuclear physics, and elementary particle physics. Knowledge of the rotation-translation group leads to a classification of crystals and quasicrystals. Knowledge of the Lorentz group leads to the construction of spinors, 4-vectors, general tensors, and classical fields. Knowledge of the Poincaré group (the Lorentz group plus translations in space and time) leads to a classification of elementary particles and the construction of quantum fields. An understanding of the invariants of the full group of space-time diffeomorphisms plays a role in general relativity. Knowledge of various other groups, including E_8, facilitates many-body theory calculations. Finally, there are the various "internal" and/or gauge symmetry groups that play an important role in our current understanding of elementary particles and the fundamental forces.

We have seen that the symplectic group is the underlying group for Hamiltonian systems. Yet, in contrast to most of the groups just mentioned, almost nothing is commonly known or readily available about the symplectic group. For example, many readers will be familiar with some aspects of the rotation group including spin (irreducible representations and how they are labeled) and how spins couple and combine (the Clebsch-Gordan series and coefficients for the rotation group). Yet few have heard or read about representations of the symplectic group, knowledge of its Clebsch-Gordan series is not widespread, and little is known in detail about its Clebsch-Gordan coefficients.

The purpose of this chapter is to describe some aspects of the finite-dimensional representations of the first few symplectic groups with the hope that this knowledge, like that for the well-studied groups, will also ultimately prove useful. [What we will actually be finding are the finite dimensional irreducible unitary representations of $usp(2n)$, which is equivalent to $sp(2n, \mathbb{R})$ over the complex field. See Sections 5.10 and 7.3.] Indeed, as a first consequence of this effort, we will find a symplectic classification of all analytic vector fields. Additional applications will be made in Chapters 29, 32, and 33.

27.1 Structure of $sp(2,\mathbb{R})$

The Lie algebra $sp(2,\mathbb{R})$ is generated by the Lie operators associated with the quadratic polynomials q^2, qp, and p^2. See Section 5.6. For present purposes it is convenient to introduce the basis polynomials

$$J_3 = -(i/4)(p^2+q^2) \tag{27.1.1}$$

$$J_\pm = (1/4)(q \pm ip)^2. \tag{27.1.2}$$

They obey the Poisson bracket rules

$$[J_3, J_\pm] = \pm J_\pm, \tag{27.1.3}$$

$$[J_+, J_-] = 2J_3. \tag{27.1.4}$$

These rules are the familiar ones for angular momentum, and indicate that the Lie algebras $so(3,\mathbb{R})$, $su(2)$, $sp(2,\mathbb{R})$, and $usp(2)$ are equivalent when one works over the complex field as in (1.1) and (1.2).

As a result of (7.3.14) through (7.3.16) we have the relations

$$: J_3 :^\dagger =: J_3 :, \tag{27.1.5}$$

$$: J_\pm :^\dagger =: J_\mp : . \tag{27.1.6}$$

Thus, $: J_3 :$ is Hermitian. Finally, consider Lie transformations of the form

$$\mathcal{M}(\theta) = \exp(-i\theta : J_3 :) = \exp[-(\theta/4) : p^2 + q^2 :]. \tag{27.1.7}$$

These transformations are both real symplectic and unitary. Indeed, they have the periodicity property

$$\mathcal{M}(\theta + 4\pi) = \mathcal{M}(\theta), \tag{27.1.8}$$

and therefore form a *maximal torus* in $Sp(2,\mathbb{R})$. See Sections 3.9 and 7.2.

We next bring the rules (1.3) and (1.4) to Cartan form. For a review of the Cartan form for a Lie algebra, see Section 5.8. Let $\boldsymbol{e}^1$ be a unit vector. For the case of $sp(2)$ it is convenient to introduce root vectors $\pm\boldsymbol{\alpha}$ by the relations

$$\pm\,\boldsymbol{\alpha} = \pm 2\boldsymbol{e}^1. \tag{27.1.9}$$

(Observe that they have length 2.) Then we have the normalization relation

$$\sum_{\boldsymbol{\mu}} (\boldsymbol{e}^1 \cdot \boldsymbol{\mu})(\boldsymbol{\mu} \cdot \boldsymbol{e}^1) = 8. \tag{27.1.10}$$

Compare with (5.8.21). Next introduce quantities c^1 and $r(\pm\boldsymbol{\alpha})$ by the relations

$$c^1 = 2J_3 = -(i/2)(p^2+q^2), \tag{27.1.11}$$

$$r(\pm\boldsymbol{\alpha}) = \sqrt{2}J_\pm = (\sqrt{2}/4)(q \pm ip)^2. \tag{27.1.12}$$

They evidently obey the Poisson bracket rules

$$[c^1, r(\boldsymbol{\mu})] = (\boldsymbol{e}^1 \cdot \boldsymbol{\mu}) r(\boldsymbol{\mu}), \tag{27.1.13}$$

$$[r(\boldsymbol{\mu}), r(-\boldsymbol{\mu})] = (\boldsymbol{e}^1 \cdot \boldsymbol{\mu}) c^1. \tag{27.1.14}$$

These are the Cartan rules for $sp(2)$. Note that there are the conjugacy relations

$$: c^1 :^\dagger =: c^1 :, \tag{27.1.15}$$

$$: r(\boldsymbol{\mu}) :^\dagger =: r(-\boldsymbol{\mu}) :, \tag{27.1.16}$$

and $: c^1 :$ is Hermitian as desired. The root vectors $\pm\boldsymbol{\alpha}$ are shown in Figure 1.1. Finally, we note the pleasing fact that for the scalar product (7.3.12) the basis elements c^1 and $r(\boldsymbol{\mu})$ all have unit norm and, indeed, are orthonormal:

$$\langle c^1, c^1 \rangle = \langle r(\boldsymbol{\mu}), r(\boldsymbol{\mu}) \rangle = 1, \tag{27.1.17}$$

$$\langle c^1, r(\boldsymbol{\mu}) \rangle = 0, \tag{27.1.18}$$

$$\langle r(\boldsymbol{\mu}), r(-\boldsymbol{\mu}) \rangle = 0. \tag{27.1.19}$$

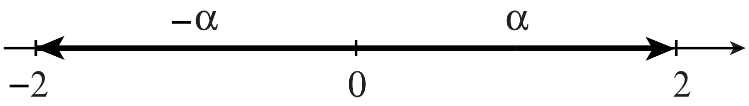


Figure 27.1.1: Root vectors for $sp(2)$.

Exercises

27.1.1. Verify (1.3) through (1.8).

27.1.2. Define J_1 and J_2 by the rules

$$J_\pm = J_1 \pm i J_2. \tag{27.1.20}$$

Verify the $su(2)$ Poisson bracket rules

$$[J_1, J_2] = i J_3, \text{ etc.} \tag{27.1.21}$$

27.1.3. Verify (1.10) and (1.13) through (1.19).

27.2 Representations of $sp(2,\mathbb{R})$

It is well known that irreducible representations of $su(2)$ are labeled by a non-negative integer or half-integer j, and vectors within a representation are labeled by an integer or half-integer m that ranges between $-j$ and j by integer steps. Let $\hat{J}_3$, $\hat{J}_\pm$ be a set of irreducible matrices (with $\hat{J}_3$ Hermitian) whose commutation rules are the same as the Poisson bracket rules (1.3) and (1.4). Then, in a representation labeled by j, there is a "highest" vector $|j\rangle$ with $m = j$ having the property

$$\hat{J}_3|j\rangle = j|j\rangle, \text{ or}$$
$$(2\hat{J}_3)|j\rangle = (2j)|j\rangle. \tag{27.2.1}$$

In the terminology of Cartan, the vector $|j\rangle$ is an eigenvector of highest weight with weight $(2j)$. [See (1.11).] It follows that the fundamental weight $\boldsymbol{\phi}^1$ for $sp(2)$ is given by the relation

$$\boldsymbol{\phi}^1 = \boldsymbol{e}^1 = \boldsymbol{\alpha}/2. \tag{27.2.2}$$

Correspondingly, the highest weight for a representation characterized by the non-negative integer n, with $n = 2j$, is given by the relation

$$\boldsymbol{w}^h = n\boldsymbol{\phi}^1\ ,\ n = 2j. \tag{27.2.3}$$

Call this representation $\Gamma(n)$.

Figure 2.1 shows the fundamental weight $\boldsymbol{\phi}^1$ along with the root vectors $\pm\boldsymbol{\alpha}$. Figure 2.2 shows the weight diagrams for the first few representations. For $su(2)$, and hence $sp(2,\mathbb{R})$, each weight (vector within a representation) has unit multiplicity. It follows that the dimension of the representation $\Gamma(n)$ is given by the relation

$$\dim\Gamma(n) = n + 1. \tag{27.2.4}$$

Sometimes we will label a representation by its dimension.

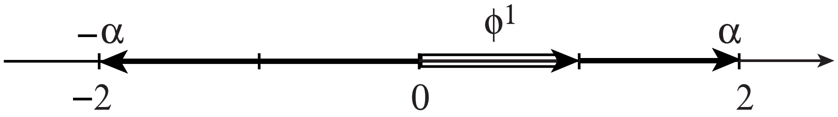


Figure 27.2.1: The fundamental weight $\boldsymbol{\phi}^1$ and the root vectors $\pm\boldsymbol{\alpha}$ for $sp(2)$.

Let $\mathcal{P}_n$ denote the set of polynomials homogeneous of degree n in the variables q, p; and let f_2 be a quadratic polynomial in q, p. Then, in view of (7.6.14), we have the relation

$$: f_2 : \mathcal{P}_n \subseteq \mathcal{P}_n. \tag{27.2.5}$$

It follows that the set of homogeneous polynomials of degree n forms a representation of $sp(2,\mathbb{R})$.

What representations occur? The study of this question is facilitated by using the map $\mathcal{A}(\theta)$ defined by the equation

$$\mathcal{A}(\theta) = \exp(-i\theta : p^2 - q^2 :). \tag{27.2.6}$$

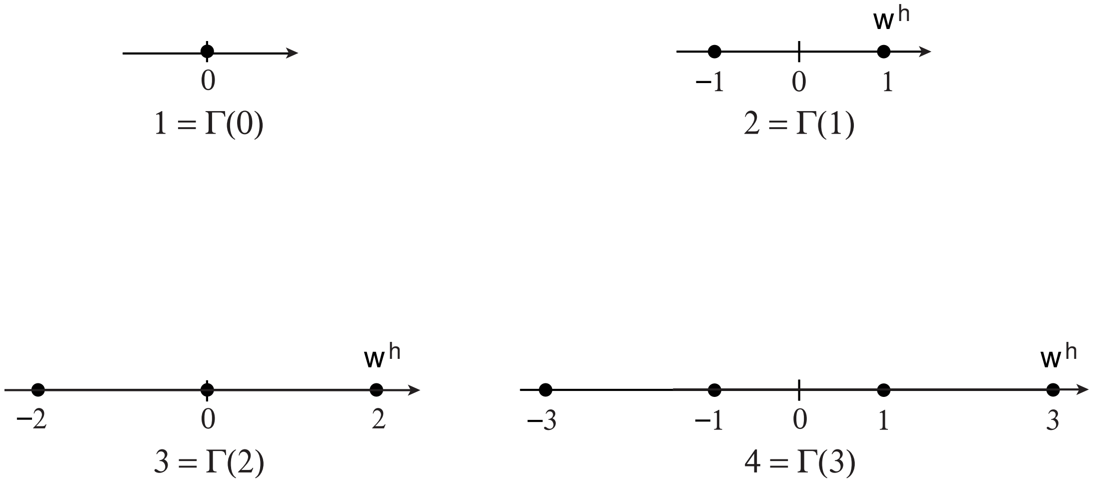


Figure 27.2.2: Weight diagrams for the $sp(2)$ repesentations $\Gamma(0)$, $\Gamma(1)$, $\Gamma(2)$, and $\Gamma(3)$.

Evidently $\mathcal{A}$ is *complex* symplectic and, in view of (7.3.14) through (7.3.16), it is also unitary. Calculation gives the results

$$\mathcal{A}(\theta)q = q\cos(2\theta) + ip\sin(2\theta), \tag{27.2.7}$$

$$\mathcal{A}(\theta)p = iq\sin(2\theta) + p\cos(2\theta). \tag{27.2.8}$$

In particular, there is the relation

$$\mathcal{A}(\pi/8)q = (1/\sqrt{2})(q + ip), \tag{27.2.9}$$

$$\mathcal{A}(\pi/8)p = i(1/\sqrt{2})(q - ip). \tag{27.2.10}$$

The map $\mathcal{A}(\pi/8)$ is the operator analog of the matrix W given by (3.9.9). See Exercise 2.2. Also, since $\mathcal{A}$ is a Lie transformation, there is the relation

$$\mathcal{A}(\pi/8)(q^r p^s) = [\mathcal{A}(\pi/8)q]^r[\mathcal{A}(\pi/8)p]^s = (1/\sqrt{2})^{r+s}(i)^s(q + ip)^r(q - ip)^s. \tag{27.2.11}$$

With the aid of $\mathcal{A}$, we define transformed Lie basis polynomials $\tilde{c}^1$, $\tilde{r}(\pm\boldsymbol{\alpha})$, by the rule

$$\tilde{c}^1 = \mathcal{A}(-\pi/8)c^1 = -qp, \tag{27.2.12}$$

$$\tilde{r}(\boldsymbol{\alpha}) = \mathcal{A}(-\pi/8)r(\boldsymbol{\alpha}) = (1/\sqrt{2})q^2, \tag{27.2.13}$$

$$\tilde{r}(-\boldsymbol{\alpha}) = \mathcal{A}(-\pi/8)r(-\boldsymbol{\alpha}) = -(1/\sqrt{2})p^2. \tag{27.2.14}$$

Since $\mathcal{A}$ is symplectic, the transformed basis polynomials obey the same Poisson bracket rules (1.13) and (1.14). See (5.4.14) and Section 6.3. Since $\mathcal{A}$ is also unitary, as is easily verified, the transformed basis polynomials also satisfy the orthonormality relations (1.17) through (1.19) and the conjugacy relations (1.15) and (1.16).

Now consider the action of the Lie operators $:\tilde{c}^1:$ and $:\tilde{r}(\pm\boldsymbol{\alpha}):$ on the general monomial $q^r p^s$. Calculation gives the results

$$:\tilde{c}^1: q^r p^s = (r - s)q^r p^s, \tag{27.2.15}$$

$$:\tilde{r}(\boldsymbol{\alpha}): q^r p^s = (\sqrt{2}) s q^{r+1} p^{s-1}, \tag{27.2.16}$$

$$:\tilde{r}(-\boldsymbol{\alpha}): q^r p^s = (\sqrt{2}) r q^{r-1} p^{s+1}. \tag{27.2.17}$$

Evidently any monomial of a given degree can be transformed into any other monomial of the same degree with the aid of $\tilde{r}(\pm\boldsymbol{\alpha})$. Therefore $sp(2)$ acts *irreducibly* on $\mathcal{P}_n$. Also q^n is the vector of highest weight in $\mathcal{P}_n$, and has the weight $n\boldsymbol{\phi}^1$,

$$:\tilde{c}^1: q^n = nq^n = (\boldsymbol{e}^1 \cdot n\boldsymbol{\phi}^1) q^n. \tag{27.2.18}$$

We conclude that $\mathcal{P}_n$ carries the representation $\Gamma(n)$. From (7.3.36) and (2.4) we find the result

$$\dim \mathcal{P}_n = N(n,2) = n+1 = \dim \Gamma(n). \tag{27.2.19}$$

The equality of $\dim \mathcal{P}_n$ and $\dim \Gamma(n)$ is to be expected from the fact that $sp(2)$ acts irreducibly on $\mathcal{P}_n$. It can be shown that $\Gamma(n)$ is self conjugate,

$$\overline{\Gamma}(n) = \Gamma(n). \tag{27.2.20}$$

See Exercise 3.7.36.

Let $\mathcal{A}(\pi/8)$ act on both sides of (2.15) through (2.17). Then, for the left side of (2.15), we find the result

$$\begin{aligned} \mathcal{A}(\pi/8): \tilde{c}^1 : q^r p^s &= \mathcal{A}(\pi/8)[\tilde{c}^1, q^r p^s] = [\mathcal{A}(\pi/8)\tilde{c}^1, \mathcal{A}(\pi/8) q^r p^s] \\ &= [c^1, \mathcal{A}(\pi/8) q^r p^s] =: c^1 : (1/\sqrt{2})^{r+s} (i)^s (q+ip)^r (q-ip)^s. \end{aligned} \tag{27.2.21}$$

For the right side we find

$$\mathcal{A}(\pi/8)(r-s) q^r p^s = (r-s)(1/\sqrt{2})^{r+s} (i)^s (q+ip)^r (q-ip)^s. \tag{27.2.22}$$

Therefore, after cancellation of common terms, (2.15) is transformed under the action of $\mathcal{A}(\pi/8)$ to the relation

$$:(p^2+q^2)/2: (q+ip)^r (q-ip)^s = i(r-s)(q+ip)^r (q-ip)^s. \tag{27.2.23}$$

Similary, (2.16) and (2.17) are transformed to the relations

$$:(q+ip)^2: (q+ip)^r (q-ip)^s = -4is(q+ip)^{r+1}(q-ip)^{s-1}, \tag{27.2.24}$$

$$:(q-ip)^2: (q+ip)^r (q-ip)^s = 4ir(q+ip)^{r-1}(q-ip)^{s+1}. \tag{27.2.25}$$

The monomials $q^r p^s$ (with $r+s=n$) obviously form a basis for $\mathcal{P}_n$. The relations (2.9) through (2.11) show that the complex polynomials $(q+ip)^r(q-ip)^s$ also form a basis for $\mathcal{P}_n$ and, with the factors $(1/\sqrt{2})^n(i^s)$, the two bases are related by the symplectic and unitary transformations $\mathcal{A}(\pm\pi/8)$. According to (2.23) the polynomials $(q+ip)^r(q-ip)^s$ are eigenfunctions of the harmonic oscillator Lie operator $:(p^2+q^2)/2:$. For this reason, they will be referred to as the *resonance* basis. The utility of the resonance basis will become clear in Chapter 23. Since it is made from Cartesian components, the monomial basis $q^r p^s$ will be referred to as the *Cartesian* basis.

Exercises

27.2.1. The Lie algebras for $su(2)$ and $sp(2)$ are equivalent over the complex field. Yet, for purposes of observing how $u(n)$ is embedded within $sp(2n)$, it is convenient to give $su(2)$ and $sp(2)$ different root vector structures. Specifically, for $su(2)$ we define two root vectors $\pm\boldsymbol{\alpha}$ by the rules

$$\pm\,\boldsymbol{\alpha} = \pm\sqrt{2}\boldsymbol{e}^1. \tag{27.2.26}$$

(Observe they have length $\sqrt{2}$). Then the $su(2)$ Lie algebra is spanned by the elements c^1, $r(\boldsymbol{\alpha})$, and $r(-\boldsymbol{\alpha})$; and the Cartan rules (1.13) and (1.14) give the results

$$[c^1, r(\pm\boldsymbol{\alpha})] = \{\boldsymbol{e}^1\cdot(\pm\boldsymbol{\alpha})\}r(\pm\boldsymbol{\alpha}) = \pm\sqrt{2}r(\pm\boldsymbol{\alpha}), \tag{27.2.27}$$

$$[r(\boldsymbol{\alpha}), r(-\boldsymbol{\alpha})] = (\boldsymbol{e}^1\cdot\boldsymbol{\alpha})c^1 = \sqrt{2}c^1. \tag{27.2.28}$$

Consider the $su(2)$ within $sp(4)$ as described in Section 5.7. Upon making the identifications $c \leftrightarrow c^1$ and $r(\pm) \leftrightarrow r(\pm\boldsymbol{\alpha})$, verify that the rules (5.7.10) and (5.7.11) are identical to (2.27) and (2.28). Show that C^1 and $R(\pm\boldsymbol{\alpha})$ as given by (5.8.8) and (5.8.11) satisfy analogous commutation rules, and therefore describe one of the $su(2)$ subgroups within $su(3)$. Show that there are two other $su(2)$ subgroups within $su(3)$ corresponding to the use of $R(\pm\boldsymbol{\beta})$ and $R(\pm\boldsymbol{\gamma})$ and suitable linear combinations of the C^j. Note that all the $su(3)$ root vectors in Figure 5.8.1 have length $\sqrt{2}$. With regard to representations of $su(2)$, call them $\Gamma(n)$, show that there is one such for each n value with $n = 0, 1, 2, \cdots$. Show that the highest weight for $\Gamma(n)$ is given by

$$\boldsymbol{w}^h = n\boldsymbol{\phi}^1 \tag{27.2.29}$$

with

$$\boldsymbol{\phi}^1 = (1/\sqrt{2})\boldsymbol{e}^1. \tag{27.2.30}$$

Draw $su(2)$ weight diagrams for the first few representations. Verify that they are similar to those for $sp(2)$, see Figure 2.2, except that the spacing between dots is $\sqrt{2}$ rather than 2. Examine the weight diagrams for $su(3)$ as shown in Figures 5.8.3 through 5.8.8. Show that the spacing between the dots in the directions of the $su(3)$ root vectors is $\sqrt{2}$. These dots describe $su(2)$ representations within $su(3)$.

27.2.2. For a $2n$-dimensional phase space, let $\mathcal{A}(\pi/8)$ be the map defined by the equation

$$\mathcal{A}(\pi/8) = \exp[-i(\pi/8) : (p_1^2 - q_1^2) + (p_2^2 - q_2^2) + \cdots + (p_n^2 - q_n^2) :]. \tag{27.2.31}$$

Show that $\mathcal{A}(\pi/8)$ has the property

$$\mathcal{A}(\pi/8)z_a = \sum_b W_{ab}z_b \tag{27.2.32}$$

where W is the matrix given by (3.9.9).

27.2.3. For first-order polynomials, and in analogy with (2.9) and (2.10), introduce the basis elements $a^\pm$ defined by the relations

$$a^+ = (1/\sqrt{2})(p + iq) = \mathcal{A}(\pi/8)p, \tag{27.2.33}$$

$$a^- = (1/\sqrt{2})(p - iq) = -i\mathcal{A}(\pi/8)q. \tag{27.2.34}$$

Show that these elements satisfy the Poisson bracket relations

$$[a^+, a^-] = \mathcal{A}(\pi/8)[p, -iq] = i. \tag{27.2.35}$$

Let H be the harmonic oscillator Hamiltonian

$$H = (\omega/2)(p^2 + q^2). \tag{27.2.36}$$

Show that H can be written in the form

$$H = \omega a^+ a^-. \tag{27.2.37}$$

Use (2.35) to verify the equations of motion

$$\dot{a}^\pm = [a^\pm, H] = \omega[a^\pm, a^+ a^-] = \pm i\omega a^\pm. \tag{27.2.38}$$

Show that they have the solution

$$a^\pm(t) = a^\pm(0)\exp(\pm i\omega t). \tag{27.2.39}$$

From this result, find $q(t)$ and $p(t)$.

27.3 Symplectic Classification of Analytic Vector Fields in Two Variables

Let $\mathcal{L}_{\boldsymbol{f}}$ be a general vector field in two variables z_1 and z_2 where $\boldsymbol{f}$ denotes the collection of 2 functions (f_1, f_2) as in Section 5.3. (The functions f_1 and f_2 may also depend on the time t, but for simplicity we will suppress this possible dependence in our notation because t only plays the role of a parameter.) Assume that f_1 and f_2 are analytic at some common point z_1^0, z_2^0. Without loss of generality we may take this point to be the origin. (If not, make a linear change of variables that sends z_1^0, z_2^0 to the origin.) Then we may decompose the Taylor expansions of the components of $\boldsymbol{f}$ into sums of homogeneous polynomials and, in so doing, decompose $\mathcal{L}_{\boldsymbol{f}}$ into a sum of vector fields of the form $\mathcal{L}_{\boldsymbol{f}^n}$ where the components of $\boldsymbol{f}^n$ are homogeneous polynomials of degree n:

$$\mathcal{L}_{\boldsymbol{f}} = \sum_{n=0}^{\infty} \mathcal{L}_{\boldsymbol{f}^n}. \tag{27.3.1}$$

The homogeneous vector fields $\mathcal{L}_{\boldsymbol{f}^n}$ can now be considered individually.

Let Σ denote the vector field

$$\Sigma = \sum_a z_a(\partial/\partial z_a) = z_1(\partial/\partial z_1) + z_2(\partial/\partial z_2) = q(\partial/\partial q) + p(\partial/\partial p). \tag{27.3.2}$$

Then, by Euler's relation, we have the result

$$\#\Sigma\#\mathcal{L}_{\boldsymbol{f}^n} = \{\Sigma, \mathcal{L}_{\boldsymbol{f}^n}\} = (n-1)\mathcal{L}_{\boldsymbol{f}^n}. \tag{27.3.3}$$

See Exercises 1.5.1 and 7.6.7. (Remark: Sometimes Σ is called the *Euler* field because of its connection with the Euler relation.) We will say that $\mathcal{L}_{\boldsymbol{f}^n}$ is homogeneous of degree $(n-1)$. In view of (5.3.3) and (5.3.17), in the special case of a Hamiltonian vector field $: f_n :$ there is the result

$$\#\Sigma\# : f_n := \{\Sigma, : f_n :\} = (n-2) : f_n : . \tag{27.3.4}$$

Thus, the vector field $: f_n :$ is homogeneous of degree $(n-2)$. Finally it is easily verified that there is a grading relation of the form

$$\{\mathcal{L}_{\boldsymbol{f}^\ell}, \mathcal{L}_{\boldsymbol{g}^m}\} = \mathcal{L}_{\boldsymbol{h}^n} \text{ with } n = \ell + m - 1. \tag{27.3.5}$$

Let f_2 be a quadratic polynomial in q, p. Then, using (3.5), we have the relation

$$\#f_2\#\mathcal{L}_{\boldsymbol{g}^m} = \{: f_2 :, \mathcal{L}_{\boldsymbol{g}^m}\} = \mathcal{L}_{\boldsymbol{h}^m}. \tag{27.3.6}$$

We draw the important conclusion that the set of homogeneous vector fields $\mathcal{L}_{\boldsymbol{g}^m}$ transforms under and forms a representation of $sp(2, \mathbb{R})$.

What irreducible representations occur? Consider first the case of Hamiltonian vector fields. In this case

$$\#f_2\# : g_m := \{: f_2 :, : g_m :\} =: [f_2, g_m] :=: (: f_2 : g_m) : . \tag{27.3.7}$$

It follows from the previous section that the Hamiltonian vector fields $: g_m :$ are transformed into each other under the action of $sp(2, \mathbb{R})$ and carry the irreducible representation $\Gamma(m)$.

What about general vector fields? Note that

$$: q := \partial/\partial p \;\text{ and }\; : p := -\partial/\partial q. \tag{27.3.8}$$

It follows that any $\mathcal{L}_{\boldsymbol{g}^0}$ is a Hamiltonian vector field, and these fields carry the representation $\Gamma(1)$. Next consider the vector fields $\mathcal{L}_{\boldsymbol{g}^1}$. They evidently form a 4-dimensional space spanned by the vector fields $z_a(\partial/\partial z_b)$ with $a = 1, 2$ and $b = 1, 2$.[1] We know that any $: h_2 :$ is such a vector field, and that these vector fields carry the representation $\Gamma(2)$, which is 3 dimensional. Also, Σ is of the form $\mathcal{L}_{\boldsymbol{g}^1}$. From (3.4) we conclude that

$$\#f_2\#\Sigma = \{: f_2 :, \Sigma\} = -\{\Sigma, : f_2 :\} = -\#\Sigma\# : f_2 := 0. \tag{27.3.9}$$

Consequently, Σ carries the representation $\Gamma(0)$. It follows that any $\mathcal{L}_{\boldsymbol{g}^1}$ can be written *uniquely* in the form

$$\begin{aligned} \mathcal{L}_{\boldsymbol{g}^1} &= \lambda_1 : q^2 : +\lambda_2 : qp : +\lambda_3 : p^2 : +\lambda_4\Sigma =: \lambda_1 q^2 + \lambda_2 qp + \lambda_3 p^2 : +\lambda_4\Sigma \\ &= : h_2 : +\lambda_4\Sigma. \end{aligned} \tag{27.3.10}$$

[1] Let $z_1, z_2, \cdots z_m$ be m variables. Consider the m^2 vector fields $z_a(\partial/\partial z_b)$. They can be shown to form a basis for the Lie algebra $g\ell(m)$. See Exercise 10.8.

The term $: h_2 :$ is a member of the representation $\Gamma(2)$, and the term $\lambda_4\Sigma$ belongs to $\Gamma(0)$. That is, the vector fields $\mathcal{L}_{\boldsymbol{g}^1}$ carry as a direct sum the representations $\Gamma(2)$ and $\Gamma(0)$. Note that

$$\dim \mathcal{L}_{\boldsymbol{g}^1} = 4 = 3 + 1 = \dim \Gamma(2) + \dim \Gamma(0), \tag{27.3.11}$$

as required.

With this background in mind, let us consider the general case $\mathcal{L}_{\boldsymbol{g}^m}$ with $m \geq 1$. Any such vector field can be written in the form

$$\mathcal{L}_{\boldsymbol{g}^m} = \sum_{a=1}^{2} g_a^m (\partial/\partial z_a) \tag{27.3.12}$$

where g_1^m and g_2^m denote two homogeneous polynomials of degree m. We have just learned that the $(\partial/\partial z_a)$ carry the representation $\Gamma(1)$, and we know from the previous section that the g_a^m carry the representation $\Gamma(m)$. It follows from the derivation property of $\#f_2\#$ that $\mathcal{L}_{\boldsymbol{g}^m}$ must carry the direct product representation $\Gamma(m) \otimes \Gamma(1)$. See Exercise 3.2. Also, in the case of $sp(2)$, we have the Clebsch-Gordan series result

$$\Gamma(m) \otimes \Gamma(1) = \Gamma(m+1) \oplus \Gamma(m-1). \tag{27.3.13}$$

This is just the $sp(2)$ analog of the familiar statement that spin $m/2$ and spin $1/2$ combine to make spin $(m+1)/2$ and spin $(m-1)/2$. Recall (5.8.33) and remember that for the purposes of the present section and previous section we have made the definition $n = 2j$.

It follows that any $\mathcal{L}_{\boldsymbol{g}^m}$ with $m \geq 1$ has the unique decomposition

$$\mathcal{L}_{\boldsymbol{g}^m} =: h_{m+1} : + \mathcal{G}^{m-1}. \tag{27.3.14}$$

Here h_{m+1} is a unique homogeneous polynomial of degree $m+1$ that is a member of the representation $\Gamma(m+1)$ and $\mathcal{G}^{m-1}$ is a unique vector field homogeneous of degree $(m-1)$ that is a member of the representation $\Gamma(m-1)$. Let us introduce the notation

$$\mathcal{H}^{m+1} =: h_{m+1} : \tag{27.3.15}$$

to denote a Hamiltonian vector field that carries the representation $\Gamma(m+1)$. Then (3.13) can be written in the form

$$\mathcal{L}_{\boldsymbol{g}^m} = \mathcal{H}^{m+1} + \ \mathcal{G}^{m-1}. \tag{27.3.16}$$

We *define* $\mathcal{G}^{m-1}$ to be the *non-Hamiltonian* part of $\mathcal{L}_{\boldsymbol{g}^m}$. What we have learned is that any homogeneous polynomial vector field in two variables can be uniquely decomposed into a Hamiltonian and a non-Hamiltonian part. We will learn subsequently that this result holds in any (even) number of variables.

In the case of two variables there is an additional step that can be made. Consider any vector field of the form $f_{m-1}\Sigma$. In view of (3.9) this vector field is a member of the representation $\Gamma(m-1)$. Thus, in the case of two variables we may write

$$\mathcal{L}_{\boldsymbol{g}^m} = : h_{m+1} : + f_{m-1}\Sigma \tag{27.3.17}$$

where both h_{m+1} and f_{m-1} are uniquely determined.

As a simple, but instructive, example of the decomposition just described, consider the case of the damped harmonic oscillator described by the equation of motion

$$\ddot{q} + 2\beta\dot{q} + q = 0. \tag{27.3.18}$$

This equation can be rewritten in the first-order form

$$\dot{q} = p, \tag{27.3.19}$$

$$\dot{p} = -(q + 2\beta p). \tag{27.3.20}$$

These equations can next be expressed in the Lie form

$$\dot{q} = \mathcal{L}q,\ \dot{p} = \mathcal{L}p \tag{27.3.21}$$

where $\mathcal{L}$ is the vector field

$$\mathcal{L} = p(\partial/\partial q) - (q + 2\beta p)(\partial/\partial p). \tag{27.3.22}$$

Evidently $\mathcal{L}$ is of the form (3.10). By comparing coefficients we find the decomposition

$$\mathcal{L} = : -(p^2 + 2\beta pq + q^2)/2 : -\beta\Sigma. \tag{27.3.23}$$

It is easily verified that the Hamiltonian $(p^2+2\beta pq+q^2)/2$ produces simple harmonic motion with a frequency ω_1 given by the relation

$$\omega_1^2 = 1 - \beta^2. \tag{27.3.24}$$

Also, in this case, the vector fields $\mathcal{H}^2$ and $\mathcal{G}^0$ commute. Since the vector field $-\beta\Sigma$ produces exponential decay, it follows that the general solution to (3.18) is of the form

$$q = Ae^{-\beta t}\sin(\omega_1 t + \phi). \tag{27.3.25}$$

What we have learned is that damping contributes both a non-Hamiltonian and a Hamiltonian part to the vector field. The non-Hamiltonian part produces exponential decay, and the Hamiltonian part shifts the frequency. For further detail, see Exercises 3.7 through 3.10.

Exercises

27.3.1. A Lie algebra is called *simple* if it has no invariant subalgebras (ideals). See Section 8.9. Show that the Lie algebra $su(2)$, and hence also $sp(2)$ and $so(3)$, is simple. Show that $su(3)$ is simple. What are the ranks of $sp(2)$, $so(3)$, $su(2)$, and $su(3)$? See Section 5.8.

27.3.2. Show that $\#f_2\#$ has the derivation property

$$\#f_2\# \sum_a g_a(\partial/\partial z_a) = \sum_a (: f_2 : g_a)(\partial/\partial z_a) + \sum_a g_a\#f_2\#(\partial/\partial z_a). \tag{27.3.26}$$

27.3.3. Compare the dimensions of both sides of (3.13).

27.3.4. Use (5.3.26), and the discussion surrounding it, as well as (3.16) and (3.17) to show that $\mathcal{G}^{m-1}$ is non-Hamiltonian.

27.3.5. Show that Σ can be written in the form

$$\Sigma = -\sum_{a,b} z_a J_{ab} : z_b : . \tag{27.3.27}$$

We know that both the z_a and the $: z_b :$ transform according to $\Gamma(1)$. But Σ carries the representation $\Gamma(0)$. It follows that the numbers J_{ab} are the Clebsch-Gordan *coefficients* that couple $\Gamma(1) \otimes \Gamma(1)'$ down to $\Gamma(0)$.

27.3.6. Consider the vector $\mathcal{L}_{\boldsymbol{g}^2}$ given by

$$\mathcal{L}_{\boldsymbol{g}^2} = q^2(\partial/\partial q). \tag{27.3.28}$$

Find h_3 and f_1 as in (3.15) for this vector field.

27.3.7. Verify (3.18) through (3.23). Verify that the vector field $-\beta\Sigma$ produces exponential decay,

$$e^{-t\beta\Sigma} q = e^{-\beta t} q, \tag{27.3.29}$$

$$e^{-t\beta\Sigma} p = e^{-\beta t} p. \tag{27.3.30}$$

For the Hamiltonian

$$H = (p^2 + 2\beta pq + q^2)/2 \tag{27.3.31}$$

make the transformation of variables

$$q = \frac{1}{\sqrt{2}}(Q - P), \tag{27.3.32}$$

$$p = \frac{1}{\sqrt{2}}(Q + P). \tag{27.3.33}$$

Verify that this transformation is symplectic, and hence H is transformed to H' with

$$H' = (1/2)[(1-\beta)P^2 + (1+\beta)Q^2]. \tag{27.3.34}$$

Also, show that Σ is unchanged by this transformation,

$$q(\partial/\partial q) + p(\partial/\partial p) = Q(\partial/\partial Q) + P(\partial/\partial P). \tag{27.3.35}$$

Next make a second transformation of variables,

$$Q = [(1-\beta)/(1+\beta)]^{1/4}\overline{q}. \tag{27.3.36}$$

$$P = [(1+\beta)/(1-\beta)]^{1/4}\overline{p}. \tag{27.3.37}$$

Verify that this transformation is also symplectic, and hence H' is transformed to H'' with

$$H'' = (1/2)(1-\beta^2)^{1/2}(\overline{p}^2 + \overline{q}^2) = (\omega_1/2)(\overline{p}^2 + \overline{q}^2). \tag{27.3.38}$$

Again show that Σ is unchanged,

$$Q(\partial/\partial Q) + P(\partial/\partial P) = \overline{q}(\partial/\partial\overline{q}) + \overline{p}(\partial/\partial\overline{p}). \tag{27.3.39}$$

(This is the *second* time that Σ is unchanged. Why must this be? See Exercise 9.4.) Evidently, in accord with previous claims, H'' produces simple harmonic motion with frequency ω_1. And, since Σ is unchanged, the new variables still exhibit the same exponential decay as that in (3.29) and (3.30).

27.3.8. The oscillator described by (3.18) is underdamped when $\beta < 1$, critically damped when $\beta = 1$, and overdamped when $\beta > 1$. Exercise 3.6 deals with the underdamped case. Carry out a smiliar analysis for the critically and overdamped cases. Hint: For the overdamped case, H' as given by (3.34) produces *hyperbolic* motion. In this case, make the transformation of variables

$$Q = [(\beta - 1)/(\beta + 1)]^{1/4}\overline{q}, \tag{27.3.40}$$

$$P = [(\beta + 1)/(\beta - 1)]^{1/4}\overline{p}. \tag{27.3.41}$$

Verify that this transformation is symplectic and hence H' is transformed to H'' with

$$H'' = (1/2)(\beta^2 - 1)^{1/2}(-\overline{p}^2 + \overline{q}^2). \tag{27.3.42}$$

Show that H'' produces growth that goes like $\exp[t(\beta^2 - 1)^{1/2}]$ as well as decay that goes as $\exp[-t(\beta^2 - 1)^{1/2}]$. For large β the growth rate of the growing term is almost as large as the decay rate in (3.29) and (3.30). They therefore nearly cancel. The net and well know result is that it takes a very long time for a highly overdamped oscillator to come to rest.

27.3.9. Find a pair of differential equations of the form

$$\dot{\overline{q}} = \cdots ,$$

$$\dot{\overline{p}} = \cdots ,$$

by expressing q, p in terms of $\overline{q}, \overline{p}$ with the aid of (3.32) through (3.37) and using (3.19) and (3.20). Find the vector field for these differential equations and decompose it into Hamiltonian and non-Hamiltonian parts. Solve the differential equations.

27.3.10. Show that H as given by (3.31) and $\mathcal{L}$ as given by (3.23) have the property

$$\mathcal{L}H = -\beta\Sigma H = -2\beta H. \tag{27.3.43}$$

Therefore, H must evolve according to the *nonoscillatory* rule

$$H = \text{ (constant) } \times e^{-2\beta t}. \tag{27.3.44}$$

Verify directly from (3.19) and (3.25) that (3.44) is, in fact, correct.

27.3.11. Let G be the function

$$G = az_1^2 + bz_1z_2 + cz_2^2 = aq^2 + bqp + cp^2. \tag{27.3.45}$$

Find the associated gradient vector field $\mathcal{L}_G$. See Exercise 5.3.7. Decompose $\mathcal{L}_G$ into Hamiltonian and non-Hamiltonian parts. Find G such that

$$\mathcal{L}_G = \Sigma. \tag{27.3.46}$$

Can a gradient vector field ever also be a Hamiltonian vector field?

27.3.12. The Van der Pol oscillator is described by the differential equation

$$\ddot{q} - 2\lambda(1 - q^2)\dot{q} + q = 0 \tag{27.3.47}$$

with $\lambda > 0$. Upon making the definition $p = \dot{q}$, show that (3.47) is produced by the vector field

$$\mathcal{L} = p(\partial/\partial q) - (q - 2\lambda p)(\partial/\partial p) - 2\lambda q^2 p(\partial/\partial p). \tag{27.3.48}$$

Evidently $\mathcal{L}$ has the homogeneous decomposition

$$\mathcal{L} = \mathcal{L}_{\boldsymbol{g}^1} + \mathcal{L}_{\boldsymbol{g}^3} \tag{27.3.49}$$

where

$$\mathcal{L}_{\boldsymbol{g}^1} = p(\partial/\partial q) - (q - 2\lambda p)(\partial/\partial p), \tag{27.3.50}$$

$$\mathcal{L}_{\boldsymbol{g}^3} = -2\lambda q^2 p(\partial/\partial p). \tag{27.3.51}$$

Verify that these homogeneous vector fields in turn have the decompositions

$$\mathcal{L}_{\boldsymbol{g}^1} =: h_2 : + \mathcal{G}^0 \tag{27.3.52}$$

with

$$h_2 = -(p^2 - 2\lambda pq + q^2)/2, \tag{27.3.53}$$

$$\mathcal{G}^0 = \lambda[q(\partial/\partial q) + p(\partial/\partial p)] = \lambda\Sigma; \tag{27.3.54}$$

$$\mathcal{L}_{\boldsymbol{g}^3} =: h_4 : + \mathcal{G}^2, \tag{27.3.55}$$

with

$$h_4 = -(\lambda/2)q^3 p, \tag{27.3.56}$$

$$\mathcal{G}^2 = -(\lambda/2)q^2\Sigma. \tag{27.3.57}$$

Show from (3.54) that the solution $q = 0$ is unstable for $\lambda > 0$. Argue that for small λ the solutions to (3.47) should be nearly those for a simple harmonic oscillator, i.e., any circle in q, p phase space. Show from energy considerations that for small λ the Van der Pol oscillator should have a limit cycle that is nearly a circle in phase space (about the origin) of radius 2, which is indeed the case. Observe that

$$\mathcal{G}^0 + \mathcal{G}^2 = \lambda(1 - q^2/2)\Sigma. \tag{27.3.58}$$

Show that for a solution of the form $q = A\sin(t + \phi)$ there is the relation

$$\langle(1 - q^2/2)\rangle = 1 - A^2/4 \tag{27.3.59}$$

where $\langle\ \rangle$ denotes time averaging. Show that there is the general operator relation

$$\{: f_2 :, (\mathcal{G}^0 + \mathcal{G}^2)\} = \lambda[: f_2 : (1 - q^2/2)]\Sigma. \tag{27.3.60}$$

By considering an operator of the form

$$\exp(-t : f_2 :)(\mathcal{G}^0 + \mathcal{G}^2)\exp(t : f_2 :), \tag{27.3.61}$$

use (3.60) to find the relation

$$\langle\mathcal{G}^0 + \mathcal{G}^2\rangle = \lambda\langle(1 - q^2/2)\rangle\Sigma. \tag{27.3.62}$$

Thus, on the limit cycle, the growth/damping due to $(\mathcal{G}^0 + \mathcal{G}^2)$ averages to zero. Consider what appears to be a generalization of the Van der Pol oscillator described by the equation

$$\ddot{q} - 2\lambda\dot{q} + 2\tau q^2\dot{q} + q = 0 \tag{27.3.63}$$

where λ and τ are positive. Verify that (3.63) can be brought to the form (3.47) by a suitable scaling of q. Verify, for small λ and τ, that (3.63) has a nearly circular limit cycle in phase space whose radius is given by the relation

$$A = 2\sqrt{(\lambda/\tau)}. \tag{27.3.64}$$

27.3.13. Suppose that f is an analytic function of the complex variable $z = x + iy$, and write the relations

$$w = f(z) \tag{27.3.65}$$

and

$$w = u(x, y) + iv(x, y). \tag{27.3.66}$$

Then, because f is assumed analytic, u and v satisfy the Cauchy-Riemann equations

$$\partial u/\partial x = \partial v/\partial y, \tag{27.3.67}$$

$$\partial u/\partial y = -\partial v/\partial x. \tag{27.3.68}$$

In terms of the phase-space variables $\{q, p\}$, consider the differential form

$$v(q, p)dq + u(q, p)dp. \tag{27.3.69}$$

According to Exercise 6.4.6, this form will be exact if there is the relation

$$\partial v/\partial p = \partial u/\partial q. \tag{27.3.70}$$

From the Cauchy-Riemann equation (3.67) we see that the relation (3.70) is in fact true, and therefore there is a function H such that

$$u(q, p) = \partial H/\partial p, \tag{27.3.71}$$

$$v(q,p) = \partial H/\partial q. \tag{27.3.72}$$

We know that any Hamiltonian gives rise to the Hamiltonian vector field $: -H :$ given by the rule

$$: -H := (\partial H/\partial p)(\partial/\partial q) - (\partial H/\partial q)(\partial/\partial p). \tag{27.3.73}$$

In view of (3.71) and (3.72), we also have the relation

$$: -H := u(q,p)(\partial/\partial q) - v(q,p)(\partial/\partial p). \tag{27.3.74}$$

Thus, any analytic function f gives rise to a Hamiltonian vector field.

Consider the differential form

$$u(q,p)dq - v(q,p)dp. \tag{27.3.75}$$

Show, using the second Cauchy-Riemann equation (3.68), that this form is also exact so that there is a function K such that

$$u(q,p) = \partial K/\partial q, \tag{27.3.76}$$

$$v(q,p) = -\partial K/\partial p. \tag{27.3.77}$$

Show that the Hamiltonian vector field $: -K :$ is given in terms of u and v by the relation

$$: -K := -v(q,p)(\partial/\partial q) - u(q,p)(\partial/\partial p). \tag{27.3.78}$$

Thus, any analytic function f also gives rise to a second Hamiltonian vector field. Show that (3.78) arises from (3.74) upon replacing f by if.

For the analytic function f given by

$$f(z) = z^2, \tag{27.3.79}$$

find the Hamiltonians H and K.

27.3.14. Review Exercise 3.13. For the analytic function f given by

$$f(z) = z^2, \tag{27.3.80}$$

consider the vector field $\mathcal{L}$ given by

$$\mathcal{L} = u(q,p)(\partial/\partial q) + v(q,p)(\partial/\partial p). \tag{27.3.81}$$

Verify that this vector field is not Hamiltonian, and decompose it into Hamiltonian and non-Hamiltonian parts. Make analogous calculations for the vector field $\mathcal{L}'$ given by

$$\mathcal{L}' = -v(q,p)(\partial/\partial q) + u(q,p)(\partial/\partial p). \tag{27.3.82}$$

27.3.15. Show that Duffing's equation (1.4.31) arises from the vector field $\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_1 + \mathcal{L}_3$ where

$$\mathcal{L}_0 = (\epsilon \cos \omega\tau)(\partial/\partial p), \tag{27.3.83}$$

$$\mathcal{L}_1 = p(\partial/\partial q) - (q + 2\beta p)(\partial/\partial p), \tag{27.3.84}$$

$$\mathcal{L}_3 = -(q^3)(\partial/\partial p). \tag{27.3.85}$$

Verify that $\mathcal{L}_0$ and $\mathcal{L}_3$ are Hamiltonian,

$$\mathcal{L}_0 =: (\epsilon \cos \omega\tau) q :, \tag{27.3.86}$$

$$\mathcal{L}_3 = - : q^4/4 : . \tag{27.3.87}$$

Using (3.22) and (3.23), decompose $\mathcal{L}_1$ into Hamiltonian and non-Hamiltonian parts.

27.4 Structure of $sp(4,\mathbb{R})$

The Lie algebra $sp(4,\mathbb{R})$ is 10 dimensional, and its Cartan subalgebra is 2 dimensional. Therefore, in the Cartan basis, there should be 8 ladder operators. They are labelled by 8 two-component root vectors consisting of 4 vectors and their negatives. We will call these 4 vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$. They are given in terms of two orthogonal unit vectors $\boldsymbol{e}^1$ and $\boldsymbol{e}^2$ by the relations

$$\boldsymbol{\alpha} = 2\boldsymbol{e}^1, \tag{27.4.1}$$

$$\boldsymbol{\beta} = \boldsymbol{e}^1 + \boldsymbol{e}^2, \tag{27.4.2}$$

$$\boldsymbol{\gamma} = 2\boldsymbol{e}^2, \tag{27.4.3}$$

$$\boldsymbol{\delta} = -\boldsymbol{e}^1 + \boldsymbol{e}^2. \tag{27.4.4}$$

The eight $sp(4)$ root vectors (the vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$ and their negatives) are shown in Figure 4.1. Note that all root vectors are of the form $(\pm\boldsymbol{e}^i \pm \boldsymbol{e}^j)$ with i, j and the signs taken independently and the zero vector omitted. Thus, there are basically two kinds of root vectors: *short* root vectors with length $\sqrt{2}$ and *long* root vectors with length 2. (And the angle between any two successive root vectors as one goes around the root diagram is 45 degrees.) They satisfy the normalization relations

$$\sum_{\boldsymbol{\mu}} (\boldsymbol{e}^i \cdot \boldsymbol{\mu})(\boldsymbol{\mu} \cdot \boldsymbol{e}^j) = 12\delta_{ij}. \tag{27.4.5}$$

Again see Section 5.8 for an analogous treatment of $su(3)$.

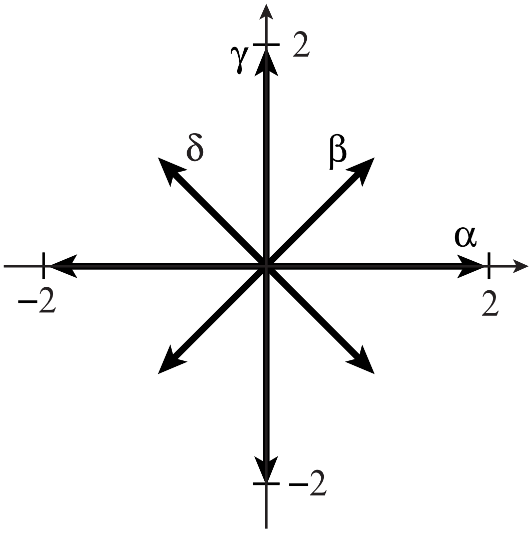


Figure 27.4.1: Root diagram showing the root vectors for $sp(4)$.

The Lie algebra $sp(4,\mathbb{R})$ is generated by the monomials $z_a z_b$ with a, b ranging from 1 to 4. See Section 5.7. For present purposes it is convenient to use as the basis for the Cartan subalgebra the polynomials

$$c^1 = -(i/2)(p_1^2 + q_1^2), \tag{27.4.6}$$

$$c^2 = -(i/2)(p_2^2 + q_2^2). \tag{27.4.7}$$

The Lie operators associated with c^1 and c^2 obviously commute. They are also Hermitian,

$$: c^j :^\dagger =: c^j : . \tag{27.4.8}$$

Finally, Lie transformations of the form

$$\begin{aligned}\mathcal{M}(\theta_1, \theta_2) &= \exp(-i\theta_1 : c^1 : -i\theta_2 : c^2 :) \\ &= \exp[-(\theta_1/2) : p_1^2 + q_1^2 : -(\theta_2/2) : p_2^2 + q_2^2 :]\end{aligned} \tag{27.4.9}$$

are real symplectic, unitary, and lie on a 2-torus which is a maximal torus in $Sp(4, \mathbb{R})$. See Sections 3.9, 5.9, and 7.2.

For the ladder operators in $sp(4)$ we use the polynomials

$$r(\pm\boldsymbol{\alpha}) = (\sqrt{2}/4)(q_1 \pm ip_1)^2, \tag{27.4.10}$$

$$r(\pm\boldsymbol{\beta}) = (1/2)(q_1 \pm ip_1)(q_2 \pm ip_2), \tag{27.4.11}$$

$$r(\pm\boldsymbol{\gamma}) = (\sqrt{2}/4)(q_2 \pm ip_2)^2, \tag{27.4.12}$$

$$r(\pm\boldsymbol{\delta}) = (i/2)(q_1 \mp ip_1)(q_2 \pm ip_2). \tag{27.4.13}$$

Their associated Lie operators obey the conjugation relations

$$: r(\boldsymbol{\mu}) :^\dagger =: r(-\boldsymbol{\mu}) : . \tag{27.4.14}$$

It is easily verified that the Cartan subalgebra and ladder operators satisfy the Poisson bracket rules

$$[c^j, c^k] = 0, \tag{27.4.15}$$

$$[c^j, r(\boldsymbol{\mu})] = (\boldsymbol{e}^j \cdot \boldsymbol{\mu}) r(\boldsymbol{\mu}), \tag{27.4.16}$$

$$[r(\boldsymbol{\mu}), r(-\boldsymbol{\mu})] = \sum_j (\boldsymbol{e}^j \cdot \boldsymbol{\mu}) c^j, \tag{27.4.17}$$

as desired. There are also the relations

$$[r(\boldsymbol{\mu}), r(\boldsymbol{\nu})] = N(\boldsymbol{\mu}, \boldsymbol{\nu}) r(\boldsymbol{\mu} + \boldsymbol{\nu}) \tag{27.4.18}$$

provided the sum $(\boldsymbol{\mu} + \boldsymbol{\nu})$ is again a root vector. All other brackets vanish. For the case of $sp(4)$, the $N(\boldsymbol{\mu}, \boldsymbol{\nu})$ have the values $\pm\sqrt{2}$. The positive N's are $N(\boldsymbol{\alpha}, \boldsymbol{\delta})$, $N(\boldsymbol{\beta}, \boldsymbol{\delta})$, $N(\boldsymbol{\beta}, -\boldsymbol{\delta})$, $N(\boldsymbol{\gamma}, -\boldsymbol{\delta})$, $N(\boldsymbol{\delta}, -\boldsymbol{\beta})$, $N(\boldsymbol{\delta}, -\boldsymbol{\gamma})$, $N(-\boldsymbol{\alpha}, \boldsymbol{\beta})$, $N(-\boldsymbol{\beta}, \boldsymbol{*})$, $N(-\boldsymbol{\beta}, *)$, $N(-\boldsymbol{\gamma}, *)$, $N(-\boldsymbol{\delta}, *)$, $N(-\boldsymbol{\delta}, *)$. We also note, as was true for sp(2), that for the scalar product (7.3.12) the basis elements c^j and $r(\boldsymbol{\mu})$ are orthonormal,

$$\langle c^j, c^j \rangle = \langle r(\boldsymbol{\mu}), r(\boldsymbol{\mu}) \rangle = 1, \tag{27.4.19}$$

$$\langle c^j, c^k \rangle = 0 \text{ for } j \neq k, \tag{27.4.20}$$

$$\langle c^j, r(\boldsymbol{\mu}) \rangle = 0, \tag{27.4.21}$$

$$\langle r(\boldsymbol{\mu}), r(\boldsymbol{\nu}) \rangle = 0 \text{ for } \boldsymbol{\mu} \neq \boldsymbol{\nu}. \tag{27.4.22}$$

At this point, we remark that it is tempting to assume that the rank of a Lie algebra equals the maximum number of mutually commuting elements. (Some texts even make this claim!) This need not be the case. For $sp(4)$, which has rank 2, it is evident that the 3 elements $:p_1^2:$, $:p_1p_2:$, and $:p_2^2:$ are mutually commuting. However, they are not Hermitian and, when exponentiated, do not even produce a torus (not to mention a maximal torus). Nor can other elements be found in the Lie algebra such that relations of the form (4.16) hold with the $:p_kp_\ell:$ playing the role of the c's. Instead, as will be evident in Section 27.5 [see (5.12), (5.14), and (5.16)], they are related by a symplectic unitary transformation to ladder operators. Therefore they do not meet the requirements to form a Cartan subalgebra.

We close this section by examining how $sp(2)$, $u(2)$, $su(2)$, and $so(2)$ reside within $sp(4)$. The presence of $sp(2)$ is evident. Comparison of Figures 1.1 and 4.1, and comparison of (1.1) and (4.10), shows that if we identify the coordinate pair q,p with the pair q_1,p_1, then the $r(\pm\boldsymbol{\alpha})$ for $sp(2)$ and $sp(4)$ agree. Also, the c^1 for $sp(2)$ given by (1.11) agrees with the c^1 given for $sp(4)$ by (4.6). We also note that there is a second $sp(2)$ within $sp(4)$ generated by $r(\pm\boldsymbol{\gamma})$ and c^2. It is identical to the $sp(2)$ of Section 27.1 if we identify the pair q,p with q_2,p_2. Thus, there is an $sp(2)$ within $sp(4)$ for each equal and opposite pair of *long* root vectors. (See also Exercise 5.5.) These root vectors have length 2 as is requred for an $sp(2)$ root vector. See Section 27.1.

The presence of $u(2)$ and $su(2)$ within $sp(4)$ is less evident. Comparison of (5.7.8) and (4.13) shows that there are the relations

$$r(\pm) = r(\mp\boldsymbol{\delta}). \tag{27.4.23}$$

Also, comparison of (5.7.4), (5.7.9) and (4.6), (4.7) gives the relations

$$b^0 = i(c^1 + c^2), \tag{27.4.24}$$

$$c = (1/\sqrt{2})(c^1 - c^2). \tag{27.4.25}$$

Thus, the $su(2)$ of Section 5.7 is associated with c and the $r(\mp\boldsymbol{\delta})$ root vectors of $sp(4)$, and also including b^0 yields $u(2)$. We also note that the spin 1 objects $h^\pm$ and h^0 of Section 5.7 given by (5.7.21) through (5.7.23) are proportional to the $sp(4)$ generators $r(\boldsymbol{\alpha})$, $r(\boldsymbol{\beta})$, and $r(\boldsymbol{\gamma})$ given by (4.10) through (4.12). Reference to Figure 4.1 shows that they should indeed transform among each other under the action of $r(\mp\boldsymbol{\delta})$ as given by (5.7.26) and (5.7.27). Moreover, we note that there is a second $su(2)$ [and a corresponding $u(2)$] within $sp(4)$ associated with the root vectors $r(\pm\boldsymbol{\beta})$. Thus, there is an $su(2)$ and a $u(2)$ within $sp(4)$ for each equal and opposite pair of *short* root vectors. These root vectors have length $\sqrt{2}$ as required for $su(2)$ root vectors. See Exercise 2.1.

Finally, there is an $so(2)$ subalgebra within $sp(4)$ whose presence is not at all obvious from looking at the $sp(4)$ root diagram. Let J_z be the quadratic polynomial defined by the equation

$$J_z = q_1p_2 - q_2p_1. \tag{27.4.26}$$

It generates rotations in the q_1,q_2 and p_1,p_2 planes:

$$\exp:\theta J_z:q_1 = q_1\cos\theta + q_2\sin\theta, \tag{27.4.27}$$

$$\exp:\theta J_z:q_2 = -q_1\sin\theta + q_2\cos\theta; \tag{27.4.28}$$

$$\exp : \theta J_z p_1 = p_1 \cos\theta + p_2 \sin\theta, \tag{27.4.29}$$

$$\exp : \theta J_z : p_2 = -p_1 \sin\theta + p_2 \cos\theta. \tag{27.4.30}$$

From (4.13) we see that it is related to elements in the Cartan basis by the equation

$$J_z = -[r(\boldsymbol{\delta}) - r(-\boldsymbol{\delta})]. \tag{27.4.31}$$

Exercises

27.4.1. Verify the Lie product rules (4.15) through (4.18).

27.4.2. Verify the relations (4.26) through (4.31). Show that J_z obeys the eigen relations

$$: J_z (q_1 \pm i q_2)^n = \mp i n (q_1 \pm i q_2)^n, \tag{27.4.32}$$

$$: J_z : (p_1 \pm i p_2)^n = \mp i n (p_1 \pm i p_2)^n. \tag{27.4.33}$$

27.4.3. Explore the properties of J_z' that is defined in analogy to J_z by the equation

$$J_z' = -[r(\boldsymbol{\beta}) - r(-\boldsymbol{\beta})]. \tag{27.4.34}$$

See Figure 4.1.

27.5 Representations of $sp(4, \mathbb{R})$

The description of representations of $sp(4)$ follows the general Cartan procedure as described for $su(3)$ in Section 5.8. For $sp(4)$, since it has rank 2, there are two fundamental weights $\boldsymbol{\phi}^1$ and $\boldsymbol{\phi}^2$. They are given by the relations

$$\boldsymbol{\phi}^1 = \boldsymbol{e}^1 = \boldsymbol{\alpha}/2, \tag{27.5.1}$$

$$\boldsymbol{\phi}^2 = \boldsymbol{e}^1 + \boldsymbol{e}^2 = \boldsymbol{\beta}, \tag{27.5.2}$$

and are shown in Figure 5.1 along with the $sp(4)$ root vectors. Thus, for $sp(4)$, every highest weight $\boldsymbol{w}^h$ is of the form

$$\boldsymbol{w}^h = m\boldsymbol{\phi}^1 + n\boldsymbol{\phi}^2 = (m+n)\boldsymbol{e}^1 + n\boldsymbol{e}^2, \tag{27.5.3}$$

where m and n are arbitrary nonnegative integers. Correspondingly, for each m, n pair, there is an irreducible representation $\Gamma(m,n)$ with highest weight $\boldsymbol{w}^h$ given by (5.3). It can be shown that the dimension of $\Gamma(m,n)$ is given by the relation

$$\dim \Gamma(m,n) = (1/6)(m+2n+3)(m+n+2)(m+1)(n+1). \tag{27.5.4}$$

See Exercise 5.1. It can also be shown that these representations are self conjugate,

$$\overline{\Gamma}(m,n) = \Gamma(m,n). \tag{27.5.5}$$

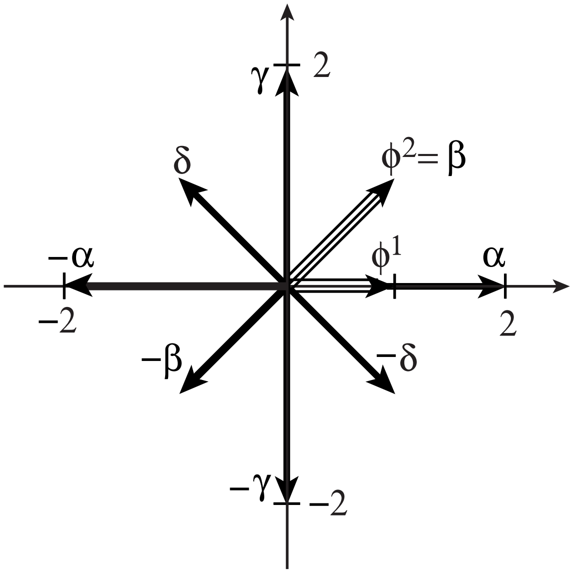


Figure 27.5.1: Fundamental weights $\boldsymbol{\phi}^1$ and $\boldsymbol{\phi}^2$ for $sp(4)$. The root vectors are also shown.

Table 27.5.1: Dimensions of Representations of $sp(4)$.

m	n	$\dim\Gamma(m,n)$	m	n	$\dim\Gamma(m,n)$
0	0	1	0	2	14
1	0	4	3	0	20
0	1	5	2	1	35
2	0	10	1	2	40
1	1	16	0	3	30

See Exercise 3.7.36. For quick reference the dimensions of the first few representations are listed in Table 5.1 above. Where there is no possibility of confusion, we will sometimes refer to a representation by its dimension.

From a knowledge of the root vectors and the highest weight it is a simple matter to construct weight diagrams for the low-dimensional representations. Figures 5.2 through 5.7 show weight diagrams for the first few representations. Inspection of these figures and reference to Table 5.1 shows that the weights must have unit multiplicities for the representations $\Gamma(0,0)$, $\Gamma(1,0)$, and $\Gamma(0,1)$. For $\Gamma(2,0)$, which is the adjoint or regular representation, the weight vector at the origin has multiplicity 2. The representation $\Gamma(1,0)$ corresponds to the representation of $sp(4)$ by 4×4 matrices of the form JS. See (5.7.27) of Section 5.7. It happens that the Lie algebras for $sp(4)$ and $so(5)$ are equivalent over the complex field. The $sp(4)$ representation $\Gamma(0,1)$, which is 5 dimensional, is related to the obvious 5×5 matrix representation of $so(5)$. See Exercise 5.4.

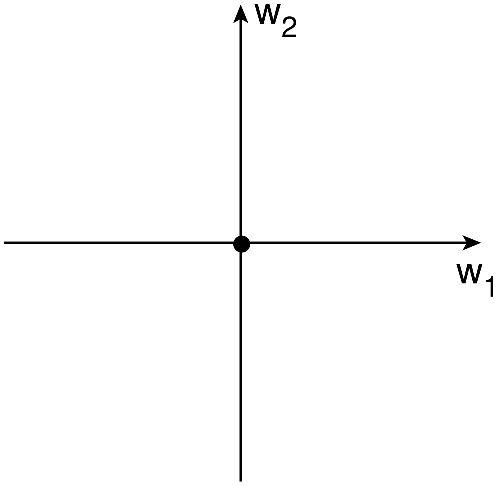


Figure 27.5.2: Weight diagram for the representation $1 = \Gamma(0,0)$.

Now, in mimicry of what was done before in Section 21.2, let $\mathcal{P}_m$ denote the set of polynomials homogeneous of degree m in the variables q_1, p_1, q_2, p_2; and let f_2 be a quadratic polynomial in these variables. Then we have the relation

$$: f_2 : \mathcal{P}_m \subseteq \mathcal{P}_m. \tag{27.5.6}$$

It follows that the set of homogeneous polynomials of degree m forms a representation of $sp(4,\mathbb{R})$.

What irreducible representations occur? The study of this question is again facilitated by a map $\mathcal{A}(\pi/8)$ defined this time by the equation

$$\mathcal{A}(\pi/8) = \exp[-i(\pi/8) : p_1^2 - q_1^2 + p_2^2 - q_2^2 :]. \tag{27.5.7}$$

As before, $\mathcal{A}$ is complex symplectic and unitary. Calculation gives the result

$$\begin{aligned}&\mathcal{A}(\pi/8)(q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2}) = \\ &(1/\sqrt{2})^{r_1+s_1+r_2+s_2}(i)^{s_1+s_2}(q_1+ip_1)^{r_1}(q_1-ip_1)^{s_1}(q_2+ip_2)^{r_2}(q_2-ip_2)^{s_2}.\end{aligned} \tag{27.5.8}$$

Evidently $\mathcal{A}(\pm\pi/8)$ again transforms between what we will again call the Cartesian and resonance bases.

With the aid of $\mathcal{A}$ we again define transformed Lie basis polynomials $\tilde{c}^j$ and $\tilde{r}(\boldsymbol{\mu})$ by the rule

$$\tilde{c}^1 = \mathcal{A}(-\pi/8)c^1 = -q_1p_1, \tag{27.5.9}$$

$$\tilde{c}^2 = \mathcal{A}(-\pi/8)c^2 = -q_2p_2, \tag{27.5.10}$$

$$\tilde{r}(\boldsymbol{\alpha}) = \mathcal{A}(-\pi/8)r(\boldsymbol{\alpha}) = (1/\sqrt{2})q_1^2, \tag{27.5.11}$$

$$\tilde{r}(-\boldsymbol{\alpha}) = \mathcal{A}(-\pi/8)r(-\boldsymbol{\alpha}) = -(1/\sqrt{2})p_1^2, \tag{27.5.12}$$

$$\tilde{r}(\boldsymbol{\beta}) = \mathcal{A}(-\pi/8)r(\boldsymbol{\beta}) = q_1q_2, \tag{27.5.13}$$

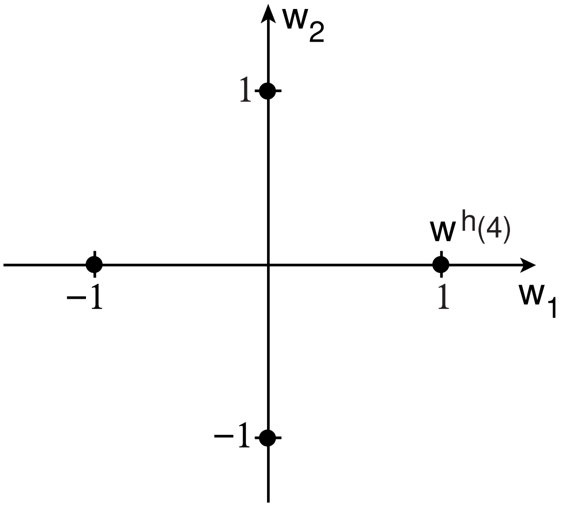


Figure 27.5.3: Weight diagram for the fundamental representation $4 = \Gamma(1,0)$.

$$\tilde{r}(-\boldsymbol{\beta}) = \mathcal{A}(-\pi/8)r(-\boldsymbol{\beta}) = -p_1p_2, \tag{27.5.14}$$

$$\tilde{r}(\boldsymbol{\gamma}) = \mathcal{A}(-\pi/8)r(\boldsymbol{\gamma}) = (1/\sqrt{2})q_2^2, \tag{27.5.15}$$

$$\tilde{r}(-\boldsymbol{\gamma}) = \mathcal{A}(-\pi/8)r(-\boldsymbol{\gamma}) = -(1/\sqrt{2})p_2^2, \tag{27.5.16}$$

$$\tilde{r}(\boldsymbol{\delta}) = \mathcal{A}(-\pi/8)r(\boldsymbol{\delta}) = p_1q_2, \tag{27.5.17}$$

$$\tilde{r}(-\boldsymbol{\delta}) = \mathcal{A}(-\pi/8)r(-\boldsymbol{\delta}) = p_2q_1. \tag{27.5.18}$$

Since $\mathcal{A}$ is symplectic, the transformed basis polynomials obey the same Poisson bracket rules (4.15) athrough (4.18). Since $\mathcal{A}$ is also unitary, the transformed basis polynomials also satisfy the orthonormality relations (4.19) through (4.22) and the conjugacy relations (4.8) and (4.14).

Now consider the actions of the Lie operators $:\tilde{c}^j:$ and $:\tilde{r}(\boldsymbol{\mu}):$ on the general monomials $q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2}$. Calculation gives the results

$$:\tilde{c}^1: q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2} = (r_1 - s_1)q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2}, \tag{27.5.19}$$

$$:\tilde{c}^2: q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2} = (r_2 - s_2)q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2}, \tag{27.5.20}$$

$$:\tilde{r}(\boldsymbol{\alpha}): q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2} = \sqrt{2}s_1q_1^{r_1+1}p_1^{s_1-1}q_2^{r_2}p_2^{s_2}, \tag{27.5.21}$$

$$:\tilde{r}(-\boldsymbol{\alpha}): q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2} = \sqrt{2}r_1q_1^{r_1-1}p_1^{s_1+1}q_2^{r_2}p_2^{s_2}, \tag{27.5.22}$$

$$:\tilde{r}(\boldsymbol{\beta}): q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2} = s_1q_1^{r_1}p_1^{s_1-1}q_2^{r_2+1}p_2^{s_2} + s_2q_1^{r_1+1}p_1^{s_1}q_2^{r_2}p_2^{s_2-1}, \tag{27.5.23}$$

$$:\tilde{r}(-\boldsymbol{\beta}): q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2} = r_1q_1^{r_1-1}p_1^{s_1}q_2^{s_2}p_2^{s_2+1} + r_2q_1^{r_1}p_1^{s_1+1}q_2^{r_2-1}p_2^{s_2}, \tag{27.5.24}$$

$$:\tilde{r}(\boldsymbol{\gamma}): q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2} = \sqrt{2}s_2q_1^{r_1}p_1^{s_1}q_2^{r_2+1}p_2^{s_2-1}, \tag{27.5.25}$$

$$:\tilde{r}(-\boldsymbol{\gamma}): q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2} = \sqrt{2}r_2q_1^{r_1}p_1^{s_1}q_2^{r_2-1}p_2^{s_2+1}, \tag{27.5.26}$$

$$:\tilde{r}(\boldsymbol{\delta}): q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2} = -r_1q_1^{r_1-1}p_1^{s_1}q_2^{r_2+1}p_2^{s_2} + s_2q_1^{r_1}p_1^{s_1+1}q_2^{r_2}p_2^{s_2-1}, \tag{27.5.27}$$

$$:\tilde{r}(-\boldsymbol{\delta}): q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2} = s_1q_1^{r_1}p_1^{s_1-1}q_2^{r_2}p_2^{s_2+1} - r_2q_1^{r_1+1}p_1^{s_1}q_2^{r_2-1}p_2^{s_2}. \tag{27.5.28}$$

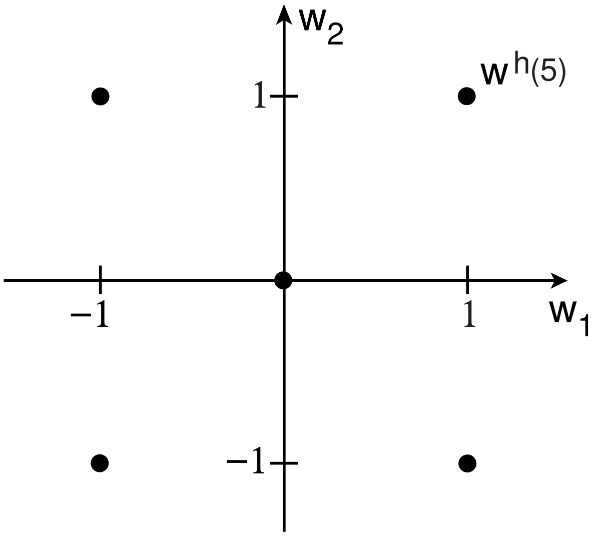


Figure 27.5.4: Weight diagram for the representation $5 = \Gamma(0, 1)$.

Evidently any monomial of a given degree can be transformed into any other monomial of the same degree with the aid of the $\tilde{r}(\boldsymbol{\mu})$. Therefore $sp(4)$ acts irreducibly on $\mathcal{P}_m$. Also, q_1^m is the vector of highest weight in $\mathcal{P}_m$, and has the weight $\boldsymbol{\omega}^h = m\boldsymbol{\phi}^1$,

$$: \tilde{c}^1 : q_1^m = mq_1^m = (\boldsymbol{e}^1 \cdot m\boldsymbol{\phi}^1)q_1^m, \tag{27.5.29}$$

$$: \tilde{c}^2 : q_1^m = 0 = (\boldsymbol{e}^2 \cdot m\boldsymbol{\phi}^1)q_1^m.$$

Upon examination of (5.3) we conclude that, in the case of 4 variables, $\mathcal{P}_m$ carries the representation $\Gamma(m, 0)$ of $sp(4)$. From (7.3.36) and (5.4) we find, as expected, the result

$$\dim \mathcal{P}_m = N(m, 4) = (1/6)(m + 3)(m + 2)(m + 1) = \dim \Gamma(m, 0). \tag{27.5.30}$$

Note that, unlike the case of 2 variables, the $\mathcal{P}_m$ for various m do not carry *all* the representations of $sp(4)$, but only the representations $\Gamma(m, 0)$.

As before, we can let $\mathcal{A}(\pi/8)$ act on both sides of (5.19) through (5.28). Doing so gives results analogous to those in (2.23) through (2.25). Consequently, as before, the $: c^j :$ and $: r(\boldsymbol{\mu}) :$ act on the resonance basis in the same way that the $: \tilde{c}^j :$ and $: \tilde{r}(\boldsymbol{\mu}) :$ act on the monomial basis.

Exercises

27.5.1. Weyl discovered that for the simple Lie algebras the dimension of a representation $\Gamma(\boldsymbol{w}^h)$ labeled by the highest weight $\boldsymbol{w}^h$ is given the formula

$$\dim \Gamma(\boldsymbol{w}^h) = \prod_{\boldsymbol{\mu}>0} [\boldsymbol{\mu} \cdot (\boldsymbol{w}^h + \boldsymbol{\mu}^+/2)]/[\boldsymbol{\mu} \cdot (\boldsymbol{\mu}^+/2)]. \tag{27.5.31}$$

Here the product is to be taken over all *positive* root vectors and $\boldsymbol{\mu}^+$ is the sum of all positive roots as in (12.49). [As was the case for weights (see Section 5.8), we define a root $\boldsymbol{\mu}$ to be

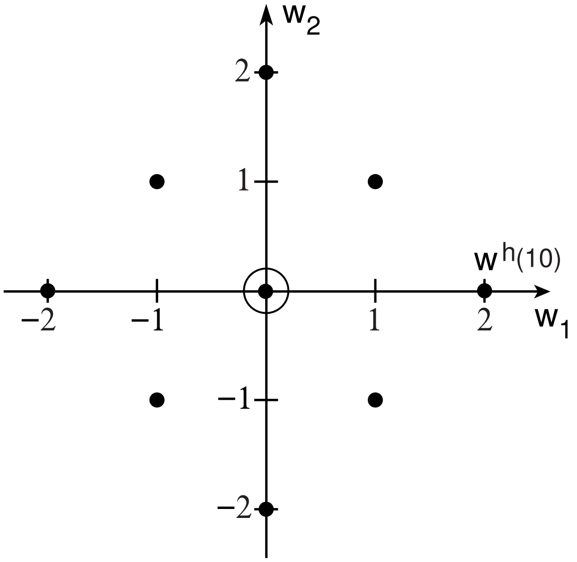


Figure 27.5.5: Weight diagram for the adjoint representation $10 = \Gamma(2,0)$. The circled weight at the origin has multiplicity 2. The other eight weights are located at the tips of the $sp(4)$ root vectors.

positive (and write $\boldsymbol{\mu} > 0$) if its first nonvanishing component is positive.] Show that the results (2.4), (5.4), and (5.8.21) for $sp(2)$, $sp(4)$, and $su(3)$ follow from Weyl's formula. If you are feeling algebraically rambunctious, verify (8.5) for $sp(6)$.

27.5.2. Verify (5.30).

27.5.3. Look at the $sp(4)$ weight diagrams shown in Figures 5.2 through 5.7. Verify that the spacing between the dots in the directions of the long $sp(4)$ root vectors is 2. These dots describe $sp(2)$ representations within $sp(4)$. See Section 21.2. Verify that the spacing between the dots in the directions of the short $sp(4)$ root vectors is $\sqrt{2}$. These dots describe $su(2)$ representations within $sp(4)$. See Exercise 2.1.

27.5.4. The goal of this exercise is to relate the Lie algebras $sp(4)$ and $so(5)$, and the Lie groups $Sp(4)$ and $SO(5)$. You already know from Exercise 3.7.31 that they have the same dimension. You also know from Section 5.10.1 that $sp(2n,\mathbb{R})$ and $usp(2n)$ are equivalent over the complex field, but not over the real field. In this exercise you will show that $usp(4)$ and $so(5,\mathbb{R})$ are isomorphic. Therefore, in so doing, you will show that that $sp(4,\mathbb{R})$ and $so(5,\mathbb{R})$ are equivalent over the complex field, but not over the real field.

Review Exercise 8.2.12. There it is shown that if $K \in su(4)$, then L given by

$$L_{\alpha\beta}(K) = -(1/2)\mathrm{tr}[K(A^{\alpha})^{\dagger}A^{\beta}] \tag{27.5.32}$$

will have the property $L \in so(6,\mathbb{R})$. Correspondingly, from the relation

$$R_{\alpha\beta}(v) = (1/4)\mathrm{tr}[v^{T}A^{\alpha}v(A^{\beta})^{\dagger}], \tag{27.5.33}$$

there will be an $R \in SO(6,\mathbb{R})$ for every $v \in SU(4)$. Now suppose that $K \in sp(4,\mathbb{C})$ as well so that $K \in usp(4)$. Then we will also have $v \in Sp(4,\mathbb{C})$ so that $v \in USp(4)$.

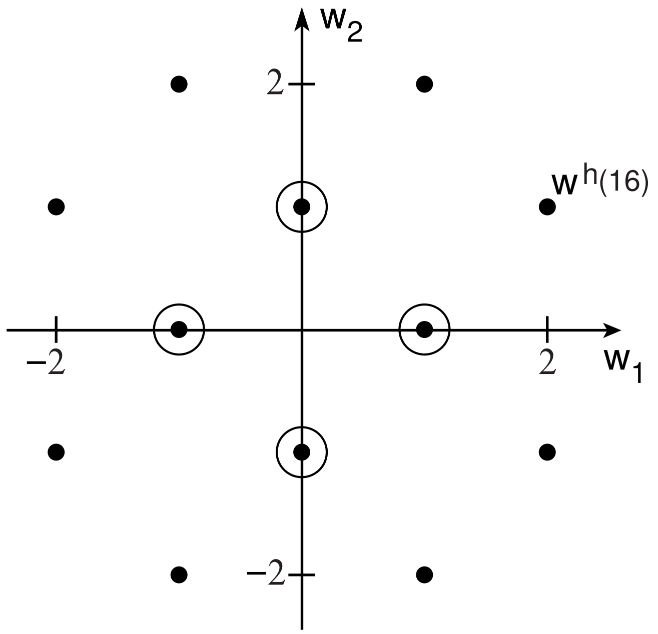


Figure 27.5.6: Weight diagram for the representation $16 = \Gamma(1,1)$. The circled weights on the inner diamond have multiplicity 2.

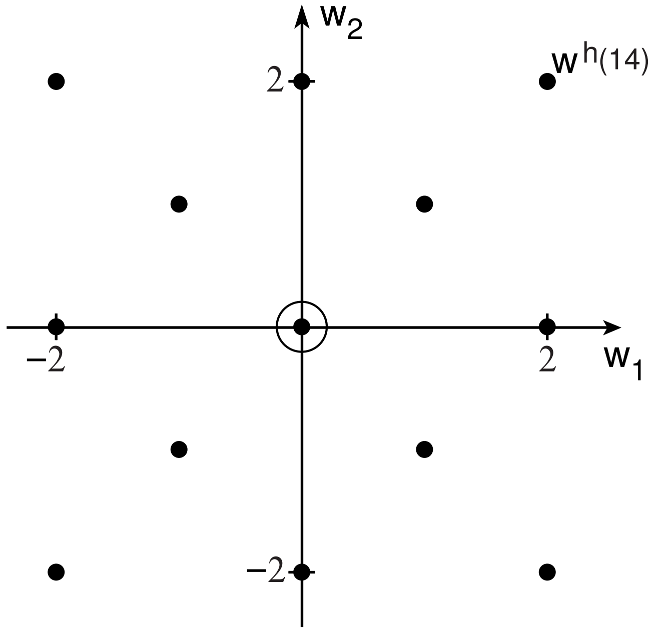


Figure 27.5.7: Weight diagram for the representation $14 = \Gamma(0,2)$. The circled weight at the origin has multiplicity 2.

Suppose we set $\alpha = 6$ in (5.33). From (8.2.98) we see that

$$A^6 = J, \tag{27.5.34}$$

and therefore, if v is symplectic, there is the relation

$$v^T A^6 v = v^T J v = J = A^6. \tag{27.5.35}$$

Consequently, show that there is the result

$$\begin{aligned} R_{6\beta}(v) &= (1/4)\text{tr}[v^T A^6 v (A^\beta)^\dagger] = (1/4)\text{tr}[v^T J v (A^\beta)^\dagger] \\ &= (1/4)\text{tr}[J(A^\beta)^\dagger] = (1/4)\text{tr}[A^6(A^\beta)^\dagger] = \delta_{6\beta}. \end{aligned} \tag{27.5.36}$$

Next set $\beta = 6$ in (5.33). Show that

$$\begin{aligned} R_{\alpha 6}(v) &= (1/4)\text{tr}[v^T A^\alpha v (A^6)^\dagger] = (1/4)\text{tr}[v^T A^\alpha v (J)^\dagger] \\ &= (1/4)\text{tr}[A^\alpha v (J)^\dagger v^T] = (1/4)\text{tr}[A^\alpha J^\dagger] \\ &= (1/4)\text{tr}[A^\alpha (A^6)^\dagger] = \delta_{\alpha 6}. \end{aligned} \tag{27.5.37}$$

Consequently show that, when $v \in USp(4)$, R has the block form

$$R(v) = \begin{pmatrix} R_{11} & R_{12} & R_{13} & R_{14} & R_{15} & 0 \\ R_{21} & R_{22} & R_{23} & R_{24} & R_{25} & 0 \\ R_{31} & R_{32} & R_{33} & R_{34} & R_{35} & 0 \\ R_{41} & R_{42} & R_{43} & R_{44} & R_{45} & 0 \\ R_{51} & R_{52} & R_{53} & R_{54} & R_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{27.5.38}$$

Let $\hat{R}$ be the 5×5 matrix

$$\hat{R} = \begin{pmatrix} R_{11} & R_{12} & R_{13} & R_{14} & R_{15} \\ R_{21} & R_{22} & R_{23} & R_{24} & R_{25} \\ R_{31} & R_{32} & R_{33} & R_{34} & R_{35} \\ R_{41} & R_{42} & R_{43} & R_{44} & R_{45} \\ R_{51} & R_{52} & R_{53} & R_{54} & R_{55} \end{pmatrix}. \tag{27.5.39}$$

We know that $R \in SO(6,\mathbb{R})$ when $v \in USp(4)$, and we have seen that R must then also have the form (5.38). Consequently, there is the result that $\hat{R} \in SO(5,\mathbb{R})$ when $v \in USp(4)$. Therefore (5.33) provides a map of $USp(4)$ into $SO(5,\mathbb{R})$ when α, β are restricted to range from 1 to 5. Verify that this map is a homomorphism,

$$\hat{R}(v_1 v_2) = \hat{R}(v_1)\hat{R}(v_2), \tag{27.5.40}$$

and that

$$\hat{R}(-v) = \hat{R}(v) \tag{27.5.41}$$

so that the homomorphism is two to one.

Finally, we should study the relation between $usp(4)$ and $so(5,\mathbb{R})$. Show from (5.32) that

$$L_{6\beta}(K) = L_{\alpha 6}(K) = 0 \text{ when } K \in usp(4). \tag{27.5.42}$$

We already know that L is real and antisymmetric when $K \in su(4)$ and hence it will be real and antisymmetric when $K \in usp(4)$ since $usp(4)$ is a subalgebra of $su(4)$. It follows that (5.32) provides a map of $usp(4)$ into $so(5,\mathbb{R})$ when α, β are restricted to range from 1 to 5. Your last task is to show that this map is an isomorphism.

As a first step, verify that $J \in usp(4)$. Consider computing $L(J)$. From (5.32) we know that

$$L_{\alpha\beta}(J) = -(1/2)\text{tr}[J(A^\alpha)^\dagger A^\beta]. \tag{27.5.43}$$

Examine the products $(A^\alpha)^\dagger A^\beta$. See (8.2.112) through (8.2.116). Observe that the products are either symmetric matrices S or antisymmetric matrices A. Recall that matrices of the form JS are traceless. Therefore we only need be concerned with those products that are antisymmetric. But in this case we only need consider those products whose results are proportional to $J = A^6$ because of the orthogonality condition (8.2.121). From (8.2.108) we see that the only products that contribute are of the form A^2A^4. Verify that

$$\begin{aligned} L_{24}(J) &= -(1/2)\text{tr}[J(A^2)^\dagger A^4] = (1/2)\text{tr}[JA^2A^4] \\ &= (1/2)\text{tr}[JA^6] = -(1/2)\text{tr}[A^6(A^6)^\dagger] = -2. \end{aligned} \tag{27.5.44}$$

Thus, show that

$$L(J) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{27.5.45}$$

Next verify that (5.32) provides a homomorphism of $usp(4)$ into $so(5,\mathbb{R})$ when α, β are restricted to range from 1 to 5. Finally, make an argument analogous to that of Exercise 8.9.19 to show that (5.32) provides an isomorphism of $usp(4)$ into $so(5,\mathbb{R})$ when α, β are restricted to range from 1 to 5.

27.5.5. In this exercise we will see that there is a particularly interesting $sp(2)$ subalgebra residing within $sp(4)$. This subalgebra is of use in the Lie algebraic treatment of (light) optical systems having axial symmetry. See Appendix X.

Consider the monomials

$$\tilde{r}(\boldsymbol{\alpha}) = (1/\sqrt{2})q_1^2, \tag{27.5.46}$$

$$\tilde{c}^1 = -q_1p_1, \tag{27.5.47}$$

$$\tilde{r}(-\boldsymbol{\alpha}) = -(1/\sqrt{2})p_1^2; \tag{27.5.48}$$

$$\tilde{r}(\boldsymbol{\gamma}) = (1/\sqrt{2})q_2^2, \tag{27.5.49}$$

$$\tilde{c}^2 = -q_2p_2, \tag{27.5.50}$$

$$\tilde{r}(-\boldsymbol{\gamma}) = -(1/\sqrt{2})p_2^2. \tag{27.5.51}$$

Note that, consistent with the symmetry of the $sp(4)$ root vector diagram shown in Figure 4.1, the the ingredients of (5.46) through (5.48) are analogous to those of (5.49) through (5.51).

Verify that there are the Poisson bracket relations

$$[\tilde{c}^1, \tilde{r}(\pm\boldsymbol{\alpha})] = \pm 2\tilde{r}(\pm\boldsymbol{\alpha})], \tag{27.5.52}$$

$$[\tilde{r}(\boldsymbol{\alpha}), \tilde{r}(-\boldsymbol{\alpha})] = 2\tilde{c}^1. \tag{27.5.53}$$

As expected, reference to (1.13) and (1.14) reveals that these are the Cartan rules for $sp(2)$. Also, evidently the ingredients of (5.49) through (5.51) obey the same rules. Finally, the ingredients of (5.46) through (5.48) are evidently in involution (commute) with the ingredients of (5.49) through (5.51).

View q_1,q_2 and p_1,p_2 as components of vectors $\boldsymbol{q}$ and $\boldsymbol{p}$ by writing

$$\boldsymbol{q} = (q_1, q_2) \text{ and } \boldsymbol{p} = (p_1, p_2). \tag{27.5.54}$$

Also, make the definitions

$$q^2 = \boldsymbol{q} \cdot \boldsymbol{q} = (q_1)^2 + (q_2)^2, \tag{27.5.55}$$

$$\boldsymbol{q} \cdot \boldsymbol{p} = q_1 p_1 + q_2 p_2, \tag{27.5.56}$$

$$p^2 = \boldsymbol{p} \cdot \boldsymbol{p} = (p_1)^2 + (p_2)^2, \tag{27.5.57}$$

$$L_+ = (1/\sqrt{2})q^2, \tag{27.5.58}$$

$$L_0 = -\boldsymbol{q} \cdot \boldsymbol{p}, \tag{27.5.59}$$

$$L_- = -(1/\sqrt{2})p^2. \tag{27.5.60}$$

Verify that there are the Poisson bracket relations

$$[L_0, L_\pm] = \pm 2L_\pm, \tag{27.5.61}$$

$$[L_+, L_{]} = 2L_0. \tag{27.5.62}$$

Evidently these relations are the Cartan rules for $sp(2)$. Is this surprising? We see from the work leading up to this point that this result is to be expected.

Finally, observe that the quantities in (5.55) through (5.57) are invariant under rotations in the $1, 2$ plane: Define the Lie operator $\mathcal{J}_z$ by the rule

$$\mathcal{J}_z =: J_z : \tag{27.5.63}$$

where J_z is defined by (4.26). See also Subsection 16.2.5.2 and (16.2.218) through (16.2.222) of Exercise 16.2.16. Verify that there are the relations

$$\mathcal{J}_z q^2 = \mathcal{J}_z(\boldsymbol{q} \cdot \boldsymbol{p}) = \mathcal{J}_z p^2 = 0. \tag{27.5.64}$$

27.6 Symplectic Classification of Analytic Vector Fields in Four Variables

For the case of analytic vector fields, and in the spirit of Section 17.3, we need to consider in this section vector fields of the form $\mathcal{L}_{\boldsymbol{g}^m}$ where the components of $\boldsymbol{g}^m$ are homogeneous polynomials of degree m in the 4 variables z_1 through z_4. Mutatis mutandis, many of the same results follow as before. With Σ defined by

$$\Sigma = \sum_a z_a(\partial/\partial z_a), \tag{27.6.1}$$

the relations (3.2) and (3.3) are still true. Also, (3.4) remains true, and (3.5) and (3.6) hold when f_2 is quadratic in z_1 through z_4. It follows that in the 4-variable case the Hamiltonian vector fields $: h_m :$ are transformed into each other under the action of $sp(4,\mathbb{R})$ and carry the irreducible representation $\Gamma(m,0)$. Also, any $\mathcal{L}_{\boldsymbol{g}^0}$ is a Hamiltonian vector field, and these fields carry the representation $\Gamma(1,0)$.

Next consider the vector fields $\mathcal{L}_{\boldsymbol{g}^1}$. They form a 16-dimensional space spanned by the vector fields $z_a(\partial/\partial z_b)$ with $a,b = 1,2,3,4$. We know that any $: h_2 :$ is such a vector field, and that these vector fields carry the representation $\Gamma(2,0)$, which is 10 dimensional. Also, Σ is of the form $\mathcal{L}_{\boldsymbol{g}^1}$ and, by (3.8), carries the 1-dimensional representation $\Gamma(0,0)$. We will see that any $\mathcal{L}_{\boldsymbol{g}^1}$ can be written uniquely in the form

$$\mathcal{L}_{\boldsymbol{g}^1} = \mathcal{H}^{2,0} + \mathcal{G}^{0,1} + \mathcal{G}^{0,0}. \tag{27.6.2}$$

Here $\mathcal{H}^{2,0}$ denotes a Hamiltonian vector field of the form

$$\mathcal{H}^{2,0} =: h_2 :, \tag{27.6.3}$$

which therefore carries the representation $\Gamma(2,0)$. $\mathcal{G}^{0,0}$ is a non-Hamiltonian vector field that is a (constant) multiple of Σ, and therefore carries the representation $\Gamma(0,0)$. Finally, $\mathcal{G}^{0,1}$ is a non-Hamiltonian vector field that carries the representation $\Gamma(0,1)$. Note from Table 5.1 that $\Gamma(0,1)$ has dimension 5 so that we have the completeness count $16 = 10 + 5 + 1$.

According to (6.3) finding a suitable basis for the vector fields in $\mathcal{H}^{2,0}$ is equivalent to finding a suitable basis for the quadratic polynomials in 4-dimensional phase space. But this has already been done: we may use the basis provided by the $sp(4)$ generators given in the Cartesian Cartan basis by (5.9) through (5.18).

For the non-Hamiltonian parts we will prove the claims just made by exhibiting suitable bases for $\mathcal{G}^{0,0}$ and $\mathcal{G}^{0,1}$. For our purposes, it is convenient to work in the Cartesian (monomial) basis and use the transformed generators defined in (5.9) through (5.18). Consider the 4 vector fields $z_a(\partial/\partial z_a)$ with $a = 1$ through 4. Evidently, they are mutually commuting. We also observe that the 3 vector fields $: \tilde{c}^1 :$, $: \tilde{c}^2 :$, and $\tilde{\mathcal{G}}^{0,0}_{0,0} = \Sigma$ are made from the $z_a(\partial/\partial z_a)$,

$$: \tilde{c}^1 := - : q_1p_1 := -p_1(\partial/\partial p_1) + q_1(\partial/\partial q_1) = z_1(\partial/\partial z_1) - z_3(\partial/\partial z_3), \tag{27.6.4}$$

$$: \tilde{c}^2 := - : q_2p_2 := -p_2(\partial/\partial p_2) + q_2(\partial/\partial q_2) = z_2(\partial/\partial z_2) - z_4(\partial/\partial z_4), \tag{27.6.5}$$

$$\tilde{\mathcal{G}}^{0,0}_{0,0} = \Sigma = z_1(\partial/\partial z_1) + z_2(\partial/\partial z_2) + z_3(\partial/\partial z_3) + z_4(\partial/\partial z_4). \tag{27.6.6}$$

As a fourth such linearly independent vector field we take the element $\tilde{\mathcal{G}}_{0,0}^{0,1}$ defined by the equation

$$\tilde{\mathcal{G}}_{0,0}^{0,1} = z_1(\partial/\partial z_1) + z_3(\partial/\partial z_3) - z_2(\partial/\partial z_2) - z_4(\partial/\partial z_4). \tag{27.6.7}$$

See Exercise 6.1. In addition we define other elements $\tilde{\mathcal{G}}_{k,\ell}^{0,1}$ by the equations

$$\begin{aligned}\tilde{\mathcal{G}}_{1,1}^{0,1} &= (1/2)\#\tilde{r}(\boldsymbol{\beta})\#\tilde{\mathcal{G}}_{0,0}^{0,1} = (1/2)\{:\tilde{r}(\boldsymbol{\beta}):,\tilde{\mathcal{G}}_{0,0}^{0,1}\}\\ &= -z_1(\partial/\partial z_4) + z_2(\partial/\partial z_3),\end{aligned} \tag{27.6.8}$$

$$\begin{aligned}\tilde{\mathcal{G}}_{-1,-1}^{0,1} &= (1/2)\#\tilde{r}(-\boldsymbol{\beta})\#\tilde{\mathcal{G}}_{0,0}^{0,1} = (1/2)\{:\tilde{r}(-\boldsymbol{\beta}):,\tilde{\mathcal{G}}_{0,0}^{0,1}\}\\ &= z_4(\partial/\partial z_1) - z_3(\partial/\partial z_2),\end{aligned} \tag{27.6.9}$$

$$\begin{aligned}\tilde{\mathcal{G}}_{-1,1}^{0,1} &= (1/2)\#\tilde{r}(\boldsymbol{\delta})\#\tilde{\mathcal{G}}_{0,0}^{0,1} = (1/2)\{:\tilde{r}(\boldsymbol{\delta}):,\tilde{\mathcal{G}}_{0,0}^{0,1}\}\\ &= -z_3(\partial/\partial z_4) - z_2(\partial/\partial z_1),\end{aligned} \tag{27.6.10}$$

$$\begin{aligned}\tilde{\mathcal{G}}_{1,-1}^{0,1} &= (1/2)\#\tilde{r}(-\boldsymbol{\delta})\#\tilde{\mathcal{G}}_{0,0}^{0,1} = (1/2)\{:\tilde{r}(-\boldsymbol{\delta}):,\tilde{\mathcal{G}}_{0,0}^{0,1}\}\\ &= z_4(\partial/\partial z_3) + z_1(\partial/\partial z_2).\end{aligned} \tag{27.6.11}$$

Since Σ commutes with all the $sp(4)$ generators, we immediately have the results

$$\#\tilde{c}^j\#\tilde{\mathcal{G}}_{0,0}^{0,0} = \{:\tilde{c}^j:,\Sigma\} = -\{\Sigma,:\tilde{c}^j:\} = 0, \tag{27.6.12}$$

$$\#r(\boldsymbol{\mu})\#\tilde{\mathcal{G}}_{0,0}^{0,0} = \{:r(\boldsymbol{\mu}):,\Sigma\} = -\{\Sigma,:r(\boldsymbol{\mu}):\} = 0. \tag{27.6.13}$$

Thus, in keeping with its labels, $\tilde{\mathcal{G}}_{0,0}^{0,0}$ carries the representation $\Gamma(0,0)$.

Direct computation shows that the five elements $\tilde{\mathcal{G}}_{k,\ell}^{0,1}$ obey the rules

$$\#\tilde{c}^j\#\tilde{\mathcal{G}}_{k,\ell}^{0,1} = \{:\tilde{c}^j:,\tilde{\mathcal{G}}_{k,\ell}^{0,1}\} = \boldsymbol{e}^j\cdot(k\boldsymbol{e}^1+\ell\boldsymbol{e}^2)\tilde{\mathcal{G}}_{k,\ell}^{0,1}. \tag{27.6.14}$$

That is why the subscripts are taken to have the k,ℓ values shown. Reference to Figure 5.4 shows that the right sides of (6.14) are the components of the weights for the representation $\Gamma(0,1)$. In particular, we see that $\tilde{\mathcal{G}}_{1,1}^{0,1}$ occupies the highest weight site $\boldsymbol{w}^h$ given by (5.3) for the representation $\Gamma(0,1)$. Therefore there should be the ladder relations

$$\#\tilde{r}(\boldsymbol{\alpha})\#\tilde{\mathcal{G}}_{1,1}^{0,1} = \#\tilde{r}(\boldsymbol{\beta})\#\tilde{\mathcal{G}}_{1,1}^{0,1} = \#\tilde{r}(-\boldsymbol{\delta})\#\tilde{\mathcal{G}}_{1,1}^{0,1} = 0. \tag{27.6.15}$$

Direct calculation shows that these relations are true. Similarly, there are the ladder relations

$$\#\tilde{r}(\pm\boldsymbol{\alpha})\#\tilde{\mathcal{G}}_{0,0}^{0,1} = 0, \tag{27.6.16}$$

$$\#\tilde{r}(\pm\boldsymbol{\gamma})\#\tilde{\mathcal{G}}_{0,0}^{0,1} = 0, \tag{27.6.17}$$

because there are no weights in $\Gamma(0,1)$, see Figure 5.4, at the sites $\pm\boldsymbol{\alpha}$, $\pm\boldsymbol{\gamma}$. Further calculation gives the relation

$$\#\tilde{r}(-\boldsymbol{\beta})\#\tilde{\mathcal{G}}_{1,1}^{0,1} = \tilde{\mathcal{G}}_{0,0}^{0,1}, \tag{27.6.18}$$

and all the other ladder relations one expects for the representation $\Gamma(0,1)$.

For the sake of comparison, consider the Hamiltonian vector fields $:\tilde{r}(\pm\boldsymbol{\beta}):$ and $:\tilde{r}(\pm\boldsymbol{\delta}):$. They belong to $\mathcal{H}^{2,0}$ and occupy the same sites as $\tilde{\mathcal{G}}^{0,1}_{\pm1,\pm1}$ and $\tilde{\mathcal{G}}^{0,1}_{\mp1,\pm1}$, respectively, in Figure 5.5. They have the form

$$:\tilde{r}(\boldsymbol{\beta}):=:q_1q_2:=z_1(\partial/\partial z_4)+z_2(\partial/\partial z_3), \tag{27.6.19}$$

$$:\tilde{r}(-\boldsymbol{\beta}):=-:p_1p_2:=z_4(\partial/\partial z_1)+z_3(\partial/\partial z_2), \tag{27.6.20}$$

$$:\tilde{r}(\boldsymbol{\delta}):=:p_1q_2:=z_3(\partial/\partial z_4)-z_2(\partial/\partial z_1), \tag{27.6.21}$$

$$:\tilde{r}(-\boldsymbol{\delta}):=:p_2q_1:=z_4(\partial/\partial z_3)-z_1(\partial/\partial z_2). \tag{27.6.22}$$

Evidently the vector fields (6.7) through (6.11) and (6.19) through (6.22) are linearly independent. However, in contrast to (6.15), we have the nonzero result

$$\#\tilde{r}(-\boldsymbol{\delta})\#:\tilde{r}(\boldsymbol{\beta}):=-(\sqrt{2}):\tilde{r}(\boldsymbol{\alpha}):. \tag{27.6.23}$$

Our proof is complete, and in the process of proof we have exhibited explicit expressions for the 5 vector fields $\tilde{\mathcal{G}}^{0,1}_{0,0}$, $\tilde{\mathcal{G}}^{0,1}_{\pm1,\pm1}$, and $\tilde{\mathcal{G}}^{0,1}_{\mp1,\pm1}$ that span $\tilde{\mathcal{G}}^{0,1}$ in the monomial basis. If desired, these vector fields can be transformed to the resonance basis with the aid of the operator

$$\hat{\mathcal{A}}(\pi/8)=\exp[-i(\pi/8)\#p_1^2-q_1^2+p_2^2-q_2^2\#]. \tag{27.6.24}$$

Of course, the results of such a transformation will again be linear combinations of $\tilde{\mathcal{G}}^{0,1}_{0,0}$, $\tilde{\mathcal{G}}^{0,1}_{\pm1,\pm1}$, and $\tilde{\mathcal{G}}^{0,1}_{\mp1,\pm1}$ because $\hat{\mathcal{A}}(\pi/8)$ is generated by an $\#f_2\#$ and we know that the set of $\mathcal{G}^{0,1}$ is transformed into itself under such transformations. For example, we have the result

$$\mathcal{G}^{0,1}_{0,0}=\hat{\mathcal{A}}(\pi/8)\tilde{\mathcal{G}}^{0,1}_{0,0}= \tag{27.6.25}$$

At this juncture we point out that there is a one-to-one correspondence between the elements $\tilde{\mathcal{G}}^{0,0}_{0,0}$ and $\tilde{\mathcal{G}}^{0,1}_{k,\ell}$ and the matrices JA of Section 4.3. We first note that, according to (4.3.3), the matrices JA are transformed into themselves under the (commutator) action of $sp(2n)$, and therefore must form a representation of $sp(2n)$. Also, taken together, the matrices JS [which generate $sp(2n)$] and the matrices JA generate $g\ell(2n)$. Similarly, the vector fields $\mathcal{H}^{2,0}$ [which generate $sp(2n)$] and the vector fields $\mathcal{G}^{0,1}$ and $\mathcal{G}^{0,0}$ span $\mathcal{L}_{\boldsymbol{g}^1}$, and it is easily verified that $\mathcal{L}_{\boldsymbol{g}^1}$ in turn generates $g\ell(2n)$. In analogy to (7.2.4), write the relations

$$\tilde{\mathcal{G}}^{0,0}_{0,0}z_c=[J\tilde{A}(0,0;0,0)z]_c=\sum_d[J\tilde{A}(0,0;0,0)]_{cd}z_d, \tag{27.6.26}$$

$$\tilde{\mathcal{G}}^{0,1}_{k,\ell}z_c=[J\tilde{A}(0,1;k,\ell)z]_c=\sum_d[J\tilde{A}(0,1;k,\ell)]_{cd}z_d. \tag{27.6.27}$$

Then, from the definitions (6.6) through (6.11), we find the results

$$\tilde{A}(0,0;0,0)=-J=\begin{pmatrix}0&0&-1&0\\0&0&0&-1\\1&0&0&0\\0&1&0&0\end{pmatrix}, \tag{27.6.28}$$

$$\tilde{A}(0,1;0,0) = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}, \tag{27.6.29}$$

$$\tilde{A}(0,1;1,1) = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{27.6.30}$$

$$\tilde{A}(0,1;-1,-1) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}, \tag{27.6.31}$$

$$\tilde{A}(0,1;-1,1) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{27.6.32}$$

$$\tilde{A}(0,1;1,-1) = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \tag{27.6.33}$$

Note that, as expected, the matrices $\tilde{A}$ are antisymmetric and span the space of 4×4 antisymmetric matrices. Also, the matrix JA, when exponentiated and with A given by any multiple of the $\tilde{A}$ in (6.28), produces a positive multiple of the identity matrix. The remaining JA, with A any linear combination of the five $\tilde{A}$ given by (6.29) through (6.33), are traceless and therefore are in $s\ell(4,\mathbb{R})$.

We now turn to the general case $\mathcal{L}_{\boldsymbol{g}^m}$ with $m \geq 1$. Any such vector field can be written in the form

$$\mathcal{L}_{\boldsymbol{g}^m} = \sum_{a=1}^{4} g_a^m(\partial/\partial z_a). \tag{27.6.34}$$

We know that the g_a^m carry the representation $\Gamma(m,0)$ and the $(\partial/\partial z_a)$ carry the representation $\Gamma(1,0)$. It follows as before from the derivation property of $\#f_2\#$ that $\mathcal{L}_{\boldsymbol{g}^m}$ must carry the direct product representation $\Gamma(m,0) \otimes \Gamma(1,0)$. In the case of $sp(4)$ there is the Clebsch-Gordan series result

$$\Gamma(m,0) \otimes \Gamma(1,0) = \Gamma(m+1,0) \oplus \Gamma(m-1,1) \oplus \Gamma(m-1,0). \tag{27.6.35}$$

Consequently, any $\mathcal{L}_{\boldsymbol{g}^m}$ with $m \geq 1$ has the unique decomposition

$$\mathcal{L}_{\boldsymbol{g}^m} = \mathcal{H}^{m+1,0} + \mathcal{G}^{m-1,1} + \mathcal{G}^{m-1,0}. \tag{27.6.36}$$

Here $\mathcal{H}^{m+1,0}$ is a Hamiltonian vector field that carries the representation $\Gamma(m+1,0)$, and $\mathcal{G}^{m-1,1}$ and $\mathcal{G}^{m-1,0}$ are non-Hamiltonian vector fields that carry the representations $\Gamma(m-1,1)$ and $\Gamma(m-1,0)$, respectively.

The vector field $\mathcal{H}^{m+1,0}$ is of the form $: h_{m+1} :$. Finding a basis for $\mathcal{G}^{m-1,0}$ is also easy. Since Σ carries the representation $\Gamma(0,0)$, we have the result that $\mathcal{G}^{m-1,0}$ is of the form

$$\mathcal{G}^{m-1,0} = f_{m-1}\Sigma, \tag{27.6.37}$$

where f_{m-1} is any homogeneous polynomial of degree $(m-1)$. Finding the basis elements for $\mathcal{G}^{m-1,1}$ in general requires some work. One may, for example, follow a procedure similar to that used to find the basis for $\mathcal{G}^{0,1}$. A more systematic procedure would be to compute and tabulate the complete set of Clebsch-Gordan coefficients for the first several representations of $sp(4)$. These coefficients could then be used for modest m to find a basis for $\Gamma(m-1,1)$ in terms of the basis elements for $\Gamma(m,0) \times \Gamma(1,0)$.

As a specific example, let us consider vector fields of the form $\mathcal{L}_{\boldsymbol{g}^2}$. This set of fields has dimension $4N(2,4) = 40$. The set of Hamiltonian vector fields $\mathcal{H}^{3,0}$ has dimension $N(3,4) = 20$. The set of non-Hamiltonian vector fields $\mathcal{G}^{1,0}$ has dimension $\dim\Gamma(1,0) = 4$. See Figure 5.3. They are spanned by the basis elements $z_a\Sigma$. The set of non-Hamiltonian vector fields $\mathcal{G}^{1,1}$ has dimension $\dim\Gamma(1,1) = 16$. See Figure 5.6. Finding a basis for them would require more work. Some further tools for this task are described in Sections 21.10 and 21.11.2.

Exercises

27.6.1. The choice of $\tilde{\mathcal{G}}^{0,1}_{0,0}$ as given in (6.7) must be made with care to assure that it is "pure" $\Gamma(0,1)$ and contains no $\Gamma(0,0)$ or $\Gamma(2,0)$ "contamination". For example, one could add any amount of Σ and $: \tilde{c}^j :$ to the selected $\tilde{\mathcal{G}}^{0,1}_{0,0}$ and still satisfy (6.14) with $k, \ell = 0$. Show that the truth of (6.18) ensures that $\tilde{\mathcal{G}}^{0,0}_{0,1}$ as defined by (6.7) has no $\Gamma(0,0)$ contamination. Show that the truth of (6.16) and (6.17) ensures that $\tilde{\mathcal{G}}^{0,1}_{0,0}$ as defined by (6.7) has no $\Gamma(2,0)$ contamination. Verify the relations (6.8) through (6.23). The basis elements given by (6.7) through (6.11) have the weights displayed in Figure 5.4. Verify that any attempt to raise or lower an element to produce one with a weight different from those shown in Figure 5.4 leads to a null result as in (6.15) through (6.18).

27.6.2. Verify (6.28) through (6.33).

27.6.3. The relation (6.35) implies the relation

$$\begin{aligned}&[\dim\Gamma(m,0)][\dim\Gamma(1,0)] = \\ &\dim\Gamma(m+1,0) + \dim\Gamma(m-1,1) + \dim\Gamma(m-1,0).\end{aligned} \tag{27.6.38}$$

Verify this relation using (5.4).

27.7 Structure of $sp(6,\mathbb{R})$

The Lie algebra $sp(6,\mathbb{R})$ is 21 dimensional, and its Cartan subalgebra is 3 dimensional. Therefore, in the Cartan basis, there should be 18 ladder operators. They are labelled by 18 three-component root vectors consisting of 9 vectors and their negatives. For convenience, we will call these 9 vectors $\boldsymbol{\alpha}^j$, $\boldsymbol{\beta}^j$, and $\boldsymbol{\gamma}^j$ where j ranges from 1 through 3. They are given in terms of three orthogonal unit vectors $\boldsymbol{e}^1$ through $\boldsymbol{e}^3$ by the relations

$$\boldsymbol{\alpha}^1 = 2\boldsymbol{e}^1, \tag{27.7.1}$$

$$\boldsymbol{\alpha}^2 = \boldsymbol{e}^1 + \boldsymbol{e}^2, \tag{27.7.2}$$

$$\boldsymbol{\alpha}^3 = \boldsymbol{e}^1 - \boldsymbol{e}^2, \tag{27.7.3}$$

$$\boldsymbol{\beta}^1 = 2\boldsymbol{e}^2, \tag{27.7.4}$$

$$\boldsymbol{\beta}^2 = \boldsymbol{e}^2 + \boldsymbol{e}^3, \tag{27.7.5}$$

$$\boldsymbol{\beta}^3 = \boldsymbol{e}^2 - \boldsymbol{e}^3, \tag{27.7.6}$$

$$\boldsymbol{\gamma}^1 = 2\boldsymbol{e}^3, \tag{27.7.7}$$

$$\boldsymbol{\gamma}^2 = \boldsymbol{e}^3 + \boldsymbol{e}^1, \tag{27.7.8}$$

$$\boldsymbol{\gamma}^3 = \boldsymbol{e}^3 - \boldsymbol{e}^1. \tag{27.7.9}$$

The 18 $sp(6)$ root vectors are shown in Figure 7.1. Note they are all of the form $(\pm\boldsymbol{e}^i \pm \boldsymbol{e}^j)$ with the signs taken independently and the zero vector omitted. They satisfy the normalization relations

$$\sum_{\boldsymbol{\mu}} (\boldsymbol{e}^i \cdot \boldsymbol{\mu})(\boldsymbol{\mu} \cdot \boldsymbol{e}^j) = 16\delta_{ij}. \tag{27.7.10}$$

The Lie algebra $sp(6,\mathbb{R})$ is generated by the monomials $z_a z_b$ with a, b ranging from 1 to 6. In analogy with the case of $sp(4,\mathbb{R})$, it is convenient to use as the basis for the Cartan subalgebra the polynomials

$$c^1 = -(i/2)(p_1^2 + q_1^2), \tag{27.7.11}$$

$$c^2 = -(i/2)(p_2^2 + q_2^2), \tag{27.7.12}$$

$$c^3 = -(i/2)(p_3^2 + q_3^2). \tag{27.7.13}$$

Their associated Lie operators are Hermitian and, when exponentiated, generate a 3-torus which is a maximal torus in $sp(6,\mathbb{R})$. For the ladder operators in $sp(6)$ we use the polynomials

$$r(\pm\boldsymbol{\alpha}^1) = (\sqrt{2}/4)(q_1 \pm ip_1)^2, \tag{27.7.14}$$

$$r(\pm\boldsymbol{\alpha}^2) = (1/2)(q_1 \pm ip_1)(q_2 \pm ip_2), \tag{27.7.15}$$

$$r(\pm\boldsymbol{\alpha}^3) = (i/2)(q_1 \pm ip_1)(q_2 \mp ip_2), \tag{27.7.16}$$

$$r(\pm\boldsymbol{\beta}^1) = (\sqrt{2}/4)(q_2 \pm ip_2)^2, \tag{27.7.17}$$

$$r(\pm\boldsymbol{\beta}^2) = (1/2)(q_2 \pm ip_2)(q_3 \pm ip_3), \tag{27.7.18}$$

$$r(\pm\boldsymbol{\beta}^3) = (i/2)(q_2 \pm ip_2)(q_3 \mp ip_3), \tag{27.7.19}$$

$$r(\pm\boldsymbol{\gamma}^1) = (\sqrt{2}/4)(q_3 \pm ip_3)^2, \tag{27.7.20}$$

$$r(\pm\boldsymbol{\gamma}^2) = (1/2)(q_3 \pm ip_3)(q_1 \pm ip_1), \tag{27.7.21}$$

$$r(\pm\boldsymbol{\gamma}^3) = (i/2)(q_3 \pm ip_3)(q_1 \mp ip_1). \tag{27.7.22}$$

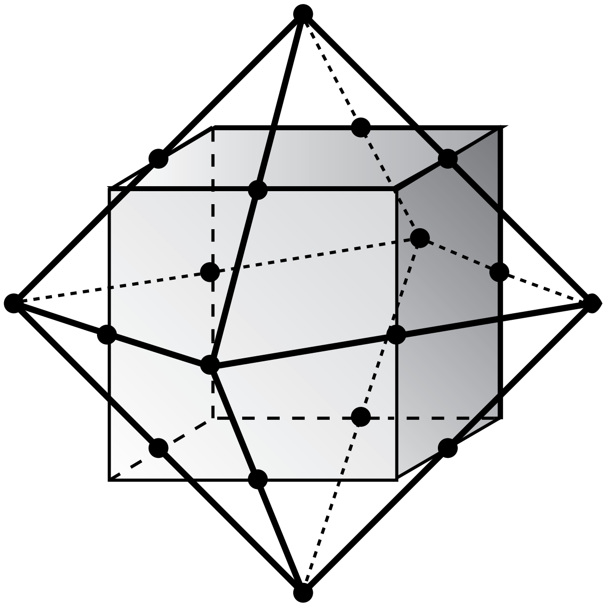


Figure 27.7.1: Root diagram showing the root vectors for $sp(6)$. The 6 tips of the long root vectors $\pm\boldsymbol{\alpha}^1$, $\pm\boldsymbol{\beta}^1$, $\pm\boldsymbol{\gamma}^1$ form the vertices of a regular octahedron. These root vectors have length 2. The remaining 12 short root vectors have length $\sqrt{2}$, and their tips lie at the midpoints of the 12 edges of the unit cube (the cube with edge 2).

Their associated Lie operators obey the standard conjugation relations (4.14). Also, the Lie algebra generated by the c^j and the $r(\boldsymbol{\mu})$ satisfy the standard rules (4.15) through (4.18). For the case of $sp(6)$, the $N(\boldsymbol{\mu},\boldsymbol{\nu})$ have the values $\pm\sqrt{2}$. The positive N's are *. As before, for the scalar product (7.3.12), the basis elements c^j and $r(\boldsymbol{\mu})$ are orthonormal and therefore satisfy the relations (4.19) through (4.22).

We close this section by examining how $sp(4)$, $su(3)$, and $so(3)$ reside within $sp(6)$. The presence of $sp(4)$ within $sp(6)$ is obvious. Comparison of (4.6) and (4.7) with (7.11) and (7.12) shows that the c^j (with $j = 1, 2$) are identical for sp(4) and sp(6). Also, comparison of (4.1) through (4.4) with (7.1) through (7.4) indicates that, apart from labeling, the root vectors of $sp(4)$ are identical to a subset of those for $sp(6)$. Therefore, there is the correspondence

$$\boldsymbol{\alpha} \leftrightarrow \boldsymbol{\alpha}^1,\ \ \boldsymbol{\beta} \leftrightarrow \boldsymbol{\alpha}^2,\ \ \boldsymbol{\gamma} \leftrightarrow \boldsymbol{\beta}^1,\ \ \boldsymbol{\delta} \leftrightarrow -\boldsymbol{\alpha}^3. \tag{27.7.23}$$

Figure 7.2 shows the $sp(6)$ root vectors of Figure 7.1 viewed from above (looking against the $\boldsymbol{e}^3$ axis). From this perspective, it is obvious that the root vectors (7.23) are arranged as required for $sp(4)$. See Figure 4.1. Finally, comparison of (4.10) through (4.13) with (7.14) through (7.17) gives the relations

$$r(\pm\boldsymbol{\alpha}) = r(\pm\boldsymbol{\alpha}^1),\ \ r(\pm\boldsymbol{\beta}) = r(\pm\boldsymbol{\alpha}^2),\ \ r(\pm\boldsymbol{\gamma}) = r(\pm\boldsymbol{\beta}^1),\ \ r(\pm\boldsymbol{\delta}) = r(\mp\boldsymbol{\alpha}^3). \tag{27.7.24}$$

The $sp(4)$ just identified within $sp(6)$ is the obvious one. Continued examination of the $sp(6)$ root diagram of Figure 7.1 indicates that there are two more $sp(4)$ subgroups within $sp(6)$ gotten from the one just described by cyclically permuting the indices $1, 2, 3$ on the variables q_1, q_2, q_3 and p_1, p_2, p_3.

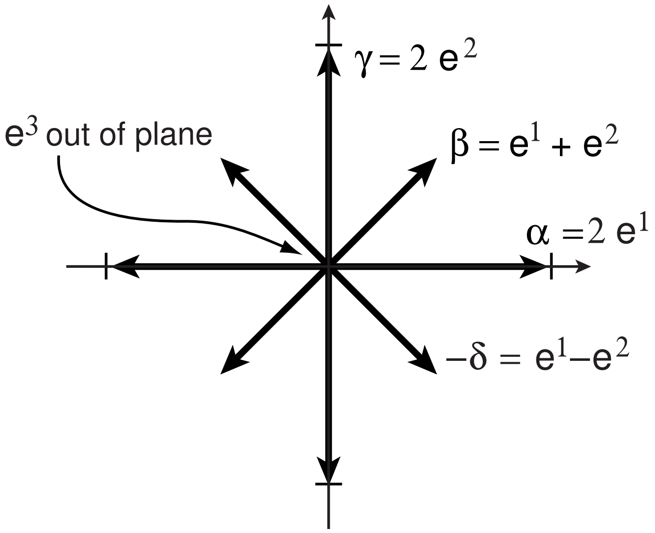


Figure 27.7.2: Top view of $sp(6)$ root vectors of Figure 7.1 showing root vectors of an $sp(4)$ subgroup. Only the $sp(6)$ root vectors in the $\boldsymbol{e}^1$, $\boldsymbol{e}^2$ plane are displayed. For clarity, all others are omitted. The vector $\boldsymbol{e}^3$ is out of the plane of the paper.

The presence of $su(3)$ and $u(3)$ within $sp(6)$ is more subtle. Consider the $sp(6)$ root vectors $\boldsymbol{\alpha}^3$, $\boldsymbol{\beta}^3$, $\boldsymbol{\gamma}^3$ given by (7.3), (7.6), and (7.9). They all have length $\sqrt{2}$. They are also linearly dependent and therefore lie in a plane,

$$\boldsymbol{\alpha}^3 + \boldsymbol{\beta}^3 + \boldsymbol{\gamma}^3 = 0. \tag{27.7.25}$$

Within this plane they radiate from the origin like spokes equally "spaced" by angles of 120°. To verify this assertion, first note that the normal to this plane is given by the vector

$$\boldsymbol{\alpha}^3 \times \boldsymbol{\beta}^3 = \boldsymbol{\beta}^3 \times \boldsymbol{\gamma}^3 = \boldsymbol{\gamma}^3 \times \boldsymbol{\alpha}^3 = \boldsymbol{e}^1 + \boldsymbol{e}^2 + \boldsymbol{e}^3. \tag{27.7.26}$$

Let $\boldsymbol{n}$ be the normal unit vector

$$\boldsymbol{n} = (\boldsymbol{e}^1 + \boldsymbol{e}^2 + \boldsymbol{e}^3)/\sqrt{3}. \tag{27.7.27}$$

Use of (*) shows that there is the relation

$$\boldsymbol{\beta}^3 = R(\boldsymbol{n}, 2\pi/3)\boldsymbol{\alpha}^3, \tag{27.7.28}$$

$$\boldsymbol{\gamma}^3 = R(\boldsymbol{n}, 2\pi/3)\boldsymbol{\beta}^3, \tag{27.7.29}$$

$$\boldsymbol{\alpha}^3 = R(\boldsymbol{n}, 2\pi/3)\boldsymbol{\gamma}^3. \tag{27.7.30}$$

Figure 7.3 shows the $sp(6)$ root vectors of Figure 7.1 viewed against the unit vector $\boldsymbol{n}$. From this perspective it is evident that the vectors $\boldsymbol{\alpha}^3$, $\boldsymbol{\beta}^3$, $\boldsymbol{\gamma}^3$ and their negatives are arranged as required for the root vectors of $su(3)$. Comparison of Figures 5.8.1 and 7.3 gives the correspondence

$$\pm\,\boldsymbol{\alpha} \leftrightarrow \pm\boldsymbol{\alpha}^3,\ \ \pm\boldsymbol{\beta} \leftrightarrow \mp\boldsymbol{\gamma}^3,\ \ \pm\boldsymbol{\gamma} \leftrightarrow \pm\boldsymbol{\beta}^3. \tag{27.7.31}$$

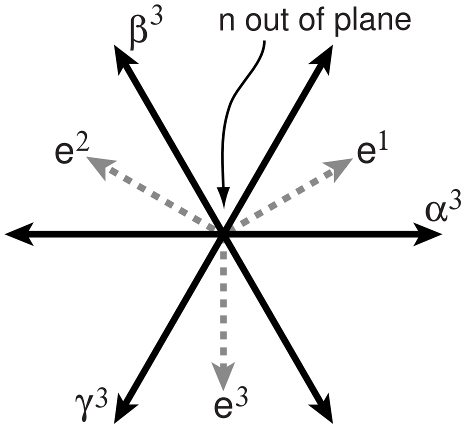


Figure 27.7.3: View against the unit vector $\boldsymbol{n}$ of the $sp(6)$ root vectors of Figure 7.1 showing root vectors of an $su(3)$ subgroup. Only the $sp(6)$ root vectors in the $\boldsymbol{\alpha}^3$, $\boldsymbol{\beta}^3$, $\boldsymbol{\gamma}^3$ plane and the $\boldsymbol{e}^1$, $\boldsymbol{e}^2$, $\boldsymbol{e}^3$ axes are displayed. For clarity, all others are omitted. The vector $\boldsymbol{n}$ is out of the plane of the paper.

Correspondingly, comparison of (7.16), (7.19), and (7.22) with (5.8.35) gives the relations

$$r(\pm\boldsymbol{\alpha}) = r(\pm\boldsymbol{\alpha}^3), \quad r(\pm\boldsymbol{\beta}) = r(\mp\boldsymbol{\gamma}^3), \quad r(\pm\boldsymbol{\gamma}) = r(\pm\boldsymbol{\beta}^3). \tag{27.7.32}$$

Finally, comparison of (5.8.5) with (7.11) through (7.13) gives the relations

$$b^0 = i(c^1 + c^2 + c^3), \quad b^3 = i(c^1 - c^2), \quad b^8 = (i/\sqrt{3})(c^1 + c^2 - 2c^3). \tag{27.7.33}$$

The ladder elements r in (7.32) combined with b^3 and b^8 in (7.33) span $su(3)$; and they all together along with b^0 span $u(3)$.

The $su(3)$ [and corresponding $u(3)$] just identified within $sp(6)$ is one of several such subgroups. Continued examination of Figure 7.1 indicates that there are three more. We know that the root vectors $\pm\boldsymbol{\alpha}^1$, $\pm\boldsymbol{\beta}^1$, $\pm\boldsymbol{\gamma}^1$ form the 6 vertices of a regular octahedron. An octahedron has 8 triangular faces consisting of 4 opposite pairs. There is an $su(3)$ set of root vectors in each plane through the origin lying between and parallel to each pair of opposite faces.

Finally, there is an $so(3)$ subalgebra within $sp(6)$ whose presence is not obvious from looking at the $sp(6)$ root diagram. The L_j defined by (5.8.89) generate simultaneous rotations in the q_1, q_2, q_3 and p_1, p_2, p_3 spaces. By (7.16), (7.19), and (7.22) they are related to elements in the Cartan basis by the equation

$$L_1 = q_2p_3 - q_3p_2 = r(\boldsymbol{\beta}^3) - r(-\boldsymbol{\beta}^3), \tag{27.7.34}$$

$$L_2 = q_3p_1 - q_1p_3 = r(\boldsymbol{\gamma}^3) - r(-\boldsymbol{\gamma}^3), \tag{27.7.35}$$

$$L_3 = q_1p_2 - q_2p_1 = r(\boldsymbol{\alpha}^3) - r(-\boldsymbol{\alpha}^3). \tag{27.7.36}$$

Note that these elements are all within $su(3)$.

Exercises

27.7.1. Verify that, with the scalar product (7.3.12), the basis elements c^j and $r(\boldsymbol{\mu})$ form an orthonormal set.

27.7.2. Verify the relations (7.28) through (7.30).

27.7.3. Verify that there are other $so(3)$ subalgebras in $sp(6)$ associated with the other $su(3)$ subalgebras in $sp(6)$.

27.8 Representations of $sp(6,\mathbb{R})$

The description of representations of $sp(6)$ follows the same general Cartan procedure as described for $su(3)$ in Section 5.8 and $sp(4)$ in Section 21.5. For $sp(6)$, since it has rank 3, there are three fundamental weights $\boldsymbol{\phi}^1$, $\boldsymbol{\phi}^2$ and $\boldsymbol{\phi}^3$. They are given by the relations

$$\boldsymbol{\phi}^1 = \boldsymbol{e}^1 = \boldsymbol{\alpha}^1/2, \tag{27.8.1}$$

$$\boldsymbol{\phi}^2 = \boldsymbol{e}^1 + \boldsymbol{e}^2 = \boldsymbol{\alpha}^2, \tag{27.8.2}$$

$$\boldsymbol{\phi}^3 = \boldsymbol{e}^1 + \boldsymbol{e}^2 + \boldsymbol{e}^3, \tag{27.8.3}$$

and are shown in Figure 8.1 along with the $sp(6)$ root vectors. Thus, for $sp(6)$, every highest weight $\boldsymbol{w}^h$ is of the form

$$\boldsymbol{w}^h = \ell\boldsymbol{\phi}^1 + m\boldsymbol{\phi}^2 + n\boldsymbol{\phi}^3 = (\ell+m+n)\boldsymbol{e}^1 + (m+n)\boldsymbol{e}^2 + n\boldsymbol{e}^3, \tag{27.8.4}$$

where ℓ, m, and n are arbitrary nonnegative integers. Correspondingly, for each ℓ, m, n triplet, there is an irreducible representation $\Gamma(\ell,m,n)$ with highest weight $\boldsymbol{w}^h$ given by (8.4). It can be shown that the dimension of $\Gamma(\ell,m,n)$ is given by the relation

$$\begin{aligned}\dim\Gamma(\ell,m,n) = {} & \frac{1}{720}(\ell+2m+2n+5)(\ell+m+2n+4)(\ell+m+n+3) \\ & \times(\ell+m+2)(m+2n+3)(m+n+2)(\ell+1)(m+1)(n+1).\end{aligned} \tag{27.8.5}$$

Again see Exercise 5.1. The representations are also self conjugate,

$$\overline{\Gamma}(\ell,m,n) = \Gamma(\ell,m,n). \tag{27.8.6}$$

See Exercise 3.7.36. For quick reference the dimensions of the first few representations are listed in Table 8.1 below. Where there is no possibility of confusion, we will sometimes refer to a representation by its dimension. Note that $\Gamma(0,1,0)$ and $\Gamma(0,0,1)$ both have dimension 14.

From a knowledge of the root vectors and the highest weight it is a simple matter to construct weight diagrams for the various low-dimensional representations. Figures 8.2 through 8.5 show weight diagrams for the first few representations. Inspection of these figures and reference to Table 8.1 shows that the weights must have unit multiplicities for the representations $\Gamma(0,0,0)$ and $\Gamma(1,0,0)$. For $\Gamma(0,1,0)$ the weight at the origin has

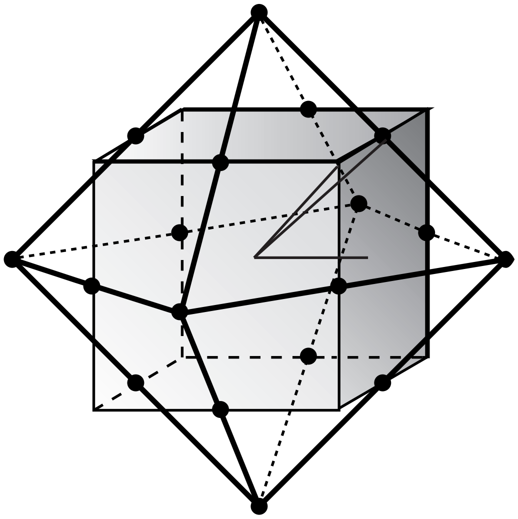


Figure 27.8.1: Fundamental weights $\boldsymbol{\phi}^1$, $\boldsymbol{\phi}^2$ and $\boldsymbol{\phi}^3$ for $sp(6)$. The root vectors are also shown.

Table 27.8.1: Dimensions of Representations of $sp(6)$.

ℓ	m	n	$\dim\Gamma(\ell,m,n)$	ℓ	m	n	$\dim\Gamma(\ell,m,n)$
0	0	0	1	3	0	0	56
1	0	0	6	2	1	0	189
0	1	0	14	2	0	1	216
0	0	1	14	1	2	0	350
2	0	0	21	1	1	1	512
1	1	0	64	1	0	2	378
1	0	1	70	0	3	0	385
0	2	0	90	0	2	1	616
0	1	1	126	0	1	2	594
0	0	2	84	0	0	3	330

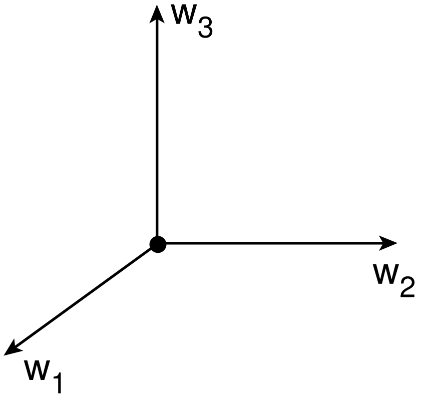


Figure 27.8.2: Weight diagram for the representation $1 = \Gamma(0,0,0)$.

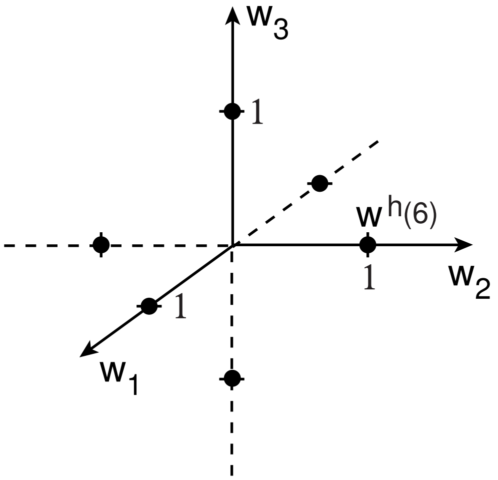


Figure 27.8.3: Weight diagram for the fundamental representation $6 = \Gamma(1,0,0)$.

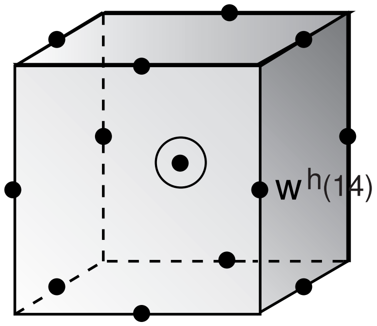


Figure 27.8.4: Weight diagram for the representation $14 = \Gamma(0,1,0)$. The circled weight at the origin has multiplicity 2. Observe from Figure 7.1 that the 12 other weights are located at the tips of the root vectors having length $\sqrt{2}$.

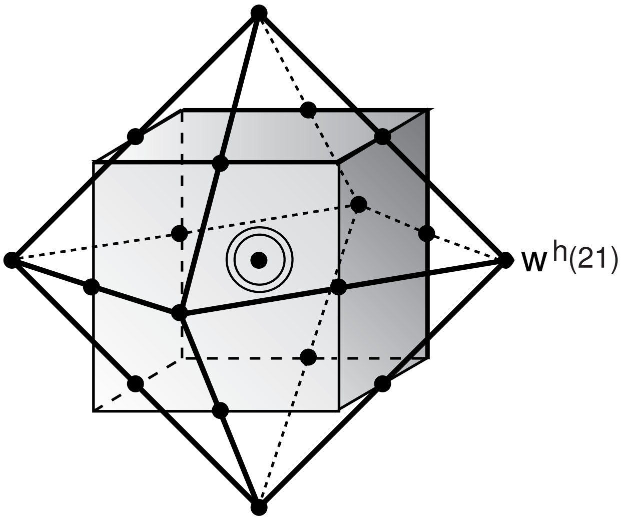


Figure 27.8.5: Weight diagram for the adjoint representation $21 = \Gamma(2,0,0)$. The doubly circled weight at the origin has multiplicity 3. The 18 other weights are located at the tips of the $sp(6)$ root vectors.

multiplicity 2; and for $\Gamma(2,0,0)$, which is the adjoint or regular representation, the weight vector at the origin has multiplicity 3.

Let $\mathcal{P}_\ell$ be the space of homogeneous polynomials of degree ℓ in the variables z_a with $a = 1,6$. Then, by arguments that are now familiar, the space $\mathcal{P}_\ell$ forms a representation of $sp(6,\mathbb{R})$. To see what representations occur we again employ a complex symplectic and unitary $\mathcal{A}(\pi/8)$ now defined by the equation

$$\mathcal{A}(\pi/8) = \exp[-i(\pi/8) : p_1^2 - q_1^2 + p_2^2 - q_2^2 + p_3^2 - q_3^2 :]. \tag{27.8.7}$$

It transforms between the Cartesian and resonance bases by the rule

$$\mathcal{A}(\pi/8)(q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2}q_3^{r_3}p_3^{s_3}) = (1/\sqrt{2})^{r_1+s_1+r_2+s_2+r_3+s_3}(i)^{s_1+s_2+s_3} \times$$
$$(q_1+ip_1)^{r_1}(q_1-ip_1)^{s_1}(q_2+ip_2)^{r_2}(q_2-ip_2)^{s_2}(q_3+ip_3)^{r_3}(q_3-ip_3)^{s_3}. \tag{27.8.8}$$

Use of $\mathcal{A}$ gives the transformed Lie basis polynomials $\tilde{c}^j$ and $\tilde{r}(\boldsymbol{\mu})$ listed below:

$$\tilde{c}^j = \mathcal{A}(-\pi/8)c^j = -q_j p_j, \tag{27.8.9}$$

$$\tilde{r}(\boldsymbol{\alpha}^1) = \mathcal{A}(-\pi/8)r(\boldsymbol{\alpha}^1) = (1/\sqrt{2})q_1^2, \tag{27.8.10}$$

$$\tilde{r}(-\boldsymbol{\alpha}^1) = \mathcal{A}(-\pi/8)r(-\boldsymbol{\alpha}^1) = -(1/\sqrt{2})p_1^2, \tag{27.8.11}$$

$$\tilde{r}(\boldsymbol{\alpha}^2) = \mathcal{A}(-\pi/8)r(\boldsymbol{\alpha}^2) = q_1 q_2, \tag{27.8.12}$$

$$\tilde{r}(-\boldsymbol{\alpha}^2) = \mathcal{A}(-\pi/8)r(-\boldsymbol{\alpha}^2) = -p_1 p_2, \tag{27.8.13}$$

$$\tilde{r}(\boldsymbol{\alpha}^3) = \mathcal{A}(-\pi/8)r(\boldsymbol{\alpha}^3) = q_1 p_2, \tag{27.8.14}$$

$$\tilde{r}(-\boldsymbol{\alpha}^3) = \mathcal{A}(-\pi/8)r(-\boldsymbol{\alpha}^3) = q_2 p_1, \tag{27.8.15}$$

$$\tilde{r}(\boldsymbol{\beta}^1) = \mathcal{A}(-\pi/8)r(\boldsymbol{\beta}^1) = (1/\sqrt{2})q_2^2, \tag{27.8.16}$$

$$\tilde{r}(-\boldsymbol{\beta}^1) = \mathcal{A}(-\pi/8)r(-\boldsymbol{\beta}^1) = -(1/\sqrt{2})p_2^2, \tag{27.8.17}$$

$$\tilde{r}(\boldsymbol{\beta}^2) = \mathcal{A}(-\pi/8)r(\boldsymbol{\beta}^2) = q_2q_3, \tag{27.8.18}$$

$$\tilde{r}(-\boldsymbol{\beta}^2) = \mathcal{A}(-\pi/8)r(-\boldsymbol{\beta}^2) = -p_2p_3, \tag{27.8.19}$$

$$\tilde{r}(\boldsymbol{\beta}^3) = \mathcal{A}(-\pi/8)r(\boldsymbol{\beta}^3) = q_2p_3, \tag{27.8.20}$$

$$\tilde{r}(-\boldsymbol{\beta}^3) = \mathcal{A}(-\pi/8)r(-\boldsymbol{\beta}^3) = q_3p_2, \tag{27.8.21}$$

$$\tilde{r}(\boldsymbol{\gamma}^1) = \mathcal{A}(-\pi/8)r(\boldsymbol{\gamma}^1) = (1/\sqrt{2})q_3^2, \tag{27.8.22}$$

$$\tilde{r}(-\boldsymbol{\gamma}^1) = \mathcal{A}(-\pi/8)r(-\boldsymbol{\gamma}^1) = -(1/\sqrt{2})p_3^2, \tag{27.8.23}$$

$$\tilde{r}(\boldsymbol{\gamma}^2) = \mathcal{A}(-\pi/8)r(\boldsymbol{\gamma}^2) = q_3q_1, \tag{27.8.24}$$

$$\tilde{r}(-\boldsymbol{\gamma}^2) = \mathcal{A}(-\pi/8)r(-\boldsymbol{\gamma}^2) = -p_3p_1, \tag{27.8.25}$$

$$\tilde{r}(\boldsymbol{\gamma}^3) = \mathcal{A}(-\pi/8)r(\boldsymbol{\gamma}^3) = q_3p_1, \tag{27.8.26}$$

$$\tilde{r}(-\boldsymbol{\gamma}^3) = \mathcal{A}(-\pi/8)r(-\boldsymbol{\gamma}^3) = q_1p_3. \tag{27.8.27}$$

Since $\mathcal{A}$ is both symplectic and unitary, the transformed basis polynomials also obey the Poisson bracket rules (4.15) through (4.18), and also satisfy the orthonormality conditions (4.19) through (4.22) and the conjugation relations (4.8) and (4.14).

When the $\tilde{c}^j$ and $\tilde{r}(\boldsymbol{\mu})$ act on the monomials $q_1^{r_1}p_1^{s_1}q_2^{r_2}p_2^{s_2}q_3^{r_3}p_3^{s_3}$, there are relations analogous to those in (5.19) through (5.28), and it is evident that $sp(6)$ acts irreducibly on $\mathcal{P}_\ell$. Also, q_1^ℓ is the vector of highest weight in $\mathcal{P}_\ell$ and has the weight $\boldsymbol{w}^h = \ell\boldsymbol{\phi}^1$,

$$:\tilde{c}^1: q_1^\ell = \ell q_1^\ell = (\boldsymbol{e}^1 \cdot \ell\boldsymbol{\phi}^1)q_1^\ell = (\boldsymbol{e}^1 \cdot \boldsymbol{w}^h)q_1^\ell, \tag{27.8.28}$$

$$:\tilde{c}^2: q_1^\ell = 0 = (\boldsymbol{e}^2 \cdot \ell\boldsymbol{\phi}^1)q_1^\ell = (\boldsymbol{e}^2 \cdot \boldsymbol{w}^h)q_1^\ell, \tag{27.8.29}$$

$$:\tilde{c}^3: q_1^\ell = 0 = (\boldsymbol{e}^3 \cdot \ell\boldsymbol{\phi}^1)q_1^\ell = (\boldsymbol{e}^3 \cdot \boldsymbol{w}^h)q_1^\ell. \tag{27.8.30}$$

It follows that $\mathcal{P}_\ell$ carries the representation $\Gamma(\ell,0,0)$. We also have, as expected, the result

$$\dim \mathcal{P}_\ell = N(\ell,6) = (1/120)(\ell+5)(\ell+4)(\ell+3)(\ell+2)(\ell+1) = \dim\Gamma(\ell,0,0). \tag{27.8.31}$$

At this point we observe that the relations (5.8.27) and (5.8.31) can be written in the form

$$\begin{aligned}\Gamma(\ell,0,0) = &\sum_{m+n=\ell}\hat{\Gamma}(m,n) \oplus \sum_{m+n=\ell-2}\hat{\Gamma}(m,n) \oplus \sum_{m+n=\ell-4}\hat{\Gamma}(m,n) \oplus \cdots \\ &\oplus \hat{\Gamma}(0,0), \text{ for } \ell \text{ even;}\end{aligned} \tag{27.8.32}$$

$$\begin{aligned}\Gamma(\ell,0,0) = &\sum_{m+n=\ell}\hat{\Gamma}(m,n) \oplus \sum_{m+n=\ell-2}\hat{\Gamma}(m,n) \oplus \sum_{m+n=\ell-4}\hat{\Gamma}(m,n) \oplus \cdots \\ &\oplus \hat{\Gamma}(1,0) \oplus \hat{\Gamma}(0,1), \text{ for } \ell \text{ odd.}\end{aligned} \tag{27.8.33}$$

[Here we have used the symbols $\hat{\Gamma}(m,n)$ to denote representations of $su(3)$ so as not to be confused with the symbols $\Gamma(m,n)$ used in Section 21.5 to denote representations of $sp(4)$.] That is, $sp(6)$ representations of the form $\Gamma(\ell,0,0)$ can be decomposed into various $\hat{\Gamma}(m,n)$

representations of its $su(3)$ subgroup, and the representations listed occur once and only once.

Finally, as before, we can let $\mathcal{A}(\pi/8)$ act on both sides of 6-variable relations analogous to (5.19) through (5.28). Doing so gives results analogous to those in (2.23) through (2.25). Consequently, as before, the $:c^j:$ and $:r(\boldsymbol{\mu}):$ act on the resonance basis in the same way that the $:\tilde{c}^j:$ and $:\tilde{r}(\boldsymbol{\mu}):$ act on the monomial basis.

Exercises

27.8.1. From (7.14) verify the relation

$$r(\boldsymbol{\alpha}^1) - r(-\boldsymbol{\alpha}^1) = i\sqrt{2}q_1p_1. \tag{27.8.34}$$

Next verify that the transformation

$$\mathcal{U}(\theta) = \exp :i\theta q_1 p_1: \tag{27.8.35}$$

is symplectic and unitary, and satisfies the relations

$$\mathcal{U}(\theta)q_1 = \exp(-i\theta)q_1, \tag{27.8.36}$$

$$\mathcal{U}(\theta)p_1 = \exp(i\theta)p_1. \tag{27.8.37}$$

See Exercise 5.4.4. As a consequence verify the relations

$$\mathcal{U}(\theta)(p_1^2 + q_1^2) = p_1^2\exp(2i\theta) + q_1^2\exp(-2i\theta), \tag{27.8.38}$$

$$\mathcal{U}(\pi/2)(p_1^2 + q_1^2) = -(p_1^2 + q_1^2), \tag{27.8.39}$$

$$\mathcal{U}(\pi/2) :c^1: \mathcal{U}^{-1}(\pi/2) = -:c^1:. \tag{27.8.40}$$

Suppose that C^j and $R(\boldsymbol{\mu})$ are any set of matrices that satisfies the commutation rules analogous to (4.15) through (4.18). That is, the C^j commute and the C^j and $R(\boldsymbol{\mu})$ satisfy the rules (5.8.12) through (5.8.14). By this definition, they provide a matrix representation of $sp(6)$. Since the relation (8.40) is purely a consequence of Lie-algebraic rules, show that there must be the matrix relation

$$U(\pi/2)C^1[U(\pi/2)]^{-1} = -C^1, \tag{27.8.41}$$

where

$$U(\pi/2) = \exp\{(\pi/\sqrt{8})[R(\boldsymbol{\alpha}^1) - R(-\boldsymbol{\alpha}^1)]\}, \tag{27.8.42}$$

$$[U(\pi/2)]^{-1} = \exp\{-(\pi/\sqrt{8})[R(\boldsymbol{\alpha}^1) - R(-\boldsymbol{\alpha}^1)]\}. \tag{27.8.43}$$

Suppose $|w_1w_2w_3\rangle$ is a vector in this representation with the property

$$C^j|w_1w_2w_3\rangle = w_j|w_1w_2w_3\rangle. \tag{27.8.44}$$

Show that the vector $[U(\pi/2)]^{-1}|\boldsymbol{w}\rangle$ has the property

$$C^1[U^{-1}(\pi/2)]^{-1}|\boldsymbol{w}\rangle = -w_1[U(\pi/2)]^{-1}|\boldsymbol{w}\rangle. \tag{27.8.45}$$

Prove that if (w_1, w_2, w_3) is a weight vector, so is $(-w_1, w_2, w_3)$. Generalize this result to show that if (w_1, w_2, w_3) is a weight vector, so are the vectors $(\pm w_1, \pm w_2, \pm w_3)$ where all ± signs are taken independently. Verify similar results for $sp(2)$ and $sp(4)$. Verify that the weight diagrams shown in Sections 21.2, 21.5, and 21.8 have this property.

27.8.2. Verify (8.31).

27.8.3. Verify by a dimension count that $\Gamma(1,0,0)$ is the fundamental representation of $sp(6)$ and $\Gamma(2,0,0)$ is the adjoint representation. Repeat analogous calculations for the cases of $sp(2)$ and $sp(4)$.

27.8.4. Work out the weight diagram for the $sp(6)$ representation $\Gamma(0,0,1)$.

27.9 Symplectic Classification of Analytic Vector Fields in Six Variables

The symplectic classification of analytic vector fields in six variables is similar to the 4-variable case. As before, it suffices to consider homogeneous vector fields. The Hamiltonian vector fields $:h_\ell:$ are transformed into each other under the action of $sp(6,\mathbb{R})$, and carry the representation $\Gamma(\ell,0,0)$. Any $\mathcal{L}_{\boldsymbol{g}^0}$ is a Hamiltonian vector field, and these fields carry the representation $\Gamma(1,0,0)$. The vector field Σ defined by (6.1) with a ranging fron 1 to 6 carries the representation $\Gamma(0,0,0)$.

The 6-dimensional analog of (6.34) shows that in this case the $\mathcal{L}_{\boldsymbol{g}^\ell}$ carry the direct product representation $\Gamma(\ell,0,0)\otimes\Gamma(1,0,0)$. For $sp(6)$ there is the Clebsch-Gordan series result

$$\Gamma(\ell,0,0)\otimes\Gamma(1,0,0)=\Gamma(\ell+1,0,0)\oplus\Gamma(\ell-1,1,0)\oplus\Gamma(\ell-1,0,0). \tag{27.9.1}$$

Consequently, any $\mathcal{L}_{\boldsymbol{g}^\ell}$ with $\ell\geq 1$ has the unique decomposition

$$\mathcal{L}_{\boldsymbol{g}^\ell}=\mathcal{H}^{\ell+1,0,0}+\mathcal{G}^{\ell-1,1,0}+\mathcal{G}^{\ell-1,0,0}. \tag{27.9.2}$$

Here $\mathcal{H}^{\ell+1,0,0}$ is a Hamiltonian vector field that carries the representation $\Gamma(\ell+1,0,0)$, and is of the form $:h_{\ell+1}:$. The quantities $\mathcal{G}^{\ell-1,1,0}$ and $\mathcal{G}^{\ell-1,0,0}$ are non-Hamiltonian vector fields that carry the representations $\Gamma(\ell-1,1,0)$ and $\Gamma(\ell-1,0,0)$, respectively. The vector fields $\mathcal{G}^{\ell-1,0,0}$ are of the form

$$\mathcal{G}^{\ell-1,0,0}=f_{\ell-1}\Sigma \tag{27.9.3}$$

where $f_{\ell-1}$ is any homogeneous polynomial of degree $(\ell-1)$, The construction of the vector fields that span $\mathcal{G}^{\ell-1,1,0}$ requires special effort.

As before we will work out the simplest case $\ell=1$ for which we have the result

$$\mathcal{L}_{\boldsymbol{g}^1}=\mathcal{H}^{2,0,0}+\mathcal{G}^{0,1,0}+\mathcal{G}^{0,0,0}. \tag{27.9.4}$$

Since $\mathcal{L}_{\boldsymbol{g}^1}$ is spanned by the vector fields $z_a(\partial/\partial z_b)$ with $a,b=1$ through 6, it has dimension 36. We know that $\mathcal{H}^{2,0,0}$ has dimension $N(2,6)=21$, and $\mathcal{G}^{0,0,0}$ has dimension 1. It follows that $\mathcal{G}^{0,1,0}$ has dimension $(36-21-1)=14$, which we know is the dimension of $\Gamma(0,1,0)$.

Similar to the case of $sp(4)$ treated in Section 21.6, finding a suitable basis for the vector fields in $\mathcal{H}^{2,0,0}$ is equivalent to finding a suitable basis for the quadratic polynomials in 6-dimensional phase space; and these basis polynomials may be taken to be the $sp(6)$ generators given in the Cartesian Cartan basis by (8.9) through (8.27).

To find a basis for the non-Hamiltonian parts we will begin with the 6 mutually commuting vector fields $z_a(\partial/\partial z_a)$ with $a = 1$ through 6. The 4 vector fields $:\tilde{c}^j:$ and $\tilde{\mathcal{G}}^{0,0,0}_{0,0,0} = \Sigma$ are made from the $z_a(\partial/\partial z_a)$,

$$:\tilde{c}^1 := z_1(\partial/\partial z_1) - z_4(\partial/\partial z_4), \tag{27.9.5}$$

$$:\tilde{c}^2 := z_2(\partial/\partial z_2) - z_5(\partial/\partial z_5), \tag{27.9.6}$$

$$:\tilde{c}^3 := z_3(\partial/\partial z_3) - z_6(\partial/\partial z_6), \tag{27.9.7}$$

$$\tilde{\mathcal{G}}^{0,0,0}_{0,0,0} = \Sigma = z_1(\partial/\partial z_1)+z_2(\partial/\partial z_2)+z_3(\partial/\partial z_3)+z_4(\partial/\partial z_4)+z_5(\partial/\partial z_5)+z_6(\partial/\partial z_6). \tag{27.9.8}$$

Let us define vector fields $\tilde{\Sigma}^j$ by the equations

$$\tilde{\Sigma}^1 = q_1(\partial/\partial q_1) + p_1(\partial/\partial p_1) = z_1(\partial/\partial z_1) + z_4(\partial/\partial z_4), \tag{27.9.9}$$

$$\tilde{\Sigma}^2 = q_2(\partial/\partial q_2) + p_2(\partial/\partial p_2) = z_2(\partial/\partial z_2) + z_5(\partial/\partial z_5), \tag{27.9.10}$$

$$\tilde{\Sigma}^3 = q_3(\partial/\partial q_3) + p_3(\partial/\partial p_3) = z_3(\partial/\partial z_3) + z_6(\partial/\partial z_6). \tag{27.9.11}$$

They are obviously independent of the $:\tilde{c}^j:$ and are also made from the $z_a(\partial/\partial z_a)$.

We already have the $:\tilde{c}^j:$ and the combination

$$\Sigma = \tilde{\Sigma}^1 + \tilde{\Sigma}^2 + \tilde{\Sigma}^3. \tag{27.9.12}$$

As the 5th and 6th such linearly independent vectors we take the elements ${}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0}$ and ${}^8\tilde{\mathcal{G}}^{0,1,0}_{0,0,0}$ defined by the equations

$${}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = \tilde{\Sigma}^1 - \tilde{\Sigma}^2 = z_1(\partial/\partial z_1) + z_4(\partial/\partial z_4) - z_2(\partial/\partial z_2) - z_5(\partial/\partial z_5), \tag{27.9.13}$$

$$\begin{aligned} {}^8\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} &= \tilde{\Sigma}^1 + \tilde{\Sigma}^2 - 2\tilde{\Sigma}^3 \\ &= z_1(\partial/\partial z_1) + z_4(\partial/\partial z_4) + z_2(\partial/\partial z_2) \\ &+ z_5(\partial/\partial z_5) - 2z_3(\partial/\partial z_3) - 2z_6(\partial/\partial z_6). \end{aligned} \tag{27.9.14}$$

Here the superscripts "3" and "8" are used to refer to the analogous diagonal structure of the Gell-Mann matrices λ^3 and λ^8 of Section 5.8. It is easily verified that these 2 vector fields obey the relations

$$\#\tilde{c}^j\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = 0, \tag{27.9.15}$$

$$\#\tilde{c}^j\#{}^8\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = 0, \tag{27.9.16}$$

and therefore are candidates for the center elements of Figure 8.4.

We define the remaining 12 vector fields that occupy the other sites of Figure 8.4 by the equations

$$\tilde{\mathcal{G}}^{0,1,0}_{1,1,0} = (1/2)\#\tilde{r}(\boldsymbol{\alpha}^2)\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = -z_1(\partial/\partial z_5) + z_2(\partial/\partial z_4), \tag{27.9.17}$$

$$\tilde{\mathcal{G}}^{0,1,0}_{-1,-1,0} = (1/2)\#\tilde{r}(-\boldsymbol{\alpha}^2)\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = -z_4(\partial/\partial z_2) + z_5(\partial/\partial z_1), \tag{27.9.18}$$

$$\tilde{\mathcal{G}}^{0,1,0}_{1,-1,0} = (1/2)\#\tilde{r}(\boldsymbol{\alpha}^3)\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = z_1(\partial/\partial z_2) + z_5(\partial/\partial z_4), \tag{27.9.19}$$

$$\tilde{\mathcal{G}}^{0,1,0}_{-1,1,0} = (1/2)\#\tilde{r}(-\boldsymbol{\alpha}^3)\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = -z_2(\partial/\partial z_1) - z_4(\partial/\partial z_5), \tag{27.9.20}$$

$$\tilde{\mathcal{G}}^{0,1,0}_{0,1,1} = \#\tilde{r}(\boldsymbol{\beta}^2)\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = -z_3(\partial/\partial z_5) + z_2(\partial/\partial z_6), \tag{27.9.21}$$

$$\tilde{\mathcal{G}}^{0,1,0}_{0,-1,-1} = \#\tilde{r}(-\boldsymbol{\beta}^2)\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = z_5(\partial/\partial z_3) - z_6(\partial/\partial z_2), \tag{27.9.22}$$

$$\tilde{\mathcal{G}}^{0,1,0}_{0,1,-1} = \#\tilde{r}(\boldsymbol{\beta}^3)\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = -z_2(\partial/\partial z_3) - z_6(\partial/\partial z_5), \tag{27.9.23}$$

$$\tilde{\mathcal{G}}^{0,1,0}_{0,-1,1} = \#\tilde{r}(-\boldsymbol{\beta}^3)\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = z_3(\partial/\partial z_2) + z_5(\partial/\partial z_6), \tag{27.9.24}$$

$$\tilde{\mathcal{G}}^{0,1,0}_{1,0,1} = \#\tilde{r}(\boldsymbol{\gamma}^2)\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = z_3(\partial/\partial z_4) - z_1(\partial/\partial z_6), \tag{27.9.25}$$

$$\tilde{\mathcal{G}}^{0,1,0}_{-1,0,-1} = \#\tilde{r}(-\boldsymbol{\gamma}^2)\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = -z_4(\partial/\partial z_3) + z_6(\partial/\partial z_1), \tag{27.9.26}$$

$$\tilde{\mathcal{G}}^{0,1,0}_{-1,0,1} = \#\tilde{r}(\boldsymbol{\gamma}^3)\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = -z_3(\partial/\partial z_1) - z_4(\partial/\partial z_6), \tag{27.9.27}$$

$$\tilde{\mathcal{G}}^{0,1,0}_{1,0,-1} = \#\tilde{r}(-\boldsymbol{\gamma}^3)\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = z_1(\partial/\partial z_3) + z_6(\partial/\partial z_4). \tag{27.9.28}$$

They obey the relations

$$\#\tilde{c}^j\#\tilde{\mathcal{G}}^{0,1,0}_{k,\ell,m} = \boldsymbol{e}^j \cdot (k\boldsymbol{e}^1 + \ell\boldsymbol{e}^2 + m\boldsymbol{e}^3), \tag{27.9.29}$$

in keeping with the sites they occupy. We note that $\tilde{\mathcal{G}}^{0,1,0}_{1,1,0}$ occupies the highest weight site $\boldsymbol{w}^h$ given by (8.4) for the representation $\Gamma(0,1,0)$. Therefore there should be the ladder relation

$$\#\tilde{r}(\boldsymbol{\alpha}^3)\#\tilde{\mathcal{G}}^{0,1,0}_{1,1,0} = 0. \tag{27.9.30}$$

Direct calculation shows that this relation is true. By comparison, the vector field $:\tilde{r}(\boldsymbol{\alpha}^2):$ has the form

$$:\tilde{r}(\boldsymbol{\alpha}^2) :=: q_1q_2 := z_1(\partial/\partial z_5) + z_2(\partial/\partial z_4). \tag{27.9.31}$$

It occupies the same 1,1,0 site in Figure 8.5,

$$\begin{aligned}\#\tilde{c}^j\# : \tilde{r}(\boldsymbol{\alpha}^2) : &= \{: \tilde{c}^j :, : \tilde{r}(\boldsymbol{\alpha}^2) :\} =: [\tilde{c}^j, \tilde{r}(\boldsymbol{\alpha}^2)] : \\ &= (\boldsymbol{e}^j \cdot \boldsymbol{\alpha}^2) : \tilde{r}(\boldsymbol{\alpha}^2) := [\boldsymbol{e}^j \cdot (\boldsymbol{e}^1 + \boldsymbol{e}^2)] : \tilde{r}(\boldsymbol{\alpha}^2) :,\end{aligned} \tag{27.9.32}$$

and is evidently linearly independent of $\tilde{\mathcal{G}}^{0,1,0}_{1,1,0}$,. It satisfies the relation

$$\#\tilde{r}(\boldsymbol{\alpha}^3)\# : \tilde{r}(\boldsymbol{\alpha}^2) := -(\sqrt{2}) : \tilde{r}(\boldsymbol{\alpha}^1) : . \tag{27.9.33}$$

It can be verified that the ${}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0}$, ${}^8\tilde{\mathcal{G}}^{0,1,0}_{0,0,0}$, and $\tilde{\mathcal{G}}^{0,1,0}_{k,\ell,m}$ satisfy all the ladder relations one expects for the representation $\Gamma(0,1,0)$. For example, there is the relation

$$(1/2)\#\tilde{r}(-\boldsymbol{\alpha}^2)\#\#\tilde{r}(\boldsymbol{\alpha}^2)\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} ={}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} \tag{27.9.34}$$

which shows that ${}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0}$ has no $\Gamma(0,0,0)$ contamination. (See Exercise 6.1.) There is also the relation

$$\#\tilde{r}(-\boldsymbol{\beta}^2)\#\#\tilde{r}(\boldsymbol{\beta}^2)\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = (1/2)({}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} -{}^8\tilde{\mathcal{G}}^{0,1,0}_{0,0,0}) \tag{27.9.35}$$

which shows that ${}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0}$ and ${}^8\tilde{\mathcal{G}}^{0,1,0}_{0,0,0}$ can be transformed into each other, and that ${}^8\tilde{\mathcal{G}}^{0,1,0}_{0,0,0}$ as well has no $\Gamma(0,0,0)$ contamination. There are also the relations

$$\#\tilde{r}(\boldsymbol{\nu})\#{}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = \#\tilde{r}(\boldsymbol{\nu})\#{}^8\tilde{\mathcal{G}}^{0,1,0}_{0,0,0} = 0 \tag{27.9.36}$$

for $\boldsymbol{\nu} = \pm\boldsymbol{\alpha}^1$, $\pm\boldsymbol{\beta}^1$, $\pm\boldsymbol{\gamma}^1$. Note that the sites $\pm2,0,0$ and $0,\pm2,0$ and $0,0,\pm2$ are empty in Figure 8.4 and occupied in Figure 8.5. Therefore (9.32) shows that ${}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0}$ and ${}^8\tilde{\mathcal{G}}^{0,1,0}_{0,0,0}$, and hence the $\tilde{\mathcal{G}}^{0,1,0}_{k,\ell,m}$, have no $\Gamma(2,0,0)$ contamination.

Finally, we observe that the results we have obtained in the monomial basis can be transformed if desired to the resonance basis with the aid of the operator

$$\hat{\mathcal{A}}(\pi/8) = \exp[-i(\pi/8)\#p_1^2 - q_1^2 + p_2^2 - q_2^2 + p_2^3 - q_3^2\#]. \tag{27.9.37}$$

Exercises

27.9.1. Work out the analogs of the relations (9.17) through (9.28) with ${}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0}$ replaced by ${}^8\tilde{\mathcal{G}}^{0,1,0}_{0,0,0}$.

27.9.2. Work out the 6×6 matrices $\tilde{A}$ corresponding to the vector fields $\tilde{\mathcal{G}}^{0,0,0}_{0,0,0}$, ${}^3\tilde{\mathcal{G}}^{0,1,0}_{0,0,0}$, ${}^8\tilde{\mathcal{G}}^{0,1,0}_{0,0,0}$, and $\tilde{\mathcal{G}}^{0,1,0}_{k,\ell,m}$ in analogy to what was done for the 4×4 case at the end of Section 21.6. Show that the matrix JA, when exponentiated and with A given by any multiple of the $\tilde{A}$ associated with $\tilde{\mathcal{G}}^{0,0,0}_{0,0,0}$, produces a positive multiple of the identity matrix. Show that the remaining JA, with A any linear combination of the fourteen $\tilde{A}$ associated with the remaining $\tilde{\mathcal{G}}$ given by (9.13) and (9.14) and (9.17) through (9.28), are traceless and therefore are in $s\ell(6,\mathbb{R})$.

27.9.3. Write and verify the $sp(6)$ analog of (6.38), as given in Exercise 6.3, using (9.1) and (8.5).

27.9.4. In Sections 21.3, 21.6, and 21.9 we learned that Σ always was invariant under the action of $sp(2)$, $sp(4)$, and $sp(6)$. The purpose of this exercise is to show that this invariance is a consequence of a more general result. Consider m-dimensional Euclidean space with coordinates

$$x = (x_1, x_2, \cdots, x_m). \tag{27.9.38}$$

Here m can be even or odd. Define a vector field Σ by the relation

$$\Sigma = \sum_{a=1}^{m} x_a \partial/\partial x_a. \tag{27.9.39}$$

Suppose that each x is sent to $\bar{x}$ under the action of some linear, but invertible, transformation M,

$$\bar{x} = Mx. \tag{27.9.40}$$

Define a transformed vector field $\bar{\Sigma}$ by the rule

$$\bar{\Sigma} = \sum_{a=1}^{m} \bar{x}_a \partial/\partial \bar{x}_a. \tag{27.9.41}$$

You are to show that

$$\bar{\Sigma} = \Sigma. \tag{27.9.42}$$

That is, Σ is invariant under the action of M. Because M is any invertible matrix, we may say that Σ is invariant under the group $GL(m, \mathbb{R})$. Since $Sp(2n, \mathbb{R})$ is a subgroup of $GL(2n, \mathbb{R})$, it follows that Σ is also invariant under the group $Sp(2n, \mathbb{R})$.

Begin by inverting (9.40),

$$x = M^{-1}\bar{x}, \tag{27.9.43}$$

and verify that this relation has the component form

$$x_a = \sum_b (M^{-1})_{ab}\bar{x}_b. \tag{27.9.44}$$

Show it follows that

$$\partial x_a/\partial \bar{x}_b = (M^{-1})_{ab}. \tag{27.9.45}$$

Verify by the chain rule that there is the relation

$$\partial/\partial \bar{x}_a = \sum_c (\partial x_c/\partial \bar{x}_a)\partial/\partial x_c = \sum_c (M^{-1})_{ca}\partial/\partial x_c. \tag{27.9.46}$$

Also, the relation (9.40) has the component form

$$\bar{x}_a = \sum_d M_{ad}x_d. \tag{27.9.47}$$

Verify, by employing (9.46) and (9.47) in (9.41), it follows that there is the relation

$$\begin{aligned}
\bar{\Sigma} &= \sum_{acd} M_{ad}(M^{-1})_{ca}\, x_d\partial/\partial x_c = \sum_{acd}(M^{-1})_{ca}M_{ad}\, x_d\partial/\partial x_c \\
&= \sum_{cd}(M^{-1}M)_{cd}\, x_d\partial/\partial x_c = \sum_{cd}\delta_{cd}\, x_d\partial/\partial x_c \\
&= \sum_c x_c\partial/\partial x_c = \Sigma.
\end{aligned} \tag{27.9.48}$$

For an infinitesimal (Lie-algebraic) version of (9.42), see Exercise 10.8.

27.9.5. Suppose an object of mass m is acted upon by a force $\boldsymbol{F}$ arising from a potential V and a velocity dependent drag force,

$$\boldsymbol{F} = -\boldsymbol{\nabla}V - 2\beta\boldsymbol{v}. \tag{27.9.49}$$

Define the particle's momentum $\boldsymbol{p}$ in the usual way

$$\boldsymbol{p} = m\boldsymbol{v} \tag{27.9.50}$$

and show that Newton's equations of motion can be written in the form

$$\dot{\boldsymbol{q}} = \boldsymbol{\mathcal{L}}\boldsymbol{q},\ \dot{\boldsymbol{p}} = \boldsymbol{\mathcal{L}}\boldsymbol{p} \tag{27.9.51}$$

where $\mathcal{L}$ is the vector field

$$\mathcal{L} = \sum_i (p_i/m)(\partial/\partial q_i) - (\partial V/\partial q_i)(\partial/\partial p_i) - (2\beta/m)p_i(\partial/\partial p_i). \tag{27.9.52}$$

Decompose $\mathcal{L}$ into Hamiltonian and non-Hamiltonian parts to find the result

$$\mathcal{L} =: -[p^2/(2m) + \beta \boldsymbol{p}\cdot\boldsymbol{q} + V] : -\beta\Sigma. \tag{27.9.53}$$

Suppose V is quadratic in the components of $\boldsymbol{q}$. Show that in this case the Hamiltonian H defined by

$$H = p^2/(2m) + \beta\boldsymbol{p}\cdot\boldsymbol{q} + V \tag{27.9.54}$$

evolves according to the rule

$$H = (\text{constant}) \times e^{-2\beta t}. \tag{27.9.55}$$

27.9.6. Consider the case of a six-dimensional phase space and all Lie operators of the form $: f_3 :$. What are the $sp(6)$ transformation properties of the $: f_3 :^\dagger$?

27.10 Scalar Product and Projection Operators for Vector Fields

Section 7.3 described a $USp(2n)$ invariant scalar product for phase-space functions. Here we will see that there is a related scalar product for vector fields, and we will find that the use of this vector-field scalar product illuminates the discussion of previous sections.

For our present purposes it is convenient to employ a vector-field basis slightly different from that used in (5.3.17) and (5.3.18). Let s_α be the various phase-space monomials indexed by α in some covenient way as in Section 7.3 or Section 32.2. Take as vector-field basis elements the quantities $\mathcal{L}_{\alpha a}$ defined by the equation

$$\mathcal{L}_{\alpha a} = s_\alpha : z_a : . \tag{27.10.1}$$

In view of the relations

$$: z_a := \sum_b J_{ab}(\partial/\partial z_b), \tag{27.10.2}$$

$$\partial/\partial z_a = -\sum_b J_{ab} : z_b :,$$

the $\mathcal{L}_{\alpha a}$ manifestly form a satisfactory basis. Now define a scalar product for these basis elements (and hence, by linearity, for all vector fields) by the rule

$$\langle \mathcal{L}_{\alpha a}, \mathcal{L}_{\beta b}\rangle = \langle s_\alpha, s_\beta\rangle\langle z_a, z_b\rangle. \tag{27.10.3}$$

Here the scalar product on the left of (10.3) is the vector-field scalar product, and the scalar products on the right are the phase-space function scalar products of Section 7.3. Evidently the vector-field scalar product defined by (10.3) is positive definite.

Suppose two vector fields $\mathcal{L}_{\boldsymbol{g}}$ and $\mathcal{L}_{\boldsymbol{h}}$ are specified as in (5.3.17). Then, in view of (10.2), use of (10.3) gives the result

$$\langle \mathcal{L}_{\boldsymbol{g}}, \mathcal{L}_{\boldsymbol{h}} \rangle = \sum_a \langle g_a, h_a \rangle. \tag{27.10.4}$$

In the case that $\boldsymbol{g}$ and $\boldsymbol{h}$ are homogeneous, there is the immediate result

$$\langle \mathcal{L}_{\boldsymbol{g}^m}, \mathcal{L}_{\boldsymbol{h}^n} \rangle = 0 \text{ when } m \neq n. \tag{27.10.5}$$

As another interesting case, suppose g_m and h_m are homogeneous phase-space polynomials of degree m. Then use of (10.3) gives the result

$$\langle : g_m :, : h_m : \rangle = m \langle g_m, h_m \rangle. \tag{27.10.6}$$

Let $: f_2 :$ be the Hamiltonian vector field associated with any quadratic polynomial f_2, and consider the corresponding adjoint operator $\# : f_2 : \#$. For the action of $\# : f_2 : \#$ on a general vector-field basis element we have the result

$$\# : f_2 : \# \mathcal{L}_{\alpha a} = (: f_2 : s_\alpha) : z_a : + s_\alpha : (: f_2 : z_a) : . \tag{27.10.7}$$

See (3.25). Now watch closely! Take the scalar product of (10.7) with the general basis vector $\mathcal{L}_{\alpha' a'}$ and manipulate the result to find the relation

$$\begin{aligned}
&\langle \# : f_2 : \# \mathcal{L}_{\alpha a}, \mathcal{L}_{\alpha' a'} \rangle = \langle [(: f_2 : s_\alpha) : z_a : + s_\alpha : (: f_2 : z_a) :], s_{\alpha'} : z_{a'} : \rangle \\
&= \langle (: f_2 : s_\alpha) : z_a :, s_{\alpha'} : z_{a'} : \rangle + \langle s_\alpha : (: f_2 : z_a) :, s_{\alpha'} : z_{a'} : \rangle \\
&= \langle : f_2 : s_\alpha, s_{\alpha'} \rangle \langle z_a, z_{a'} \rangle + \langle s_\alpha, s_{\alpha'} \rangle \langle : f_2 : z_a, z_{a'} \rangle \\
&= \langle s_\alpha, : f_2 :^\dagger s_{\alpha'} \rangle \langle z_a, z_{a'} \rangle + \langle s_\alpha, s_{\alpha'} \rangle \langle z_a, : f_2 :^\dagger z_{a'} \rangle \\
&= \langle s_\alpha : z_a :, (: f_2 :^\dagger s_{\alpha'}) : z_{a'} : \rangle + \langle s_\alpha : z_a :, s_{\alpha'} : (: f_2 :^\dagger z_{a'}) : \rangle \\
&= \langle s_\alpha : z_a :, [(: f_2 :^\dagger s_{\alpha'}) : z_{a'} : + s_{\alpha'} : (: f_2 :^\dagger z_{a'}) :] \rangle \\
&= \langle s_\alpha : z_a :, \# : f_2 :^\dagger \# s_{\alpha'} : z_{a'} : \rangle = \langle \mathcal{L}_{\alpha a}, \# :: f_2 :^\dagger \# \mathcal{L}_{\alpha' a'} : \rangle.
\end{aligned} \tag{27.10.8}$$

Here we have used (7.3.15). And, in view of (7.3.15) and (10.8), we have found the beautiful result

$$\# : f_2 : \#^\dagger = \# : f_2 :^\dagger \#. \tag{27.10.9}$$

In mimicry of (7.3.31), define the analogous operator $\hat{\mathcal{M}}$, which acts on vector fields, by the rule

$$\hat{\mathcal{M}} = \exp(\# : f_2^c : \#) \exp(i \# : f_2^a : \#). \tag{27.10.10}$$

As a consequence of (8.1.11), operators of the form (10.10) give a realization of the group $USp(2n)$ acting on the space of vector fields. Moreover, in view of (7.3.26), (7.3.30), and (10.9), we have the result

$$\hat{\mathcal{M}}^\dagger = \hat{\mathcal{M}}^{-1}. \tag{27.10.11}$$

It follows that the vector-field scalar product defined by (10.3) is also $USp(2n)$ invariant. Finally, as a special case, we see that the operators $\hat{\mathcal{A}}$ defined by (6.24) and (9.37) are symplectic and unitary.

From Section 9 we know that the general homogeneous polynomial vector field (in 6 dimensions) has the decomposition (9.2). Since each term in the decomposition has different $sp(6)$ [and therefore $usp(6)$] transformation properties, we might expect that the different terms in the decomposition would be mutually orthogonal. This is indeed the case. That is, for the scalar product (10.3) and the decomposition (9.2), there are the relations

$$\begin{aligned}\langle \mathcal{H}^{\ell+1,0,0}, \mathcal{G}^{\ell-1,1,0}\rangle &= \langle \mathcal{H}^{\ell+1,0,0}, \mathcal{G}^{\ell-1,0,0}\rangle \\ &= \langle \mathcal{G}^{\ell-1,1,0}, \mathcal{G}^{\ell-1,0,0}\rangle = 0. \end{aligned} \tag{27.10.12}$$

Note that all the vector fields in (10.12) have the same degree of homogeneity. If the degrees are different, the vector fields are automatically orthogonal by (10.5).

The relations (10.12) will be proved in Subsection 11.2 by group-theoretic methods. Here we will begin to describe a complementary result. Section 9 showed that the decomposition (9.2) exists. However, given a specific vector field $\mathcal{L}_{\boldsymbol{g}^\ell}$, the only method proposed for finding $\mathcal{H}^{\ell+1,0,0}$, $\mathcal{G}^{\ell-1,1,0}$, and $\mathcal{G}^{\ell-1,0,0}$ was to construct in detail the bases for these spaces and then match coefficients. Fortunately, there is a more direct approach that accomplishes major aspects of this task. Part of this approach is described below, and the remainder will be described in Subsection 11.2.

Here we will show that there are linear projection operators $\mathcal{P}^H$ and $\mathcal{P}^G$ that can be described explicitly and that act on vector fields $\mathcal{L}_{\boldsymbol{g}^\ell}$ to yield the results

$$\mathcal{P}^H \mathcal{L}_{\boldsymbol{g}^\ell} = \mathcal{H}^{\ell+1,0,0}, \tag{27.10.13}$$

$$\mathcal{P}^G \mathcal{L}_{\boldsymbol{g}^\ell} = \mathcal{G}^{\ell-1,1,0} + \mathcal{G}^{\ell-1,0,0} = \mathcal{G}_{\ell+1}. \tag{27.10.14}$$

They also have the properties

$$(\mathcal{P}^H)^2 = \mathcal{P}^H,\ (\mathcal{P}^G)^2 = \mathcal{P}^G, \tag{27.10.15}$$

$$\mathcal{P}^H \mathcal{P}^G = \mathcal{P}^G \mathcal{P}^H = 0, \tag{27.10.16}$$

$$\mathcal{P}^H + \mathcal{P}^G = \mathcal{I}. \tag{27.10.17}$$

Here $\mathcal{I}$ denotes the identity operator. Finally, the $\mathcal{H}^{\ell+1,0,0}$ and $\mathcal{G}_{\ell+1}$ defined by (10.13) and (10.14) satisfy

$$\langle \mathcal{H}^{\ell+1,0,0}, \mathcal{G}_{\ell+1}\rangle = 0. \tag{27.10.18}$$

We will also show directly that

$$\langle \mathcal{H}^{\ell+1,0,0}, \mathcal{G}^{\ell-1,1,0}\rangle = \langle \mathcal{H}^{\ell+1,0,0}, \mathcal{G}^{\ell-1,0,0}\rangle = 0. \tag{27.10.19}$$

We remark that the relations (10.13) and (10.14) are sufficient to carry out the decomposition required for factorizing general maps as will be done in Section 26.1.

We will first define the projection operators, and then show that they possess the advertised properties. Suppose we are given $\mathcal{L}_{\boldsymbol{g}^\ell}$ and hence $\boldsymbol{g}^\ell(z)$. Then, in the spirit of (7.6.24), we *define* the homogeneous polynomial $h_{\ell+1}$ by the rule

$$h_{\ell+1} = -[1/(\ell+1)] \sum_{ab} g_a^\ell(z) J_{ab} z_b. \tag{27.10.20}$$

Note that $h_{\ell+1}$ depends *linearly* on the g_a^ℓ. We now define $\mathcal{P}^H$ by the rule

$$\mathcal{P}^H \mathcal{L}_{\boldsymbol{g}^\ell} = \mathcal{H}^{\ell+1,0,0} =: h_{\ell+1} : . \tag{27.10.21}$$

Let us compute the action of $: h_{\ell+1} :$ on z_c. We find the intermediate result

$$\begin{aligned} : h_{\ell+1} : z_c &= [h_{\ell+1}, z_c] = -[1/(\ell+1)] \sum_{ab} J_{ab} [g_a^\ell z_b, z_c] \\ &= -[1/(\ell+1)] \sum_{ab} J_{ab} (g_a^\ell [z_b, z_c] + z_b [g_a^\ell, z_c]) \\ &= -[1/(\ell+1)] \{ \sum_{ab} g_a^\ell J_{ab} J_{bc} + \sum_{ab} J_{ab} z_b [g_a^\ell, z_c] \} \\ &= [1/(\ell+1)] \{ g_c^\ell + \sum_{ab} J_{ab} z_b [z_c, g_a^\ell] \}. \end{aligned} \tag{27.10.22}$$

Here we have used (1.7.10) and (3.1.3). Next write the tautology

$$[z_c, g_a^\ell] = [z_a, g_c^\ell] - A_{ac} \tag{27.10.23}$$

where

$$A_{ac} = [z_a, g_c^\ell] - [z_c, g_a^\ell]. \tag{27.10.24}$$

Note that A_{ac} is antisymmetric under the interchange of indices. Insertion of (10.23) into (10.22) gives the further result

$$: h_{\ell+1} : z_c = [1/(\ell+1)] \{ g_c^\ell + \sum_{ab} J_{ab} z_b [z_a, g_c^\ell] - \sum_{ab} J_{ab} z_b A_{ac} \}. \tag{27.10.25}$$

The center term on the right of (10.25) can be evaluated,

$$\begin{aligned} \sum_{ab} J_{ab} z_b [z_a, g_c^\ell] &= \sum_{ab} J_{ab} z_b : z_a : g_c^\ell = -\sum_{ab} J_{ba} z_b : z_a : g_c^\ell \\ &= \Sigma g_c^\ell = \ell g_c^\ell. \end{aligned} \tag{27.10.26}$$

Here (7.6.50), (3.1), and (3.26) have been used. Therefore (10.25) can be rewritten in the form

$$g_c^\ell =: h_{\ell+1} : z_c + [1/(\ell+1)] \sum_{ab} J_{ab} z_b A_{ac}. \tag{27.10.27}$$

The second term on the right can be manipulated further. Use the antisymmetry of J and the fact that a, b are dummy summation indices to write

$$[1/(\ell+1)] \sum_{ab} J_{ab} z_b A_{ac} = -[1/(\ell+1)] \sum_{ab} z_b J_{ba} A_{ac} = -[1/(\ell+1)] \sum_{ab} z_a J_{ab} A_{bc}. \tag{27.10.28}$$

As a result of this manipulation (10.27) can be rewritten in the form

$$g_c =: h_{\ell+1} : z_c - [1/(\ell+1)] \sum_{ab} z_a J_{ab} A_{bc}. \tag{27.10.29}$$

Correspondingly, $\mathcal{L}_{\boldsymbol{g}^\ell}$ can be written in the form

$$\mathcal{L}_{\boldsymbol{g}^\ell} =: h_{\ell+1} : + \mathcal{L}_{{}^G\boldsymbol{g}^\ell} \tag{27.10.30}$$

where

$${}^G g_c^\ell = -[1/(\ell+1)] \sum_{ab} z_a J_{ab} A_{bc}. \tag{27.10.31}$$

Note that ${}^G\boldsymbol{g}^\ell$ is linear in $\boldsymbol{g}^\ell$ since A is linear in $\boldsymbol{g}^\ell$. We now define the projection operator $\mathcal{P}^G$ by the rule

$$\mathcal{P}^G \mathcal{L}_{\boldsymbol{g}^\ell} = \mathcal{G}^{\ell-1,1,0} + \mathcal{G}^{\ell-1,0,0} = \mathcal{G}_{\ell+1} = \mathcal{L}_{{}^G\boldsymbol{g}^\ell}, \tag{27.10.32}$$

which is simply a rewriting of the relation

$$\mathcal{G}_{\ell+1} = \mathcal{L}_{\boldsymbol{g}^\ell} - : h_{\ell+1} : . \tag{27.10.33}$$

With the projection operator definitions (10.21) and (10.32), the relation (10.30) shows that (10.17) holds by construction.

It remains to be shown that the projection operators have the advertised properties. Suppose the $\boldsymbol{g}^\ell$ corresponding to $: h_{\ell+1} :$ is used in (10.31) to compute ${}^G\boldsymbol{g}^\ell$. According to (7.6.7) the matrix A given by (10.24) vanishes in this case. Consequently, ${}^G\boldsymbol{g}^\ell$ is zero. This observation verifies the second assertion in (10.16). Conversely, suppose ${}^G\boldsymbol{g}^\ell$ is used in (10.20) to compute $h_{\ell+1}$. We first observe that (10.31) can be rewritten in the form

$${}^G g_a^\ell = -[1/(\ell+1)] \sum_{cd} z_c J_{cd} A_{da}. \tag{27.10.34}$$

Consequently, we have the result

$$\begin{aligned} h_{\ell+1} &= -[1/(\ell+1)] \sum_{ab} {}^G g_a^\ell J_{ab} z_b \\ &= [1/(\ell+1)^2] \sum_{abcd} z_c J_{cd} A_{da} J_{ab} z_b \\ &= [1/(\ell+1)^2] \sum_{bc} z_c (JAJ)_{cb} z_b. \end{aligned} \tag{27.10.35}$$

However, since both J and A are antisymmetric, it follows that

$$(JAJ)^T = J^T (A)^T J^T = -JAJ. \tag{27.10.36}$$

Therefore the right side of (10.35) vanishes by antisymmetry and we find

$$h_{\ell+1} = 0. \tag{27.10.37}$$

This observation verifies the first assertion in (10.16). Finally, with the aid of (10.16) and (10.17), we find that

$$(\mathcal{P}^H)^2 = \mathcal{P}^H(\mathcal{I} - \mathcal{P}^G) = \mathcal{P}^H - \mathcal{P}^H\mathcal{P}^G = \mathcal{P}^H. \tag{27.10.38}$$

This calculation and its counterpart for $\mathcal{P}^G$ verify (10.15).

To verify (10.18) we first calculate that

$$\begin{aligned} : h_{\ell+1} : z_c &= [h_{\ell+1}, z_c] = \sum_{de} (\partial h_{\ell+1}/\partial z_d) J_{de} (\partial z_c/\partial z_e) \\ &= \sum_{de} (\partial h_{\ell+1}/\partial z_d) J_{de} \delta_{ce} = -\sum_d J_{cd} (\partial h_{\ell+1}/\partial z_d). \end{aligned} \qquad (27.10.39)$$

Consequently, we find using (10.4), (10.34), and (10.39) the intermediate result

$$\begin{aligned} \langle \mathcal{G}_{\ell+1}, \mathcal{H}^{\ell+1,0,0} \rangle &= \langle \mathcal{L}_{G\boldsymbol{g}^\ell}, : h_{\ell+1} : \rangle \\ &= [1/(\ell+1)] \sum_{abcd} J_{ab} J_{cd} \langle z_a A_{bc}, (\partial h_{\ell+1}/\partial z_d) \rangle \\ &= [1/(\ell+1)] \sum_{abcd} J_{ab} J_{cd} \langle A_{bc}, (\partial^2 h_{\ell+1}/\partial z_a \partial z_d)) \rangle. \end{aligned} \qquad (27.10.40)$$

Here, in the last line, (7.3.14) has also been used. Define tensors T^1 and T^2 by the rules

$$T^1_{abcd} = J_{ab} J_{cd}, \qquad (27.10.41)$$

$$T^2_{abcd} = \langle A_{bc}, (\partial^2 h_{\ell+1}/\partial z_a \partial z_d) \rangle. \qquad (27.10.42)$$

From the antisymmetry of J the tensor T^1 has the symmetry property

$$T^1_{dcba} = J_{dc} J_{ba} = J_{ab} J_{cd} = T^1_{abcd}. \qquad (27.10.43)$$

From the antisymmetry of A and the symmetry of $(\partial^2 h_{\ell+1}/\partial z_a \partial z_d)$ the tensor T^2 has the symmetry property

$$\begin{aligned} T^2_{dcba} &= \langle A_{cb}, (\partial^2 h_{\ell+1}/\partial z_d \partial z_a) \rangle \\ &= -\langle A_{bc}, (\partial^2 h_{\ell+1}/\partial z_a \partial z_d) \rangle = -T^2_{abcd}. \end{aligned} \qquad (27.10.44)$$

It follows that

$$\sum_{abcd} J_{ab} J_{cd} \langle A_{bc}, (\partial^2 h_{\ell+1}/\partial z_a \partial z_d) \rangle = \sum_{abcd} T^1_{abcd} T^2_{abcd} = -\sum_{abcd} T^1_{dcba} T^2_{dcba} = 0, \qquad (27.10.45)$$

and consequently

$$\langle \mathcal{G}_{\ell+1}, \mathcal{H}^{\ell+1,0,0} \rangle = 0. \qquad (27.10.46)$$

To verify (10.19) suppose that $\mathcal{L}_{\boldsymbol{g}^\ell}$ is the vector field $\mathcal{G}^{\ell-1,0,0}$ given in (9.3). Then we have the relation

$$g^\ell_c = f_{\ell-1} z_c. \qquad (27.10.47)$$

Consequently, from (10.4), (10.39), and (10.47), we find the result

$$\begin{aligned} \langle \mathcal{G}^{\ell-1,0,0}, \mathcal{H}^{\ell+1,0,0} \rangle &= \langle \mathcal{L}_{\boldsymbol{g}^\ell}, : h_{\ell+1} : \rangle = -[1/(\ell+1)] \sum_{cd} J_{cd} \langle z_c f_{\ell-1}, (\partial h_{\ell+1}/\partial z_d) \rangle \\ &= -[1/(\ell+1)] \sum_{cd} J_{cd} \langle f_{\ell+1}, (\partial^2 h_{\ell+1}/\partial z_c \partial z_d \rangle = 0. \end{aligned} \qquad (27.10.48)$$

Here we have again employed (7.3.14), and used the antisymmetry of J and the symmetry of $(\partial^2 h_{\ell+1}/\partial z_c \partial z_d)$ to infer that the sum in (10.47) vanishes. Finally, in view of (10.46), (10.48), and the definition of $\mathcal{G}_{\ell+1}$ as given in the second part of (10.14) or, equivalently, in (10.33), we conclude that both statements in (10.19) are correct.

Let us evaluate the scalar products between vector fields associated with linear transformations. Let F be any $2n \times 2n$ matrix, possibly complex, and use it to define a vector field $\mathcal{L}_{\boldsymbol{f}^1}$ by the rule

$$\mathcal{L}_{\boldsymbol{f}^1} = \sum_{ab} (JF)_{ab} z_b (\partial/\partial z_a). \tag{27.10.49}$$

Then, comparison of (5.3.17) and (10.49) gives the relation

$$f_a^1 = \sum_b (JF)_{ab} z_b. \tag{27.10.50}$$

From (10.49) we also find the result

$$\mathcal{L}_{\boldsymbol{f}^1} z_c = \sum_d (JF)_{cd} z_d, \tag{27.10.51}$$

which is analogous to (6.26) and (6.27).

Let G be a second $2n \times 2n$ matrix, and use it to define the vector field $\mathcal{L}_{\boldsymbol{g}^1}$. Now compute the scalar product between $\mathcal{L}_{\boldsymbol{f}^1}$ and $\mathcal{L}_{\boldsymbol{g}^1}$. Doing so gives the result

$$\begin{aligned} \langle \mathcal{L}_{\boldsymbol{f}^1}, \mathcal{L}_{\boldsymbol{g}^1} \rangle &= \sum_a \langle f_a^1, g_a^1 \rangle \\ &= \sum_{abc} \langle (JF)_{ab} z_b, (JG)_{ac} z_c \rangle \\ &= \sum_{abc} [(JF)_{ab}]^* (JG)_{ac} \langle z_b, z_c \rangle \\ &= \sum_{abc} [(JF)^\dagger]_{ba} (JG)_{ac} \delta_{bc} = \text{tr } [(JF)^\dagger JG] \\ &= \text{tr } [F^\dagger J^\dagger JG] = \text{tr } (F^\dagger G). \end{aligned} \tag{27.10.52}$$

Here a "*" denotes complex conjugation, and use has been made of (3.1.6). Note that this scalar product is the same as that in (4.4.16).

Suppose S is a real symmetric matrix. Use it to define a quadratic polynomial h_2 as in (7.2.3),

$$h_2 = -(1/2) \sum_{de} S_{de} z_d z_e. \tag{27.10.53}$$

Then, from (7.2.4), there is an associated vector field $\mathcal{L}_{\boldsymbol{s}^1}$ given by the relation

$$\mathcal{L}_{\boldsymbol{s}^1} =: h_2 := \sum_{ab} (JS)_{ab} z_b (\partial/\partial z_a) = \sum_a s_a^1 (\partial/\partial z_a), \tag{27.10.54}$$

with

$$s_a^1 = \sum_b (JS)_{ab} z_b. \tag{27.10.55}$$

Define the Hamiltonian vector field $\mathcal{H}^2$ by writing

$$\mathcal{H}^2 =: h_2 :, \tag{27.10.56}$$

and let $\mathcal{G}_2$ be any non-Hamiltonian vector field of the form (10.49) with F being any real antisymmetric matrix A. Compare with (6.26) and (6.27). Then, using (10.52), we find the result

$$\langle \mathcal{G}_2, \mathcal{H}^2 \rangle = \text{tr }(A^T S) = -\text{tr }(AS) = 0. \tag{27.10.57}$$

Here we have used the easily proved fact that the trace of the product of an antisymmetric and a symmetric matrix is always zero. See (4.4.86). We observe that (10.57) is a special case of the general result (10.46).

We close this section with a further study of first-degree vector fields in $2n$ variables. Consider again the relation (10.49) and decompose F, which we now assume to be real, into symmetric and antisymmetric parts by writing

$$F = S^F + A^F. \tag{27.10.58}$$

That is, we write

$$\mathcal{L}_{\boldsymbol{f}} = \sum_{ab} (JF)_{ab} z_b (\partial/\partial z_a) = \sum_{ab} [J(S^F + A^F)]_{ab} z_b (\partial/\partial z_a) = \mathcal{L}_{\boldsymbol{f}^S} + \mathcal{L}_{\boldsymbol{f}^A} \tag{27.10.59}$$

with

$$\mathcal{L}_{\boldsymbol{f}^S} = \sum_{ab} (JS^F)_{ab} z_b (\partial/\partial z_a) \tag{27.10.60}$$

and

$$\mathcal{L}_{\boldsymbol{f}^A} = \sum_{ab} (JA^F)_{ab} z_b (\partial/\partial z_a). \tag{27.10.61}$$

We will now verify directly, as expected, that $\mathcal{L}_{\boldsymbol{f}^S}$ is a Hamiltonian vector field and $\mathcal{L}_{\boldsymbol{f}^A}$ is a non-Hamiltonian vector field. In particular, we will show that there are the relations

$$\mathcal{P}^H \mathcal{L}_{\boldsymbol{f}^S} = \mathcal{L}_{\boldsymbol{f}^S}, \tag{27.10.62}$$

$$\mathcal{P}^G \mathcal{L}_{\boldsymbol{f}^A} = \mathcal{L}_{\boldsymbol{f}^A}. \tag{27.10.63}$$

Exercises

27.10.1. Show that

$$: f : \ = \sum_a (\partial f/\partial z_a) : z_a : . \tag{27.10.64}$$

27.10.2. Verify (10.4).

27.10.3. Verify (10.6). Hint: You may use brute force or (10.49), (7.3.14), and (7.6.50).

27.10.4. Let $\mathcal{L}_{f_{\ell-1}}$ be the vector field corresponding to (9.3), and let $\mathcal{L}_{f'_{\ell-1}}$ be a second such field. Show that

$$\langle \mathcal{L}_{f_{\ell-1}}, \mathcal{L}_{f'_{\ell-1}} \rangle = (\ell - 1 + 2n)\langle f_{\ell-1}, f'_{\ell-1} \rangle \tag{27.10.65}$$

where $(2n)$ is the phase-space dimension.

27.10.5. Let $h_{\ell+1}$ be any homogeneous polynomial of degree $(\ell+1)$. Find the $\boldsymbol{g}^\ell$ in the $\mathcal{L}_{\boldsymbol{g}^\ell}$ that equals $: h_{\ell+1} :$. Insert this $\boldsymbol{g}^\ell$ in (10.20) and verify that the $h_{\ell+1}$ so produced agrees with the original $h_{\ell+1}$. You have again verified the first result in (10.15).

27.10.6. Consider vector fields that are in $\mathcal{G}^{\ell-1,0,0}$ and therefore can be written in the form (10.47). Insert these $\boldsymbol{g}^\ell$ into (10.20) and show that the $h_{\ell+1}$ they produce vanish.

27.10.7. Verify that the vector fields $\mathcal{H}^{2,0,0}$, $\mathcal{G}^{0,1,0}$, and $\mathcal{G}^{0,0,0}$ found explicitly in Section 9 satisfy (10.12).

27.10.8. Let $z_1, z_2, \cdots z_m$ be m variables. Consider the m^2 vector fields $\mathcal{L}_{ab}$ defined by the rule

$$\mathcal{L}_{ab} = z_a(\partial/\partial z_b). \tag{27.10.66}$$

Show that these vector fields obey the commutation rules

$$\{\mathcal{L}_{ab}, \mathcal{L}_{cd}\} = \delta_{bc}\mathcal{L}_{ad} - \delta_{ad}\mathcal{L}_{cb}. \tag{27.10.67}$$

Let A be an $m \times m$ matrix. Associate with each such matrix the vector field $\mathcal{L}^A$ defined by the rule

$$\mathcal{L}^A = \sum_{ab} A_{ab}\mathcal{L}_{ab}. \tag{27.10.68}$$

Show that there is the relation

$$\{\mathcal{L}^A, \mathcal{L}^B\} = \mathcal{L}^C \tag{27.10.69}$$

where

$$C = \{A, B\}. \tag{27.10.70}$$

That is, verify that the $\mathcal{L}_{ab}$ yield a basis for the general linear group Lie algebra $g\ell(m)$.

Define a vector field Σ by the relation

$$\Sigma = \mathcal{L}^I = \sum_a \mathcal{L}_{aa} \tag{27.10.71}$$

where I is the identity matrix. Verify that there is the relation

$$\{\mathcal{L}_{ab}, \Sigma\} = 0. \tag{27.10.72}$$

You have shown that Σ is invariant under $g\ell(m)$.

27.10.9. Review Exercise 10.8 above. Show that the vector fields spanned by the elements

$$\mathcal{L}_{ab} = z_a(\partial/\partial z_b) - z_b(\partial/\partial z_a) \tag{27.10.73}$$

yield a basis for the Lie algebra $so(m)$. For the cases $m = 2n = 2$, $m = 2n = 4$, and $m = 2n = 6$ decompose these elements into Hamiltonian and non-Hamiltonian parts. Show that in each case the Hamiltonian parts span a Lie algebra, and identify these Lie algebras. What can be said about the non-Hamiltonian parts?

27.10.10. Use the machinery of this section to find the decompositions (3.51) through (3.56).

27.10.11. Consider the matrices $\tilde{A}$ given by (6.28) through (6.33). Relate them to the matrices C^0 through C^3 and E^1 and E^2 given by (4.3.137) through (4.3.140) and (4.3.145) and (4.3.146). Also relate them to the matrices A^1 through A^6 given by (8.2.87) through (8.2.92). Verify directly that the trace of the product of any two different $\tilde{A}$ matrices vanishes. Relate this fact to the assertion (10.12) and the relation (10.52). Show that all matrices of the form $J\tilde{A}(0,1;*,*)$ are traceless, and therefore all matrices of the form $\exp[J\tilde{A}(0,1;*,*)]$ have determinant +1.

27.11 Products and Casimir Operators

In this section we will explore the properties of products of entities when each entity taken by itself has well-defined properties under the action of the symplectic group. For example, if f_ℓ, g_m, h_n, $\cdots$ are homogeneous polynomials, we could ask about the transformation properties of the product $[(f_\ell)(g_m)(h_n)\cdots]$. Or, we could ask about the transformation properties of the product of Lie operators $[:f_\ell::g_m::h_n:\cdots]$. As a third example, we could ask about the properties of the product of adjoint operators $[\#f_\ell\#\#g_m\#\#h_n\#\cdots]$. The first case, the transformation properties of the product $[(f_\ell)(g_m)(h_n)\cdots]$, is simple because the polynomials f_ℓ, g_m, h_n, $\cdots$ can be multiplied together to yield some net polynomial, and the transformation properties of this polynomial are already known. The remaining two cases require more work.

27.11.1 The Quadratic Casimir Operator

We will find that a question of particular interest, and also the simplest, is to determine the transformation properties of the two-element products $[:f_2::g_2:]$ and $[\#f_2\#\#g_2\#]$. In the case of $sp(2)$, we know that $:f_2:$ and $:g_2:$ (and $\#f_2\#$ and $\#g_2\#$) each carry the rerpresentation $\Gamma(2)$, and therefore the product carries the representation $\Gamma(2)\otimes\Gamma(2)$. Also, there is the Clebsch-Gordan series result

$$\Gamma(2)\otimes\Gamma(2)=\Gamma(0)\oplus\Gamma(2)\oplus\Gamma(4). \tag{27.11.1}$$

(This is just the familiar statement for $su(2)$ or $sp(2)$ that spin 1 and spin 1 combine to make spin 0, 1, and 2.)

In the case of $sp(4)$, the corresponding representation for each factor is $\Gamma(2,0)$; and the corresponding Clebsch-Gordan series result is known from group theory to be

$$\Gamma(2,0)\otimes\Gamma(2,0)=\Gamma(0,0)\oplus\Gamma(0,1)\oplus\Gamma(0,2)\oplus\Gamma(2,0)\oplus\Gamma(2,1)\oplus\Gamma(4,0). \tag{27.11.2}$$

Finally, in the case of $sp(6)$, the representation for each factor is $\Gamma(2,0,0)$; and the corresponding Clebsch-Gordan series result is

$$\begin{aligned}&\Gamma(2,0,0)\otimes\Gamma(2,0,0)=\\&\Gamma(0,0,0)\oplus\Gamma(0,1,0)\oplus\Gamma(0,2,0)\oplus\Gamma(2,0,0)\oplus\Gamma(2,1,0)\oplus\Gamma(4,0,0).\end{aligned} \tag{27.11.3}$$

Observe from (11.1) through (11.3) that in each case the *identity* representation [the representations $\Gamma(0)$, $\Gamma(0,0)$, and $\Gamma(0,0,0)$] occurs once and only once. (Strictly speaking, we can only conclude that there is the potential for the identity representation to occur. The sought after quantity may in fact vanish. See Exercise 11.7.) Consequently, there must be some combination of quantities of the form $[: f_2 :: g_2 :]$, or of the form $[\#f_2\#\#g_2\#]$, that is *invariant* (commutes with all generators) under the action of the symplectic group. Moreover, this combination is unique up to an overall multiplicative constant. This combination is called the *Casimir* operator for the symplectic group or symplectic Lie algebra. More particularly, it is called the *quadratic* Casimir operator since it is composed of two factors.

Now that we know that a quadratic Casimir operator exists (and is unique), the problem is to find it explicitly. In effect, what we must do is find the Clebsch-Gordan *coefficients* that produce the identity representations in the series (11.1) through (11.3). We will work up to this task by stages.

Suppose L is a Lie algebra with basis elements B_1, B_2, $\cdots$. Then, as in Section 3.7, the basis elements satisfy Lie product rules of the form

$$[B_\alpha, B_\beta] = \sum_\gamma c^\gamma_{\alpha\beta} B_\gamma. \tag{27.11.4}$$

Here $[,]$ denotes the Lie product (however realized) and the quantities $c^\gamma_{\alpha\beta}$ are the structure constants that specify L.

Next, suppose R is a *realization* of L in terms of $m \times m$ matrices.[2] Then, for each basis element B_α, there will be an associated matrix $\hat{B}_\alpha$, and these matrices will obey the commutation rules

$$\{\hat{B}_\alpha, \hat{B}_\beta\} = \sum_\gamma c^\gamma_{\alpha\beta} \hat{B}_\gamma \tag{27.11.5}$$

with the same structure constants as in (11.4). See Section 3.7.

Since a Lie algebra is a vector space, it is natural to consider the possibility of introducing some kind of scalar product among the elements of L. Suppose B and B' are any two elements in L, and let (B, B') denote their scalar product. Then, by linearity, there is the result

$$(B, B') = \sum_{\alpha\alpha'} b^\alpha (b')^{\alpha'} (B_\alpha, B_{\alpha'}) \tag{27.11.6}$$

where the b^α and $(b')^{\alpha'}$ are the components of B and B',

$$B = \sum_\alpha b^\alpha B_\alpha, \tag{27.11.7}$$

$$B' = \sum_{\alpha'} (b')^{\alpha'} B_{\alpha'}. \tag{27.11.8}$$

[Note that we have taken the scalar product to be *linear* (no complex conjugation), in both the components b^α and $(b')^{\alpha'}$ rather than antilinear (complex conjugation) in one and linear

[2] In this context we use the term *realization* rather than *representation* because in this chapter we wish, for the most part, to use the term *representation* only in the specific/technical sense of referring to some $\Gamma(\cdots)$.

in the other as in (7.3.12).] The relation (11.6) can be rewritten in the form

$$(B, B') = \sum_{\alpha\alpha'} b^{\alpha}(b')^{\alpha'} g_{\alpha\alpha'} \tag{27.11.9}$$

where $g_{\alpha\alpha'}$ is defined by writing

$$g_{\alpha\alpha'} = (B_{\alpha}, B_{\alpha'}). \tag{27.11.10}$$

In view of (11.9), the quantities $g_{\alpha\alpha'}$ may be regarded as the entries in some kind of *metric tensor*, and the scalar product between any two elements in L is specified once the entries $g_{\alpha\alpha'}$ are specified.

In principle, the entries $g_{\alpha\alpha'}$ may be defined at will. However, it is advantageous to define $g_{\alpha\alpha'}$ in a way that involves some properties of the Lie algebra L and has certain desired features. A way to do this is to define $g_{\alpha\alpha'}$ with the aid of the realization R by writing

$$(B_{\alpha}, B_{\alpha'})_R = g^R_{\alpha\alpha'} = \text{ tr } (\hat{B}_{\alpha}\hat{B}_{\alpha'}). \tag{27.11.11}$$

Here we have written the sub and superscript R to indicate that the realization R has been used. See Section 4.4, equation (4.4.39), for an analogous construction.

In the case of *sp*(2), for example, suppose we use as a basis the matrices B^0, F, and G associated with the quadratic phase-space polynomials b^0, f, and g as described in Section 5.6. Then we find the result

$$g^F = \begin{array}{cccc} \backslash & b^0 & f & g \\ b^0 & -2 & 0 & 0 \\ f & 0 & 2 & 0 \\ g & 0 & 0 & 2 \end{array}. \tag{27.11.12}$$

Here we have employed the notation g^F to indicate that for the realization R we have used the 2×2 *fundamental* or defining representation. Also, we have labeled the entries in g^F by the associated quadratic phase-space polynomials. Suppose instead of the basis B^0, F, and G we use the Cartan basis of Section 21.1. The 2×2 matrices associated with these basis elements are easily found. See Exercise 11.1. Using these matrices gives the result

$$g^F = \begin{array}{cccc} \backslash & c^1 & r(+\boldsymbol{\alpha}) & r(-\boldsymbol{\alpha}) \\ c^1 & 2 & 0 & 0 \\ r(+\boldsymbol{\alpha}) & 0 & 0 & 2 \\ r(-\boldsymbol{\alpha}) & 0 & 2 & 0 \end{array} \tag{27.11.13}$$

where we have now labeled the entries by the polynomials associated with the Cartan basis elements.

We observe that g^F is symmetric as is desired for a metric tensor. This symmetry property is true in general for any realization R,

$$g^R_{\alpha\alpha'} = g^R_{\alpha'\alpha}, \tag{27.11.14}$$

because the trace operation has the permutation symmetry property (3.6.124). Indeed, by linearity and symmetry, we have the results

$$(B, B')_R = \text{ tr } (\hat{B}\hat{B}') = \text{ tr } (\hat{B}'\hat{B}) = (B', B)_R.$$

Analogous calculations can be carried out for the cases of $sp(4)$ and $sp(6)$ using the fundamental matrix representations of Sections 5.7 and 5.8. Here it is convenient to introduce additional notation. According to Section 5.5, associated with any quadratic polynomial f there is an associated Hamiltonian matrix JS^f. For any two quadratic polynomials f and g. let us make the definitions

$$\langle f, g\rangle_F = (f, g)_F = (JS^f, JS^g)_F = \text{tr}(JS^f JS^g). \tag{27.11.15}$$

Then, if we use the Cartan bases of Sections 21.4 and 21.7 and the fundamental representations we find, both for $sp(4)$ and $sp(6)$, results analogous to (11.13) that can be written in the general form

$$\langle c^j, c^k\rangle_F = 2\delta_{jk}, \tag{27.11.16}$$

$$\langle c^j, r(\boldsymbol{\mu})\rangle_F = 0, \tag{27.11.17}$$

$$\langle r(\boldsymbol{\mu}), r(\boldsymbol{\nu})\rangle_F = 0, \text{ if } \boldsymbol{\mu} \neq -\boldsymbol{\nu}, \tag{27.11.18}$$

$$\langle r(\boldsymbol{\mu}), r(-\boldsymbol{\mu})\rangle_F = 2. \tag{27.11.19}$$

Note that, as they stand, the relations (11.13) and (11.16) through (11.19) or, equivalently (11.11) with $R = F$, define a scalar product for the elements in the various $sp(2n, \mathbb{R})$ Lie algebras. If we identify the Lie elements with their associated quadratic polynomials in the phase space variables z using relations of the form (5.5.1) and (5.5.3), then we have in effect also defined a scalar product $\langle f, g\rangle_F$ among quadratic polynomials. However, unlike the scalar product of Section 7.3, this scalar product is only defined for *quadratic* polynomials. Exercises 11.1 and 11.3 examine the relation between these two scalar products.

The definition (11.11) has a further desirable property beyond symmetry that is less obvious. Let C be some element in L. Use it to *transform* any basis element B_α into the element B_α^{tr} by the rule

$$\begin{aligned} B_\alpha^{\text{tr}} &= \exp(\epsilon : C :)B_\alpha = B_\alpha + \epsilon : C : B_\alpha + (\epsilon^2/2!) : C :^2 B_\alpha + \cdots \\ &= B_\alpha + \epsilon[C, B_\alpha] + (\epsilon^2/2!)[C, [C, B_\alpha]] + \cdots . \end{aligned} \tag{27.11.20}$$

Here $: C :$ is a differential operator in the case that the Lie product is a Poisson bracket. Otherwise it is simply the *adjoint* operator defined by the property

$$: C : B_\alpha = [C, B_\alpha]. \tag{27.11.21}$$

See (3.7.31) and (5.3.2). The matrix analog of (11.20) in the realization R is the transformation

$$\begin{aligned} \hat{B}_\alpha^{\text{tr}} &= \exp(\epsilon\#\hat{C}\#)\hat{B}_\alpha = \hat{B}_\alpha + \epsilon\#\hat{C}\#\hat{B}_\alpha + (\epsilon^2/2!)\#C\#^2 B_\alpha + \cdots \\ &= \hat{B}_\alpha + \epsilon\{\hat{C}, \hat{B}_\alpha\} + (\epsilon^2/2!)\{\hat{C}, \{\hat{C}, \hat{B}_\alpha\}\} + \cdots . \end{aligned} \tag{27.11.22}$$

At this point we invoke the relation

$$\exp(\epsilon \# \hat{C} \#)\hat{B}_\alpha = \exp(\epsilon\hat{C})\hat{B}_\alpha \exp(-\epsilon\hat{C}), \tag{27.11.23}$$

which is the matrix analog of (8.2.5) and derived in the same way. It follows that (11.22) can be rewritten in the form

$$\hat{B}_\alpha^{\rm tr} = \exp(\epsilon\hat{C})\hat{B}_\alpha \exp(-\epsilon\hat{C}). \tag{27.11.24}$$

Consequently, from (11.11) and (11.15), we have the result

$$\begin{aligned}(B_\alpha^{\rm tr}, B_{\alpha'}^{\rm tr})_R &= {\rm tr}(\hat{B}_\alpha^{\rm tr}\hat{B}_{\alpha'}^{\rm tr}) \\ &= {\rm tr}[\exp(\epsilon\hat{C})\hat{B}_\alpha \exp(-\epsilon\hat{C})\exp(\epsilon\hat{C})\hat{B}_{\alpha'}\exp(-\epsilon\hat{C})] \\ &= {\rm tr}[\exp(\epsilon\hat{C})\hat{B}_\alpha\hat{B}_{\alpha'}\exp(-\epsilon\hat{C})] \\ &= {\rm tr}[\exp(-\epsilon\hat{C})\exp(\epsilon\hat{C})\hat{B}_\alpha\hat{B}_{\alpha'}] = {\rm tr}(\hat{B}_\alpha\hat{B}_{\alpha'}) \\ &= (B_\alpha, B_{\alpha'})_R. \end{aligned} \tag{27.11.25}$$

Here we have again used standard properties of the trace operation. See Exercise 3.6.7.[3] But we know that objects of the form $[\exp(\epsilon : C :)]$ correspond to Lie group elements generated by the Lie algebra L. Therefore, (11.25) shows that the scalar product (11.11) has the remarkable property that it is *invariant* under the action of the group.

The relation (11.25) displays group invariance in *finite* (group) form. It is also instructive to view group invariance in *infinitesimal* (Lie-algebraic) form. This is easily done by retaining only the first two terms in (11.20) and equating powers of ϵ. Doing so in (11.25) gives the result

$$([C, B_\alpha], B_{\alpha'})_R + (B_\alpha, [C, B_{\alpha'}])_R = 0. \tag{27.11.26}$$

Upon setting $C = B_{\alpha''}$ and relabeling indices, (11.26) takes the beautifully symmetric, if less illuminating, forms

$$([B_\alpha, B_{\alpha'}], B_{\alpha''})_R = (B_\alpha, [B_{\alpha'}, B_{\alpha''}])_R, \tag{27.11.27}$$

$$(B_\alpha, [B_{\alpha'}, B_{\alpha''}])_R = (B_{\alpha'}, [B_{\alpha''}, B_\alpha])_R = (B_{\alpha''}, [B_\alpha, B_{\alpha'}])_R.$$

Have you ever encountered relations like (11.26) and (11.27) before? You have. See Exercise 11.8.

The invariance relation (11.25) has implications for the metric tensor g^R. Since the relation (11.20) involves only Lie products and sums, we know from (11.4) that there are transformation coefficients $U_{\alpha\beta}$ such that (11.20) can be rewritten in the form

$$B_\alpha^{\rm tr} = \sum_\beta U_{\alpha\beta}B_\beta. \tag{27.11.28}$$

Inserting this relation into (11.25) gives the results

$$\sum_{\beta\beta'} U_{\alpha\beta}U_{\alpha'\beta'}(B_\beta, B_{\beta'})_R = (B_\alpha, B_{\alpha'})_R, \tag{27.11.29}$$

[3] We are also in the uncomfortable position of using the symbols *tr* to stand both for *transformed* and *trace*. Some flexibility of mind is sometimes required.

or, with the aid of (11.10),

$$\sum_{\beta\beta'} U_{\alpha\beta} U_{\alpha'\beta'} g^R_{\beta\beta'} = g^R_{\alpha\alpha'}. \tag{27.11.30}$$

If we view the quantities $U_{\alpha\beta}$ and $g^R_{\alpha\alpha'}$ as entries in matrices, the relation (11.30) can be written in the compact form

$$U g^R U^T = g^R. \tag{27.11.31}$$

Finally, we know that U is invertible. [Simply change the sign of ϵ in (11.20).] Suppose that g^R is also invertible. We will soon see that it is for the symplectic Lie algebra. [Indeed, g^R can be shown to be invertible for all simple Lie algebras.] Then (11.31) can also be rewritten in the form

$$(U^T)^{-1}(g^R)^{-1}U^{-1} = (g^R)^{-1}, \tag{27.11.32}$$

or

$$U^T(g^R)^{-1}U = (g^R)^{-1}. \tag{27.11.33}$$

We are ready to construct the quadratic Casimir operator. Following the usual procedure, we define a metric tensor $g_R^{\alpha\alpha'}$ with raised indices by the rule

$$g_R^{\alpha\alpha'} = [(g^R)^{-1}]_{\alpha\alpha'}. \tag{27.11.34}$$

Suppose $\hat{B}_\alpha$ is any set of linear operators that obey (11.5), but do not necessarily belong to the reaization R used to define g^R. They might, for example, be differential operators or matrices belonging to some other realization. We define the associated quadratic Casimir operator $\mathcal{C}_2$ by the rule

$$\mathcal{C}_2 = \sum_{\alpha\alpha'} g_R^{\alpha\alpha'} \hat{B}_\alpha \hat{B}_{\alpha'}. \tag{27.11.35}$$

We must now show that $\mathcal{C}_2$ has the desired properties. Suppose $\exp(\epsilon\#\hat{C}\#)$ is applied to both sides of (11.35). Here $\hat{C}$ is some linear combination of the $\hat{B}_\alpha$. Doing so, and making use of (11.22) and the isomorphism property (8.2.14), gives the result

$$\mathcal{C}_2^{\rm tr} = \exp(\epsilon\#\hat{C}\#)\mathcal{C}_2 = \sum_{\alpha\alpha'} g_R^{\alpha\alpha'} \hat{B}^{\rm tr}_\alpha \hat{B}^{\rm tr}_{\alpha'}. \tag{27.11.36}$$

But we know that

$$\hat{B}^{\rm tr}_\alpha = \sum_\beta U_{\alpha\beta}\hat{B}_\beta \tag{27.11.37}$$

since only the structure constants $c^\gamma_{\alpha\beta}$ are involved in the computation of U. See (11.5) and (11.22). Therefore, (11.36) can be rewritten in the form

$$\mathcal{C}_2^{\rm tr} = \sum_{\alpha\alpha'\beta\beta'} g_R^{\alpha\alpha'} U_{\alpha\beta} U_{\alpha'\beta'} \hat{B}_\beta \hat{B}_{\beta'}. \tag{27.11.38}$$

Also, when written in expanded form, (11.33) and (11.34) yield the relation

$$\sum_{\alpha\alpha'} g_R^{\alpha\alpha'} U_{\alpha\beta} U_{\alpha'\beta'} = g_R^{\beta\beta'}. \tag{27.11.39}$$

Consequently, we have result

$$\mathcal{C}_2^{\mathrm{tr}} = \sum_{\beta\beta'} g_R^{\beta\beta'} \hat{B}_\beta \hat{B}_{\beta'} = \mathcal{C}_2. \tag{27.11.40}$$

That is, $\mathcal{C}_2$ is invariant under group action.

As a special case of (11.40), set $\hat{C} = \hat{B}_{\alpha''}$ and equate powers of ϵ in (11.36) and (11.40). Doing so gives the infinitesimal result

$$\#\hat{B}_{\alpha''}\#\mathcal{C}_2 = \{\hat{B}_{\alpha''}, \mathcal{C}_2\} = 0. \tag{27.11.41}$$

That is, all Lie generators commute with $\mathcal{C}_2$. Put yet another way, the raised metric tensor entries $g_R^{\alpha\alpha'}$ are the Clebsch-Gordan coefficients that couple together two copies of the representation associated with the B_α (the adjoint representation) to form the identity representation.

There is still another way of looking at our result. Since the commutator is antisymmetric, the relation (11.41) also states that $\mathcal{C}_2$ commutes with all Lie generators. And from this result, by the linearity and derivation properties of the commutator, we conclude that $\mathcal{C}_2$ commutes with *all* products and sums of products of Lie generators.

27.11.2 Applications of the Quadratic Casimir Operator

Before continuing on to a discussion of higher-order Casimir operators, let us pause to make use of the quadratic Casimir operator for the symplectic group. For our discussion we will use the fundamental representation. Examination of (11.13) and (11.16) through (11.19) shows that in this case

$$(g^F)^2 = 4I \tag{27.11.42}$$

and hence

$$g_F^{\alpha\alpha'} = (1/4) g^F_{\alpha\alpha'}. \tag{27.11.43}$$

Therefore, in view of (11.16) through (11.19) and up to a normalization which we choose for convenience, the quadratic Casimir for the symplectic group is given by the relation

$$\mathcal{C}_2 = \sum_j (C^j)^2 + \sum_{\boldsymbol{\mu}} R(-\boldsymbol{\mu}) R(\boldsymbol{\mu}). \tag{27.11.44}$$

Here the elements C^j and $R(\boldsymbol{\mu})$ are some kind of linear operators or matrices that obey commutation rules analogous to (4.15) through (4.18).

We will soon apply $\mathcal{C}_2$ to the highest weight state $|\boldsymbol{w}^h\rangle$ in some representation Γ. Before doing so, it is useful to rewrite $\mathcal{C}_2$ in a form that is convenient for this purpose. As was the case for weights (see Section 5.8), we define a root $\boldsymbol{\mu}$ to be *positive* if its first nonvanishing component is positive. For example, in the case of $sp(4)$, the roots $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $-\boldsymbol{\delta}$ are positive. See Figure 4.1. Note that if $\boldsymbol{\mu}$ is positive, then $-\boldsymbol{\mu}$ is not positive. Conversely, if $\boldsymbol{\mu}$ is not positive, then $-\boldsymbol{\mu}$ is positive. Thus, half the root vectors are positive, and the other half (their negatives) are not. With this definition in mind, we may rewrite (11.44) in the form

$$\mathcal{C}_2 = \sum_j (C^j)^2 + \sum_{\boldsymbol{\mu}>0} [R(\boldsymbol{\mu}) R(-\boldsymbol{\mu}) + R(-\boldsymbol{\mu}) R(\boldsymbol{\mu})]. \tag{27.11.45}$$

Here the notation $\boldsymbol{\mu} > 0$ indicates that $\boldsymbol{\mu}$ is positive. Next write the simple identity

$$R(\boldsymbol{\mu})R(-\boldsymbol{\mu}) + R(-\boldsymbol{\mu})R(\boldsymbol{\mu}) = 2R(-\boldsymbol{\mu})R(\boldsymbol{\mu}) + \{R(\boldsymbol{\mu}), R(-\boldsymbol{\mu})\}. \tag{27.11.46}$$

But, by (4.17), we have the relation

$$\{R(\boldsymbol{\mu}), R(-\boldsymbol{\mu})\} = \sum_j (\boldsymbol{e}^j \cdot \boldsymbol{\mu}) C^j. \tag{27.11.47}$$

Consequently, (11.45) can be rewritten in the form

$$\mathcal{C}_2 = \sum_j (C^j)^2 + \sum_{\boldsymbol{\mu}>0} \sum_j (\boldsymbol{e}^j \cdot \boldsymbol{\mu}) C^j + 2 \sum_{\boldsymbol{\mu}>0} R(-\boldsymbol{\mu})R(\boldsymbol{\mu}). \tag{27.11.48}$$

There is one final simplification. Define $\boldsymbol{\mu}^+$ to be the sum of all positive roots,

$$\boldsymbol{\mu}^+ = \sum_{\boldsymbol{\mu}>0} \boldsymbol{\mu}. \tag{27.11.49}$$

With this definition, $\mathcal{C}_2$ takes the form

$$\mathcal{C}_2 = \sum_j [(C^j)^2 + (\boldsymbol{e}^j \cdot \boldsymbol{\mu}^+) C^j] + 2 \sum_{\boldsymbol{\mu}>0} R(-\boldsymbol{\mu})R(\boldsymbol{\mu}). \tag{27.11.50}$$

We are ready to apply $\mathcal{C}_2$ to $|\boldsymbol{w}^h\rangle$. First observe that

$$R(\boldsymbol{\mu})|\boldsymbol{w}^h\rangle = 0 \text{ if } \boldsymbol{\mu} > 0. \tag{27.11.51}$$

Were this not so, $|\boldsymbol{w}^h\rangle$ would not be an eigenvector of the C^j with highest weight. [See (5.8.16).] Now the virtue of writing $\mathcal{C}_2$ in the form (11.50) is apparent. Also, we have the relations

$$\sum_j (C^j)^2 |\boldsymbol{w}^h\rangle = (\boldsymbol{w}^h \cdot \boldsymbol{w}^h)|\boldsymbol{w}^h\rangle, \tag{27.11.52}$$

$$\sum_j (\boldsymbol{e}^j \cdot \boldsymbol{\mu}^+) C^j |\boldsymbol{w}^h\rangle = (\boldsymbol{\mu}^+ \cdot \boldsymbol{w}^h)|\boldsymbol{w}^h\rangle. \tag{27.11.53}$$

It follows that $|\boldsymbol{w}^h\rangle$ is an eigenvector of $\mathcal{C}_2$ having eigenvalue $\lambda(\boldsymbol{w}^h, \boldsymbol{\mu}^+)$,

$$\mathcal{C}_2 |\boldsymbol{w}^h\rangle = \lambda(\boldsymbol{w}^h, \boldsymbol{\mu}^+)|\boldsymbol{w}^h\rangle, \tag{27.11.54}$$

with

$$\lambda(\boldsymbol{w}^h, \boldsymbol{\mu}^+) = (\boldsymbol{w}^h \cdot \boldsymbol{w}^h) + (\boldsymbol{\mu}^+ \cdot \boldsymbol{w}^h). \tag{27.11.55}$$

There is one last observation: We know that every state in a representation can be obtained by suitable linear combinations of products of ladder operators and Cartan subalgebra operators and constants applied to the highest-weight state. Also, $\mathcal{C}_2$ commutes with all these operations. It follows that *all* the vectors in an irreducible representation are eigenvectors of $\mathcal{C}_2$ with the *same* common eigenvalue $\lambda(\boldsymbol{w}^h, \boldsymbol{\mu}^+)$.

For future use, let us work out explicitly the eigenvalues of $\mathcal{C}_2$ for general representations in the cases of $sp(2)$, $sp(4)$, and $sp(6)$. Begin with $sp(2)$. In this case

$$\boldsymbol{w}^h = n\boldsymbol{\phi}^1 = n\boldsymbol{e}^1, \tag{27.11.56}$$

and

$$\boldsymbol{\mu}^+ = \boldsymbol{\alpha} = 2\boldsymbol{e}^1. \tag{27.11.57}$$

Consequently, for the representation $\Gamma(n)$, $\mathcal{C}_2$ has the eigenvalue

$$\lambda(\boldsymbol{w}^h, \boldsymbol{\mu}^+) = n^2 + 2n = n(n+2) = 4j(j+1). \tag{27.11.58}$$

We also note that $\mathcal{C}_2$ has the explicit form

$$\begin{aligned} \mathcal{C}_2 &= (C^1)^2 + R(-\boldsymbol{\alpha})R(\boldsymbol{\alpha}) + R(\boldsymbol{\alpha})R(-\boldsymbol{\alpha}) \\ &= 4\hat{J}_3^2 + 2\hat{J}_-\hat{J}_+ + 2\hat{J}_+\hat{J}_- \\ &= 4(\hat{J}_1^2 + \hat{J}_2^2 + \hat{J}_3^2). \end{aligned} \tag{27.11.59}$$

Here the quantities $\hat{J}_\pm$ and $\hat{J}_3$ (or $\hat{J}_1$ to $\hat{J}_3$) are some kind of linear operators or matrices that obey commutation rules analogous to (1.3), (1.4), or (1.21). The results (11.58) and (11.59) are those expected for $su(2)$. In particular, the quadratic Casimir operator is proportional to the square of the angular momentum, which is known to commute with the $\hat{J}_k$.

For the case of $sp(4)$,

$$\boldsymbol{w}^h = m\boldsymbol{\phi}^1 + n\boldsymbol{\phi}^2 = (m+n)\boldsymbol{e}^1 + n\boldsymbol{e}^2, \tag{27.11.60}$$

and

$$\boldsymbol{\mu}^+ = \boldsymbol{\alpha} + \boldsymbol{\beta} + \boldsymbol{\gamma} + (-\boldsymbol{\delta}) = 4\boldsymbol{e}^1 + 2\boldsymbol{e}^2. \tag{27.11.61}$$

Consequently, for the representation $\Gamma(m,n)$, $\mathcal{C}_2$ has the eigenvalue

$$\lambda(\boldsymbol{w}^h, \boldsymbol{\mu}^+) = m^2 + 2mn + 2n^2 + 4m + 6n. \tag{27.11.62}$$

For the case of $sp(6)$,

$$\boldsymbol{w}^h = \ell\boldsymbol{\phi}^1 + m\boldsymbol{\phi}^2 + n\boldsymbol{\phi}^3 = (\ell+m+n)\boldsymbol{e}^1 + (m+n)\boldsymbol{e}^2 + n\boldsymbol{e}^3, \tag{27.11.63}$$

and

$$\boldsymbol{\mu}^+ = \boldsymbol{\alpha}^1 + \boldsymbol{\alpha}^2 + \boldsymbol{\alpha}^3 + \boldsymbol{\beta}^1 + \boldsymbol{\beta}^2 + \boldsymbol{\beta}^3 + \boldsymbol{\gamma}^1 + \boldsymbol{\gamma}^2 + (-\boldsymbol{\gamma}^3) = 6\boldsymbol{e}^1 + 4\boldsymbol{e}^2 + 2\boldsymbol{e}^3. \tag{27.11.64}$$

Consequently, for the representation $\Gamma(\ell,m,n)$, $\mathcal{C}_2$ has the eigenvalue

$$\lambda(\boldsymbol{w}^h, \boldsymbol{\mu}^+) = \ell^2 + 2m^2 + 3n^2 + 2\ell m + 2\ell n + 4mn + 6\ell + 10m + 12n. \tag{27.11.65}$$

With this background, we are prepared to use the quadratic Casimir operator to prove the orthogonality relations (10.12). Our main tool will be the adjoint Lie operator version of $\mathcal{C}_2$ defined by writing

$$\mathcal{C}_2 = \sum_j (\# : c^j : \#)^2 + \sum_{\boldsymbol{\mu}} \# : r(-\boldsymbol{\mu}) : \#\# : r(\boldsymbol{\mu}) : \#. \tag{27.11.66}$$

From (4.8), (4.14), and (10.9) we find the conjugation relations

$$[(\# : c^j : \#)^2]^\dagger = [(\# : c^j : \#)^\dagger]^2 = (\# : c^j :^\dagger \#)^2 = (\# : c^j : \#)^2, \tag{27.11.67}$$

$$\begin{aligned} &[(\# : r(-\boldsymbol{\mu}) : \#\# : r(\boldsymbol{\mu}) : \#]^\dagger = [\# : r(\boldsymbol{\mu}) : \#]^\dagger [\# : r(-\boldsymbol{\mu}) : \#]^\dagger \\ &= [\# : r(\boldsymbol{\mu}) :^\dagger \#][\# : r(-\boldsymbol{\mu}) :^\dagger \#] = [\# : r(-\boldsymbol{\mu}) : \#\# : r(\boldsymbol{\mu}) : \#]. \end{aligned} \tag{27.11.68}$$

It follows that $\mathcal{C}_2$ is Hermitian,

$$\mathcal{C}_2^\dagger = \mathcal{C}_2. \tag{27.11.69}$$

Apply $\mathcal{C}_2$ to any element of the vector fields $\mathcal{H}^{\ell+1,0,0}$, $\mathcal{G}^{\ell-1,1,0}$, and $\mathcal{G}^{\ell-1,0,0}$. Based on (11.65), we find (writing in short-hand form) the results

$$\mathcal{C}_2\mathcal{H}^{\ell+1,0,0} = [(\ell+1)^2 + 6(\ell+1)]\mathcal{H}^{\ell+1,0,0}, \tag{27.11.70}$$

$$\begin{aligned} \mathcal{C}_2\mathcal{G}^{\ell-1,1,0} &= [(\ell-1)^2 + 2 + 2(\ell-1) + 6(\ell-1) + 10]\mathcal{G}^{\ell-1,1,0} \\ &= [(\ell-1)^2 + 8(\ell-1) + 12]\mathcal{G}^{\ell-1,1,0}, \end{aligned} \tag{27.11.71}$$

$$\mathcal{C}_2\mathcal{G}^{\ell-1,0,0} = [(\ell-1)^2 + 6(\ell-1)]\mathcal{G}^{\ell-1,0,0}. \tag{27.11.72}$$

Next consider matrix elements of the form $\langle \mathcal{H}^{\ell+1,0,0}, \mathcal{C}_2\mathcal{G}^{\ell-1,1,0}\rangle$. From (11.71) we have the result

$$\langle \mathcal{H}^{\ell+1,0,0}, \mathcal{C}_2\mathcal{G}^{\ell-1,1,0}\rangle = [(\ell-1)^2 + 8(\ell-1) + 12]\langle \mathcal{H}^{\ell+1,0,0}, \mathcal{G}^{\ell-1,1,0}\rangle. \tag{27.11.73}$$

However, using (11.69) and (11.70), these matrix elements also satisfy the relation

$$\begin{aligned} &\langle \mathcal{H}^{\ell+1,0,0}, \mathcal{C}_2\mathcal{G}^{\ell-1,1,0}\rangle = \langle \mathcal{C}_2^\dagger\mathcal{H}^{\ell+1,0,0}, \mathcal{G}^{\ell-1,1,0}\rangle = \langle \mathcal{C}_2\mathcal{H}^{\ell+1,0,0}, \mathcal{G}^{\ell-1,1,0}\rangle \\ &= [(\ell+1)^2 + 6(\ell+1)]\langle \mathcal{H}^{\ell+1,0,0}, \mathcal{G}^{\ell-1,1,0}\rangle. \end{aligned} \tag{27.11.74}$$

By combining (11.73) and (11.74) we find the result

$$(2\ell+2)\langle \mathcal{H}^{\ell+1,0,0}, \mathcal{G}^{\ell-1,1,0}\rangle = 0. \tag{27.11.75}$$

In similar fashion we find the results

$$(4\ell+12)\langle \mathcal{H}^{\ell+1,0,0}, \mathcal{G}^{\ell-1,0,0}\rangle = 0, \tag{27.11.76}$$

$$(2\ell+10)\langle \mathcal{G}^{\ell-1,1,0}, \mathcal{G}^{\ell-1,0,0}\rangle = 0. \tag{27.11.77}$$

We observe that none of the quantities in parentheses on the left sides of (11.75) through (11.77) vanish for $\ell \geq 0$. Therefore all the scalar products of the form (10.12) vanish as advertised.

There is a related use of the quadratic Casimir operator that is also important. In Section 21.10 it was shown that there is an operator $\mathcal{P}^{\mathcal{G}}$ that projects out the non-Hamiltonian part of a general homogeneous vector field $\mathcal{L}_{\boldsymbol{g}^\ell}$,

$$\mathcal{P}^{\mathcal{G}}\mathcal{L}_{\boldsymbol{g}^\ell} = \mathcal{G}_{\ell+1}, \tag{27.11.78}$$

and an explicit procedure was given for finding $\mathcal{G}_{\ell+1}$. See (10.20) and (10.33). Now we will show how $\mathcal{C}_2$ can be used to decompose $\mathcal{G}_{\ell+1}$ into its separate parts $\mathcal{G}^{\ell-1,1,0}$ and $\mathcal{G}^{\ell-1,0,0}$,

$$\mathcal{G}_{\ell+1} = \mathcal{G}^{\ell-1,1,0} + \mathcal{G}^{\ell-1,0,0}. \tag{27.11.79}$$

Suppose $\mathcal{C}_2$ is applied to both sides of (11.79). Then, from (11.71) and (11.72), we find the result

$$\begin{aligned}\mathcal{C}_2\mathcal{G}_{\ell+1} &= \mathcal{C}_2\mathcal{G}^{\ell-1,1,0} + \mathcal{C}_2\mathcal{G}^{\ell-1,0,0}\\ &= [(\ell-1)^2 + 8(\ell-1) + 12]\mathcal{G}^{\ell-1,1,0} + [(\ell-1)^2 + 6(\ell-1)]\mathcal{G}^{\ell-1,0,0}.\end{aligned} \tag{27.11.80}$$

The two relations (11.79) and (11.80) can be solved for $\mathcal{G}^{\ell-1,1,0}$ and $\mathcal{G}^{\ell-1,0,0}$ to give the explicit results

$$\mathcal{G}^{\ell-1,1,0} = \{2\ell+10\}^{-1}\{\mathcal{C}_2 - [(\ell-1)^2 + 6(\ell-1)]\}\mathcal{G}_{\ell+1}, \tag{27.11.81}$$

$$\mathcal{G}^{\ell-1,0,0} = -\{2\ell+10\}^{-1}\{\mathcal{C}_2 - [(\ell-1)^2 + 8(\ell-1) + 12]\}\mathcal{G}_{\ell+1}. \tag{27.11.82}$$

Thus, given any homogeneous vector field $\mathcal{L}_{\boldsymbol{g}^\ell}$, we have an explicit procedure for finding its Hamiltonian part $\mathcal{H}^{\ell+1,0,0}$ and its non-Hamiltonian parts $\mathcal{G}^{\ell-1,1,0}$ and $\mathcal{G}^{\ell-1,0,0}$. Of course, this is not the full story. For some purposes we would like to have a complete set of basis vectors for the spaces $\mathcal{H}^{\ell+1,0,0}$, $\mathcal{G}^{\ell-1,1,0}$, and $\mathcal{G}^{\ell-1,0,0}$. This was done for the case $\ell = 1$ in Section 21.9, and we would like to have analogous results for all ℓ, or for at least the first few values of ℓ (say $\ell = 2$ through 6 or so). The spaces $\mathcal{H}^{\ell+1,0,0}$ and $\mathcal{G}^{\ell-1,0,0}$ are relatively easy to handle because the elements of $\mathcal{H}^{\ell+1,0,0}$ are of the form $: h_{\ell+1} :$ and the elements of $\mathcal{G}^{\ell-1,0,0}$ are of the form (9.3). In both cases one is working with homogeneous polynomials and must find suitable basis polynomials that correspond to the various weights in the weight diagrams for the representations $\Gamma(\ell+1,0,0)$ and $\Gamma(\ell-1,0,0)$. This is relatively straightforward except for the problem of finding additional labels and associated properties when the weights have multiplicities higher than 1. Handling the space $\mathcal{G}^{\ell-1,1,0}$ is more difficult. In this case it would be helpful to have explicit knowledge of the Clebsch-Gordan coefficients of $sp(6)$ [and $sp(4)$] for at least the relatively low-dimensional representations.

27.11.3 Higher-Order Casimir Operators

Before leaving the subject of Casimir operators, something should be said about cubic and higher-order Casimirs. For simplicity, only the cubic case will be considered, but the generalization to higher orders should be evident.

As before, we work with some realization R, and let $\hat{B}_\alpha$ denote the basis elements of the Lie algebra in this realization. Then, in analogy to (11.11), we define a rank *three* tensor ${}_3g^R_{\alpha\alpha'\alpha''}$ by writing

$${}_3g^R_{\alpha\alpha'\alpha''} = \text{ tr } (\hat{B}_\alpha\hat{B}_{\alpha'}\hat{B}_{\alpha''}). \tag{27.11.83}$$

In view of (11.24) and the properties of the trace, we have the relation

$$\text{tr } (\hat{B}^{\text{tr}}_\alpha\hat{B}^{\text{tr}}_{\alpha'}\hat{B}^{\text{tr}}_{\alpha''}) = \text{tr } (\hat{B}_\alpha\hat{B}_{\alpha'}\hat{B}_{\alpha''}). \tag{27.11.84}$$

From this relation and (11.28) we deduce that ${}_3g^R$ has the property

$$\sum_{\beta\beta'\beta''} U_{\alpha\beta}U_{\alpha'\beta'}U_{\alpha''\beta''}{}_3g^R_{\beta\beta'\beta''} = {}_3g^R_{\alpha\alpha'\alpha''}, \tag{27.11.85}$$

which is the analog of (11.30).

Next use g_R to raise the indices on ${}_3g^R$ by the rule

$${}_3g_R^{\alpha\alpha'\alpha''} = g_R^{\alpha\gamma}g_R^{\alpha'\gamma'}g_R^{\alpha''\gamma''}\,{}_3g^R_{\gamma\gamma'\gamma''}, \tag{27.11.86}$$

Here, and in what follows, we have and will use the summation convention. The raised tensor ${}_3g_R$ has the property

$${}_3g_R^{\alpha\alpha'\alpha''}U_{\alpha\beta}U_{\alpha'\beta'}U_{\alpha''\beta''} ={}_3 g_R^{\beta\beta'\beta''}, \tag{27.11.87}$$

which is the analog of (11.33). [Note that in (11.85) the summation is over the second indices in the U's, while in (11.87) it is over the first indices.] To verify (11.87), simply compute. From (11.86) we have

$${}_3g_R^{\alpha\alpha'\alpha''}U_{\alpha\beta}U_{\alpha'\beta'}U_{\alpha''\beta''} = g_R^{\alpha\gamma}U_{\alpha\beta}g^{\alpha'\gamma'}U_{\alpha'\beta'}g^{\alpha''\gamma''}U_{\alpha''\beta''}{}_3g^R_{\gamma\gamma'\gamma''}. \tag{27.11.88}$$

However, by changing indices, (11.85) can be written in the form

$${}_3g^R_{\gamma\gamma'\gamma''} = U_{\gamma\delta}U_{\gamma'\delta'}U_{\gamma''\delta''}{}_3g^R_{\delta\delta'\delta''}. \tag{27.11.89}$$

Now substitute (11.89) in (11.88) to get the result

$$\begin{aligned} {}_3g_R^{\alpha\alpha'\alpha''}U_{\alpha\beta}U_{\alpha'\beta'}U_{\alpha''\beta''} &= g_R^{\alpha\gamma}U_{\alpha\beta}U_{\gamma\delta}g_R^{\alpha'\gamma'}U_{\alpha'\beta'}U_{\gamma'\delta'}g_R^{\alpha''\gamma''}U_{\alpha''\beta''}U_{\gamma''\delta''}{}_3g^R_{\delta\delta'\delta''} \\ &= g_R^{\beta\delta}g_R^{\beta'\delta'}g_R^{\beta''\delta''}\,{}_3g^R_{\delta\delta'\delta''} = {}_3g_R^{\beta\beta'\beta''}, \end{aligned} \tag{27.11.90}$$

as claimed. Here repeated use has been made of (11.30).

We are now ready to construct the third-order Casimir operator. As before, suppose $\hat{B}_\alpha$ is any set of linear operators that obey (11.5), but do not necessarily belong to the realization used to define g^R, ${}_3g^R$, and hence ${}_3g_R$. We define the associated cubic Casimir operator $\mathcal{C}_3$ by writing

$$\mathcal{C}_3 = {}_3g_R^{\alpha\alpha'\alpha''}\hat{B}_\alpha\hat{B}_{\alpha'}\hat{B}_{\alpha''}. \tag{27.11.91}$$

As a consequence of (11.87), this operator also has the invariance property

$$\mathcal{C}_3^{\rm tr} = \exp(\epsilon\#C\#)\mathcal{C}_3 = \mathcal{C}_3, \tag{27.11.92}$$

and hence

$$\#\hat{B}_\delta\#\mathcal{C}_3 = \{\hat{B}_\delta, \mathcal{C}_3\} = 0 \text{ for all } \delta. \tag{27.11.93}$$

Indeed, using (11.37) and (11.87), we find the result

$$\begin{aligned} \mathcal{C}_3^{\rm tr} &= {}_3g_R^{\alpha\alpha'\alpha''}\hat{B}_\alpha^{\rm tr}\hat{B}_{\alpha'}^{\rm tr}\hat{B}_{\alpha''}^{\rm tr} \\ &= {}_3g_R^{\alpha\alpha'\alpha''}U_{\alpha\beta}U_{\alpha'\beta'}U_{\alpha''\beta''}\hat{B}_\beta\hat{B}_{\beta'}\hat{B}_{\beta''} \\ &= {}_3g_R^{\beta\beta'\beta''}\hat{B}_\beta\hat{B}_{\beta'}\hat{B}_{\beta''} = \mathcal{C}_3. \end{aligned} \tag{27.11.94}$$

Thus, as anticipated, $\mathcal{C}_3$ also commutes with all Lie generators, and therefore also with all products and sums of products of generators. A special case of this result is that $\mathcal{C}_2$ and $\mathcal{C}_3$ commute.

Finally, we remark that it can be shown that a rank-k simple Lie algebra has k functionally independent Casimir operators. Moreover, the eigenvalues of these Casimir operators, when acting on any vector in an irreducible representation, can be used to determine the representation. For example, in the case of $sp(6)$, the three Casimir operators $\mathcal{C}_2$, $\mathcal{C}_4$, and $\mathcal{C}_6$ are functionally independent and can be used to determine the values of ℓ, m, n in $\Gamma(\ell, m, n)$.

Exercises

27.11.1. Verify (11.12) using (11.11) and the matrices B^0, F, and G given by (5.6.7), (5.6.13), and (5.6.14). Find the 2×2 matrices associated with the elements associated with c^1 and $r(\pm\boldsymbol{\alpha})$ given by (1.11) and (1.12). Use these matrices to verify (11.13). Find g^F for $sp(4)$ using the fundamental representation given by (5.7.42).

For *quadratic* polynomials, using (5.5.1) and (5.5.2) and the correspondences (5.5.3) and (5.5.4), we have made the definition $\langle f, g\rangle_F = (JS^f, JS^g)_F = \text{tr}\,(JS^f JS^g)$. Compare $\langle f, g\rangle$ and $\langle f, g\rangle_F$. See Exercise 7.3.8. Verify the general result $\langle f^a, f^c\rangle_F = (JS^{fa}, JS^{fc})_F = 0$. Use (3.8.14), (3.8.22), (7.2.3), (7.2.4), and (7.3.53); and employ the notation S^{fa} and S^{fc} to denote the parts of S^f that anticommute and commute with J, respectively. See Exercise 7.3.10 for the analogous result $\langle f_2^a, f_2^c\rangle = 0$. For a pair of quadratic polynomials f and g, make the decompositions $f = f^a + f^c$ and $g = g^a + g^c$. Verify the relation $\langle f, g\rangle_F = 2\langle f^a, g^a\rangle - 2\langle f^c, g^c\rangle$. As a special case there is the relation $\langle f, f\rangle_F = 2\langle f^a, f^a\rangle - 2\langle f^c, f^c\rangle$, which shows that the form $\langle f, f\rangle_F$ is neither positive nor negative definite. This is to be expected because $Sp(2n, \mathbb{R})$ is not compact. We also have the relation

$$\langle f^c, f^c\rangle_F = (JS^{fc}, JS^{fc})_F = -2\langle f^c, f^c\rangle < 0,$$

and we know that the JS^{fc} generate $U(n)$, the maximal compact subgroup of $Sp(2n, \mathbb{R})$. See the last comment in Exercise 11.3 below.

27.11.2. Verify the relations (11.16) through (11.19).

27.11.3. Review Exercise 11.1 above. This exercise further explores the relation between the Lie-algebraic metric for the Lie algebra $usp(2n)$ and the $USp(2n)$ invariant scalar product. See Sections 5.10 and 7.3. Consider the $sp(2n)$ Lie-algebraic metric given by (11.16) through (11.19). Define elements Σ and Δ by the rules

$$\Sigma(\boldsymbol{\mu}) = (1/\sqrt{2})[r(\boldsymbol{\mu}) + r(-\boldsymbol{\mu})], \tag{27.11.95}$$

$$\Delta(\boldsymbol{\mu}) = (1/\sqrt{2})[r(\boldsymbol{\mu}) - r(-\boldsymbol{\mu})]. \tag{27.11.96}$$

Evidently, for $\boldsymbol{\mu} > 0$, the elements $\Sigma(\boldsymbol{\mu})$ and $\Delta(\boldsymbol{\mu})$ span the same space as the elements $r(\boldsymbol{\mu})$ and $r(-\boldsymbol{\mu})$. Using (11.16) through (11.19) verify (with $\boldsymbol{\mu}, \boldsymbol{\nu} > 0$) the relations

$$(c^j, \Sigma(\boldsymbol{\mu}))_F = (c^j, \Delta(\boldsymbol{\mu}))_F = 0, \tag{27.11.97}$$

$$(\Sigma(\boldsymbol{\mu}), \Delta(\boldsymbol{\nu}))_F = 0, \tag{27.11.98}$$

$$(\Sigma(\boldsymbol{\mu}), \Sigma(\boldsymbol{\nu}))_F = 2\delta_{\boldsymbol{\mu}\boldsymbol{\nu}}, \tag{27.11.99}$$

$$(\Delta(\boldsymbol{\mu}), \Delta(\boldsymbol{\nu}))_F = -2\delta_{\boldsymbol{\mu}\boldsymbol{\nu}}. \tag{27.11.100}$$

Show that $:\Sigma:$ and $:\Delta:$ obey the conjugacy relations

$$:\Sigma(\boldsymbol{\mu}):^{\dagger} = :\Sigma(\boldsymbol{\mu}):, \tag{27.11.101}$$

$$:\Delta(\boldsymbol{\mu}):^{\dagger} = -:\Delta(\boldsymbol{\mu}):. \tag{27.11.102}$$

Consider as a basis set the elements ic^j, $i\Sigma(\boldsymbol{\mu})$, and $\Delta(\boldsymbol{\mu})$ with $\boldsymbol{\mu} > 0$. Call these elements b_α. Show that their associated Lie operators are all anti-Hermitian,

$$:b_\alpha:^{\dagger} = -:b_\alpha:. \tag{27.11.103}$$

Consequently, they form a basis for $usp(2n)$ and, when exponentiated, generate $USp(2n)$. See Section 7.3. Verify that these elements satisfy the relation

$$(b_\alpha, b_\beta)_F = -2\delta_{\alpha\beta} = -2\langle b_\alpha, b_\beta\rangle. \tag{27.11.104}$$

Here the Lie-algebraic scalar product on the left is that given by (11.16) through (11.19), and the scalar product on the right is that given by (4.19) through (4.22) and arises from the construction of Section 7.3. Since we have been working over the complex field, we know that the Lie-algebraic scalar product is invariant under $Sp(2n, C)$. See (11.25). It is therefore also invariant under $USp(2n)$ because $USp(2n)$ is a subgroup of $Sp(2n, C)$. The relation (11.104) is consistent with this invariance because we already know from Section 7.3 that the scalar product on the right is invariant under $USp(2n)$.

One last comment: From the discussion of Section 5.10 we know that $USp(2n)$ is compact. Inspection of (11.104) shows that the Lie-algebraic metric for $usp(2n)$ is negative definite. It can be shown that for any simple Lie algebra the Lie-algebraic metric is negative definite if and only if the corresponding Lie group is compact. See also (11.109) and (11.110) for the cases of the compact groups $SU(2)$ and $SO(3, \mathbb{R})$.

27.11.4. The relation (11.30) displays the invariance of the metric tensor under finite group action. Show that (11.27) describes this same invariance in infinitesimal form. In particular, use (11.27) to produce the relation

$$c^{\gamma}_{\alpha\alpha'} g^R_{\gamma\alpha''} = c^{\gamma}_{\alpha'\alpha''} g^R_{\gamma\alpha}. \tag{27.11.105}$$

Use the metric tensor g^R to *lower* the upper index on the structure constants by the rule

$$c_{\alpha\alpha'\alpha''} = c^{\gamma}_{\alpha\alpha'} g^R_{\gamma\alpha''}. \tag{27.11.106}$$

Show that the lowered structure constants are completely antisymmetric (antisymmetric under the interchange of any pair of adjacent indices).

27.11.5. Find the analog of the formulas (11.81) and (11.82) for the case of 4-dimensional phase space.

27.11.6. In the case of a 6-dimensional phase space, consider the vector field $\mathcal{L}_{\boldsymbol{g}^2}$ given by the relation

$$\mathcal{L}_{\boldsymbol{g}^2} = (q_1)^2\partial/\partial q_1. \tag{27.11.107}$$

Using the methods of Sections 21.10 and 21.11.2, decompose $\mathcal{L}_{\boldsymbol{g}^2}$ into Hamiltonian and non-Hamiltonian parts $\mathcal{H}^{3,0,0}$, $\mathcal{G}^{1,1,0}$, and $\mathcal{G}^{1,0,0}$.

27.11.7. For the case of a 2-dimensional phase space we know that quadratic functions f_2 and g_2 carry the representation $\Gamma(2)$. Therefore, from (11.1), we might naively expect that the product f_2g_2 might contain the identity representation $\Gamma(0)$. In analogy with (11.59), if the identity representation does occur, it should be a multiple of the polynomial $(J_1^2+J_2^2+J_3^2)$ where the J_j are given by (1.1), (1.2), and (1.20). Show that in fact there is the relation

$$J_1^2 + J_2^2 + J_3^2 = 0. \tag{27.11.108}$$

Thus, in this case, the sought after quantity actually vanishes. In retrospect, this is to be expected because we know from Section 21.2 that quartic polynomials in two variables, of which all polynomials of the form f_2g_2 are examples, carry only the representation $\Gamma(4)$. Note that $\Gamma(4)$ also occurs in the Clebsch-Gordan series (11.1).

27.11.8. The purpose of this exercise is to explore the consequences of the relations (11.27) in the case of $su(2)$, or equivalently, $so(3,\mathbb{R})$. In the case of $su(2)$, suppose we employ the realization provided by the K^α matrices of Exercise 3.7.30. Verify the scalar product results

$$(K^\alpha, K^\beta)_F = \text{ tr } (K^\alpha K^\beta) = (-i/2)^2 \text{ tr } (\sigma^\alpha\sigma^\beta) = -(1/2)\delta_{\alpha\beta} = -(1/2)\boldsymbol{e}_\alpha \cdot \boldsymbol{e}_\beta. \tag{27.11.109}$$

Show that in the cases of $su(2)$ and $so(3,\mathbb{R})$ there are the related results

$$(K^\alpha, K^\beta)_K = (L^\alpha, L^\beta)_F = \text{ tr } (L^\alpha L^\beta) = -2\delta_{\alpha\beta} = -2\boldsymbol{e}_\alpha \cdot \boldsymbol{e}_\beta. \tag{27.11.110}$$

Here the subscript K stands for *Killing* in anticipation of the next section. Next show, using the notation of Exercise 3.2.27, that there are the relations

$$(\boldsymbol{a}\cdot\boldsymbol{K}, \boldsymbol{b}\cdot\boldsymbol{K})_F = -(1/2)\boldsymbol{a}\cdot\boldsymbol{b}, \tag{27.11.111}$$

$$(\boldsymbol{a}\cdot\boldsymbol{K}, \boldsymbol{b}\cdot\boldsymbol{K})_K = (\boldsymbol{a}\cdot\boldsymbol{L}, \boldsymbol{b}\cdot\boldsymbol{L})_F = -2\boldsymbol{a}\cdot\boldsymbol{b}. \tag{27.11.112}$$

Now examine the first relation in (11.27). In the case of $su(2)$, and using the fundamental representation, it reads

$$(\{K_\alpha, K_\beta\}, K_\gamma)_F = (K_\alpha, \{K_\beta, K_\gamma\})_F. \tag{27.11.113}$$

Show that multiplying the left side of (11.113) by the quantity $a_\alpha b_\beta c_\gamma$, and summing over α, β, and γ, yield the result

$$\sum_{\alpha\beta\gamma} a_\alpha b_\beta c_\gamma (\{K_\alpha, K_\beta\}, K_\gamma)_F = (\{\boldsymbol{a}\cdot\boldsymbol{K}, \boldsymbol{b}\cdot\boldsymbol{K}\}, \boldsymbol{c}\cdot\boldsymbol{K})_F = -(1/2)(\boldsymbol{a}\times\boldsymbol{b})\cdot\boldsymbol{c}. \tag{27.11.114}$$

Show that multiplying the right side of (11.113) by the quantity $a_\alpha b_\beta c_\gamma$, and summing over α, β, and γ, yield the result

$$\sum_{\alpha\beta\gamma} a_\alpha b_\beta c_\gamma (K_\alpha, \{K_\beta, K_\gamma\})_F = (\boldsymbol{a}\cdot\boldsymbol{K}, \{\boldsymbol{b}\cdot\boldsymbol{K}, \boldsymbol{c}\cdot\boldsymbol{K}\})_F = -(1/2)\boldsymbol{a}\cdot(\boldsymbol{b}\times\boldsymbol{c}). \quad (27.11.115)$$

You have shown, in the case of $su(2)$, that (11.113) is equivalent to the familiar statement about the interchange of the dot and the cross in three-dimensional vector algebra,

$$(\boldsymbol{a}\times\boldsymbol{b})\cdot\boldsymbol{c} = \boldsymbol{a}\cdot(\boldsymbol{b}\times\boldsymbol{c}). \quad (27.11.116)$$

Carry out the analogous calculation for both the cases of $su(2)$ and $so(3,\mathbb{R})$ using (11.110) to again arrive at the conclusion (11.116). Finally show, in the cases of $su(2)$ and $so(3, R$, that the second relation in (11.27) is equivalent to the relation

$$\boldsymbol{a}\cdot(\boldsymbol{b}\times\boldsymbol{c}) = \boldsymbol{b}\cdot(\boldsymbol{c}\times\boldsymbol{a}) = \boldsymbol{c}\cdot(\boldsymbol{a}\times\boldsymbol{b}). \quad (27.11.117)$$

27.11.9. Show that the first relation in (11.27) can be rewritten in the form

$$(- : B_{\alpha'} : B_\alpha,\ B_{\alpha''})_R = (B_\alpha,\ \ : B_{\alpha'} : B_{\alpha''})_R. \quad (27.11.118)$$

Here we have used the more compact notation

$$: B_{\alpha'} := \text{ad } B_{\alpha'}. \quad (27.11.119)$$

Compare (3.7.71), (5.3.2), and (11.21). Show that (11.118) implies the relation

$$: B_{\alpha'} :^{\dagger} = - : B_{\alpha'} : \quad (27.11.120)$$

with respect to the inner product $(\ ,\ )_R$. That is, $: B_{\alpha'} :$ is anti-Hermitian with respect to this inner product.

27.11.10. Exercise on the Casimir operator for $SO(4,\mathbb{R})$.

27.12 The Killing Form

Section 21.11.1 introduced the concept of a scalar product for the elements of a Lie algebra and defined a metric tensor with the aid of a realizaton R. An important special case of this construction is the Killing form. The *Killing form* is simply the metric tensor g^R in the case that the realizaton R is the *adjoint* representation. See the end of Section 3.7 to review the definition of the adjoint representation. Let us call this tensor g^K in honor of Killing. Then, using (11.11) and (3.7.56), we find the result

$$g^K_{\alpha\alpha'} = \text{ tr } (\hat{B}_\alpha \hat{B}_{\alpha'}) = \sum_{\mu\nu} (\hat{B}_\alpha)_{\mu\nu} (\hat{B}_{\alpha'})_{\nu\mu} = \sum_{\mu\nu} c^\mu_{\alpha\nu} c^\nu_{\alpha'\mu}. \quad (27.12.1)$$

As (12.1) shows, the Killing form (metric tensor) has the advantage that it is constructed directly in terms of the structure constants. It is therefore directly available without further study of the Lie algebra.[4] By contrast, the g^F that we have been using for $sp(2n)$

[4] Suppose a Lie algebra L is specified by presenting its structure constants. Then g^K is computable using (12.1). It can be shown that g^K is invertible if and only if L is semisimple.

is constructed from a knowledge of the fundamental $2n \times 2n$ matrix representation. For most groups these matrices are usually much smaller than the matrices for the adjoint representation, and therefore, *if known*, far easier to use. For example, in the case of $sp(6)$, the fundamental representation involves 6×6 matrices, and the adjoint representation involves 21×21 matrices. [However, it turns out that for $E_8(248)$ the lowest dimensional representation *is* the adjoint representation, and 248×248 matrices are required.]

It can be shown in general for a *simple* Lie algebra that g^K and g^F are proportional. [For example, the identity representation occurs once and only once in the $sp(2n)$ Clebsch-Gordan series (11.1) through (11.3).] From (11.16) we know that for the fundamental representation of $sp(2n)$ there is the relation

$$(\hat{C}^1, \hat{C}^1)_F = 2. \tag{27.12.2}$$

And for the adjoint representation of $sp(2n)$ there is the relation

$$(\hat{C}^1, \hat{C}^1)_K = \sum_{\boldsymbol{\mu}} (\boldsymbol{e}^1 \cdot \boldsymbol{\mu})(\boldsymbol{\mu} \cdot \boldsymbol{e}^1) = 4n + 4. \tag{27.12.3}$$

See Exercise 12.1. It follows that g^K and g^F are related by the equation

$$g^K_{\alpha\alpha'} = (2n+2) g^F_{\alpha\alpha'}. \tag{27.12.4}$$

Still a bit more can be said. As before, let g^R be the metric tensor obtained using the realization R as in (11.11). Then there is a relation of the form

$$g^R_{\alpha\alpha'} = \tau(R) g^F_{\alpha\alpha'}. \tag{27.12.5}$$

where $\tau(R)$ is a *positive* proportionality constant that depends on the realization.[5] According to (12.4), τ has the value $(2n+2)$ for the adjoint representation and the Lie algebra $sp(2n)$.

Exercises

27.12.1. Verify (12.2). Using (3.7.56), show that in the adjoint representation the matrix $\hat{C}^1$ is diagonal and has as its diagonal entries ℓ zeroes (where ℓ is the rank of the Lie algebra) and the numbers $(\boldsymbol{e} \cdot \boldsymbol{\mu})$ where $\boldsymbol{\mu}$ ranges over all the root vectors. Next show that $(\hat{C}^1, \hat{C}^1)_K$ has the value

$$(\hat{C}^1, \hat{C}^1)_K = \text{tr } [(\hat{C}^1)^2] = \sum_{\boldsymbol{\mu}} (\boldsymbol{e}^1 \cdot \boldsymbol{\mu})(\boldsymbol{\mu} \cdot \boldsymbol{e}^1). \tag{27.12.6}$$

Finally, given that the root vectors for $sp(2n)$ are all combinations of the form $\pm \boldsymbol{e}^j \pm \boldsymbol{e}^k$ with the signs taken independently, verify (12.3).

27.12.2. Review Exercise 3.7.30. The 2×2 matrices K^α and the 3×3 matrices L^α displayed there provide the fundamental and adjoint representations of $su(2)$, respectively. Use these matrices to construct metric tensors for $su(2)$. Show that

$$g^F_{\alpha\alpha'} = -(1/2)\delta_{\alpha\alpha'} \tag{27.12.7}$$

[5] We remark that a relation of the form (12.5) holds among the *irreducible* representations of any *simple* Lie algebra. It need not hold in general. See Exercise 12.2.

and

$$g^K_{\alpha\alpha'} = -2\delta_{\alpha\alpha'}. \tag{27.12.8}$$

Note that g^F and g^K are proportional with a positive proportionality constant, as expected because $su(2)$ is simple. Note also that they are negative definite because $SU(2)$ is compact. (Recall that it has the topology of S^3.) Finally, since the Lie algebras $so(3, R$ and $su(2)$ are the same, these g^F and g^K also provide metric tensors for $so(3, \mathbb{R})$.

27.12.3. Use the matrices (4.4.31) through (4.4.34) as a basis for $g\ell(2, \mathbb{R})$ and show that in this case g^F is given by (4.4.40). Find the adjoint representation for $g\ell(2, \mathbb{R})$, which will be a set if 4 × 4 matrices, compute the Killing form g^K, and show that it is *singular*. Thus, (12.5) does not hold in this case. Show that $g\ell(2, \mathbb{R})$ is not *simple* and also not *semisimple*. See also the discussion at the end of Section 3.7.

27.12.4. Using the $sp(4)$ matrices (5.7.42) and matrices of the form JA with the antisymmetric matrices A given by (6.28) through (6.33), find g^F for $g\ell(4, \mathbb{R})$.

27.13 Enveloping Algebra

So far we have been exploring the properties of products of entities when each entity by itself has well-defined properties under the action of some group (in our case, the symplectic group). These entities were either polynomials in some variables on which the group acted, or Lie operators, or adjoint Lie operators. (They could also be matrices or other linear operators. As shown in Section 26.*, they could also be moments of a particle distribution.) They were not necessarily in the Lie algebra of the group, but they had the two properties that they could be multiplied together (multiplication was defined) and they transformed in some systematic way under the action of the Lie algebra.

For some (perhaps mathematical) purposes it is useful to explore what can be done by working directly and abstractly with only the Lie algebra itself rather than various concrete entities such as polynomials, Lie operators, adjoint Lie operators, etc. But now there is a problem because, for an abstract Lie algebra, there is no meaning for the "ordinary" product of any two elements in the Lie algebra. That is, if A and B are elements in a Lie algebra L, the Lie product $[A, B]$ is defined, but there is no meaning to the product AB. In particular, there is no meaning to an associative product ABC such that $(AB)C = A(BC)$. This apparent obstacle can be overcome by a clever construction. We will see that in many ways the *tensor* product can be used to play the role of an ordinary product.

Since L is a vector space, it *is* meaningful to talk about tensor products of the space with itself. For example, if the elements B_α are a basis for L, we may consider the space of all linear combinations of tensor products of the form $B_\alpha \otimes B_{\alpha'}$. We will call this vector space L^2, and write

$$L^2 = L \otimes L. \tag{27.13.1}$$

Similarly, we may consider the space of all linear combinations of tensor products of the form $B_\alpha \otimes B_{\alpha'} \otimes B_{\alpha''}$. Note that for a tensor product there is the associative property

$$B_\alpha \otimes (B_{\alpha'} \otimes B_{\alpha''}) = (B_\alpha \otimes B_{\alpha'}) \otimes B_{\alpha''}. \tag{27.13.2}$$

We will call this vector space L^3, and write

$$L^3 = L \otimes L \otimes L. \tag{27.13.3}$$

It is now obvious how to define still higher-order tensor product spaces L^4, L^5, $\cdots$.

We would also like to define L^1 and L^0. For L^1 we take the Lie algebra itself. That is, L^1 is the vector space consisting of all linear combinations of the B_α. What about L^0? Since L is a vector space, there must be some associated field of scalars (say the complex numbers) with *unit* element 1. Let L^0 be the vector space of all linear combinations (scalar multiples) of 1. Evidently L^0 is a one-dimensional vector space, and consists of the field of scalars associated with L.

Now watch closely! Having defined the vector spaces L^n, we define the vector space $\mathcal{T}$ to be the direct sum of all these vector spaces,

$$\mathcal{T} = L^0 \oplus L^1 \oplus L^2 \oplus L^3 + \cdots . \tag{27.13.4}$$

Suppose $\mathcal{A}$ and $\mathcal{B}$ are any two elements in $\mathcal{T}$. As a simple example, suppose they are of the form

$$\mathcal{A} = a + bB_\alpha = a1 + bB_\alpha, \tag{27.13.5}$$

$$\mathcal{B} = c + dB_\beta = c1 + dB_\beta. \tag{27.13.6}$$

Let us compute their tensor product. Doing so gives the result

$$\mathcal{A} \otimes \mathcal{B} = (a1 + bB_\alpha) \otimes (c1 + dB_\beta) = a1 \otimes c1 + a1 \otimes dB_\beta + bB_\alpha \otimes c1 + bB_\alpha \otimes dB_\beta. \tag{27.13.7}$$

Let us make the obvious rules

$$1 \otimes 1 = 1,\ 1 \otimes B_\beta = B_\beta,\ B_\alpha \otimes 1 = B_\alpha,\ \text{etc.}, \tag{27.13.8}$$

which can be summarized abstractly by writing

$$L^0 \otimes L^0 = L^0,\ L^0 \otimes L^n = L^n,\ L^n \otimes L^0 = L^n. \tag{27.13.9}$$

Then we find the result

$$\mathcal{A} \otimes \mathcal{B} = ac + adB_\beta + bcB_\alpha + bdB_\alpha \otimes B_\beta. \tag{27.13.10}$$

We know that, by construction, $\mathcal{T}$ is a vector space. We now see that it can be also be viewed as an *associative* algebra with the operation of multiplication taken to be the tensor product. The standard nomenclature is to call $\mathcal{T}$ the *tensor algebra* of L.

Our next step is to give $\mathcal{T}$ a Lie-algebraic structure. Suppose again that $\mathcal{A}$ and $\mathcal{B}$ are any two elements in $\mathcal{T}$. We define their commutator by the rule

$$\{\mathcal{A}, \mathcal{B}\} = \mathcal{A} \otimes \mathcal{B} - \mathcal{B} \otimes \mathcal{A}. \tag{27.13.11}$$

It is obvious that this commutator has the desired antisymmetry property (3.7.41). Let us check the Jacobi condition. We find the results

$$\{\mathcal{A}, \{\mathcal{B}, \mathcal{C}\}\} = \mathcal{A} \otimes \mathcal{B} \otimes \mathcal{C} - \mathcal{B} \otimes \mathcal{C} \otimes \mathcal{A} - \mathcal{A} \otimes \mathcal{C} \otimes \mathcal{B} + \mathcal{C} \otimes \mathcal{B} \otimes \mathcal{A}. \tag{27.13.12}$$

$$\{\mathcal{B},\{\mathcal{C},\mathcal{A}\}\} = \mathcal{B}\otimes\mathcal{C}\otimes\mathcal{A} - \mathcal{C}\otimes\mathcal{A}\otimes\mathcal{B} - \mathcal{B}\otimes\mathcal{A}\otimes\mathcal{C} + \mathcal{A}\otimes\mathcal{C}\otimes\mathcal{B}. \tag{27.13.13}$$

$$\{\mathcal{C},\{\mathcal{A},\mathcal{B}\}\} = \mathcal{C}\otimes\mathcal{A}\otimes\mathcal{B} - \mathcal{A}\otimes\mathcal{B}\otimes\mathcal{C} - \mathcal{C}\otimes\mathcal{B}\otimes\mathcal{A} + \mathcal{B}\otimes\mathcal{A}\otimes\mathcal{C}. \tag{27.13.14}$$

Inspection shows that if (13.12) through (13.14) are summed, all the terms on the right cancel in pairs to give the desired result

$$\{\mathcal{A},\{\mathcal{B},\mathcal{C}\}\} + \{\mathcal{B},\{\mathcal{C},\mathcal{A}\}\} + \{\mathcal{C},\{\mathcal{A},\mathcal{B}\}\} = 0. \tag{27.13.15}$$

Thus, $\mathcal{T}$ has been made into a Lie algebra with the Lie product taken to be the tensor product commutator (13.11).

We continue our exploration by defining adjoint operators in the standard way. Suppose $\mathcal{C}$ is some element in $\mathcal{T}$ and let $\mathcal{A}$ be any element in $\mathcal{T}$. Then we define the adjoint operator $\#\mathcal{C}\#$, which maps $\mathcal{T}$ into itself, by the rule

$$\#\mathcal{C}\#\mathcal{A} = \{\mathcal{C},\mathcal{A}\}. \tag{27.13.16}$$

We claim that $\#\mathcal{C}\#$ is a *derivation.* To see this, compute $\#\mathcal{C}\#(\mathcal{A}\otimes\mathcal{B})$ to find the result

$$\begin{aligned}
\#\mathcal{C}\#(\mathcal{A}\otimes\mathcal{B}) &= \{\mathcal{C},\mathcal{A}\otimes\mathcal{B}\} = \mathcal{C}\otimes\mathcal{A}\otimes\mathcal{B} - \mathcal{A}\otimes\mathcal{B}\otimes\mathcal{C} \\
&= \mathcal{C}\otimes\mathcal{A}\otimes\mathcal{B} - \mathcal{A}\otimes\mathcal{C}\otimes\mathcal{B} + \mathcal{A}\otimes\mathcal{C}\otimes\mathcal{B} \\
&- \mathcal{A}\otimes\mathcal{B}\otimes\mathcal{C} + \mathcal{A}\otimes\mathcal{C}\otimes\mathcal{B} - \mathcal{A}\otimes\mathcal{C}\otimes\mathcal{B} \\
&= \{\mathcal{C},\mathcal{A}\}\otimes\mathcal{B} + \mathcal{A}\otimes\{\mathcal{C},\mathcal{B}\} \\
&= (\#C\#\mathcal{A})\otimes\mathcal{B} + \mathcal{A}\otimes(\#C\#\mathcal{B}).
\end{aligned} \tag{27.13.17}$$

Note that (13.17) may be viewed as a rule that tells one how to compute the commutator of $\mathcal{C}$ with a product if one already knows how to compute the commutator of $\mathcal{C}$ with the individual elements that form the product. As a further example of such a rule, consider a commutator of the form $\{\mathcal{A}\otimes\mathcal{B},\mathcal{C}\otimes\mathcal{D}\}$, which is a commutator of two products. By using adjoint operator notation and the result (13.17) we find the relation

$$\begin{aligned}
\{\mathcal{A}\otimes\mathcal{B},\mathcal{C}\otimes\mathcal{D}\} &= \#\mathcal{A}\otimes\mathcal{B}\#(\mathcal{C}\otimes\mathcal{D}) \\
&= (\#\mathcal{A}\otimes\mathcal{B}\#\mathcal{C})\otimes\mathcal{D} + \mathcal{C}\otimes(\#\mathcal{A}\otimes\mathcal{B}\#\mathcal{D}) \\
&= \{\mathcal{A}\otimes\mathcal{B},\mathcal{C}\}\otimes\mathcal{D} + \mathcal{C}\otimes\{\mathcal{A}\otimes\mathcal{B},\mathcal{D}\} \\
&= -\{\mathcal{C},\mathcal{A}\otimes\mathcal{B}\}\otimes\mathcal{D} - \mathcal{C}\otimes\{\mathcal{D},\mathcal{A}\otimes\mathcal{B}\} \\
&= -[\#\mathcal{C}\#(\mathcal{A}\otimes\mathcal{B})]\otimes\mathcal{D} - \mathcal{C}\otimes[\#\mathcal{D}\#(\mathcal{A}\otimes\mathcal{B})] \\
&= -(\#\mathcal{C}\#\mathcal{A})\otimes\mathcal{B}\otimes\mathcal{D} - \mathcal{A}\otimes(\#\mathcal{C}\#\mathcal{B})\otimes\mathcal{D} \\
&-\mathcal{C}\otimes(\#\mathcal{D}\#\mathcal{A})\otimes\mathcal{B} - \mathcal{C}\otimes\mathcal{A}\otimes(\#\mathcal{D}\#\mathcal{B}).
\end{aligned} \tag{27.13.18}$$

We see that (13.18) gives a rule for finding the commutator of two products if we know how to compute the commutators of the constituents.

So far we have only made use of the vector-space structure of L. Let us now also employ its Lie product structure. With the results of the last paragraph still fresh in mind, we observe that the constitutents of any element in $\mathcal{T}$ are ultimately the B_α. Therefore, any commutator in $\mathcal{T}$ can ultimately be reduced to commutators among the B_α. At this point we would like to use the Lie product structure of L to stipulate these commutators by the rule

$$\{B_\alpha, B_\beta\} = B_\alpha \otimes B_\beta - B_\beta \otimes B_\alpha = [B_\alpha, B_\beta] = \sum_\gamma c^\gamma_{\alpha\beta} B_\gamma. \tag{27.13.19}$$

The motivation for this move is that $\mathcal{T}$, because it contains $L^1 = L$ as a subspace, would then also contain a copy of L as a *Lie subalgebra*. Evidently (13.19) (which may be viewed as a *reduction* rule that reduces higher-order tensor products resulting from commutators to lower-order tensor products) is compatible with the antisymmetry and Jacobi properties of the commutator because of (3.7.44) and (3.7.45). There is therefore some hope that it can be enforced *consistently*. [Here is an example question of consistency: We can enforce the condition (13.19) and then multiply or commute. Or, we may first multiply or commute and then enforce the condition. Do these two procedures give the same result?] But hope is not enough when proof is required.

The standard way to show that (13.19) can be invoked consistently is to construct the *enveloping* algebra. Let $O_{\alpha\beta}$ denote the element

$$O_{\alpha\beta} = \{B_\alpha, B_\beta\} - [B_\alpha, B_\beta]. \tag{27.13.20}$$

Next, let O denote the set of all linear combinations of the $O_{\alpha\beta}$,

$$O = \text{ set of all elements in } \mathcal{T} \text{ of the form } \sum_{\alpha\alpha'} b_\alpha b'_{\alpha'} O_{\alpha\alpha'}. \tag{27.13.21}$$

Finally, let $\mathcal{O}$ be the set of all elements formed by tensor multiplying all of O on both the left and right by *all* elements in $\mathcal{T}$ and forming all linear combinations of such elements,

$$\begin{aligned} \mathcal{O} = \; & \text{set of all linear combinations of elements in } \mathcal{T} \\ & \text{of the form } \mathcal{A} \otimes O \otimes \mathcal{B} \text{ for all } \mathcal{A}, \mathcal{B} \in \mathcal{T}. \end{aligned} \tag{27.13.22}$$

Evidently $\mathcal{O}$ is a linear vector space since, by construction, all linear combinations of elements in $\mathcal{O}$ are again in $\mathcal{O}$. The set of $\mathcal{O}$ is also an associative algebra that is *invariant* under tensor multiplication on either the left or right side by any element in $\mathcal{T}$,

$$\mathcal{A}' \otimes \mathcal{O} \in \mathcal{O}, \; \mathcal{O} \otimes \mathcal{B}' \in \mathcal{O} \text{ for all } \mathcal{A}', \mathcal{B}' \in \mathcal{T}. \tag{27.13.23}$$

That is so because all multiplications have also already occurred in the definition (13.22). For this reason, $\mathcal{O}$ is called the *two-sided ideal* generated by the $O_{\alpha\beta}$. (Recall that in Section 8.9 an ideal was defined in the Lie-algebraic context to be the set of all elements invariant under the Lie product. Here the concept is the same except that the product is tensor multiplication from the left and the right.)

Suppose we use the set $\mathcal{O}$ to set up an equivalence relation among all elements in $\mathcal{T}$. We will say that two elements $\mathcal{A}$ and $\mathcal{A}'$ in $\mathcal{T}$ are equivalent if their difference (recall that both $\mathcal{A}$ and $\mathcal{A}'$ are vectors, and therefore can be added and subtracted) is in the set $\mathcal{O}$,

$$\mathcal{A} \sim \mathcal{A}' \Leftrightarrow (\mathcal{A} - \mathcal{A}') \in \mathcal{O}. \tag{27.13.24}$$

It is easily verified, by a discussion analogous to that in Section 8.9, that (13.24) does indeed define an equivalence relation. Next, this equivalence relation can be used to set up equivalence classes. By the standard arguments, when this is done, any element in $\mathcal{O}$ will be in the equivalence class $\{0\}$ that contains the zero vector in $\mathcal{T}$,

$$\mathcal{B} \in \mathcal{O} \Leftrightarrow \{\mathcal{B}\} = \{0\}. \tag{27.13.25}$$

[For the analogous Lie-algebraic case, see (8.9.3).] Finally, let $\mathcal{E}$ be the quotient space of $\mathcal{T}$ with respect to $\mathcal{O}$,

$$\mathcal{E} = \mathcal{T}/\mathcal{O}. \tag{27.13.26}$$

Since $\mathcal{T}$ is an associative algebra and $\mathcal{O}$ is an ideal in $\mathcal{T}$, the quotient space $\mathcal{E}$ will also be an associative algebra. It is called the *enveloping algebra* of L. (It is also often called the *universal* enveloping algebra because it can be shown to be unique up to an isomorphism.)

We note that because of (13.25), all elements in $\mathcal{O}$, (that is, all elements in $\mathcal{T}$ that contain $O_{\alpha\beta}$) are automatically replaced by 0 in $\mathcal{E}$. This is equivalent to enforcing the condition (13.19) in $\mathcal{E}$. And, because $\mathcal{E}$ is an associative algebra, we have verified that this condition can be enforced consistently.

Soon we will use the enveloping algebra to construct Casimir operators. To do so, it is useful to first explore further the property of adjoint operators $\#\mathcal{C}\#$. Since $\#\mathcal{C}\#$ is a derivation, recall (13.17), the operator $\exp(\epsilon\#\mathcal{C}\#)$ is an isomorphism for tensor multiplication,

$$[\exp(\epsilon\#\mathcal{C}\#)](\mathcal{A} \otimes \mathcal{B}) = \{[\exp(\epsilon\#\mathcal{C}\#)]\mathcal{A}\} \otimes \{[\exp(\epsilon\#\mathcal{C}\#)]\mathcal{B}\}. \tag{27.13.27}$$

Indeed, if $\mathcal{F}$ is any element of $\mathcal{T}$ composed of tensor products of the B_α, we have the result

$$\begin{aligned}&[\exp(\epsilon\#\mathcal{C}\#)]\mathcal{F}(B_{\alpha_1}, B_{\alpha_2}, B_{\alpha_3}\cdots) = \\ &\mathcal{F}([\exp(\epsilon\#\mathcal{C}\#)]\mathcal{B}_{\alpha_1}, [\exp(\epsilon\#\mathcal{C}\#)]\mathcal{B}_{\alpha_2}, [\exp(\epsilon\#\mathcal{C}\#)]\mathcal{B}_{\alpha_3}\cdots).\end{aligned} \tag{27.13.28}$$

See Section 8.2 for the standard arguments justifying this result.

Now suppose $\mathcal{C} = C$ where C is some element in L as in (11.20). Then we have the results

$$\#C\#B_\alpha = \{C, B_\alpha\} = [C, B_\alpha] =: C : B_\alpha. \tag{27.13.29}$$

Here we have used (13.19). It follows from (11.20), (11.28), and (13.29) that in this case there is the relation

$$[\exp(\epsilon\#\mathcal{C}\#)]B_\alpha = \exp(\epsilon : C :)B_\alpha = U_{\alpha\beta}B_\beta, \tag{27.13.30}$$

where we have again used the summation convention. Correspondingly, in this case (13.28) can be rewritten in the form

$$[\exp(\epsilon\#\mathcal{C}\#)]\mathcal{F}(B_{\alpha_1}, B_{\alpha_2}, B_{\alpha_3}\cdots) = \mathcal{F}(U_{\alpha_1\beta_1}B_{\beta_1}, U_{\alpha_2\beta_2}B_{\beta_2}, U_{\alpha_3\beta_3}B_{\beta_3}\cdots). \tag{27.13.31}$$

We are now ready to discuss Casimir operators. In analogy with (11.35) we now define the quadratic Casimir operator $\mathcal{C}_2$ to be the quantity

$$\mathcal{C}_2 = \sum_{\alpha\alpha'} g_R^{\alpha\alpha'} B_\alpha \otimes B_{\alpha'}. \tag{27.13.32}$$

Note that in this context $\mathcal{C}_2$ is not first of all an "operator", but rather is an element in the enveloping algebra $\mathcal{E}$. In analogy with (11.38), let us see how it transforms. We find the result

$$\begin{aligned}\mathcal{C}_2^{\rm tr} &= \exp(\epsilon\#\mathcal{C}\#)\mathcal{C}_2 = \sum_{\alpha\alpha'} g_R^{\alpha\alpha'}(U_{\alpha\beta}B_\beta)\otimes(U_{\alpha'\beta'}B_{\beta'}) \\ &= \sum_{\alpha\alpha'} g_R^{\alpha\alpha'}U_{\alpha\beta}U_{\alpha'\beta'}B_\beta\otimes B_{\beta'} \\ &= \sum_{\beta\beta'}(\sum_{\alpha\alpha'} g_R^{\alpha\alpha'}U_{\alpha\beta}U_{\alpha'\beta'})B_\beta\otimes B_{\beta'} \\ &= \sum_{\beta\beta'} g_R^{\beta\beta'}B_\beta\otimes B_{\beta'} = \mathcal{C}_2. \end{aligned} \tag{27.13.33}$$

Here we have used (13.31) and (11.39). We see that $\mathcal{C}_2$ is again invariant. Also, the infinitesimal version of (13.33) with $\mathcal{C} = C = B_{\alpha''}$ gives the result

$$\#B_{\alpha''}\#\mathcal{C}_2 = \{B_{\alpha''}, \mathcal{C}_2\} = 0. \tag{27.13.34}$$

We see that $\mathcal{C}_2$ commutes with all the elements in L. Moreover, since everything in the enveloping algebra is constructed from elements in L, it follows that $\mathcal{C}_2$ commutes with all the elements of the enveloping algebra,

$$\{\mathcal{C}_2, \mathcal{E}\} = 0. \tag{27.13.35}$$

At this point we pause to note that we might use (13.19) to rearrange [in analogy to (11.46) and (11.47)] the terms in $\mathcal{C}_2$ as given by (13.32) to get an expression for $\mathcal{C}_2$ analogous to (11.50). What would happen if we then compute $\{B_{\alpha''}, \mathcal{C}_2\}$ using the rearranged $\mathcal{C}_2$? According to our previous discussion about consistency, $\mathcal{O}$ is an ideal thereby guaranteeing that the quotient space $\mathcal{T}/\mathcal{O} = \mathcal{E}$ is an associative algebra. Therefore the result should be (and is indeed) the same.

The construction of higher-order Casimir operators proceeds in a similar fashion. For example, the analog of (11.86) is

$$\mathcal{C}_3 =_3 g_R^{\alpha\alpha'\alpha''}B_\alpha\otimes B_{\alpha'}\otimes B_{\alpha''}. \tag{27.13.36}$$

Again, it is an element in the enveloping algebra. It too commutes with all the elements in L, and therefore also with all of $\mathcal{E}$. In general, in the context of the present discussion, we may define a Casimir operator to be any element in the enveloping algebra that commutes with *all* the elements in the enveloping algebra. Put another way, the Casimir operators form the *center* of the enveloping algebra.

Exercises

27.13.1. From (13.4) it follows that $\mathcal{T}$, the tensor algebra of L, can be expressed in the form

$$\mathcal{T} = \mathcal{T}^0 \oplus \mathcal{T}^1 \oplus \mathcal{T}^2 \oplus \mathcal{T}^3 \oplus \cdots, \tag{27.13.37}$$

where

$$\mathcal{T}^n = L^n. \tag{27.13.38}$$

Suppose that L has dimension k. Show that each subspace $\mathcal{T}^n$ then has dimension

$$\dim \mathcal{T}^n = k^n. \tag{27.13.39}$$

The algebra $\mathcal{E}$, the enveloping algebra of L, can also be decomposed in the form

$$\mathcal{E} = \mathcal{E}^0 \oplus \mathcal{E}^1 \oplus \mathcal{E}^2 \oplus \mathcal{E}^3 \oplus \cdots , \tag{27.13.40}$$

where each subspace $\mathcal{E}^n$ of "degree" n is spanned by the tensor products of n basis elements $(B_{\alpha_1} \otimes B_{\alpha_2} \otimes B_{\alpha_3} \otimes \cdots B_{\alpha_n})$. However, in the case of $\mathcal{E}$, the relation (13.19) can be used to rearrange the basis elements in each $\mathcal{E}^n$ so that the subscripts have a definite standard ordering. For example, we may arrange them in ascending order,

$$\begin{aligned}&\text{rearranged } (B_{\alpha_1} \otimes B_{\alpha_2} \otimes B_{\alpha_3} \otimes \cdots B_{\alpha_n}) = (B_{\beta_1} \otimes B_{\beta_2} \otimes B_{\beta_3} \otimes \cdots B_{\beta_n})\\ &\text{with } \beta_1 \leq \beta_2 \leq \beta_3 \leq \cdots \beta_n.\end{aligned} \tag{27.13.41}$$

In the rearrangement process various terms of lower degree may be generated, but they simply feed down to $\mathcal{E}^{n-1}$, etc. It follows that the various terms in some standard ordering, say that shown on the right side of (13.41), *span* $\mathcal{E}^n$. It can be shown that they are linearly independent as well, and therefore form a *basis* for $\mathcal{E}^n$. This basis is called the *Poincaré-Birkhoff-Witt* basis. Show that the dimension of $\mathcal{E}^n$ is given by the relation

$$\dim \mathcal{E}^n = N(n, k) \tag{27.13.42}$$

with $N(n, k)$ given by (7.3.40). Hint: Once a standard ordering has been established, the counting of basis elements is the same as counting monomials.

27.13.2. Suppose R is a realization of some Lie algebra L. Thus, if the B_α form a basis of L, there are associated matrices $\hat{B}_\alpha$ in the realization R. Let $\mathcal{R}$ be the *linear* map that sends the B_α to the $\hat{B}_\alpha$,

$$\mathcal{R}(B_\alpha) = \hat{B}_\alpha. \tag{27.13.43}$$

(Note that since both L and the set of $m \times m$ matrices are vector spaces, it makes sense to talk about a linear map that sends one into the other.) Then, by the definition of a realization, we have the relation

$$\begin{aligned}\mathcal{R}([B_\alpha, B_\beta]) &= \mathcal{R}(\sum_\gamma c^\gamma_{\alpha\beta} B_\gamma) = \sum_\gamma c^\gamma_{\alpha\beta} \hat{B}_\gamma = \{\hat{B}_\alpha, \hat{B}_\beta\}\\ &= \{\mathcal{R}(B_\alpha), \mathcal{R}(B_\beta)\}.\end{aligned} \tag{27.13.44}$$

Next, let us extend the definition of $\mathcal{R}$ to have it act on any basis element in the tensor algebra $\mathcal{T}$ by the rule

$$\begin{aligned}\mathcal{R}(B_{\alpha_1} \otimes B_{\alpha_2} \otimes B_{\alpha_3} \cdots) &= \mathcal{R}(B_{\alpha_1})\mathcal{R}(B_{\alpha_2})\mathcal{R}(B_{\alpha_3}) \cdots\\ &= \hat{B}_{\alpha_1}\hat{B}_{\alpha_2}\hat{B}_{\alpha_3} \cdots .\end{aligned} \tag{27.13.45}$$

Show that $\mathcal{R}$ sends any element of the two-sided ideal $\mathcal{O}$ to the zero matrix,

$$\mathcal{R}(O_{\alpha\beta}) = 0,\ \mathcal{R}(O) = 0,\ \mathcal{R}(\mathcal{O}) = 0. \tag{27.13.46}$$

Thus, the image of any element in $\mathcal{T}$ under the action of $\mathcal{R}$ depends only on the equivalence class to which the element belongs, and we may equally well view $\mathcal{R}$ as acting on $\mathcal{T}/\mathcal{O} = \mathcal{E}$. Show that $\mathcal{R}$ sends the Casimir operator (13.32) defined in the enveloping algebra context to the Casimir operator (11.35) defined in the representation context. Show that in general anything that it discovered about Casimir operators in the enveloping algebra context is immediately transferable to the realization context, and vice versa.

27.14 The Symplectic Lie Algebras $sp(8)$ and Beyond

The previous sections in this chapter have treated the cases of $sp(2)$, $sp(4)$, and $sp(6)$. The Lie algebraic structure of all the $sp(2n)$, for example root vectors and fundamental weight vectors, is also known. In particular, for $sp(2n)$, a representation is characterized by n non-negative integers $k_1, k_2, \cdots k_n$ and may be denoted by the symbols $\Gamma(k_1, k_2, \cdots k_n)$. Homogeneous polynomials of degree ℓ in the $2n$ components of z again carry representations of $sp(2n)$, and for these representations there is the result

$$k_1 = \ell, \tag{27.14.1}$$

$$k_j = 0 \text{ for } j = 2, 3, \cdots n. \tag{27.14.2}$$

There is also an analogous Clebsch-Gordon series result of the form (9.1) where all entries in $\Gamma(k_1, k_2, \cdots k_n)$ are zero save for the first two,

$$\begin{aligned}&\Gamma(\ell, 0, 0, \cdots) \otimes \Gamma(1, 0, 0, \cdots)\\&= \Gamma(\ell + 1, 0, 0, \cdots) \oplus \Gamma(\ell - 1, 1, 0, \cdots)\\&\ \oplus \Gamma(\ell - 1, 0, 0, \cdots).\end{aligned} \tag{27.14.3}$$

Thus, the symplectic classification of all analytic vector fields in any (even) dimension is in principle known. Moreover, any $\mathcal{L}_{\boldsymbol{g}^\ell}$ with $\ell \geq 1$ has the unique decomposition

$$\mathcal{L}_{\boldsymbol{g}^\ell} = \mathcal{H}^{\ell+1,0,0,\cdots} + \mathcal{G}^{\ell-1,1,0,\cdots} + \mathcal{G}^{\ell-1,0,0,\cdots}. \tag{27.14.4}$$

Here $\mathcal{H}^{\ell+1,0,0,\cdots}$ is a Hamiltonian vector field that carries the representation $\Gamma(\ell+1, 0, 0, \cdots)$, and is of the form $:h_{\ell+1}:$. The quantities $\mathcal{G}^{\ell-1,1,0,\cdots}$ and $\mathcal{G}^{\ell-1,0,0,\cdots}$ are non-Hamiltonian vector fields that carry the representations $\Gamma(\ell - 1, 1, 0, \cdots)$ and $\Gamma(\ell - 1, 0, 0, \cdots)$, respectively.

Exercises

27.14.1. Show that any $2n \times 2n$ matrix that commutes with all $sp(2n)$ matrices (in the fundamental representation) must be a multiple of I. Show that any $2n \times 2n$ matrix that commutes with all $Sp(2n)$ matrices (in the fundamental representation) must be a multiple of I.

27.15 Momentum Maps and Casimirs

The previous sections, among other things, have shown how to decompose an analytic vector field $\mathcal{L}_{\boldsymbol{g}}$ into Hamiltonian and non-Hamiltonian parts. Here we address a somewhat more restricted question. Suppose we are given a vector field $\mathcal{L}_{\boldsymbol{g}}$, and also know that it came from some Hamiltonian h so that there is in principle the relation

$$\mathcal{L}_{\boldsymbol{g}} =: h : . \tag{27.15.1}$$

That is, $\mathcal{L}_{\boldsymbol{g}}$ is a Hamiltonian vector field. We then say that there is a *momentum map* μ that sends $\mathcal{L}_{\boldsymbol{g}}$ to h,

$$\mu(\mathcal{L}_{\boldsymbol{g}}) = h. \tag{27.15.2}$$

Here the name *momentum* is associated with the fact that in some simple examples the resulting h is some kind of momentum such as linear or angular momentum.

In this section we will develop/review what is required for $\mathcal{L}_{\boldsymbol{g}}$ to be Hamiltonian, and then see how to determine h in terms of $\boldsymbol{g}$. We will also see how momentum maps are related to integrals of motion and, when there are several integrals of motion, how to construct from them integrals of motion that are in involution.

27.15.1 Momentum Maps and Conservation Laws

Why might one be interested in momentum maps? Given a vector field $\mathcal{L}_{\boldsymbol{g}}$, we may define a family of maps $\mathcal{M}(\tau)$, not to be confused with momentum maps, by the rule

$$\mathcal{M}(\tau) = \exp(\tau\mathcal{L}_{\boldsymbol{g}}). \tag{27.15.3}$$

[For a discussion of some of the properties of general Lie operators (general vector fields) and their associated Lie transformations, see Exercises 5.3.10 and 5.4.14.] The maps $\mathcal{M}(\tau)$ send phase space into itself according to the relation

$$\bar{z}(\tau) = \mathcal{M}(\tau)z, \tag{27.15.4}$$

and they evidently form a one-parameter group. Now suppose the motion of some system is governed by some Hamiltonian $H(z,t)$ and suppose that this Hamiltonian is invariant under the action of $\mathcal{M}(\tau)$,

$$\mathcal{M}(\tau)H(z,t) = H(z,t). \tag{27.15.5}$$

From this invariance/symmetry relation and (15.3) we conclude that

$$\mathcal{L}_{\boldsymbol{g}}H(z,t) = 0. \tag{27.15.6}$$

But, if (15.1) holds, (15.6) can be rewritten as

$$\mathcal{L}_{\boldsymbol{g}}H(z,t) =: h : H = [h,H] = 0, \tag{27.15.7}$$

and we see that h is an integral of motion. Thus, the existence of symmetry and a momentum map implies the existence of an integral of motion (a *conserved quantity* or *conservation law*) and vice versa.[6]

[6] Observe that the assumed symmetry described by (15.5) is a *continuous* symmetry. The condition (15.5) is supposed to hold over a continuous range of τ. We also remark that *Emmy Noether* (1882-1935) was the first to explore in detail the connection between continuous symmetries and conservation laws.

How can one test a vector field to see if it is Hamiltonian and therefore there is a momentum map? We have already seen that the maps $\mathcal{M}(\tau)$ form a one-parameter group. Suppose we now require that these maps be symplectic for all τ. That is, we require that

$$[\bar{z}_a(\tau), \bar{z}_b(\tau)] = J_{ab} \text{ for all } \tau. \tag{27.15.8}$$

Put another way, we require that each transformation $\mathcal{M}(\tau)$ *preserve* the symplectic structure of phase space. For small τ we have the result

$$\mathcal{M}(\tau) = \mathcal{I} + \tau \mathcal{L}_{\boldsymbol{g}} + O(\tau^2). \tag{27.15.9}$$

It follows that

$$\bar{z}(\tau)_a = \mathcal{M}(\tau) z_a = z_a + \tau \mathcal{L}_{\boldsymbol{g}} z_a + O(\tau)^2 = z_a + \tau g_a + O(\tau)^2. \tag{27.15.10}$$

Here we have used the result

$$\mathcal{L}_{\boldsymbol{g}} z_a = g_a. \tag{27.15.11}$$

Upon employing (15.10) we find the result

$$\begin{aligned}[\bar{z}_a(\tau), \bar{z}_b(\tau)] &= [z_a, z_b] + \tau\{[z_a, g_b] + [g_a, z_b]\} + O(\tau^2) \\ &= J_{ab} + \tau\{[z_a, g_b] + [g_a, z_b]\} + O(\tau^2).\end{aligned} \tag{27.15.12}$$

Enforcing (15.8) and equating powers of τ give the result

$$[z_a, g_b] + [g_a, z_b] = 0. \tag{27.15.13}$$

We have seen this condition before in Lemma 6.2 of Section 7.6. There we learned that (15.13) implies and is implied by the relation

$$g_a =: h : z_a, \tag{27.15.14}$$

which, in view of (15.11), is equivalent to the relation (15.1). Thus, (15.13) is a necessary and sufficient condition for $\mathcal{L}_{\boldsymbol{g}}$ to be a Hamiltonian vector field.

We also learned how to construct h. It is given, up to an additive constant, by the relation

$$h(z) = -\int_P^z \sum_{cd} g_c(z') J_{cd}\, dz'_d \tag{27.15.15}$$

where P is *any* path ending at the point z. A convenient path is that which connects the origin and z by a straight line,

$$z'(\lambda) = \lambda z \text{ with } \lambda \in [0, 1]. \tag{27.15.16}$$

For this path we find the explicit result

$$h(z) = -\sum_{cd} z_d J_{cd} \int_0^1 d\lambda g_c(\lambda z). \tag{27.15.17}$$

This relation specifies h in terms of the g_c, and hence in terms of $\mathcal{L}_{\boldsymbol{g}}$. It therefore provides the map μ described in (15.2). Note that this specification is equivalent to the *definition* (10.20) of the Hamiltonian part of a general homogeneous vector field.

At this point we remark that in some circumstances one initially has a transformation or a group of transformations whose action is only on the position coordinates q. Such transformations can always be extended to symplectic actions on phase space, and thus in this case the existence of Hamiltonian vector fields is guaranteed. Recall Exercise 6.5.2.

As a first example of the momentum map process, consider (for a 6-dimensional phase space) the case where

$$\begin{aligned} &g_1(z) = 1, \\ &g_a(z) = 0,\ a \neq 1, \end{aligned} \tag{27.15.18}$$

so that

$$\mathcal{L}_{\boldsymbol{g}} = \partial/\partial z_1 = \partial/\partial q_1. \tag{27.15.19}$$

It is easily verified that

$$(\mathcal{L}_{\boldsymbol{g}})^n z = 0 \text{ for } n \geq 2, \tag{27.15.20}$$

from which it follows that

$$\bar{z}_a(\tau) = \exp(\tau\mathcal{L}_{\boldsymbol{g}})z_a = z_a + \tau\delta_{a,1}. \tag{27.15.21}$$

That is, $\mathcal{L}_{\boldsymbol{g}}$ generates translations in phase space along the $z_1 = q_1$ axis. Evidently, the g_a specified by (15.18) satisfy (15.13) so that $\mathcal{L}_{\boldsymbol{g}}$ is a Hamiltonian vector field. Finally, the formula (15.17) for h is easily evaluated to give the result

$$h(z) = -z_4 = -p_1, \tag{27.15.22}$$

the negative of the first component of the linear momentum. Thus, invariance under translation implies the conservation of linear momentum, and vice versa.

As a second example, suppose that

$$\begin{aligned} \mathcal{L}_{\boldsymbol{g}} &= z_1\partial/\partial z_2 - z_2\partial/\partial z_1 + z_4\partial/\partial z_5 - z_5\partial/\partial z_4 \\ &= (q_1\partial/\partial q_2 - q_2\partial/\partial q_1) + (p_1\partial/\partial p_2 - p_2\partial/\partial p_1). \end{aligned} \tag{27.15.23}$$

For this example the nonzero g_a are given by the relations

$$\begin{aligned} g_1(z) &= -z_2, \\ g_2(z) &= z_1, \\ g_4(z) &= -z_5, \\ g_5(z) &= z_4. \end{aligned} \tag{27.15.24}$$

With the $g_a(z)$ in view, it is easily checked that (15.13) holds. It therefore makes sense to continue on to compute h. The integrals appearing on the right side of (15.17) are easy to evaluate because the $h_a(z)$ are homogeneous of degree one. Doing so gives the results

$$\int_0^1 d\lambda g_c(\lambda z) = \int_0^1 d\lambda\lambda g_c(z) = (1/2)g_c(z). \tag{27.15.25}$$

Finally, employing (15.25) in (15.17) gives the result

$$h(z) = -(z_1z_5 - z_2z_4) = -(q_1p_2 - q_2p_1). \tag{27.15.26}$$

From the second line of (15.23) we recognize $\mathcal{L}_{\boldsymbol{g}}$ as the generator of simultaneous rotations in the q_1, q_2 and p_1, p_2 planes, and from (15.26) we see that h is the negative of the third component of the angular momentum. Thus, invariance under rotation implies the conservation of angular momentum, and vice versa.

27.15.2 Use of Casimirs

In general, if a Hamiltonian is invariant under the action of some n-dimensional group that preserves symplectic structure, there will be n associated integrals of motion. These integrals, however, need not be in mutual involution. Think, for example, of the components of angular momentum. Their Poisson bracket Lie algebra provides a realization of $su(2)$ [or, equivalently, $so(3,\mathbb{R})$], and they are therefore not in involution. Generally integrals will be in involution if, and only if, the corresponding Lie operators commute. Recall (5.3.14). In this subsection we will explore briefly how Casimirs can sometimes be used to construct integrals that are in involution.

Suppose that a Hamiltonian H is indeed invariant under the action of some n-dimensional group that preserves symplectic structure, and therefore there are n associated integrals of motion. Call these integrals h^α. Let C be any function of these integrals,

$$C = C(h^1, h^2, \cdots, h^n). \tag{27.15.27}$$

Then we know form Exercise 5.2.4 that C will also be an integral of motion. Our goal will be to construct a C such that it is functionally independent of any one of the h^α, but is also in involution with any of them.

We know that the h^α will form a Lie algebra with the Poisson bracket serving as a Lie product. That is, there will be relations of the form

$$[h^\alpha, h^\beta] = \sum_\gamma c^\gamma_{\alpha\beta} h^\gamma. \tag{27.15.28}$$

(Note that these relations are consistent with Poisson's theorem that states that the Poisson bracket of two integrals of motion is again an integral of motion. See Exercise 5.2.3.) The structure constants $c^\gamma_{\alpha\beta}$ can be used to construct a Killing metric tensor $g^K_{\alpha\alpha'}$, and from $g^K_{\alpha\alpha'}$, assuming it is invertible, we can construct $g_K^{\alpha\alpha'}$. With $g_K^{\alpha\alpha'}$ in hand, we can define the function C_2 by the rule

$$C_2 = \sum_{\alpha\alpha'} g_K^{\alpha\alpha'} h^\alpha h^{\alpha'}. \tag{27.15.29}$$

In analogy to the calculations carried out for the quadratic Casimir operator $\mathcal{C}_2$ in Section 21.11.1, it is easily verified that C_2 is in involution with the h^α. It can happen that C_2 vanishes identically. See Exercise 21.11.7. But, if C_2 does not vanish, we have found two functionally independent integrals in involution, namely C_2 and any one of the h^α.

At this point we might go on to find additional integrals C_3, etc. constructed in analogy to the higher-order Casimir operators. They will also be in involution and, if these integrals are nonvanishing and functionally independent, we will have found additional nontrivial integrals. In general, assuming the Lie algebra in question is simple, we may hope to find as many integrals in involution as there are labels necessary to specify a representation for the Lie algebra and to specify vectors within a representation.

Let us apply this construction to the rotation group example of the previous subsection. Suppose that H has the three integrals

$$h^1 = q_2p_3 - q_3p_2, \tag{27.15.30}$$

$$h^2 = q_3p_1 - q_1p_3, \tag{27.15.31}$$

$$h^3 = q_1p_2 - q_2p_1. \tag{27.15.32}$$

They form an $su(2)$ Lie algebra,

$$[h^1, h^2] = h^3, \text{ etc.} \tag{27.15.33}$$

The metric tensor for $su(2)$ is given in Exercise 21.12.2. It follows, after a convenient renormalization, that we may take for C_2 the quantity

$$C_2 = (h^1)^2 + (h^2)^2 + (h^3)^2, \tag{27.15.34}$$

which is the square of the angular momentum.

Exercises

27.15.1. Verify that (15.5) implies (15.6), and conversely.

27.15.2. Suppose (15.13) holds so that $\mathcal{L}_{\boldsymbol{g}}$ is a Hamiltonian vector field. Suppose also that $\mathcal{L}_{\boldsymbol{g}}$ can be decomposed into a sum of homogeneous parts as in (3.1). (This will certainly be possible if $\mathcal{L}_{\boldsymbol{g}}$ is analytic.) Show that then (10.20) and (15.17) are equivalent.

27.15.3. Verify that use of (15.18) in (15.17) does yield (15.22). Verify that the g_a given by (15.18) and the h given by (15.22) do indeed satisfy (15.1).

27.15.4. Verify, by evaluating the effect of $\exp(\tau\mathcal{L}_{\boldsymbol{g}})$ on phase space, that $\mathcal{L}_{\boldsymbol{g}}$ as given by (15.23) does indeed generate simultaneous rotations in the q_1, q_2 and p_1, p_2 planes. Verify (15.24) through (15.26). Verify that the g_a given by (15.24) and the h given by (15.26) do indeed satisfy (15.1).

27.15.5. Verify the relations (15.33). Verify that C_2 and any one of the h^α are in involution.

Bibliography

Group and Lie Algebra Theory

See also the Lie Group Theory sections of the Bibliographies for Chapters 3 and 5.

[1] A.O. Barut and R. Raczka, *Theory of Group Representations and Applications*, World Scientific (1986).

[2] N. Bourbaki, *Lie Groups and Lie Algebras, Elements of Mathematics, Chapters 1-3*, Springer-Verlag (1989).

[3] T. Brocker and T.T. Dieck, *Representations of Compact Lie Groups*, Springer-Verlag (1985).

[4] R. Cahn, *Semi-Simple Lie Algebras and their Representations*, Dover (2006).

[5] J-Q. Chen, *Group Representation Theory for Physicists*, World Scientific (1989).

[6] A.R. Edmonds, *Angular Momentum in Quantum Mechanics*, Princeton University Press(1957).

[7] M.E. Rose, *Elementary Theory of Angular Momentum*, John Wiley and Sons (1957).

[8] L. Biedenharn and J. Louck, *Angular Momentum in Quantum Physics: Theory and Application*, Cambridge University Press (2009).

[9] W. Fulton and J. Harris, *Representation Theory, A First Course*, Corrected third printing, Springer-Verlag (1996).

[10] H. Georgi, *Lie Algebras in Particle Physics*, Perseus Books (1999).

[11] R. Goodman and N.R. Wallach, *Representations and Invariants of the Classical Groups*, Cambridge University Press (1998).

[12] R. Goodman and N.R. Wallach, *Symmetry, Representations, and Invariants*, Springer (2009).

[13] J.E. Humphreys, *Introduction to Lie Algebras and Representation Theory*, Springer-Verlag (1972).

[14] J.E. Humphreys, *Representations of Semisimple Lie Algebras in the BBG Category $\mathcal{O}$*, American Mathematical Society (2008).

[15] N. Jacobson, *Lie Algebras*, Interscience Publishers (1962).

[16] A.W. Knapp, *Lie Groups Beyond an Introduction*, Second Edition, Birkhäuser (2005).

[17] A.W. Knapp, *Representation Theory of Semisimple Groups, An Overview Based on Examples*, Princeton (1986).

[18] D.H. Sattinger and O.L. Weaver, *Lie Groups and Algebras with Applications to Physics, Geometry, and Mechanics*, Springer-Verlag (1986).

[19] V.S. Varadarajan, *Lie Groups, Lie Algebras, and Their Representations*, Springer-Verlag (1984).

[20] J.E. Campbell, *Introductory Treatise on Lie's Theory of Finite Continuous Transformation Groups*, Chelsea Publishing (1903 and 1966).

[21] H. Weyl, *The Classical Groups: Their Invariants and Representations*, Princeton University Press (1946).

[22] B.G. Wybourne, *Classical Groups for Physicists*, John Wiley and Sons (1974).

[23] W. Rossmann, *Lie Groups, An Introduction Through Linear Groups*, Oxford (2002).

[24] A. Baker, *Matrix Groups: An Introduction to Lie Group Theory*, Springer (2006).

[25] J.G.F. Belinfante and B. Kolman, *A Survey of Lie Groups and Lie Algebras with Applications and Computational Methods*, Society for Industrial and Applied Mathematics (1972).

[26] P. Di Francesco, P. Mathieu, and D. Sénéchal, *Conformal Field Theory*, Springer (1997). See Chapter 13.

[27] D. Montgomery and L. Zippin, *Topological Transformation Groups*, Interscience (1955).

[28] P. Tondeur, *Introduction to Lie Groups and Transformation Groups*, Lecture Notes in Mathematics **7**, Springer-Verlag (1965).

[29] P. Cvitanović, *Group Theory, Birdtracks, Lie's, and Exceptional Groups*, Princeton (2008).

[30] J.-P. Serre, *Lie Algebras and Lie Groups: 1964 Lectures Given at Harvard University*, Springer (2005).

[31] J.-P. Serre and G. Jones, *Complex Semisimple Lie Algebras*, Springer (2001).

[32] J. Dieudonné, *Special Functions and Linear Representations of Lie Groups*, American Mathematical Society (1980).

[33] J. Dieudonné, *Treatise on Analysis*, Volumes 10-IV and 10-V in the series Pure and Applied Mathematics, Academic Press (1974 and 1977).

[34] Zhong-Qi Ma, *Group Theory for Physicists*, World Scientific (2007).

[35] F. Iachello, *Lie Algebras and Applications*, 2nd edition, Springer (2015).

[36] A. Zee, *Group Theory in a Nutshell for Physicists*, Princeton University Press (2016).

[37] B. Hall, *Lie Groups, Lie Algebras, and Representations: an Elementary Introduction*, 2nd edition, Springer (2015).

[38] A. Baker, *Matrix Groups: An Introduction to Lie Group Theory*, Springer (2002).

[39] K. Erdmann and M. Wildon, *Inroduction to Lie Algebras*, Springer (2011).

[40] H. Pollatsek, *Lie Groups: A Problem-Oriented Introduction via Matrix Groups*, Mathematical Association of America (2009).

[41] Robert Hermann has published several books on various applications of Lie theory as well as others that provide commentary on various papers of Lie and related work of other mathematicians of his time period. For a list of some of them, see the Wikipedia Web site http://en.wikipedia.org/wiki/Robert_Hermann_(mathematician).

Overview and History of the Theory of Lie Algebras and Lie Groups

[42] R. Howe, “A Century of Lie Theory,” *American Mathematical Society Centennial Proceedings, Vol. 2: Mathematics into the Twenty-first Century*, Providence, R.I., American Mathematical Society, pp. 101-320 (1992).

[43] T. Hawkins, *Emergence of the Theory of Lie Groups: An Essay in the History of Mathematics 1869-1926*, Springer (2000).

[44] A. Borel, *Essays in the History of Lie Groups and Algebraic Groups*, American Mathematical Society (2001).

[45] A. Stubhaug, *The Mathematician Sophus Lie*, Springer-Verlag (2002).

[46] W. Schmid, “Poincaré and Lie Groups”, *Bulletin Amer. Math. Soc.* (N.S.) **6** pp. 175-186 (1982). This article can also be found in *The Mathematical Heritage of Henri Poincaré, Proceedings of Symposia in Pure Mathematics of the American Mathematical Society* **39**, Parts 1 and 2, F. Browder, Edit., American Mathematical Society (1983).

Decomposing Vector Fields

[47] D. Lewis and J.E. Marsden, “The Hamiltonian-Dissipative Decomposition of Normal Forms of Vector Fields”, *Proc. of the Conf. on Bifurcation Theory and its Num. An.*, pp. 51-78, Xi’an Jaitong Univ. Press (1989).

Enveloping Algebras and Casimir Operators

See also the references above on Group and Lie Algebra Theory.

[48] J. Dixmier, *Enveloping Algebras*, North-Holland Publishing (1977).

[49] R. Campoamor-Stursberg, "A new matrix method for the Casimir operators of the Lie algebras $wsp(N, R)$ and $Isp(2N, R)$", *J. Phys. A: Math. Gen.* **38**, p. 4187 (2005).

[50] K. Maurischat, "Casimir operators for symplectic groups", arXiv:1011.4777v1 [math.NT] (2010).

Combinatorics of $Sp(2n, \mathbb{C})$ Representations

[51] Sheila Sundaram, *On the Combinatorics of Representations of* $Sp(2n, C)$, Ph. D. thesis, Massachusetts Institute of Technology (1986).

Computer Programs

The programs listed below cover all the simple Lie groups, and are extremely useful. They were employed, for examples, to check Table 8.1, to compute weights and multiplicities for various representations, and to produce Clebsch-Gordan series such as (9.1).

[52] W.G. McKay, J. Patera, and D.W. Rand, *SimpLie User's Manual*, Macintosh Software for Representations of Simple Lie Algebras, Centre de recherches mathématiques, Université de Montréal, C.P. 6128-A, Montréal, QC H3C 3J7, Canada (1990).

[53] M.A.A. van Leeuwen, A.M. Cohen, and B. Lisser, *Lie Manual describing LIE Version 2.1.* LIE is a software package for Lie group theoretical computations developed by the Computer Algebra Group of CWI (Centrum voor Wiskunde en Informatica, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands). For further information see the Web site http://www-math.univ-poitiers.fr/~maavl/LiE/.

[54] The Computational Algebra Group within the University of Sydney School of Mathematics and Statistics has produced the program CAYLEY and its replacement MAGMA. They provide a mathematically rigorous environment for computing with algebraic, number-theoretic, combinatoric, and geometric objects. For further information, see the Web site http://magma.maths.usyd.edu.au/magma/

[55] There is a program called *Schur* for calculating properties of Lie groups and symmetric functions. For further information, see the Web site http://schur.sourceforge.net

Momentum Maps (sometimes called Moment Maps)

[56] V. Guillemin and S. Sternberg, *Symplectic Techniques in Physics*, Cambridge (1984).

[57] J.-M. Souriau, *Structure of Dynamical Systems: a Symplectic View of Physics*, Birkhäuser (1997).

[58] J.E. Marsden amd T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Second Edition, Springer (1999).

[59] D. D. Holm, *Geometric Mechanics, Part I: Dynamics and Symmetry*, Imperial College Press, World Scientific (2008).

[60] D. D. Holm, T. Schmah, C. Stoica, and D. C. P. Ellis, *Geometric Mechanics, from Finite to Infinite Dimensions*, Oxford (2009).

[61] A. Cannas da Silva, *Lectures on Symplectic Geometry*, Corrected second printing, Springer-Verlag (2008).

[62] D. Neuenschwander, *Emmy Noether's Wonderful Theorem*, Johns Hopkins University Press (2011).

[63] Y. Kosmann-Schwarzbach, *The Noether Theorems, Invariance and Conservation Laws in the Twentieth Century*, B. Schwarzbach, Translator, Springer (2011).

[64] P. J. Olver, Review of Y. Kosmann-Schwarzbach's Springer (2011) book "The Noether Theorems, Invariance and Conservation Laws in the Twentieth Century", *Bulletin of the American Mathematical Society* **50** 161 (2012).

[65] P. J. Olver, *Applications of Lie Groups to Differential Equations*, Springer-Verlag (1993).

Chapter 28

Numerical Study of Stroboscopic Duffing Map

28.1 Introduction

This chapter continues the study of the Duffing equation begun in Section 1.4.3. Recall that we are interested in the behavior of the system governed by the differential equation

$$\ddot{q} + 2\beta\dot{q} + q + q^3 = -\epsilon \sin \omega\tau, \tag{28.1.1}$$

or its equivalent first-order equation pair

$$\begin{aligned} \dot{q} &= p, \\ \dot{p} &= -2\beta p - q - q^3 - \epsilon \sin \omega\tau. \end{aligned} \tag{28.1.2}$$

Because the right sides of (1.1) and (1.2) are periodic with period

$$T = 2\pi/\omega, \tag{28.1.3}$$

we were able to define stroboscopic times

$$\tau^n = nT, \tag{28.1.4}$$

and were able to reduce the study of the long-term behavior of the driven Duffing oscillator to the study of its associated stroboscopic map $\mathcal{M}$ under repeated iteration.

As indicated in Subsection 1.4.3, the driven Duffing oscillator is expected to display an enormously rich behavior that varies widely with the parameter values β, ϵ, and ω. Consequently, even providing an overview of what can happen requires considerable work, and even then we shall be able to discuss only some of its complexity.

Our analysis will parallel that for the logistic map as done in Section 1.2.1. We will find the fixed points of $\mathcal{M}$ for a small value of the driving strength ϵ, and track them in q, p space as the driving frequency ω is varied thereby producing a Feigenbaum/bifurcation diagram. Subsequently we will gradually increase the value of ϵ all the while observing the Feigenbaum/bifurcation diagram for $\mathcal{M}$ as a function of ω. For simplicity, we will hold the damping parameter β at the constant value $\beta = 0.1$.[1]

[1] Of course, one can also make Feigenbaum diagrams in which some other parameter, say ϵ, is varied while

28.2 Review of Simple Harmonic Oscillator Behavior

But first suppose that the q^3 term in (1.1) or (1.2) were missing. Then we know how to solve the differential equation, which is just that of a driven damped simple harmonic oscillator. The solution would consist of a particular solution plus any solution of the homogeneous equation. The particular solution, call it $q_f(\tau)$, is given by the relation

$$q_f(\tau) = -A(\beta,\omega)\epsilon\sin(\omega\tau+\phi) \tag{28.2.1}$$

where

$$A(\beta,\omega) = 1/\sqrt{(1-\omega^2)^2+(2\beta\omega)^2} \tag{28.2.2}$$

and

$$\phi(\beta,\omega) = -\text{Arctan}[(2\beta\omega)/(1-\omega^2)]. \tag{28.2.3}$$

Differentiating (2.1) gives the related result

$$p_f(\tau) = -\omega A(\beta,\omega)\epsilon\cos(\omega\tau+\phi). \tag{28.2.4}$$

Evidently $q_f(\tau)$ and $p_f(\tau)$ are periodic in τ with period T and therefore, as the subscript f is intended to convey, the phase-space point $\{q_f(0), p_f(0)\}$ is a *fixed* point of the stroboscopic map $\mathcal{M}$ for the driven damped simple harmonic oscillator. Moreover, if $\beta > 0$, then all solutions of the homogeneous equation are exponentially damped as $\tau \to \infty$, and therefore $\{q_f(0), p_f(0)\}$ is a stable (and unique) attracting fixed point. We may therefore make the identification

$$\{q_\infty, p_\infty\} = \{q_f(0), p_f(0)\}. \tag{28.2.5}$$

Figures 2.1 and 2.2 display $A(\beta,\omega)$ and $\phi(\beta,\omega)$ as a function of ω for the case $\beta = .1$, and Figures 2.3 and 2.4 show q_∞ and p_∞ as functions of ω (for the case $\beta = 0.1$ and $\epsilon = .15$), and Figure 2.5 shows them both.[2] As expected, there is resonant behavior in the vicinity of $\omega = 1$ since the coefficient of q in (1.1) is unity.[3] Also note that both q_∞ and p_∞ approach zero when ω either goes to zero or to infinity. See Exercise 2.1.

the others, including ω, are held fixed. We choose to vary ω because so doing brings resonance behavior to the fore.

[2] The value $\beta = .1$ for the damping coefficient corresponds to a quality factor $Q \simeq 4.95$. See Exercise 2.2.

[3] It was the desire for q_∞ to exhibit a resonance-like peak as a function of ω that dictated the choice (1.4.28) for ψ.

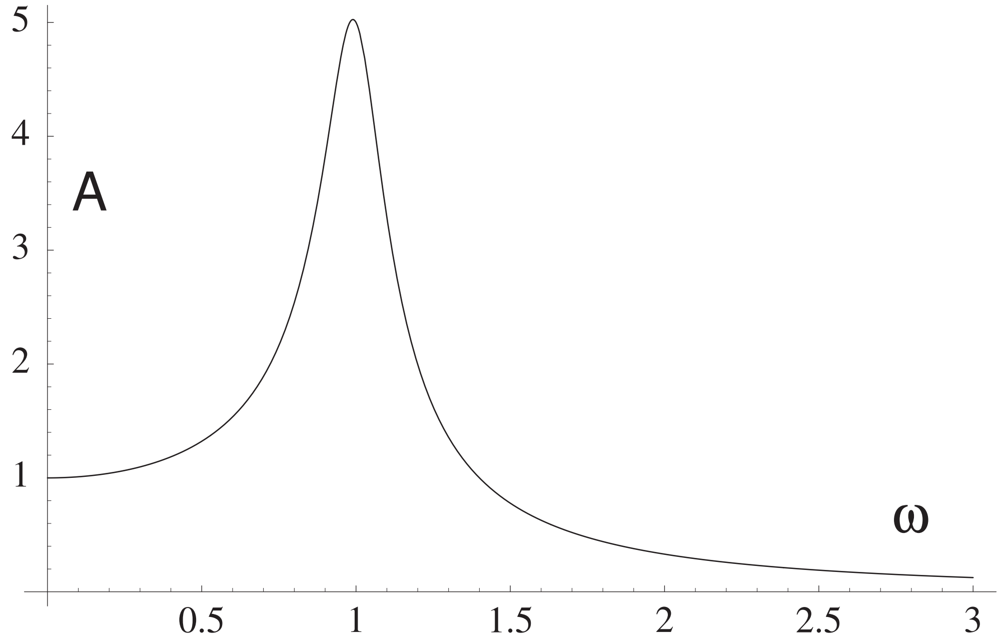


Figure 28.2.1: The quantity $A(\beta, \omega)$ as a function of ω (for the case $\beta = 0.1$).

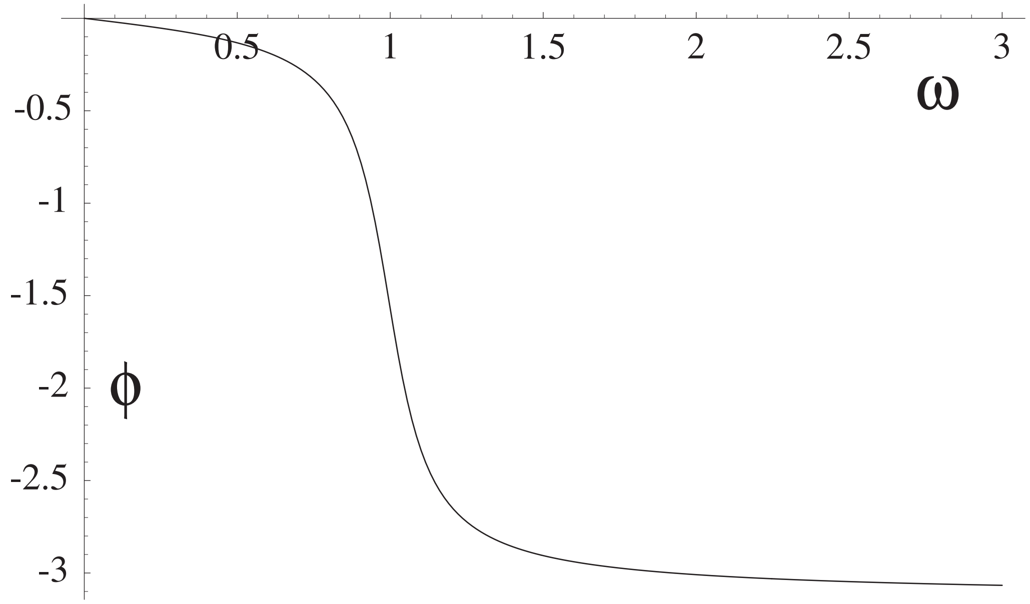


Figure 28.2.2: The quantity $\phi(\beta, \omega)$ as a function of ω (for the case $\beta = 0.1$).

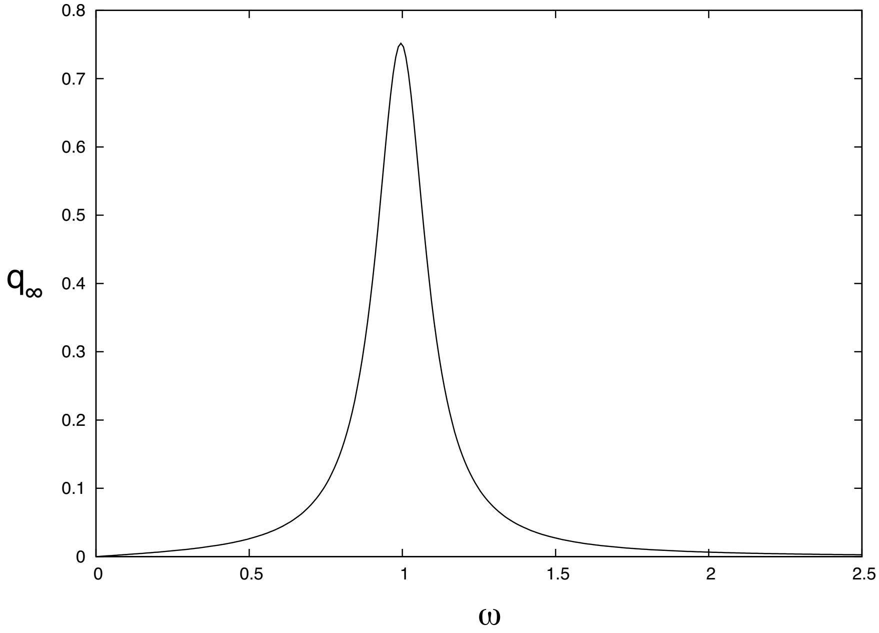


Figure 28.2.3: Feigenbaum diagram showing limiting values q_∞ as a function of ω (when $\beta = 0.1$ and $\epsilon = .15$) for the stroboscopic driven damped simple harmonic oscillator map.

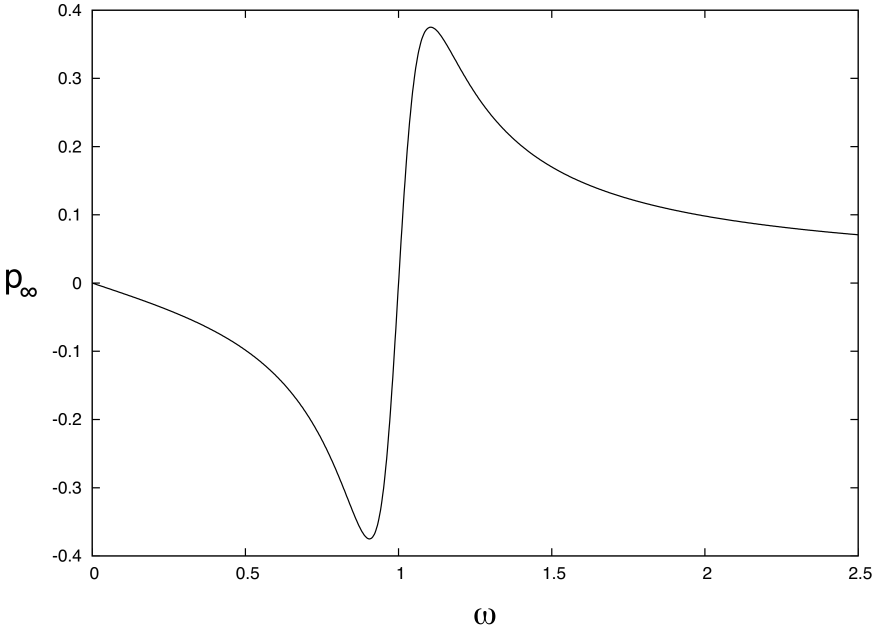


Figure 28.2.4: Feigenbaum diagram showing limiting values p_∞ as a function of ω (when $\beta = 0.1$ and $\epsilon = .15$) for the stroboscopic driven damped simple harmonic oscillator map.

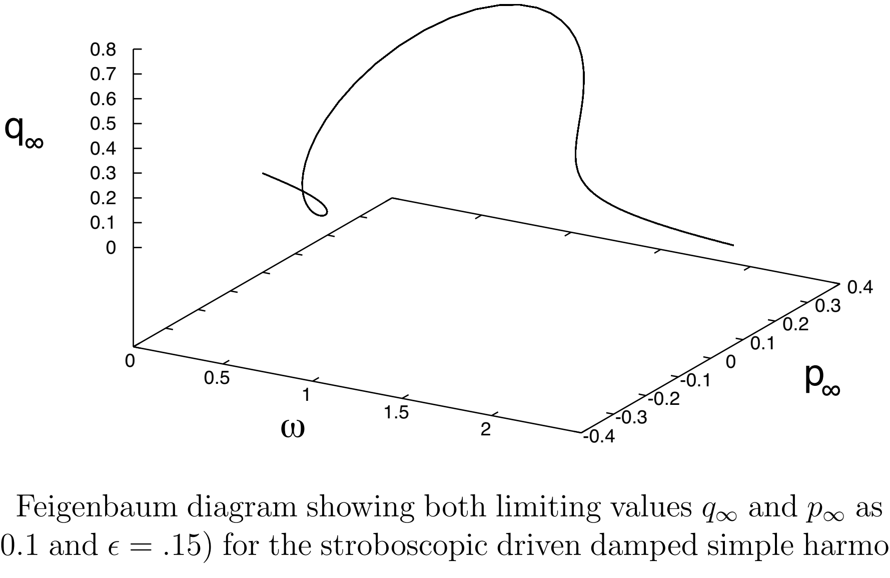


Figure 28.2.5: Feigenbaum diagram showing both limiting values q_∞ and p_∞ as a function of ω (when $\beta = 0.1$ and $\epsilon = .15$) for the stroboscopic driven damped simple harmonic oscillator map.

Exercises

28.2.1. Show, using (2.1), (2.4), and (2.5), that

$$q_\infty = -A(\beta,\omega)\epsilon\sin\phi, \tag{28.2.6}$$

$$p_\infty = -\omega A(\beta,\omega)\epsilon\cos\phi. \tag{28.2.7}$$

Show that using (2.2) and (2.3) in (2.6) and (2.7) gives the equivalent results

$$q_\infty = 2\beta\omega\epsilon/[(\omega^2-1)^2+(2\beta\omega)^2], \tag{28.2.8}$$

$$p_\infty = \omega(\omega^2-1)\epsilon/[(\omega^2-1)^2+(2\beta\omega)^2]. \tag{28.2.9}$$

Determine the behavior of q_∞ and p_∞ as ω either goes to zero or goes to infinity.

28.2.2. The *quality factor* Q of a damped harmonic oscillator is defined by the relation

$$Q = \omega_R/(2\beta) \tag{28.2.10}$$

where ω_R is the resonant frequency. For the normalization used in (1.1),

$$\omega_R^2 = 1-2\beta^2. \tag{28.2.11}$$

Show that $Q \simeq 4.95$ when $\beta = 0.1$.

28.3 Behavior for Small Driving when Nonlinearity is Included

If the driving strength ϵ is small enough and the damping coefficient β is large enough, then we expect $q(\tau)$ to be small, and therefore the q^3 term in (1.1) can indeed be neglected, at least in zeroth approximation. Figure 3.1 shows q_∞ as a function of ω for the case $\beta = 0.1$ and $\epsilon = .15$ when the q^3 term in (1.1) is *retained*, and Figure 3.2 shows both q_∞ and p_∞. Now we are dealing with the stroboscopic Duffing map, and the results shown were obtained by numerical integration. Evidently these figures resemble their simple harmonic oscillator counterparts, Figures 2.3 and 2.5. In particular, there is only *one* fixed point for each value of ω and its basin is the entire q, p plane. (Consequently there are no fixed points for powers of $\mathcal{M}$ apart from the fixed point of $\mathcal{M}$ itself.) Note, however, the appearance of some structure near the value $\omega = 1/3$, and that the resonance peak in q_∞ near $\omega = 1$ is reduced in amplitude and slightly tipped toward the right.

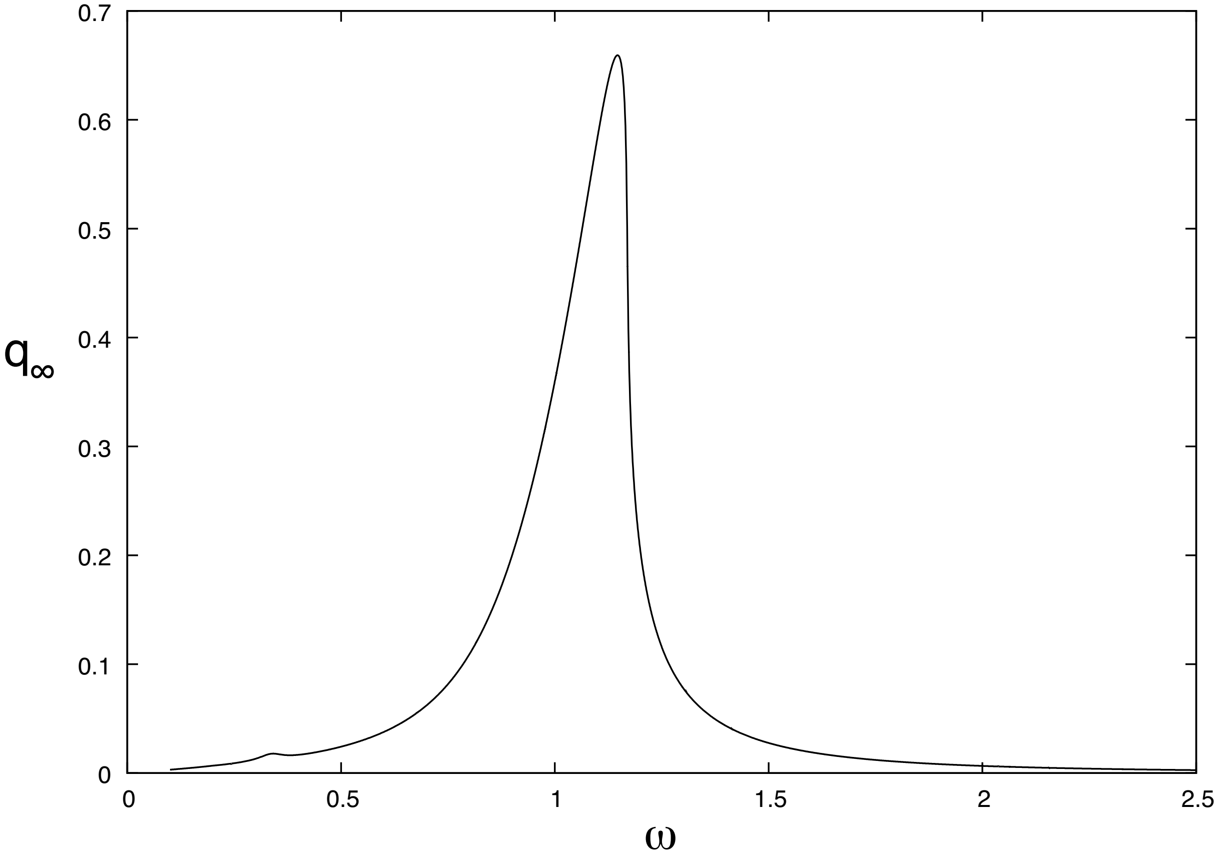


Figure 28.3.1: Feigenbaum diagram showing limiting values q_∞ as a function of ω (when $\beta = 0.1$ and $\epsilon = .15$) for the stroboscopic Duffing map.

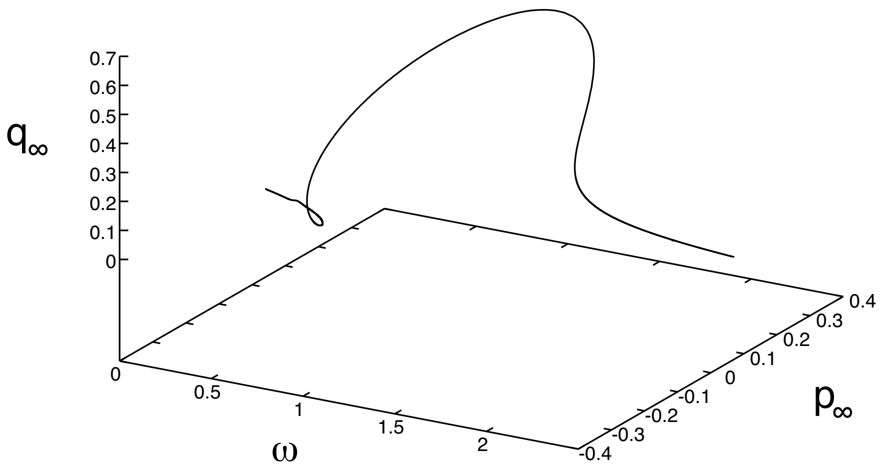


Figure 28.3.2: Feigenbaum diagram showing both limiting values q_∞ and p_∞ as a function of ω (when $\beta = 0.1$ and $\epsilon = .15$) for the stroboscopic Duffing map.

28.4 What Happens Initially When the Driving Is Increased?

28.4.1 Saddle-Node (Blue-Sky) Bifurcations

We have found the Feigenbaum diagram of the stroboscopic Duffing map for small driving strength ϵ. As promised, let us now increase ϵ to see what occurs. Figures 4.1 and 4.2 show results for the case $\epsilon = 1.5$. Evidently the height of the resonance peak has grown in response to the increased driving and has taken on a more complicated structure. And the feature originally near $\omega = 1/3$ has become a clearly defined *subresonant* peak. These peaks have also moved to larger values of ω. This is to be expected since the natural frequency of an oscillator having a hard spring increases with amplitude. Moreover, additional features now appear to the left of those already recognized.

Most striking, for $\omega \in (1.8\cdots, 2.7\cdots)$, there are *three* fixed points in place of the *single* fixed point originally present for the case of less driving. Two of these fixed points are stable and the third, whose coordinates as a function of ω are shown as a red line, is unstable. (How the unstable fixed point can be found is described in Section 29.4.) What happens is that, as ω is increased from small values, a pair of fixed points, one unstable and one stable, is 'born' near $\omega = 1.8\cdots$. This is sometimes called a *saddle-node* bifurcation. (The term *saddle* denotes a particular kind of unstable fixed point, and the term *node* denotes a particular kind of stable fixed point.) It is also called a *blue sky* bifurcation since these fixed points seem to appear out of nowhere, i.e. out of the blue. (They actually come out of the complex domain).[4] Then, as ω is further increased, the unstable fixed point moves to meet and 'annihilate' the original fixed point at $\omega = 2.7\cdots$ in an *inverse* saddle-node (or blue sky) bifurcation thereby leaving behind only the stable fixed point born near $\omega = 1.8\cdots$. All this behavior can be understood on topological grounds. See Section 29.5.

[4] For an example of a blue sky bifurcation in the case of the one-dimensional logistic (quadratic) map, see the end of Exercise 1.2.7.

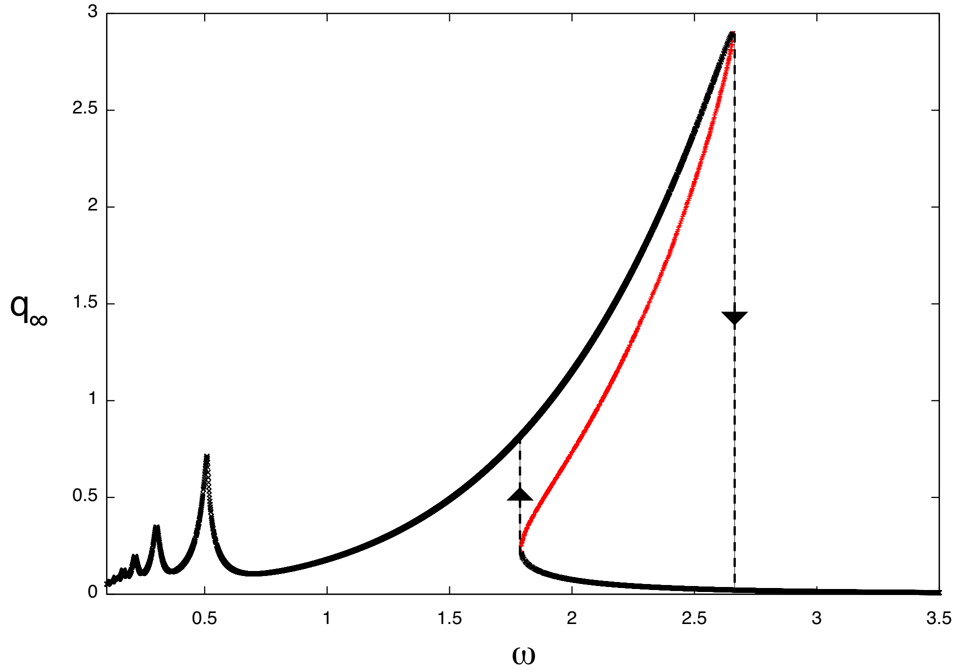


Figure 28.4.1: Feigenbaum/bifurcation diagram showing limiting values q_∞ as a function of ω (when $\beta = 0.1$ and $\epsilon = 1.5$) for the stroboscopic Duffing map. Also shown, in red, is the trail of the unstable fixed point. Finally, jumps in the steady-state amplitude are illustrated by vertical dashed lines at $\omega \simeq 1.8$ and $\omega \simeq 2.6$.

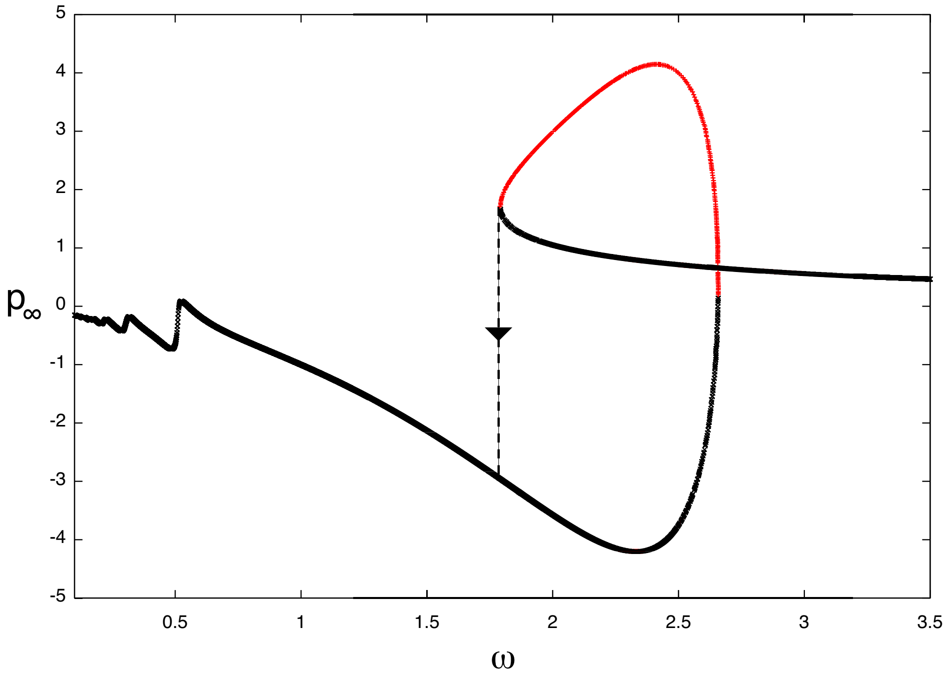


Figure 28.4.2: Feigenbaum/bifurcation diagram showing limiting values of p_∞ as a function of ω (when $\beta = 0.1$ and $\epsilon = 1.5$) for the stroboscopic Duffing map. Also shown, in red, is the trail of the unstable fixed point. Finally, a downward jump in the steady-state value p_∞ at $\omega \simeq 1.8$ is illustrated by a vertical dashed line. There is also an upward jump between the two black curves at $\omega \simeq 2.6$, but this feature is too small to be easily indicated by a second vertical dashed line.

28.4.2 Basins

Figure 4.3 shows the basins of attraction for the two stable fixed points when $\omega = 2.25$. The stable fixed points have the locations

$$w^1 = (q_\infty, p_\infty) = (0.04247237, 0.84035059) \text{ (green)} \tag{28.4.1}$$

and

$$w^2 = (q_\infty, p_\infty) = (1.68001491, -4.14472685) \text{ (red)}. \tag{28.4.2}$$

The unstable fixed point has the location

$$w^3 = (q_\infty, p_\infty) = (1.32261, 3.88274). \tag{28.4.3}$$

Since these two basins together comprise the entire q, p plane, there are no other stable fixed points of $\mathcal{M}$. Moreover, there are no fixed points for powers of $\mathcal{M}$ apart from the fixed points of $\mathcal{M}$ itself. Finally, it can be shown that, unlike the complex logistic map, the basin boundaries are smooth. This is because in this case there are no homoclinic points. See Section 29.6 and Figure 29.6.8. Thus, in this parameter regime, the long-term behavior of the driven Duffing oscillator is relatively simple.[5]

What is the actual motion for the periodic orbits associated with these fixed points? Figure 4.4 shows $q(\tau)$ for the two stable fixed points, and Figure 4.5 shows $q(\tau)$ for the unstable fixed point, all for the case $\omega = 2.25$. At this point one can make two interesting observations.

28.4.3 Symmetry

The first observation is that if $q(\tau)$ is a solution (periodic or otherwise) to Duffing's equation, then so is $\bar{q}(\tau)$ with

$$\bar{q}(\tau) = -q(\tau - \pi/\omega). \tag{28.4.4}$$

Note that, in view of (1.3), there is the relation

$$\pi/\omega = T/2 \tag{28.4.5}$$

so that (4.4) can also be written in the form

$$\bar{q}(\tau) = -q(\tau - T/2). \tag{28.4.6}$$

This property is an example of what is sometimes called *equivariance*, and occurs in this case because the left side of (1.1) is odd in q and does not explicitly contain the time. See Exercise 4.1. Thus, given a solution q of Duffing's equation, use of (4.6) produces a related solution $\bar{q}$. In principle this solution may be different, but it could also be the same as the original one. Inspection of Figures 4.4 and 4.5 reveals that

$$\bar{q}(\tau) = q(\tau) \tag{28.4.7}$$

[5] We remark that there would be no attractors, and consequently no basins, in the zero damping limit $\beta \to 0$, for then the system would be Hamiltonian, and we have learned in Subsections 3.4 and 6.4 that Hamiltonian systems have neither attractors or repellers. That is one reason why the long-term behavior of most Hamiltonian systems is so complicated.

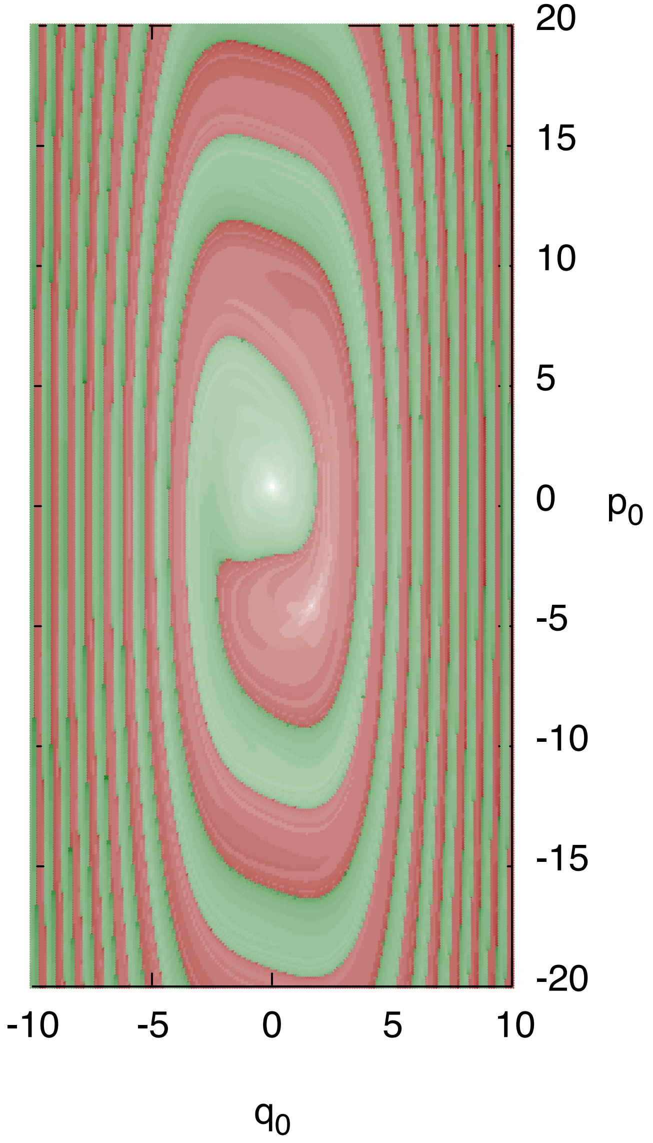


Figure 28.4.3: Basins of attraction for the two stable fixed points (when $\omega = 2.25$, $\beta = 0.1$, and $\epsilon = 1.5$) for the stroboscopic Duffing map. Green points are in the basin of the attracting fixed point w^1 and red points are in the basin of the attracting fixed point w^2. There is also an unstable fixed point w^3. See Figures 29.6.7 and 29.6.8.

for all three periodic orbits shown. Therefore in this case each solution is sent into itself under the 'barring' operation. It can be verified that the same is true for all the periodic solutions associated with all the fixed points found so far.

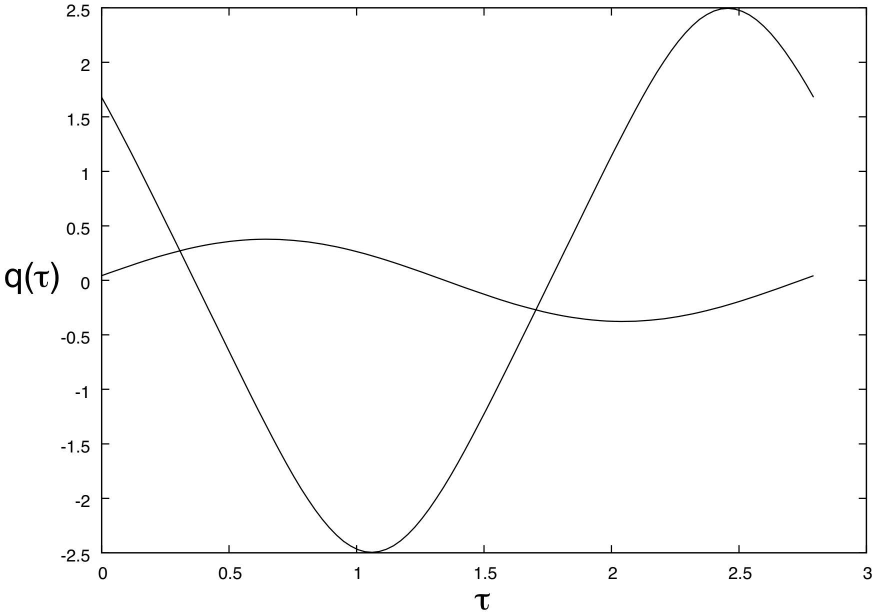


Figure 28.4.4: Stable periodic orbits $q(\tau)$ (when $\omega = 2.25$, $\beta = 0.1$, and $\epsilon = 1.5$) for the Duffing equation.

28.4.4 Amplitude Jumps

The second observation is that the two stable periodic orbits shown in Figure 4.4 have *different* amplitudes. With reference to Figures 4.1 and 4.2, suppose that $\omega \simeq 1.5$ and that the Duffing oscillator has settled down to the periodic orbit associated with the attracting fixed point q_∞, p_∞. (For this value of ω there is only one fixed point, and it is attracting.) Next imagine slowly increasing ω (*slowly* means in a time large compared to the time required to settle down to the attracting periodic orbit). Then the Duffing oscillator will essentially remain on the periodic orbit associated with the value of q_∞ shown as the upper curve in Figure 4.1. This will continue to be the case until ω reaches the value $\omega \simeq 2.6$, at which value the stable fixed point merges with the unstable fixed point and they mutually annihilate. What has happened is that the basin of attraction of this stable fixed point has shrunk to zero. When this occurs, the oscillator orbit finds itself in the basin of the other remaining stable fixed point and is rapidly attracted to the periodic orbit associated with that stable fixed point. Moreover, as Figure 4.1 suggests and Figure 4.4 confirms, the amplitude of oscillation associated with this new periodic orbit is considerably less than that associated with the old. Thus, the Duffing oscillator exhibits an *amplitude jump* (to an appreciably *lower* value) as ω is increased beyond $\omega \simeq 2.6$. This amplitude jump is illustrated by the vertical dashed lines at $\omega \simeq 2.6$ in Figures 4.1 and 4.2.

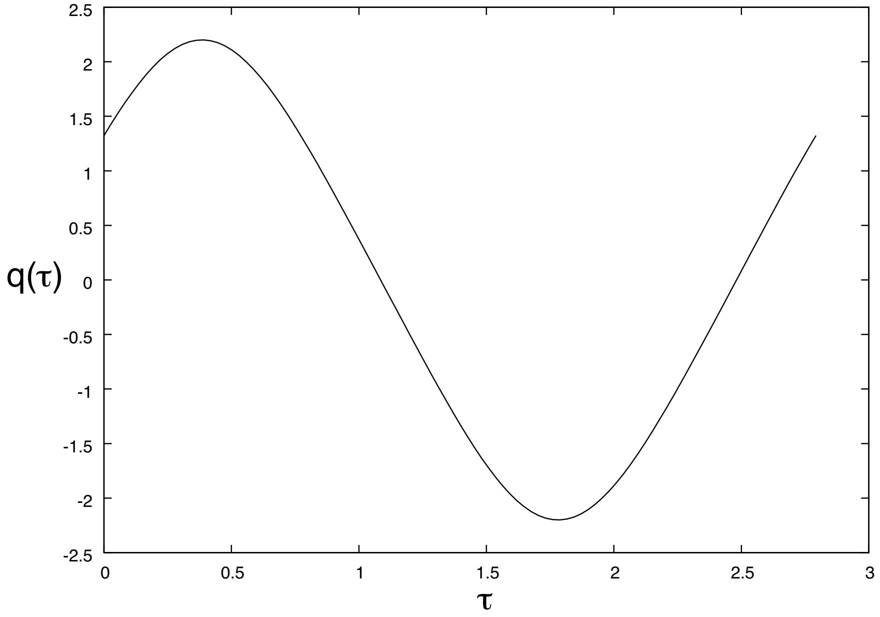


Figure 28.4.5: Unstable periodic orbit $q(\tau)$ (when $\omega = 2.25$, $\beta = 0.1$, and $\epsilon = 1.5$) for the Duffing equation.

28.4.5 Hysteresis

Next suppose that $\omega > 2.6$, say $\omega \simeq 3$, and that the Duffing oscillator has settled down to the periodic orbit associated with the attracting fixed point q_∞, p_∞. (For this value of ω there is again only one fixed point, and it is attracting.) Now slowly decrease ω. Then the Duffing oscillator will essentially remain on the periodic orbit associated with the value of q_∞ shown as the lower curve in Figure 4.1. This will continue to be the case until ω reaches the value $\omega \simeq 1.8$, at which value the stable fixed point in question merges with the unstable fixed point and they mutually annihilate. What has happened is that the basin of attraction of this stable fixed point has now shrunk to zero. When this occurs, the oscillator orbit finds itself in the basin of the other remaining stable fixed point and is rapidly attracted to the periodic orbit associated with that stable fixed point. Now the amplitude of oscillation will jump to a *larger* value. This amplitude jump is also illustrated by vertical dashed lines at $\omega \simeq 1.8$ in Figures 4.1 and 4.2. Note that the ω values for the two amplitude jumps are different. Thus, the steady-state amplitude of the Duffing oscillator exhibits *hysteresis* as ω is slowly varied back and forth over the range in which saddle-node bifurcations occur.

Exercises

28.4.1. Show that the left side of (1.1) changes sign under the replacement of q by $-q$. Show that the right side of (1.1) changes sign under the replacement of τ by $(\tau - \pi/\omega)$. Verify that if $q(\tau)$ is a solution to Duffing's equation, then so is $\bar{q}(\tau)$ as given by (4.4) or (4.6). Let $\bar{\bar{q}}$ denote the result of applying the barring operation to $\bar{q}$. Show that if q is a solution associated with a fixed point of $\mathcal{M}$, and therefore is a periodic solution with period

T, then $\overline{\overline{q}}= q$. Verify that the harmonic oscillator solution (2.1) satisfies (4.7).

28.5 Pitchfork Bifurcations and Symmetry

Let us continue to increase ϵ. Figure 5.1 shows that a qualitatively new feature appears when ϵ is near 2.2: a *bubble* is formed *between* the major resonant peak (the one that has saddle-node bifurcated) and the subresonant peak immediately to its left. To explore the nature of this bubble, let us make ϵ still larger, which, we anticipate, will result in the bubble becoming larger. Figures 5.2 and 5.3 show Feigenbaum diagrams in the case $\epsilon = 5.5$. Now the major resonant peak and the subresonant peak have moved to larger ω values. Correspondingly, the bubble between them has also moved to larger ω values. Moreover, it is larger, yet another smaller bubble has formed, and the subresonant peak between them has also undergone a saddle-node bifurcation. For future use, we will call the major resonant peak the *first* or *leading* saddle-node bifurcation, and we will call the subresonant peak between the two bubbles the *second* saddle-node bifurcation, etc. Also, we will call the bubble just to the left of the first saddle-node bifurcation the *first* or *leading* bubble, and the next bubble will be called the *second* bubble, etc.

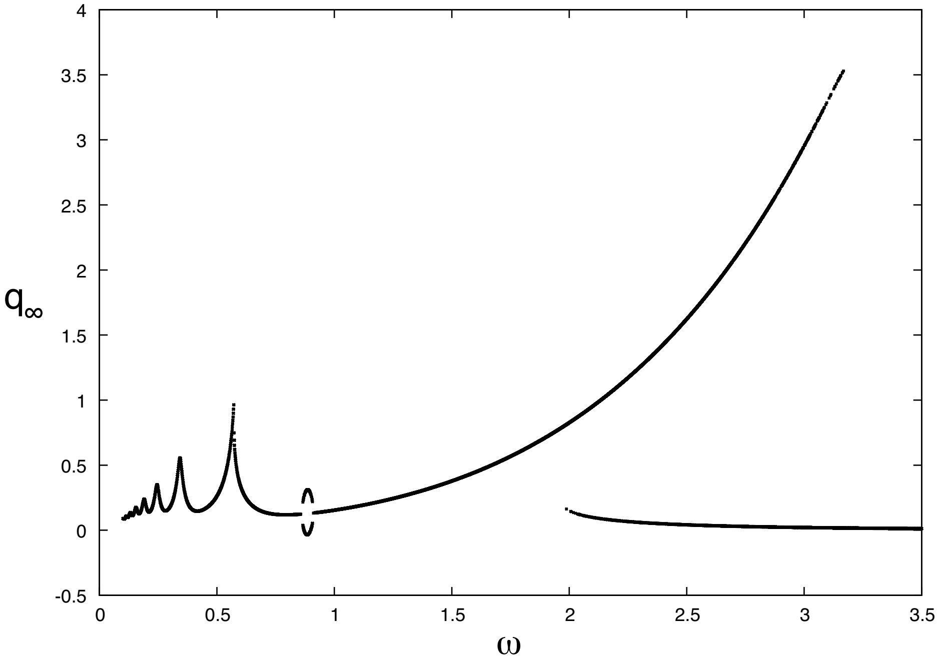


Figure 28.5.1: Feigenbaum diagram showing limiting values q_∞ as a function of ω (when $\beta = 0.1$ and $\epsilon = 2.2$) for the stroboscopic Duffing map. It displays that a bubble has now formed at $\omega \approx .8$.

Figure 5.4 shows the larger (leading) bubble in Figure 5.2 in more detail and with the addition of red lines indicating the trails of unstable fixed points. It reveals that the bubble describes the *simultaneous* bifurcation of a single fixed point into three fixed points. Two of these fixed points are stable and the third, whose q coordinate as a function of ω is shown as

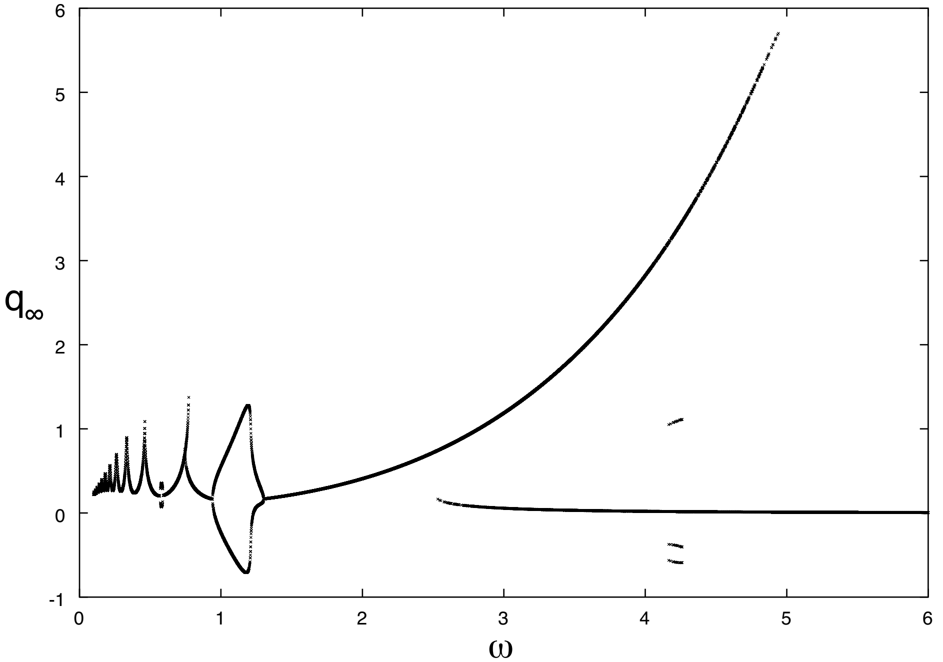


Figure 28.5.2: Feigenbaum diagram showing limiting values q_∞ as a function of ω (when $\beta = 0.1$ and $\epsilon = 5.5$) for the stroboscopic Duffing map. The first bubble has grown, a second smaller bubble has formed to its left, and the sub-resonant peak between them has saddle-node bifurcated to become the second saddle-node bifurcation.

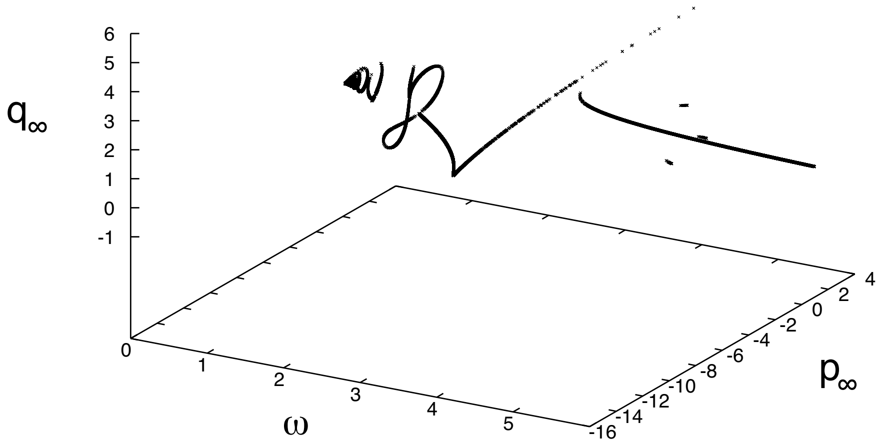


Figure 28.5.3: Feigenbaum diagram showing both limiting values q_∞ and p_∞ as a function of ω (when $\beta = 0.1$ and $\epsilon = 5.5$) for the stroboscopic Duffing map.

a red line, is unstable. What happens is that, as ω is increased, a *single* stable fixed point becomes a *triplet* of fixed points, two of which are stable and one of which is unstable. This is called a *pitchfork* bifurcation. Then, as ω is further increased, these three fixed points again merge, in an inverse pitchfork bifurcation, to form what is again a single stable fixed point. This behavior can also be understood on topological grounds. Again see Section 29.5.

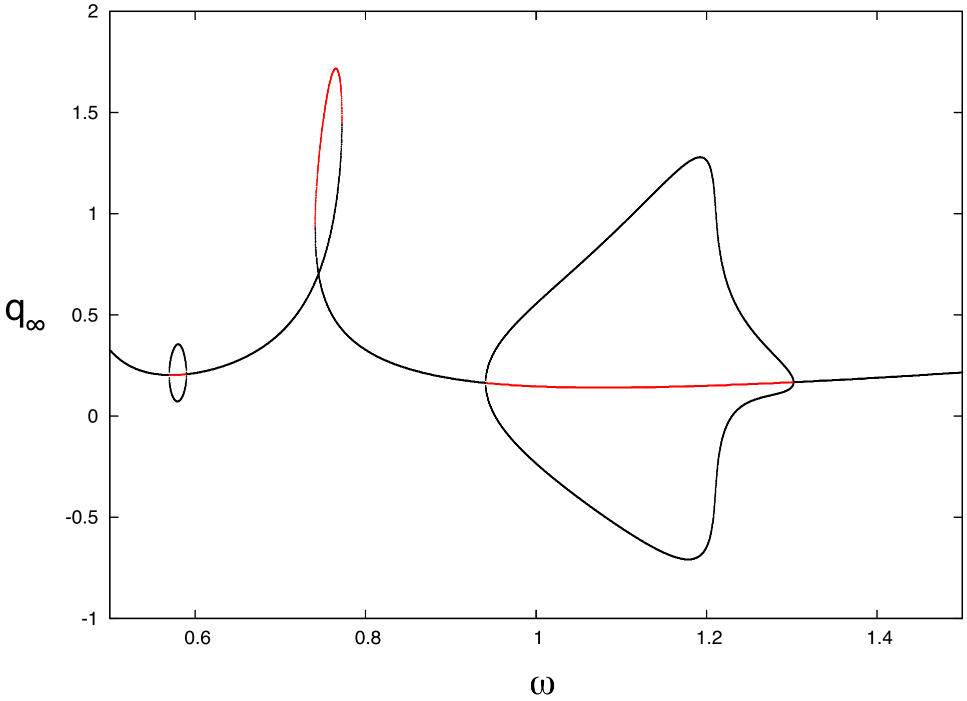


Figure 28.5.4: An enlargement of Figure 5.2 with the addition of red lines indicating the trails of unstable fixed points.

As a side comment, we remark that a pitchfork bifurcation could better be called a pitchfork *trifurcation*. Unlike a saddle-node bifurcation, is this case all three fixed points appear where once there was only one. True pitchfork bifurcations are rare, and only occur in the presence of symmetry, in this case the equivariance symmetry described earlier. However, it can happen, particularly in the case of near symmetry, that a saddle-node bifurcation occurs very close to another stable fixed point so that from a distance what appears to be happening is a pitchfork bifurcation. As an example of near symmetry, the left side of (1.1) could be modified (perturbed) to contain an additional term of the form δq^2 where δ is small. Figure 5.5 illustrates how Figure 5.4 is modified when the term $0.02q^2$ is added to the left side of (1.1). Evidently the pitchfork bifurcation becomes a saddle-node bifurcation. A pair of fixed points, one stable and one unstable, is born in the vicinity of $\omega = 1$, and they move as ω is increased. However, unlike the case of Figure 4.1, they then annihilate *each other* near $\omega = 1.3$ rather than the unstable fixed point moving up to the other fixed point so that this pair is mutually annihilated. Note also that the perturbation destroys the small bubble that was near $\omega = .6$ (which was also a pitchfork bifurcation before the perturbation was introduced).

To continue with the case of the pitchfork bifurcation, and as we did in the case of a

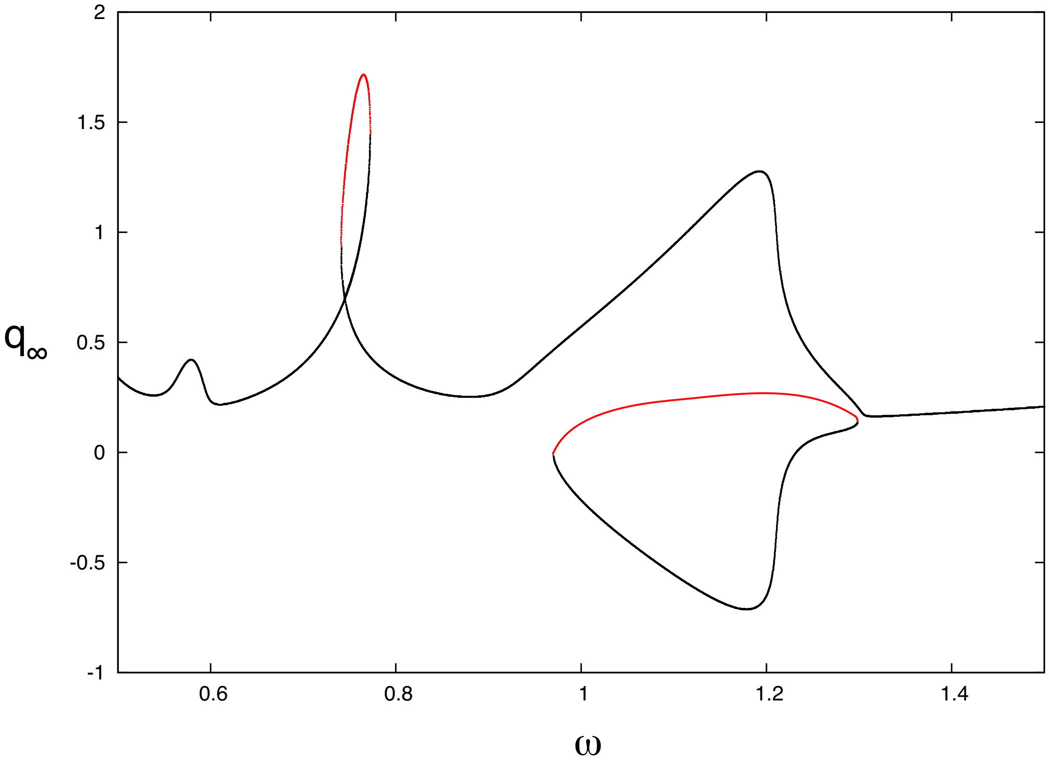


Figure 28.5.5: Transformation of a pitchfork bifurcation into a saddle-node bifurcation due to the inclusion of the symmetry breaking perturbation $0.02q^2$. Also shown as red lines are the trails of unstable fixed points. Note, however, that the stable-unstable pair of fixed points born at $\omega \approx 1$ self annihilates at $\omega \approx 1.3$ rather than the unstable fixed point annihilating the other stable fixed point as happens in Figure 4.1.

saddle-node bifurcation, let us plot the three periodic orbits associated with the three fixed points. Figure 5.6 displays $q(\tau)$ for the two stable fixed points, and Figure 5.7 displays $q(\tau)$ for the unstable fixed point, all for the case $\omega = 1.1$. The stable fixed points have the locations

$$(q_\infty, p_\infty) = (0.942055303, -0.792682910) \tag{28.5.1}$$

and

$$(q_\infty, p_\infty) = (-0.55292184, -1.72277791). \tag{28.5.2}$$

The unstable fixed point has the location

$$(q_\infty, p_\infty) = (0.140706, -1.05507). \tag{28.5.3}$$

In this case inspection shows that the unstable periodic orbit is sent into itself under the barring operation, as before. However, unlike the saddle-node case, the two stable orbits are *interchanged* under the barring operation.[6] Also, since the amplitudes of the two stable periodic oscillations are the same as a result of their being interchanged under the barring operation (see Figure 5.6), there are no amplitude jumps associated with pitchfork bifurcations.

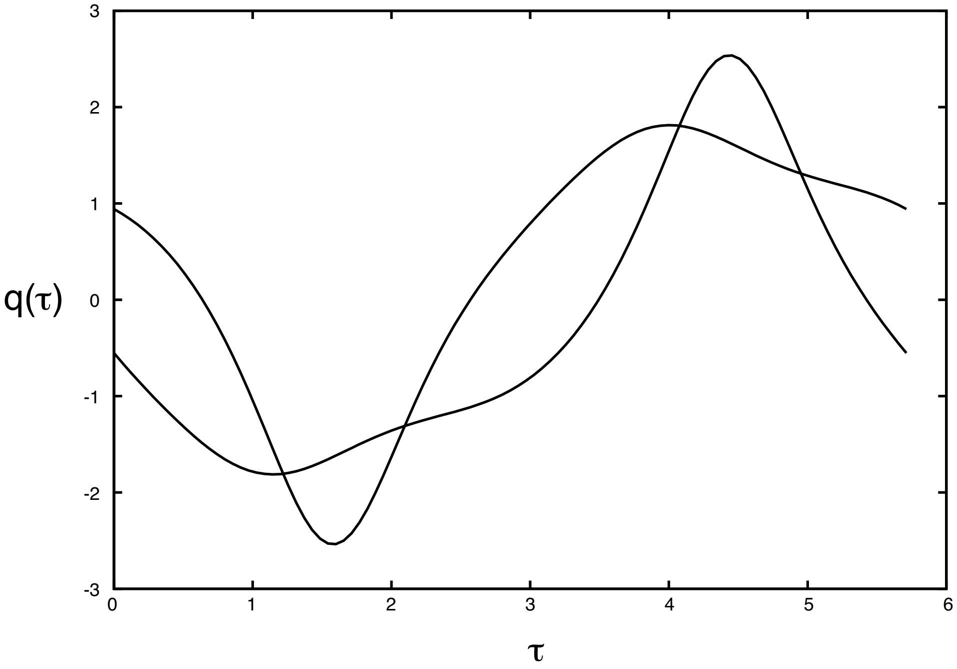


Figure 28.5.6: Stable periodic orbits $q(\tau)$ (when $\omega = 1.1$, $\beta = 0.1$, and $\epsilon = 5.5$) for the Duffing equation.

[6] The change in the nature of periodic Duffing orbits at a pitchfork bifurcation is sometimes described as *dynamical spontaneous symmetry breaking.* For small ϵ values and all ω values, all periodic orbits have the symmetry property of being invariant (sent into themselves) under the barring operation. For larger ϵ values, as ω is varied, some periodic orbits appear that no longer have this symmetry, even though the underlying equations of motion retain the same symmetry for all values of ϵ and ω.

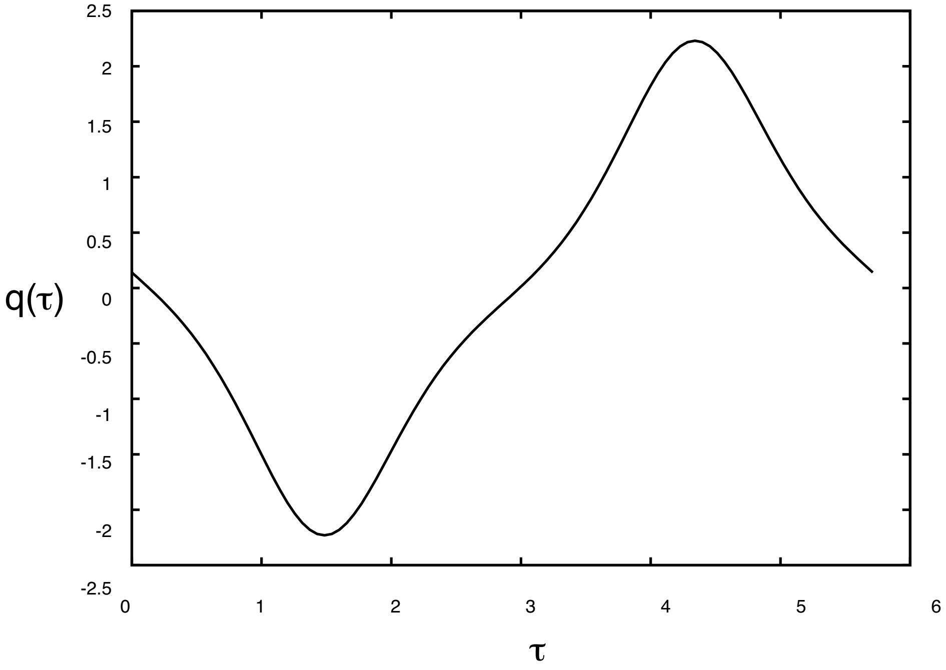


Figure 28.5.7: Unstable periodic orbit $q(\tau)$ (when $\omega = 1.1$, $\beta = 0.1$, and $\epsilon = 5.5$) for the Duffing equation.

28.6 Period Tripling Bifurcations and Fractal Basin Boundaries

Close examination of Figures 5.2 and 5.3 shows that something also happens in the vicinity of $\omega = 4.15$: Three attracting fixed points of $\mathcal{M}^3$ appear and then again vanish as ω is varied. Although these points are fixed points of $\mathcal{M}^3$, they are not fixed points of $\mathcal{M}$, and hence are *period-three* fixed points of $\mathcal{M}$. They correspond to solutions that do not have period T, but rather are periodic with period $3T$. Solutions that are not periodic with the drive period T, but are periodic with a period that is some integer multiple of T, are said to be *subharmonic*.[7]

Figure 6.1 shows an enlargement of that portion of Figure 5.2 where period tripling occurs. The period-three fixed points are shown in green and the period-one fixed points, both stable and unstable and unstable, are shown in red. What happens is that $\mathcal{M}^3$ exhibits saddle-node (blue-sky) bifurcations so that $\mathcal{M}^3$ fixed points are born (and subsequently annihilate) in pairs. Three of each pair, those that are attracting, are shown in Figures 5.2 and 5.3. Figure 6.1 shows the period-one fixed points in red (two stable and one unstable) and the period-three fixed points in green. Inspection of the green features suggests that there are six fixed points of $\mathcal{M}^3$ that occur as stable-unstable pairs. Figures 6.2 through

[7] If a periodic solution has period nT, it has fundamental frequency ω/n. Correspondingly, such a solution is called $1/n$ subharmonic. For the case being discussed here, $n = 3$. It is sometimes stated in the literature that for the driven Duffing equation there is no subharmonic corresponding to the case $n = 2$, which would be the case of period *doubling*. However, we will eventually see that, for sufficiently strong driving, period doubling does occur.

6.4 confirm this analysis. They show each pair of saddle and node fixed points of $\mathcal{M}^3$ as functions of ω.

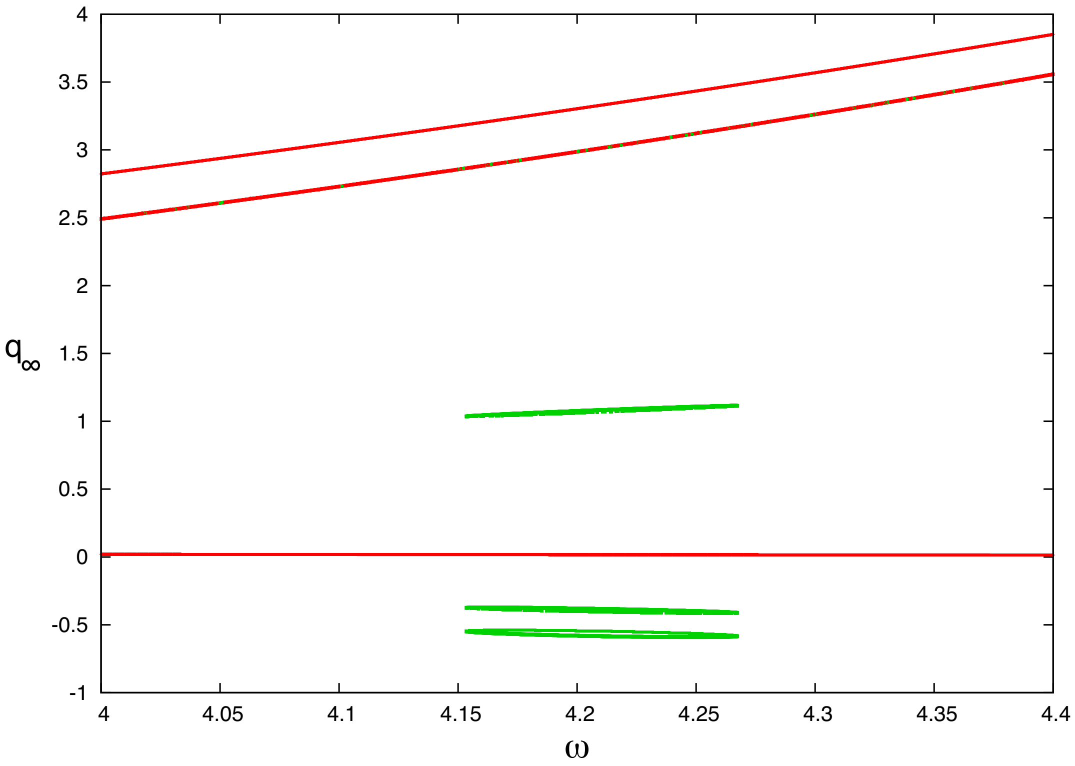


Figure 28.6.1: An enlargement of Figure 5.2 showing, for $\mathcal{M}$, the period-one fixed points in red (two stable and one unstable) and the stable-unstable pairs of period-three fixed points in green.

What can be said about the basin structure of $\mathcal{M}$ and $\mathcal{M}^3$ in this case? Let us set

$$\omega = 4.21, \tag{28.6.1}$$

which is a convenient value roughly midway between the birth and annihilation values of ω for the period-three fixed points in Figures 6.1 through 6.4. Numerical study shows that for this value of ω there are the following attracting fixed points:

$$w^1 = (0.01666814, 1.38706838) \text{ (white)}, \tag{28.6.2}$$

$$w^2 = (3.32944854, -15.41028862) \text{ (blue)}; \tag{28.6.3}$$

$$z^1 = (1.08279489, 1.40756189) \text{ (red)}, \tag{28.6.4}$$

$$z^2 = \mathcal{M}z^1 = (-0.58378622, 0.30474951) \text{ (green)}, \tag{28.6.5}$$

$$z^3 = \mathcal{M}z^2 = (-0.38267261, 2.82716098) \text{ (yellow)}. \tag{28.6.6}$$

The points w^1 and w^2 are the attracting fixed points of $\mathcal{M}$. Of course, they will also be attracting fixed points of $\mathcal{M}^3$. The points z^1, z^2, z^3 are the attracting fixed points of $\mathcal{M}^3$,

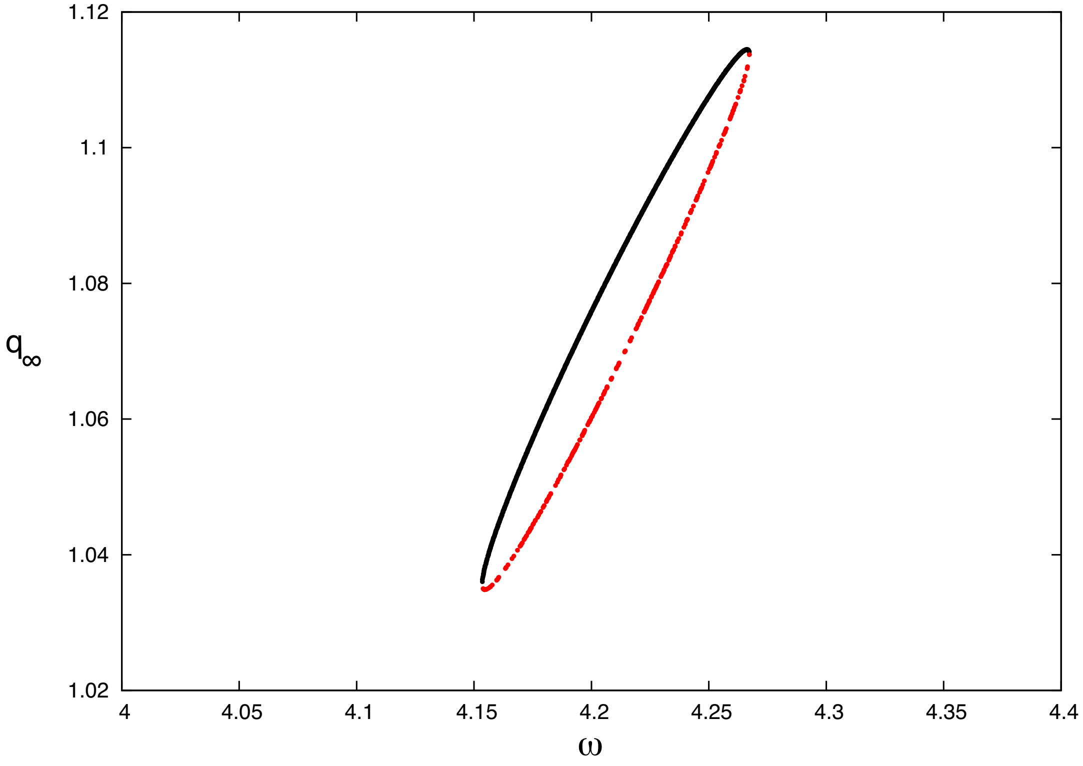


Figure 28.6.2: A blue-sky bifurcation that produces, and then subsequently destroys, a pair of stable (black) and unstable (red) period-three fixed points. These points correspond to the upper green feature shown in Figure 6.1.

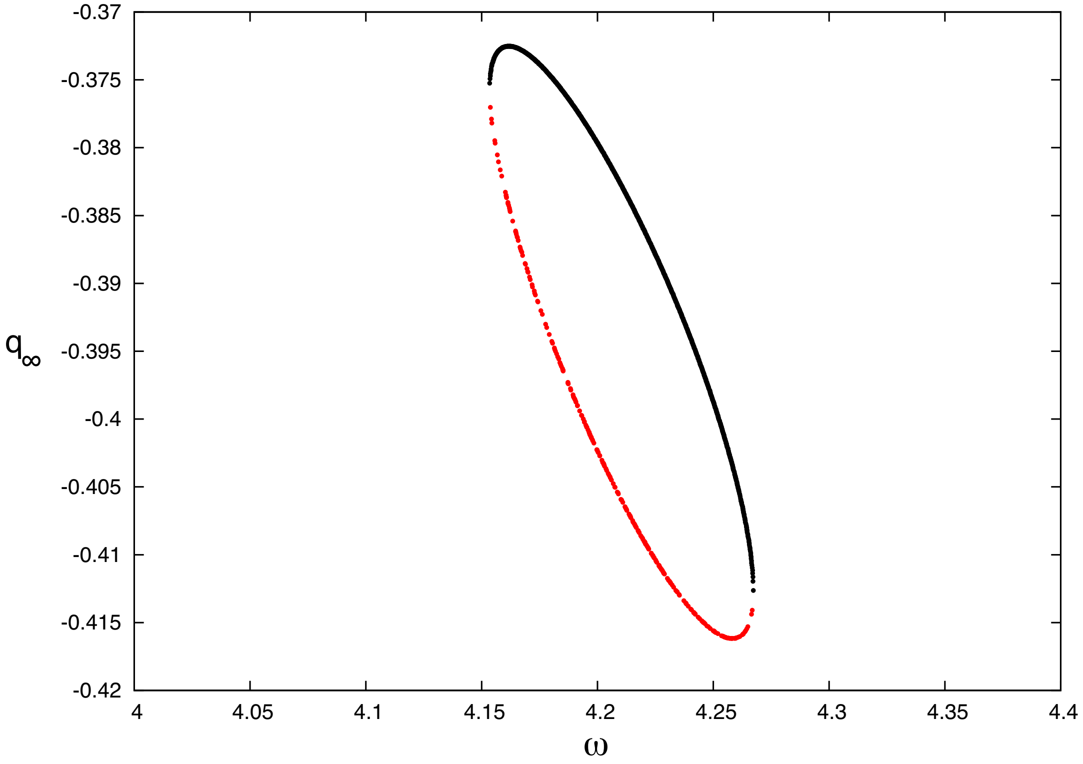


Figure 28.6.3: A blue-sky bifurcation that produces, and then subsequently destroys, a pair of stable (black) and unstable (red) period-three fixed points. These points correspond to the center green feature shown in Figure 6.1.

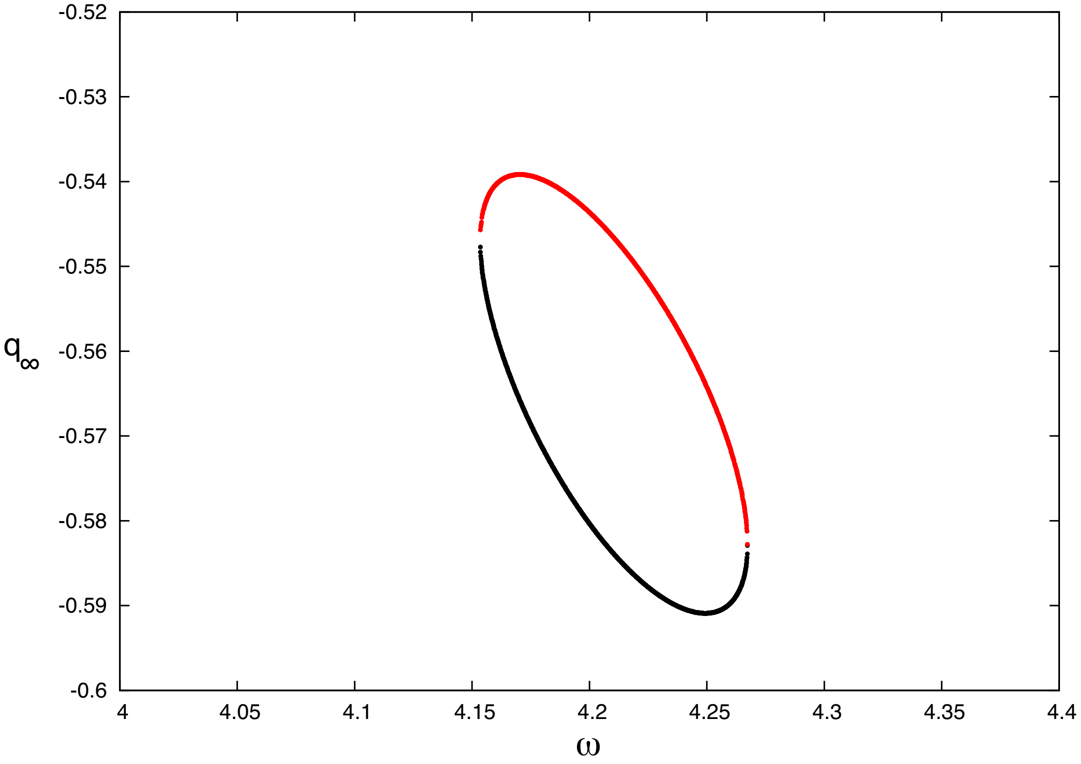


Figure 28.6.4: A blue-sky bifurcation that produces, and then subsequently destroys, a pair of stable (black) and unstable (red) period-three fixed points. These points correspond to the bottom green feature shown in Figure 6.1.

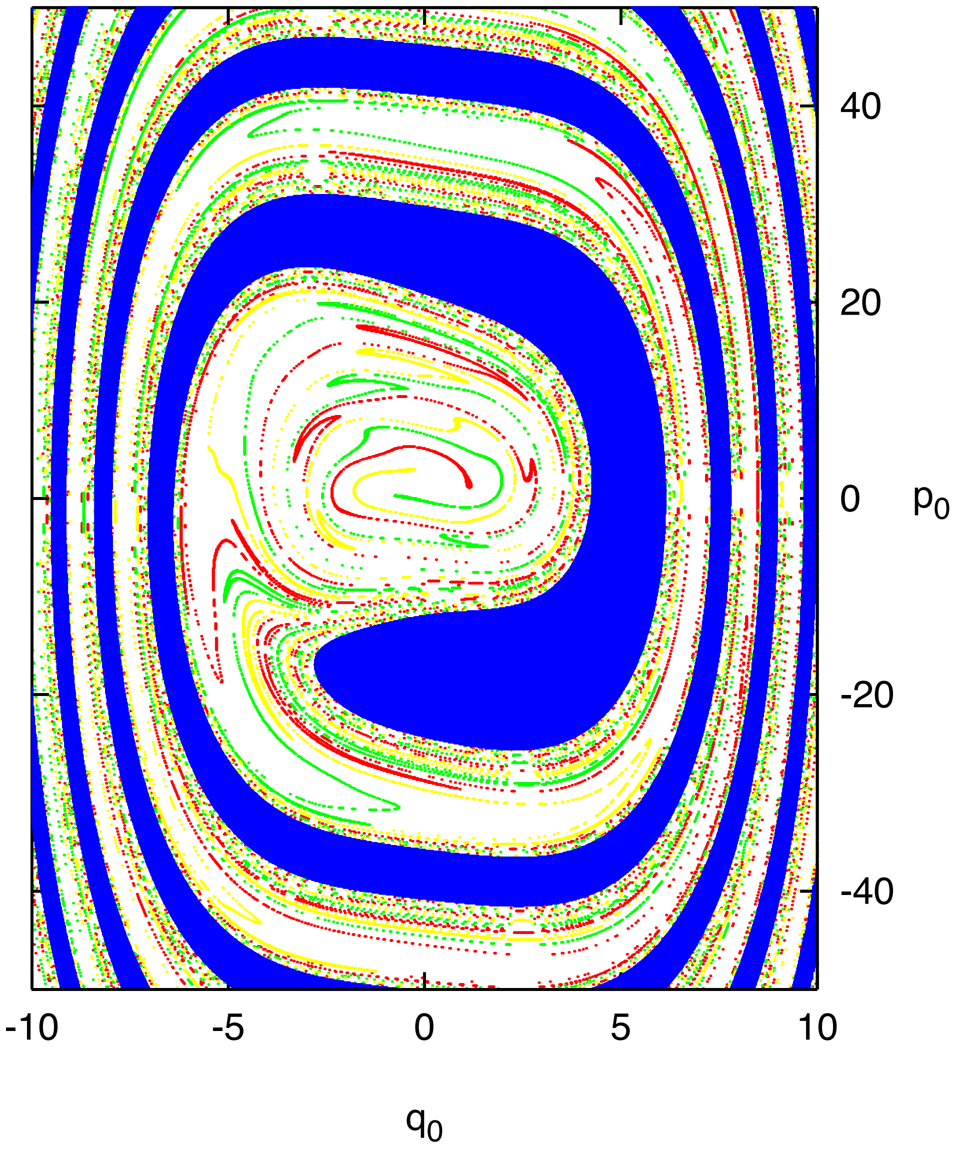


Figure 28.6.5: Basins, using the map $\mathcal{M}^3$ and with $\omega = 4.21$, for the period-one attracting fixed points w^1 (white) and w^2 (blue), and the period-three attracting fixed points z^1 (red), z^2 (green), and z^3 (yellow).

and hence attracting period-three fixed points of $\mathcal{M}$. Figure 6.5 displays, using the map $\mathcal{M}^3$, the basins for all these attracting fixed points.

When viewed from a distance, the basins of w^1 and w^2 in Figure 6.5 look much like those in Figure 4.3. However, closer inspection of the white basin, that of w^1, reveals that it contains within it the basins of the period-three fixed points z^1, z^2, and z^3. Moreover, the basins of the period-three fixed points consist of principal components (which contain the period-three fixed points z^1, z^2, and z^3 themselves) plus what appear to be an infinite number of disconnected pieces. (Recall that all white points are in the basin of w^1.) Finally, the pieces of the period-three basins crowd ever more closely together (but still separated by ever smaller white areas) in the vicinity of the boundary of the w^2 (blue) basin so that this basin-boundary structure becomes fractal. These features are seen even more clearly in Figure 6.6, which is an enlargement (with different q, p scales) of Figure 6.5 in the vicinity of the points w^1, z^1, z^2, and z^3. In this figure the fixed points w^1 and z^1, z^2, and z^3 themselves are shown as small black dots. Because the basin-boundary structure is fractal, the final fate of an orbit launched in the vicinity of the basin boundary (which of the five attracting fixed points w^1, w^2, z^1, z^2, and z^3 it eventually approaches) depends very sensitively on the initial conditions.

28.7 Asymptotic ω Behavior

Examination of all the Feigenbaum/bifurcation diagrams produced so far for the Duffing oscillator shows that their behavior is consistent with the hypothesis

$$\lim_{\omega\to 0} q_\infty = 0, \quad \lim_{\omega\to 0} p_\infty = 0, \tag{28.7.1}$$

$$\lim_{\omega\to \infty} q_\infty = 0, \quad \lim_{\omega\to \infty} p_\infty = 0. \tag{28.7.2}$$

That is, in the limits $\omega \to 0$ or ∞, there is a single attracting fixed point and its basin is the entire q, p plane. Correspondingly, for each value of ϵ (and β) there is only a finite range of ω values that is of interest. In the case of Figures 5.2 and 5.3, for example, extension of the ω range to smaller and larger values shows that the advertised small and large ω asymptotic behavior has already set in so that no new features appear beyond those already seen.

To explore the $\omega \to 0$ limit, rewrite (1.1) in the form

$$q + q^3 = -\epsilon \sin \omega\tau - [\ddot{q} + 2\beta\dot{q}]. \tag{28.7.3}$$

If ω is small, we may expect that $q(\tau)$ will be slowly varying and therefore $\dot{q}$ and $\ddot{q}$ will be small. As an illustration of this expectation, Figure 7.1 displays the quantity $[\ddot{q} + 2\beta\dot{q}]$ as a function of τ for the periodic solution when $\omega = .01$ (and $\beta = .1$ and $\epsilon = 5.5$), the ω value associated with the left end of Figure 5.2. Note that this quantity is small, and numerical calculations verify that asymptotically it goes to zero linearly in ω as ω goes to zero.[8] If this

[8] We remark that the 'wiggles' (ringing) in Figure 7.1 are real. They also appear in $q(\tau)$. If Duffing's equation is linearized around the Ansatz (7.4), then $[\ddot{q} + 2\beta\dot{q}]$ appears as a driving term of the linearized equation. The wiggles are the (damped) response to the sharp peaks in the driving term. Examination of Figure 7.1 reveals that the wiggles occur just to the right of the peaks at $\tau = 0$ and $\tau \simeq 300$.

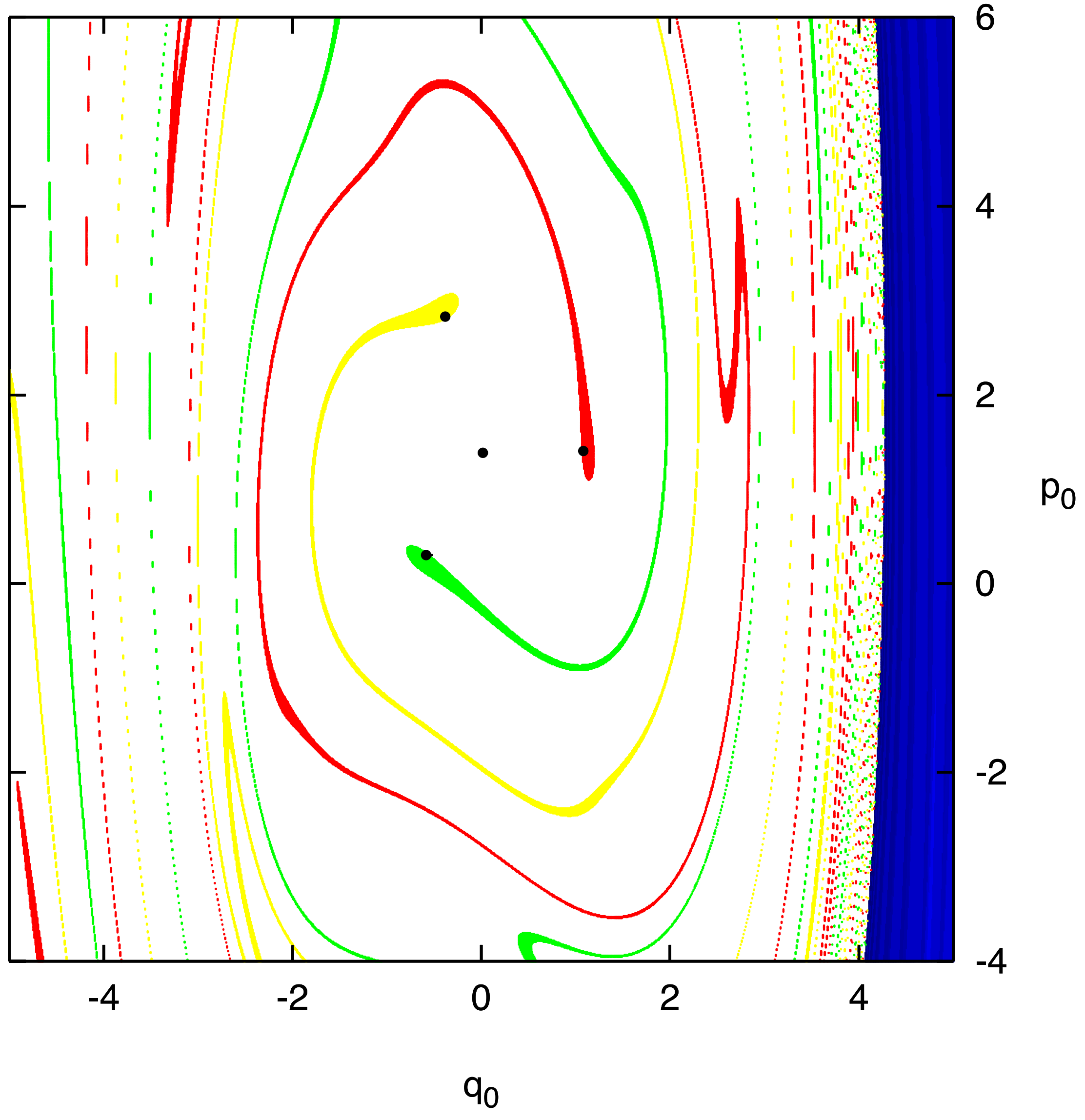


Figure 28.6.6: An enlargement of a portion of Figure 6.5. The fixed points w^1 and z^1, z^2, and z^3 themselves are shown as small black dots. The small black dot at the center of the figure is the fixed point w^1. Three small black dots near the ends of the red, green, and yellow filaments surround w^1. These are the $\mathcal{M}^3$ fixed points z^1, z^2, and z^3, respectively. The principal components of the period-three basins contain the fixed points z^1, z^2, and z^3. Note the crowding of the red, green, and yellow pieces of the period-three basins against the blue basin of w^2 (but still separated by ever smaller white areas) thereby making this basin boundary structure fractal.

quantity is neglected on the right side of (7.3), this relation becomes

$$q + q^3 = -\epsilon \sin \omega\tau \tag{28.7.4}$$

and consequently there is the result

$$q(0) + [q(0)]^3 = 0 \tag{28.7.5}$$

with the solution

$$q(0) = q_\infty = 0. \tag{28.7.6}$$

Moreover, if (7.4) holds, then there is also the relation

$$\dot{q} + 3q^2\dot{q} = -\omega\epsilon \cos \omega\tau, \tag{28.7.7}$$

from which, using (7.6), it follows that

$$\dot{q}(0) = -\omega\epsilon. \tag{28.7.8}$$

From (7.8), in turn, we infer that

$$\lim_{\omega\to 0} \dot{q}(0) = \lim_{\omega\to 0} p_\infty = 0. \tag{28.7.9}$$

Thus, (7.1) is correct.

To explore the $\omega \to \infty$ limit, rewrite (1.1) in the form

$$\ddot{q} = -\epsilon \sin \omega\tau - [2\beta\dot{q} + q + q^3]. \tag{28.7.10}$$

If ω is very large, we may expect that $q(\tau)$ will be rapidly varying and therefore $\ddot{q}$ will be very large compared to the other terms on the left side of (1.1). Correspondingly, the other terms will be small in the large ω limit. As an illustration of this expectation, Figure 7.2 displays the quantity $[2\beta\dot{q} + q + q^3]$ as a function of τ for the periodic solution when $\omega = 15$ (and $\beta = .1$ and $\epsilon = 5.5$), an ω value beyond the right end of Figure 5.2. We see that this quantity is indeed small, and further numerical work reveals that it vanishes as ω goes to infinity. If this quantity is neglected on the right side of (7.10), this relation becomes

$$\ddot{q} = -\epsilon \sin \omega\tau \tag{28.7.11}$$

with the solution

$$\dot{q}(\tau) = (1/\omega)\epsilon \cos \omega\tau, \tag{28.7.12}$$

$$q(\tau) = (1/\omega^2)\epsilon \sin \omega\tau. \tag{28.7.13}$$

From (7.13) we conclude that

$$\lim_{\omega\to\infty} q_\infty = \lim_{\omega\to\infty} q(0) = 0. \tag{28.7.14}$$

And from (7.12) we conclude that

$$\lim_{\omega\to\infty} p_\infty = \lim_{\omega\to\infty} \dot{q}(0) = 0. \tag{28.7.15}$$

We see that (7.2) also holds.

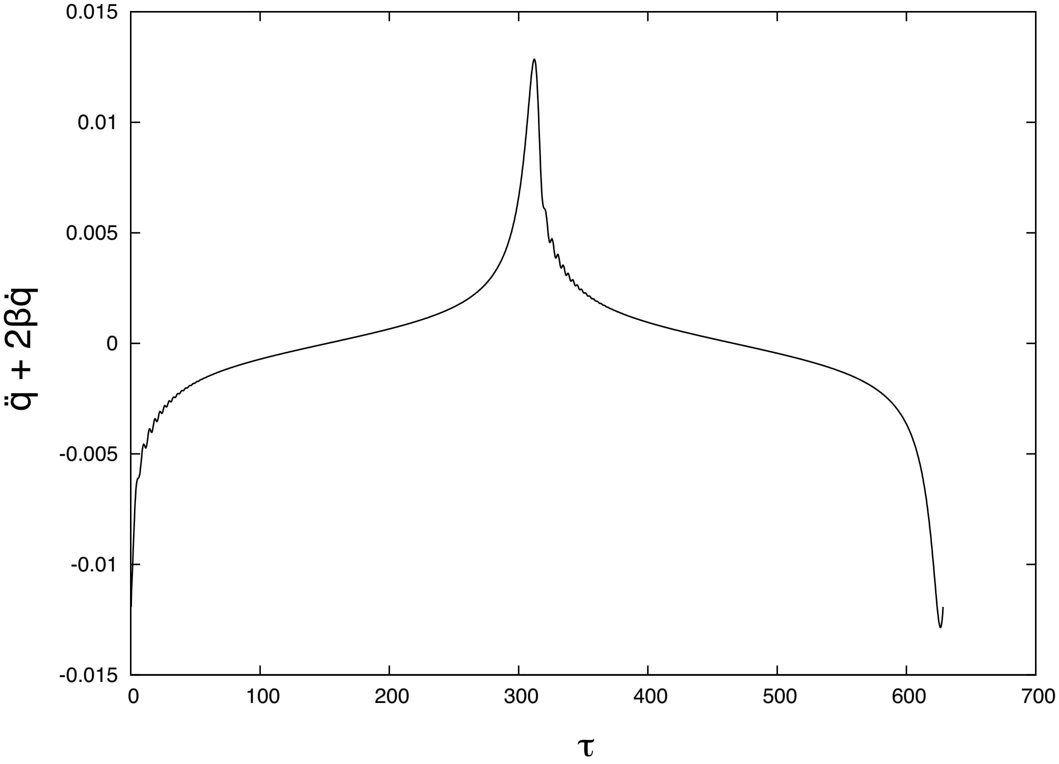


Figure 28.7.1: The quantity $[\ddot{q} + 2\beta\dot{q}]$ as a function of τ for the periodic solution when $\omega = .01$ (and $\beta = .1$ and $\epsilon = 5.5$).

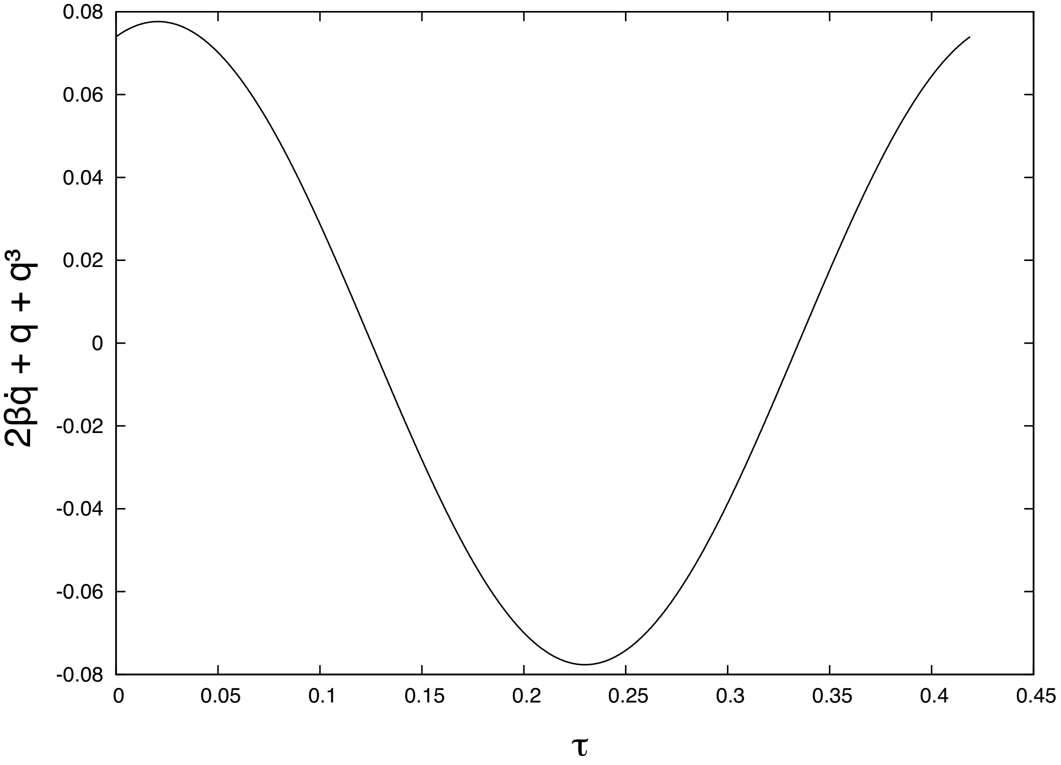


Figure 28.7.2: The quantity $[2\beta\dot{q} + q + q^3]$ as a function of τ for the periodic solution when $\omega = 15$ (and $\beta = .1$ and $\epsilon = 5.5$).

28.8 Period Doubling Cascade

We end our study of the Duffing equation by increasing ϵ from its earlier value $\epsilon = 5.5$ to larger values. Based on our experience so far, we might anticipate that the Feigenbaum diagram would become ever more complicated. That is indeed the case. Figure 8.1 displays q_∞ obtained numerically, when $\beta = 0.1$ and $\epsilon = 22.125$, as a function of ω for the range $\omega \in (0, 12)$. Evidently the behavior of the attractors for the stroboscopic Duffing map, which is what is shown in Figure 8.1, is extremely complicated. There are now a great many fixed points of $\mathcal{M}$ itself and various powers of $\mathcal{M}$. For small values of ω, and as in Figures 4.1, 5.1, and 5.2, there are many resonant peaks and numerous saddle-node and pitchfork bifurcations. For larger values of ω there are more complicated bifurcations. In this figure, and some subsequent figures, the coloring scheme is chosen to guide the eye in following bifurcation trees with colors changing when the period changes. Points with period one are colored red, and points of very high or no discernible period are colored black.

Let us begin by describing the more mundane features of the diagram. As already mentioned, at the left end of the diagram there is a series of saddle-node bifurcations as before, and the first one has moved to larger ω values so that it now occurs over the range $\omega \in (4, 10)$. Also, now more numerous, there are again bubbles for small values of ω. And the first bubble has also moved to larger ω values so that it now ends near $\omega = 2$. For the right end of the diagram, numerical study indicates that there are no new structures beyond $\omega = 12$ so that the asymptotic behavior (7.14) and (7.15) sets in for ω values larger than those shown.

There are also many new features. First, there are numerous higher-period fixed points that appear and disappear through blue-sky bifurcations. Some of them have been color coded in the figure. Moreover, some of the trails of the higher-period fixed points have little bubbles, and some of these little bubbles have an infinite number of sub-bubbles within them. That is, some of the higher-period fixed points (most evidently, those of period three) seem to have complete period doubling cascades with chaotic behavior at the end of the cascade. Figure 8.2 shows this behavior in further detail. It is too complicated to be studied further here.

What we do wish to note in Figure 8.2 is that what we have been calling the the first and second bubbles have within them the *beginnings* of period doubling cascades. Recall that each of these bubbles consists of three period-one fixed points. One of them is unstable, and hence invisible in a Feigenbaum diagram. The other two are stable, and their trails as ω is varied form the bubble. We see that these cascades *do not complete* but rather, after several period doublings, each cascade ceases and then successively undoes itself by a sequence of mergers to ultimately result in what is again a single stable period-one fixed point. This behavior is similar to that exhibited by the simple map described in Appendix J.

Suppose the value of ϵ is increased still further. Figure 8.3 shows the Feigenbaum diagram when $\epsilon = 25$. It looks similar to Figure 8.1 for $\epsilon = 22.125$. For example, there are again no new structures beyond $\omega = 12$ so that the asymptotic behavior (7.14) and (7.15) sets in for ω values larger than those shown. However, Figure 8.4, which is an enlargement of Figure 8.3, shows that the Feigenbaum cascades in the first and second bubbles now go to

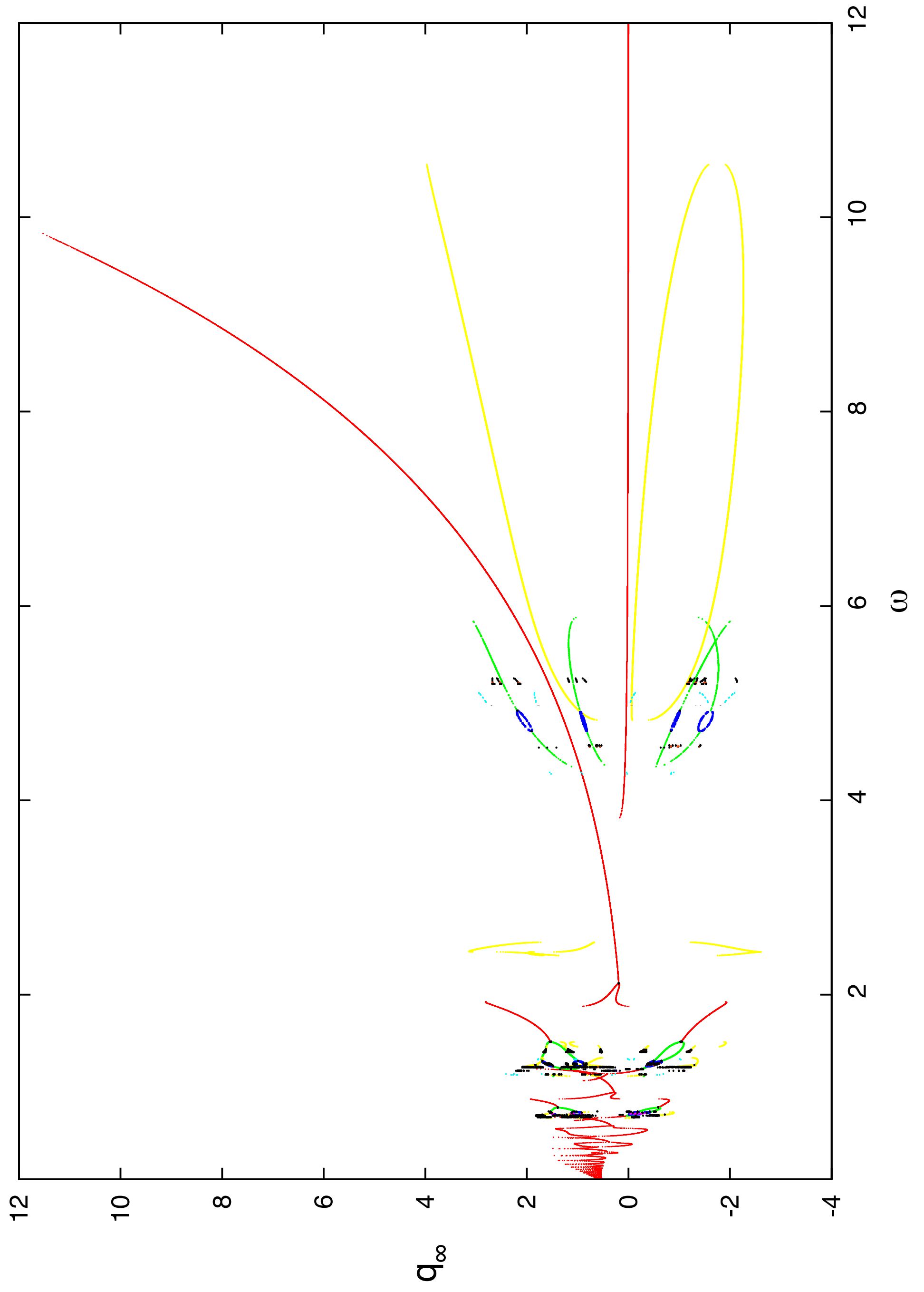


Figure 28.8.1: Feigenbaum diagram showing limiting values q_∞ as a function of ω (when $\beta = 0.1$ and $\epsilon = 22.125$) for the stroboscopic Duffing map.

completion.[9] Then, as ω is further increased, each complete period-doubling cascade again undoes itself. Note also that there appears to be a window of stability (near $\omega = 1.33$) within the completed cascade in the first bubble.

We close this lengthy discussion with a further brief study of some aspects of the period doubling cascade in the first bubble. Figure 8.5 shows an enlargement of part of this cascade. Specifically, it shows the beginnings of the period doubling cascades that occur in the bubble. The bubble has already formed due to the pitchfork bifurcation at the value $\omega \simeq 1.2284$, a value somewhat smaller than the ω values shown, and for which $(q_\infty, p_\infty) \simeq (.3982, 2.332)$. See Figure 8.4. Within the bubble there are two period-doubling cascades that begin when $\omega \simeq 1.268$: a lower one for which $(q_\infty, p_\infty) \simeq (-.228, 1.802)$, and an upper one for which $(q_\infty, p_\infty) \simeq (1.131, 2.215)$. It can be shown that the period-doubling cascades begin at the same ω values in both the upper and lower trails because of equivariance symmetry. See Sections 29.12.2 and 29.12.3. Unfortunately the view of the upper period-doubling bifurcation is somewhat complicated by the simultaneous appearance of a wisp of the trail of the stable stable fixed point associated with the second saddle-node bifurcation (the one mostly to the left of the first bubble) that appears to overlay the period-doubling bifurcation. This is an accident of our plotting scheme that happens to occur when $\epsilon = 25$. It would not overlay the period-doubling bifurcation if we had made a Feigenbaum diagram showing p_∞ versus ω instead of q_∞ versus ω. It also does not overlay the period-doubling bifurcation for other values of ϵ. Examine Figure 8.2 (for which $\epsilon = 22.125$) in the vicinity $\omega \approx 1.22$ and $q_\infty \approx 1.2$.

Let us now examine in more detail the period-doubling cascade that occurs in the upper part of the first bubble. See Figure 8.6. Evidently, for the smaller driving frequencies and in this region of phase space, there is a single period-one fixed point corresponding to an attractor. As the frequency is increased there is an infinite cascade of period doublings, and the motion appears chaotic by $\omega \simeq 1.29$. The resemblance between Figure 8.6 and Figure 1.2.4 for the logistic map is quite striking. In particular, numerical studies indicate that the frequencies ω_j at which successive bifurcations occur behave in a way analogous to (1.2.14) with (to within numerical accuracy) Feigenbaum's value of δ. Of course, as Figure 8.3 illustrates, the Duffing stroboscopic map is vastly more complicated than the logistic map, and its behavior resembles that of the logistic map only in a limited parameter range and only in a limited region of phase space. Note also that the logistic map acts on a one-dimensional space while the Duffing stroboscopic map acts on a two-dimensional space. Figure 1.2.4 tells the full story for the logistic map. By contrast, Figure 8.6 for the Duffing stroboscopic map is a projection onto the q axis of points in the two-dimensional q, p space. For full information one would need a figure made in the style of Figures 3.2 and 5.3.[10]

[9] Further numerical study indicates that the cascade in the first bubble is complete by the time $\epsilon = 22.25$ while that in the second bubble remains incomplete. Shortly thereafter the cascade in the second bubble also completes. Finally, numerical study reveals that the cascades associated with the higher-period blue-sky fixed points do complete for some ϵ values less that 22.125.

[10] See Figure 29.7.5 for such a figure in the case of the damped Hénon map.

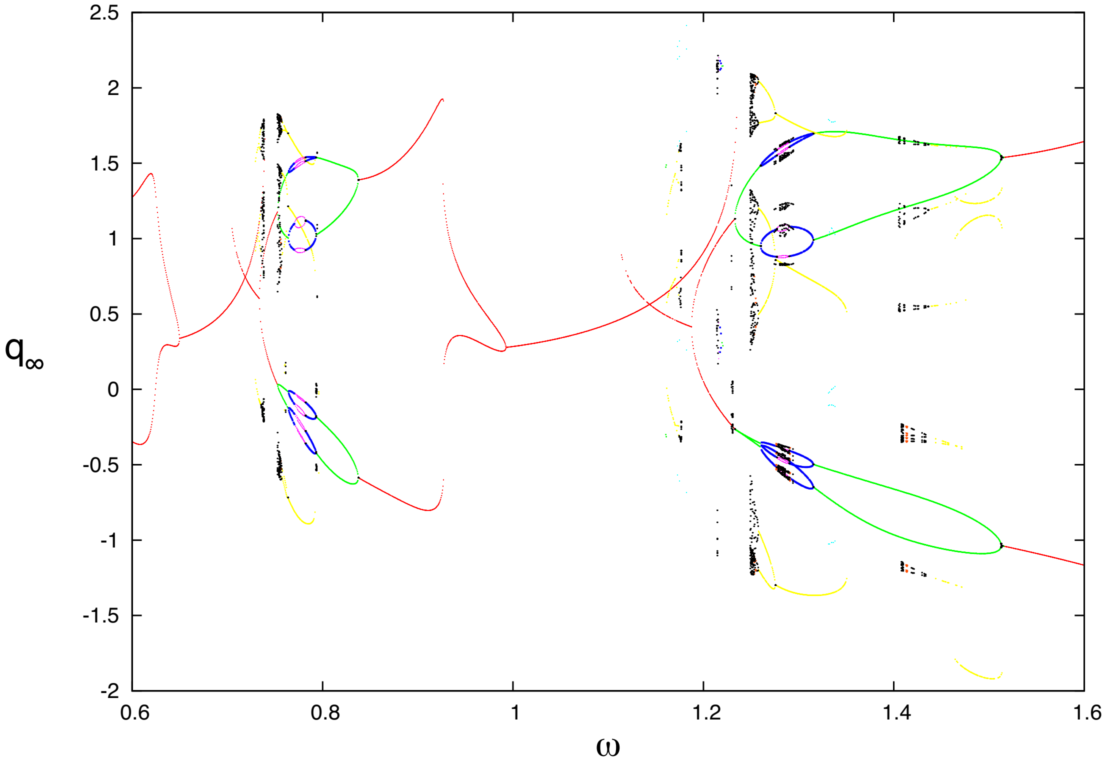


Figure 28.8.2: Enlargement of a portion of Figure 8.1 displaying limiting values of q_∞ as a function of ω (when $\beta = 0.1$ and $\epsilon = 22.125$) for the stroboscopic Duffing map. It shows part of the first bubble at the far right, the second bubble, and part of a third bubble at the far left. Examine the first and second bubbles. Each initially consists of two stable period-one fixed points. Each also contains the beginnings of period-doubling cascades. These cascades do not complete, but rather cease and then undo themselves by successive mergings to again result in a pair of stable period-one fixed points. There are also many higher-period fixed points and their associated cascades.

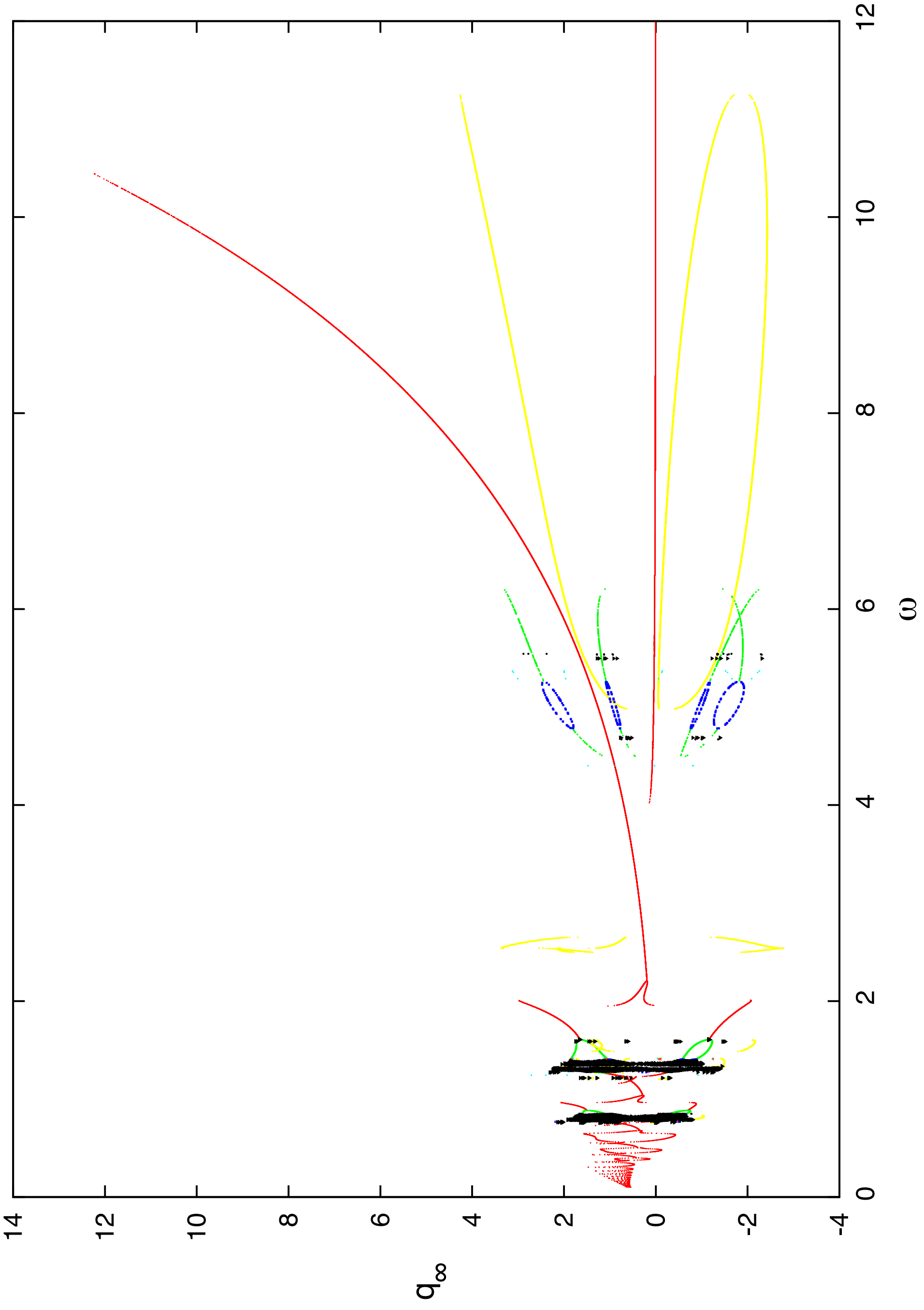


Figure 28.8.3: Feigenbaum diagram showing limiting values q_∞ as a function of ω (when $\beta = 0.1$ and $\epsilon = 25$) for the stroboscopic Duffing map.

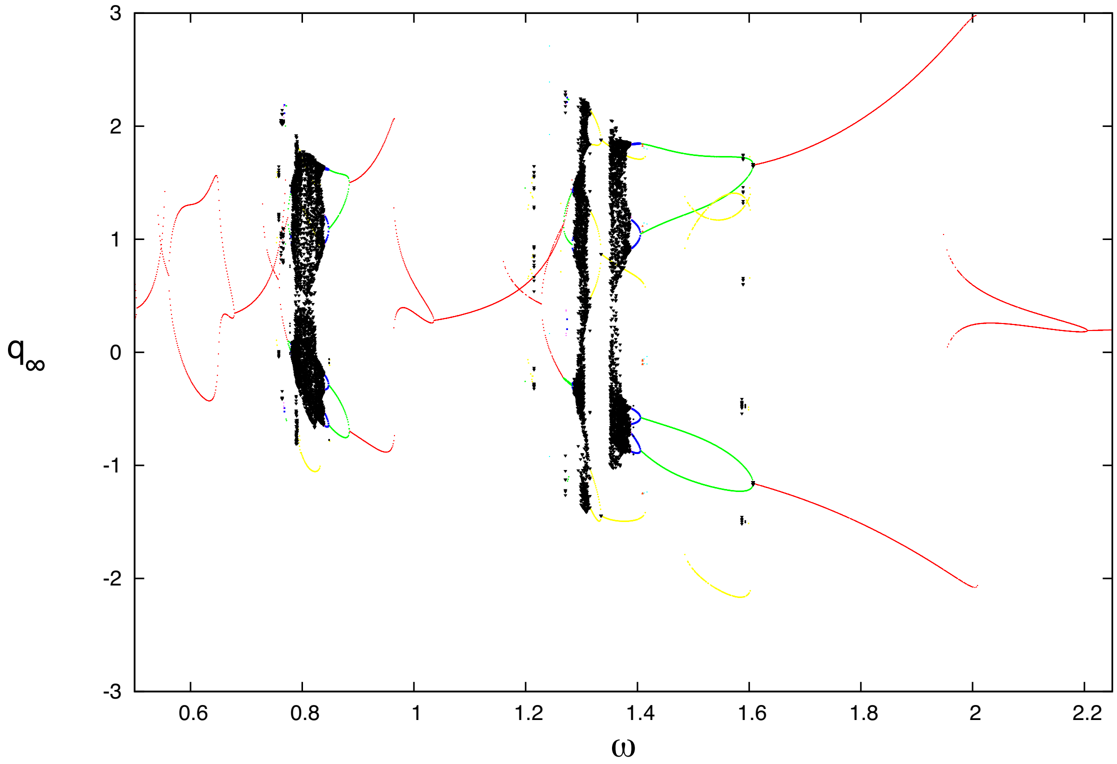


Figure 28.8.4: Enlargement of a portion Figure 8.3 showing the first, second, and third bubbles. The period-doubling cascades in each of the first and second bubbles now complete. Then they undo themselves as ω is further increased. There is no period doubling in the third bubble when $\epsilon = 25$.

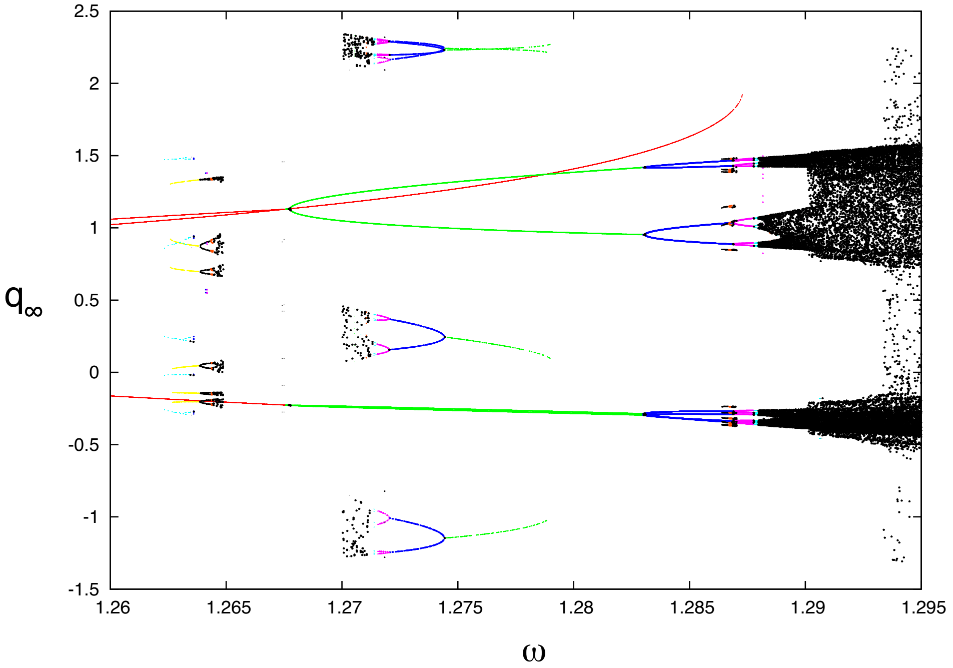


Figure 28.8.5: Detail of part of the first bubble in Figure 8.4 showing upper and lower infinite period-doubling cascades. Part of the trail of the stable fixed point associated with the second saddle-node bifurcation accidentally appears to overlay the upper period doubling bifurcation. Finally, associated with higher-period fixed points, there are numerous cascades and followed by successive mergings.

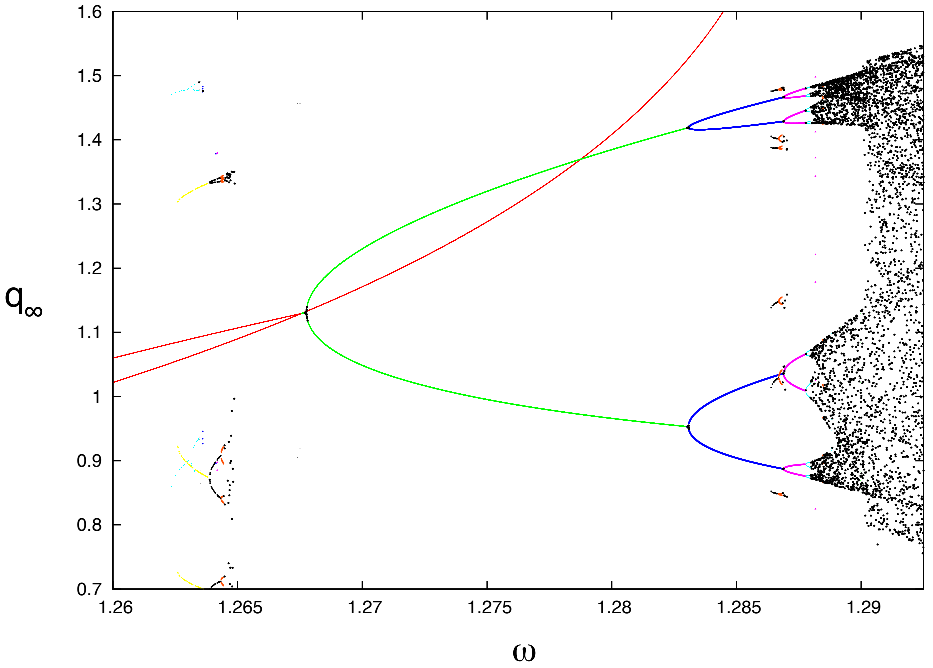


Figure 28.8.6: Detail of part of the upper cascade in Figure 8.5 showing an infinite period-doubling cascade, followed by chaos, for what was initially a stable period-one fixed point.

28.9 Strange Attractor

Finally, Figure 9.1 shows both the q_∞ and p_∞ values associated with Figure 8.6 when $\omega = 1.2902$ (and $\beta = .1$ and $\epsilon = 25$). Evidently the set of points q_∞, p_∞ for ω values just beyond the end of the cascade is quite complicated. By construction the set is an *attractor*. That is, points nearby this set are brought ever closer to the set under repeated action of $\mathcal{M}$. Moreover, numerical evidence suggests that this set has an infinite number of points and that the action of $\mathcal{M}$ on points in this set is to move them about within the set in a very complicated way. Finally, the set appears to be fractal. That is, it displays self similarity under repeated magnification. Therefore it may be an instance of what is called a *strange attractor*. For example, Figure 9.2 shows an enlargement of part of Figure 9.1. Repeated enlargement is expected to show a continued self-similar structure. For an instance of a strange attractor in the case of the damped Hénon map, see Sections 29.7 and 29.9. For more about the Duffing stroboscopic map, see Section 29.12. Finally, we warn the reader that there is no universal agreement among authors about the meaning of the adjective *strange* when applied to attractors. Some simply mean that the attractor has an infinite number of points. Some take fractal behavior to be the defining feature of what it means to be strange. Others require a sensitive dependence on initial conditions. Still others require what is technically called nonuniformly hyperbolic behavior.

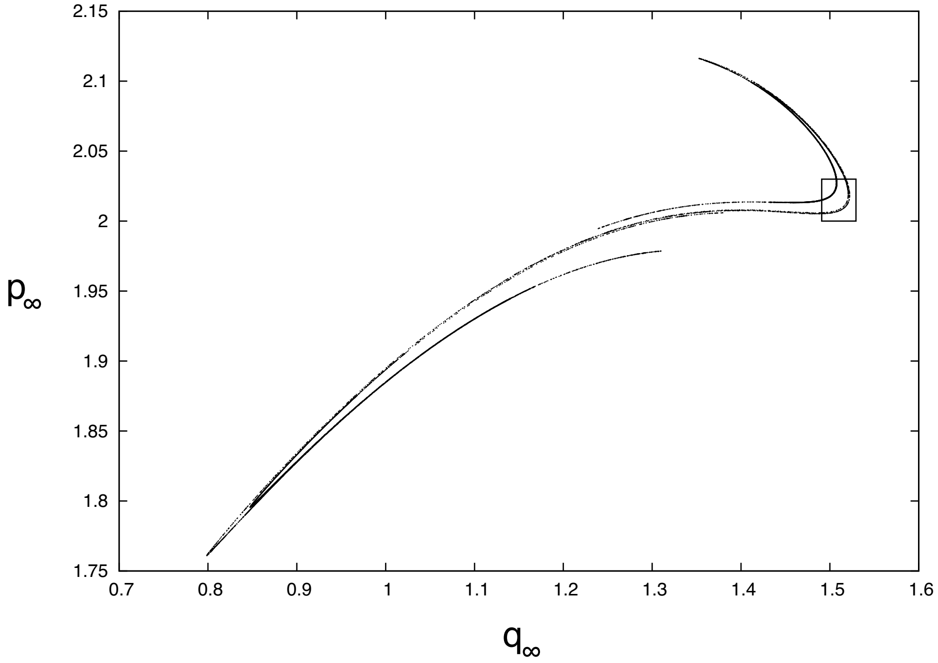


Figure 28.9.1: Limiting values of q_∞, p_∞ for the stroboscopic Duffing map when $\omega = 1.2902$ (and $\beta = .1$ and $\epsilon = 25$). They appear to lie on a strange attractor.

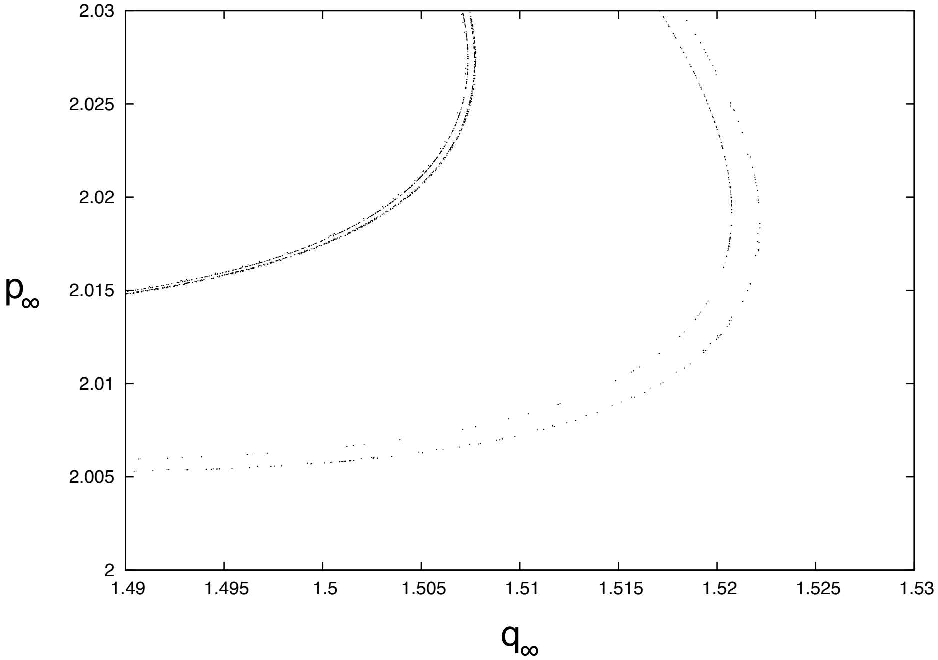


Figure 28.9.2: Enlargement of boxed portion of Figure 9.1 illustrating the beginning of self-similar fractal structure.

28.10 Acknowledgment

Dobrin Kaltchev made major contributions to the work of this chapter.

Bibliography

[1] G. Duffing, *Erzwungene Schwingungen bei veränderlicher Eigenfrequenz und ihre technische Bedeutung*, Braunschweig, Druck und Verlag von Friedr. Vieweg und Sohn (1918).

[2] I. Kovacic and M. Brennan, Edit., *The Duffing Equation: Nonlinear Oscillators and their Behaviour*, Wiley (2011).

[3] V. Barger and M. Olsson, *Classical Mechanics: A Modern Perspective*, Second Edition, McGraw Hill (1995). Some of our figures and some of our discussion of Duffing's equation were motivated by Sections 11.1 through 11.4 of this book. However, we found somewhat different numerical results. Presumably this is because, for the parameter values of interest, the long-term integration of Duffing's equation requires unusual care. For our calculations we used a 10th order Adams integrator started with 6th order Runge Kutta. Integrations were performed for several different step sizes in both double and quadruple (64 and 128 bit) precision to assure satisfactory convergence. See Chapter 2.

[4] C. Olson and M. Olsson, "Dynamical symmetry breaking and chaos in Duffing's equation", *Am. J. Phys.* **59**, p. 907 (1991).

[5] U. Parlitz and W. Lauterborn, "Superstructure in the bifurcation set of the Duffing equation", *Physics Letters* **107A**, p. 351 (1985).

[6] K. Becker and R. Seydel, "A Duffing equation with more than 20 branch points", *Lecture Notes in Mathematics* **878**, p. 98, E. Allgower et al., edit., Springer-Verlag (1981).

[7] I. Kyprianidis, "Dynamics of a nonlinear electrical oscillator described by Duffing's equation", Aristotle University of Thessaloniki (Greece) Physics Department Report, apparently unpublished but available on the Web.

[8] J. M. T. Thompson and H. B. Stewart, *Nonlinear Dynamics and Chaos*, Second Edition, John Wiley (2002).

[9] R. Van Dooren and M. De Groote, "Numerical Evidence of Feigenbaum's δ in Nonlinear Oscillations", *Journal of Computational Physics* **105**, 173-177 (1993).

[10] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag (1983).

[11] D. Jordan and P. Smith, *Nonlinear Ordinary Differential Equations*, Oxford University Press (1979).

[12] D. Ruelle, "What Is a Strange Attractor?", *Notices of the American Mathematical Society* **53**, p. 764, August 2006.

[13] Duffing equation Web sites. Any search engine will find several Web sites devoted to the Duffing equation.

Chapter 29

General Maps

Most of this book is devoted to the use of Lie methods for *symplectic* maps. However, Lie methods can also be used for general maps. This chapter treats general maps in an even number of variables and exploits the decomposition of general analytic vector fields (in an even number of variables) into Hamiltonian and non-Hamiltonian parts as described in Chapter 27. Section 1 describes the Lie factorization of general maps. Section 2 classifies all two-dimensional quadratic maps, and Section 3 studies, as an example, the Lie factorization of the general two-dimensional quadratic map. Sections 4 through 6 describe various concepts that are useful for studying maps including fixed points, the Poincaré index, stable and unstable manifolds, homoclinic points, and homoclinic tangles. Section 7 introduces the general Hénon map. Section 8 presents a preliminary study of the general Hénon map by finding and characterizing its fixed points, producing expansions about them, and factorizing the map about them. Section 9 describes period doubling and strange attractors for the general Hénon map, and Section 10 attempts to find integrals. Section 11 describes and studies quadratic symplectic maps in more than two dimensions. Section 12 obtains and studies Taylor approximations to the stroboscopic Duffing map. A final section discusses the expected analytic behavior of fixed points of a map and the eigenvalues of the linear part of the map.

29.1 Lie Factorization of General Maps

Section 7.6 showed that (modulo questions of convergence) any analytic symplectic map can be written in factorized product form. The purpose of this section is to explore what can be done for general analytic maps (in an even number of variables). That is, we will consider the group of all (analytic) diffeomorphisms. We will find that a general map can be written as a product of two factors. One factor is a product of exponentials of non-Hamiltonian Lie operators, and the second factor is an analytic symplectic map.

We begin by copying the expansion (7.6.1) for $\mathcal{M}$,

$$\bar{z}_a = \sum_b L_{ab} z_b + \sum_{bc} T_{abc} z_b z_c + \sum_{bcd} U_{abcd} z_b z_c z_d + \cdots , \qquad (29.1.1)$$

but now no longer require that $\mathcal{M}$ be symplectic. First of all the linear (matrix) part of $\mathcal{M}$, which we now write as L, need not be a symplectic matrix. However, we will require that L

have the symplectic polar decomposition

$$L = RQ. \tag{29.1.2}$$

Here R is symplectic and Q is a J-symmetric matrix that can be written in the form

$$Q = \exp(JA) \tag{29.1.3}$$

where A is antisymmetric. Sufficient conditions on L for (1.2) and (1.3) to hold were established in Section 4.3.

In (24.10.49) employ for F the matrix A of (1.3) and call the result $\mathcal{G}_2$,

$$\mathcal{G}_2 = \mathcal{L}_{\boldsymbol{g}} \tag{29.1.4}$$

with

$$g_b = \sum_d (JA)_{bd} z_d. \tag{29.1.5}$$

From the work of Section 24.10 we know that $\mathcal{G}_2$ is a a non-Hamiltonian vector field with the action

$$\mathcal{G}_2 z_b = \sum_d (JA)_{bd} z_d. \tag{29.1.6}$$

Recall (1.4) and see (24.10.51). It follows that there are the associated relations

$$(\mathcal{G}_2)^m z_b = \sum_d [(JA)^m]_{bd} z_d, \tag{29.1.7}$$

$$[\exp(\mathcal{G}_2)] z_b = \sum_d [\exp(JA)]_{bd} z_d = \sum_d Q_{bd} z_d. \tag{29.1.8}$$

Also, since R is symplectic, there are polynomials $f_2^a(z)$ and $f_2^c(z)$ associated with R. See Section 7.6. By using these polynomials and (1.8) we find the results

$$[\exp(\mathcal{G}_2)\exp(:f_2^c:)\exp(:f_2^a:)] z_b = [\exp(\mathcal{G}_2)] \sum_d R_{bd} z_d$$
$$= \sum_d R_{bd} [\exp(\mathcal{G}_2)] z_d = \sum_{de} R_{bd} Q_{de} z_e = \sum_e L_{be} z_e, \tag{29.1.9}$$

$$[\exp(-:f_2^a:)\exp(-:f_2^c:)\exp(-\mathcal{G}_2)] z_b = \sum_e (L^{-1})_{be} z_e. \tag{29.1.10}$$

These results can be written more compactly in the form

$$\exp(\mathcal{G}_2)\exp(:f_2^c:)\exp(:f_2^a:) z = Lz, \tag{29.1.11}$$

$$\exp(-:f_2^a:)\exp(-:f_2^c:)\exp(-\mathcal{G}_2) z = L^{-1} z. \tag{29.1.12}$$

Finally, since all the factors on the left side of (1.12) are Lie transformations, we have the result

$$\exp(-:f_2^a:)\exp(-:f_2^c:)\exp(-\mathcal{G}_2) h(z) = h(L^{-1}z), \tag{29.1.13}$$

for any function h.

Now, in the spirit of Section 7.6, apply $[\exp(-:f_2^a:)\exp(-:f_2^c:)\exp(-\mathcal{G}_2)]$ to both sides of (1.1). Doing so, and making use of (1.12) and (1.13), give the result

$$\exp(-:f_2^a:)\exp(-:f_2^c:)\exp(-\mathcal{G}_2)\overline{z}_b = z_b + r_b(>1). \tag{29.1.14}$$

Here, as before, the notation $r_b(>m)$ denotes *any* “remainder” series consisting of terms of degree higher than m. To proceed further, suppose the remainder terms $r_b(>1)$ are decomposed into second degree terms $g_b(2;z)$ and higher degree terms by writing the relations

$$r_b(>1) = g_b(2;z) + r_b(>2). \tag{29.1.15}$$

With this notation, we may rewrite (1.14) in the form

$$\exp(-:f_2^a:)\exp(-:f_2^c:)\exp(-\mathcal{G}_2)\overline{z}_b = z_b + g_b(2;z) + r_b(>2). \tag{29.1.16}$$

Let $\mathcal{L}_{\boldsymbol{g}^2}$ be the vector field defined by the equation

$$\mathcal{L}_{\boldsymbol{g}^2} = \sum_b g_b(2;z)(\partial/\partial z_b) = \boldsymbol{g}(2;z)\cdot\boldsymbol{\partial}. \tag{29.1.17}$$

Apply $\exp(-\mathcal{L}_{\boldsymbol{g}^2})$ to both sides of (1.16),

$$\begin{aligned}&\exp(-\mathcal{L}_{\boldsymbol{g}^2})\exp(-:f_2^a:)\exp(-:f_2^c:)\exp(-\mathcal{G}_2)\overline{z}_b = \\ &\exp(-\mathcal{L}_{\boldsymbol{g}^2})z_b + \exp(-\mathcal{L}_{\boldsymbol{g}^2})g_b(2;z) + \exp(-\mathcal{L}_{\boldsymbol{g}^2})r_b(>2).\end{aligned} \tag{29.1.18}$$

In analogy with (7.6.14) there is the general relation

$$(\mathcal{L}_{\boldsymbol{g}^m})f_n \in \mathcal{P}_{m+n-1}. \tag{29.1.19}$$

From this relation we deduce the results

$$\exp(-\mathcal{L}_{\boldsymbol{g}^2})z_b = z_b - (\mathcal{L}_{\boldsymbol{g}^2})z_b + (1/2)(\mathcal{L}_{\boldsymbol{g}^2})^2 z_b + \cdots = z_b - g_b(2;z) + r_b(>2), \tag{29.1.20}$$

$$\exp(-\mathcal{L}_{\boldsymbol{g}^2})g_b(2;z) = g_b(2;z) - (\mathcal{L}_{\boldsymbol{g}^2})g_b(2;z) + \cdots = g_b(2;z) + r_b(>2), \tag{29.1.21}$$

$$\exp(-\mathcal{L}_{\boldsymbol{g}^2})r_b(>2) = r_b(>2) - (\mathcal{L}_{\boldsymbol{g}^2})r_b(>2) + \cdots = r_b(>2) + r_b(>3). \tag{29.1.22}$$

Consequently, (1.18) can be rewritten in the form

$$\exp(-\mathcal{L}_{\boldsymbol{g}^2})\exp(-:f_2^a:)\exp(-:f_2^c:)\exp(-\mathcal{G}_2)\overline{z}_b = z_b + r_b(>2). \tag{29.1.23}$$

Now decompose the remainder terms $r_b(>2)$ into third degree terms $g_b(3,z)$ and higher degree terms by writing the relations

$$r_b(>2) = g_b(3;z) + r_b(>3). \tag{29.1.24}$$

Finally, substitute this decomposition into (1.23) to get the result

$$\exp(-\mathcal{L}_{\boldsymbol{g}^2})\exp(-:f_2^a:)\exp(-:f_2^c:)\exp(-\mathcal{G}_2)\overline{z}_b = z_b + g_b(3;z) + r_b(>3). \tag{29.1.25}$$

Clearly the process that led from (1.16) to (1.25) can be repeated at will. Consequently, there exist uniquely defined vector fields $\mathcal{L}_{\boldsymbol{g}^2}, \mathcal{L}_{\boldsymbol{g}^3}, \cdots$ such that for any n there is a relation of the form

$$\begin{aligned}&\exp(-\mathcal{L}_{\boldsymbol{g}^n})\exp(-\mathcal{L}_{\boldsymbol{g}^{n-1}})\cdots\exp(-\mathcal{L}_{\boldsymbol{g}^2})\times\\ &\exp(-:f_2^a:)\exp(-:f_2^c:)\exp(-\mathcal{G}_2)\bar{z}_b = z_b + r_b(>n).\end{aligned} \tag{29.1.26}$$

Now rewrite (1.26) in the form

$$\bar{z}_b = \exp(\mathcal{G}_2)\exp(:f_2^c:)\exp(:f_2^a:)\exp(\mathcal{L}_{\boldsymbol{g}^2})\cdots\exp(\mathcal{L}_{\boldsymbol{g}^n})z_b + r_b(>n), \tag{29.1.27}$$

and let $n \to \infty$. Then, if the remainder term tends to zero, we obtain the result

$$\bar{z} = \mathcal{M}z \tag{29.1.28}$$

with $\mathcal{M}$ expressed as the product

$$\mathcal{M} = \exp(\mathcal{G}_2)\exp(:f_2^c:)\exp(:f_2^a:)\exp(\mathcal{L}_{\boldsymbol{g}^2})\exp(\mathcal{L}_{\boldsymbol{g}^3})\cdots. \tag{29.1.29}$$

Otherwise the result is true only formally. In this latter case the infinite product (1.29) is also not convergent.

There is still more that can be done. From (24.14.3) we know that each vector field $\mathcal{L}_{\boldsymbol{g}^\ell}$ in (1.29) can be written in the form

$$\mathcal{L}_{\boldsymbol{g}^\ell} =: h_{\ell+1} : + \mathcal{G}^{\ell-1,1,0,\cdots} + \mathcal{G}^{\ell-1,0,0,\cdots} =: h_{\ell+1} : + \mathcal{G}''_{\ell+1}. \tag{29.1.30}$$

Here we have lumped the non-Hamiltonian parts together and simply called the result $\mathcal{G}''_{\ell+1}$. Note that the methods of Section 24.10 are adequate for this purpose since they allow us to find $h_{\ell+1}$ and $\mathcal{G}''_{\ell+1}$, and we do not need to further decompose $\mathcal{G}''_{\ell+1}$ into $\mathcal{G}^{\ell-1,1,0,\cdots}$ and $\mathcal{G}^{\ell-1,0,0,\cdots}$. With the aid of the BCH series (3.7.33) and (3.7.39) [or, equivalently the Zassenhaus series (8.8.1) and (8.8.2)] as applied to vector fields, and in view of the grading relation (24.3.5), we may rewrite the product $[\exp(\mathcal{L}_{\boldsymbol{g}^2})\exp(\mathcal{L}_{\boldsymbol{g}^3})\cdots]$ in the form

$$\exp(\mathcal{L}_{\boldsymbol{g}^2})\exp(\mathcal{L}_{\boldsymbol{g}^3})\cdots = \exp(\mathcal{G}'_3)\exp(\mathcal{G}'_4)\cdots\exp(:f_3:)\exp(:f_4):\cdots. \tag{29.1.31}$$

Here use has been made of the decomposition (1.30). We see that all non-Hamiltonian vector fields have been brought to the left, and all Hamiltonian vector fields have been brought to the right. Because of the grading relation (24.3.5), only a finite number of vector field commutators need be evaluated to compute each $\mathcal{G}'_m$ and f_m. Here we have dropped one prime from the $\mathcal{G}''_m$ and changed h_m to f_m to indicate that both are generally changed as a result of the various commutators involved.

As a result of the refactorization (1.31) the map $\mathcal{M}$ as given by (1.29) can be rewritten in the form

$$\begin{aligned}\mathcal{M} = \;&\exp(\mathcal{G}_2)\exp(:f_2^c:)\exp(:f_2^a:)\exp(\mathcal{G}'_3)\exp(\mathcal{G}'_4)\cdots\times\\ &\exp(:f_3:)\exp(:f_4:)\cdots.\end{aligned} \tag{29.1.32}$$

The terms $[(\mathcal{G}_3')\exp(\mathcal{G}_4')\cdots]$ can be moved past the terms $[\exp(:f_2^c:)\exp(:f_2^a:)]$ with the aid of the relations

$$\begin{aligned}&\exp(:f_2^c:)\exp(:f_2^a:)\exp(\mathcal{G}_3')\exp(\mathcal{G}_4')\cdots =\\ &\exp(:f_2^c:)\exp(:f_2^a:)\exp(\mathcal{G}_3')\exp(\mathcal{G}_4')\cdots\exp(-:f_2^a:)\exp(-:f_2^c:)\times\\ &\exp(:f_2^c:)\exp(:f_2^a:),\end{aligned}\tag{29.1.33}$$

and

$$\begin{aligned}&\exp(:f_2^c:)\exp(:f_2^a:)\exp(\mathcal{G}_3')\exp(\mathcal{G}_4')\cdots\exp(-:f_2^a:)\exp(-:f_2^c:) =\\ &\exp(\mathcal{G}_3)\exp(\mathcal{G}_4)\cdots\,.\end{aligned}\tag{29.1.34}$$

Here we have used the fact that the non-Hamiltonian vector fields are transformed among themselves under the action of $sp(2n,\mathbb{R})$, and have dropped the prime from each $\mathcal{G}_m'$ to indicate the result of this transformation. Putting everything together gives the final result

$$\begin{aligned}\mathcal{M} = \;&\exp(\mathcal{G}_2)\exp(\mathcal{G}_3)\exp(\mathcal{G}_4)\cdots\times\\ &\exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_3:)\exp(:f_4:)\cdots\,.\end{aligned}\tag{29.1.35}$$

We see that, as promised, $\mathcal{M}$ has been written as a product of two factors. The first factor is a product of exponentials of non-Hamiltonian vector fields, and the second is the by now familiar general (origin preserving) symplectic map.

The last item to discuss is the inclusion of translations. That is, suppose the map $\mathcal{M}$ is of the general form (7.7.7). In that case, in view of the work of Section 7.1, there is the simple modification

$$\begin{aligned}\mathcal{M} = \;&\exp(\mathcal{G}_2)\exp(\mathcal{G}_3)\exp(\mathcal{G}_4)\cdots\times\\ &[\exp(:f_2^c:)\exp(:f_2^a:)\exp(:f_3:)\exp(:f_4:)\cdots]\exp(:g_1:).\end{aligned}\tag{29.1.36}$$

Exercises

29.1.1. For the complex logistic map (1.2.29) in the two-dimensional real form given by (1.2.102) and (1.2.103) make the identifications

$$x = q,\tag{29.1.37}$$

$$y = p,\tag{29.1.38}$$

so that in terms of q and p the map takes the form

$$q_{n+1} = \alpha q_n - \beta p_n - \alpha(q_n^2 - p_n^2) + 2\beta q_n p_n,\tag{29.1.39}$$

$$p_{n+1} = \beta q_n + \alpha p_n - \beta(q_n^2 - p_n^2) - 2\alpha q_n p_n.\tag{29.1.40}$$

Make a polar decomposition of the control parameter

$$\gamma = \alpha + i\beta\tag{29.1.41}$$

by writing

$$\gamma = \rho \exp(i\phi). \tag{29.1.42}$$

Show that in terms of the parameters ρ, ϕ the linear part of the map about the origin (which is a fixed point) takes the form

$$R = \rho \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}. \tag{29.1.43}$$

29.1.2. Forest map.

29.2 Classification of General Two-Dimensional Quadratic Maps

The general quadratic map, even in the case of only two variables, is a complicated object. We know, for example, that the complex logistic map produces fractal sets in the mapping plane and the fantastically complicated Mandelbrot set in the control plane. See Section 1.2 and Exercise 1.2.2. We also suspect that the Hénon map, another two-dimensional quadratic map, exhibits even more complicated behavior. Again see Section 1.2. In this section we will make a preliminary classification of general two-dimensional quadratic maps. What we will do is make an *affine* transformation on both z and $\bar{z}$ (the same transformation on each), and consider two maps *equivalent* if one can be transformed into the other by such a change of variables. (An affine transformation is a translation followed by a linear transformation. These transformations evidently form a group called the affine group. Moreover, we observe that if a map is quadratic, then it will remain quadratic under any change of variables that is an affine transformation.) Indeed, the two maps will be *conjugate* as described in Chapter 19.

Let $\mathcal{N}$ denote the nonlinear quadratic map

$$\begin{aligned} \bar{q} &= q + b_1 q^2 + 2b_2 qp + b_3 p^2 = q + (z, Bz), \\ \bar{p} &= p + c_1 q^2 + 2c_2 qp + c_3 p^2 = p + (z, Cz). \end{aligned} \tag{29.2.1}$$

Here we have used the notation $z = (q, p)$ and

$$\begin{aligned} B &= \begin{pmatrix} b_1 & b_2 \\ b_2 & b_3 \end{pmatrix}, \\ C &= \begin{pmatrix} c_1 & c_2 \\ c_2 & c_3 \end{pmatrix}. \end{aligned} \tag{29.2.2}$$

Let $\mathcal{R}$ denote the linear map

$$\begin{aligned} \bar{\bar{q}} &= r\bar{q} + s\bar{p}, \\ \bar{\bar{p}} &= t\bar{q} + u\bar{p}. \end{aligned} \tag{29.2.3}$$

Finally, let $\mathcal{T}$ denote the translation map

$$\bar{\bar{\bar{q}}} = \bar{\bar{q}} + d,$$

$$\overset{\equiv}{p}=\overset{=}{p}+e. \tag{29.2.4}$$

Then we speculate that the *general quadratic map*, which we will denote as $\mathcal{M}_{gq}$ can be written in the form

$$\mathcal{M}_{gq} = \mathcal{TRN}. \tag{29.2.5}$$

Here we have employed the usual mathematical ordering convention as described in the beginning of Section 8.3. Carrying out the operations indicated in (2.5) shows that $\mathcal{M}_{gq}$ has the explicit representation given by the relations below.

Action of $\mathcal{M}_{gq}$:

$$\overset{\equiv}{q}= d + rq + sp + (z, B'z),$$

$$\overset{\equiv}{p}= e + tq + up + (z, C'z), \tag{29.2.6}$$

where B' and C' are the symmetric matrices

$$B' = rB + sC,$$

$$C' = tB + uC. \tag{29.2.7}$$

We see that $\mathcal{M}_{gq}$ will in fact be the general quadratic map in two variables provided B' and C' can be *any* two symmetric matrices. Let R be the matrix defined by the equation

$$R = \begin{pmatrix} r & s \\ t & u \end{pmatrix}, \tag{29.2.8}$$

and let V be the vector with *matrix* entries B and C,

$$V = \begin{pmatrix} B \\ C \end{pmatrix}, \tag{29.2.9}$$

Define V' analogously. With this notation (2.7) can be rewritten in the form

$$V' = RV. \tag{29.2.10}$$

This relation can be inverted if R is invertible ($\det R \neq 0$),

$$V = R^{-1}V'. \tag{29.2.11}$$

Evidently if R is invertible, we can always find symmetric matrices B and C such that B' and C' are any desired symmetric matrices: we simply use the matrices B and C given by (2.11). Note also that R is the Jacobian matrix of $\mathcal{M}_{gq}$ at the origin,

$$R = M_{gq}(0). \tag{29.2.12}$$

See Exercise 1.4.6 and Section 6.1. Finally we remark that, driven by a shortage of symbols, in this section and in some subsequent sections we have used $\mathcal{R}$ and R to denote *general* linear maps and matrices rather than *symplectic* linear maps and matrices as was our previous convention.

For economy of notation, let us rewrite $\mathcal{M}_{gq}$, as given by (2.6), in the form

Action of $\mathcal{M}_{gq}$:

$$\bar{q} = d + rq + sp + (z, Bz),$$
$$\bar{p} = e + tq + up + (z, Cz), \tag{29.2.13}$$

where, again, B and C are arbitrary symmetric matrices. At this point, if not before, it is evident that $\mathcal{M}_{gq}$ is specified by 12 parameters: d and e for the translation part, r through u for the linear part, and the two symmetric matrices B and C for the nonlinear part. We now seek to simplify $\mathcal{M}_{gq}$, as given by (2.13), by a suitable change of variables, and we will allow this change of variables to be parameter dependent. We will then learn that 6 of the dimensions in the 12-parameter space may be viewed as associated with the choice of variables, and 6 may be viewed as being intrinsic to $\mathcal{M}_{gq}$ itself. And, as a result of suitably changing variables, we will find a *transformed* map $\mathcal{M}_{gq}^{tr}$ that depends on 6 parameters.

Begin with a *displacement* (translation) transformation of variables by writing

$$q = q' + \alpha,$$
$$p = p' + \beta; \tag{29.2.14}$$
$$\bar{q} = \bar{q}' + \alpha,$$
$$\bar{p} = \bar{p}' + \beta. \tag{29.2.15}$$

We also express (2.14) and (2.15) in more compact form by writing

$$z = z' + \gamma \ , \ \ \bar{z} = \bar{z}' + \gamma \tag{29.2.16}$$

where γ is a two-vector with components α and β. Under this change of variables (2.13) takes the form

$$\bar{q}' = -\alpha + d + r\alpha + s\beta + (\gamma, B\gamma) + rq' + sp' + 2(z', B\gamma) + (z', Bz'),$$
$$\bar{p}' = -\beta + e + t\alpha + u\beta + (\gamma, C\gamma) + tq' + up' + 2(z', C\gamma) + (z', Cz'). \tag{29.2.17}$$

Suppose we now require that the transformed map given by (2.17) have no *constant* terms. That is, the transformed map should have no translation part, and therefore should send the origin into itself. This requirement leads to the equations

$$\alpha = d + r\alpha + s\beta + (\gamma, B\gamma), \tag{29.2.18}$$

$$\beta = e + t\alpha + u\beta + (\gamma, C\gamma). \tag{29.2.19}$$

We see, as might have been expected, that we have made the equivalent requirement that γ be a *fixed* point of $\mathcal{M}_{gq}$. That is, γ is a solution of the equations

$$q = d + rq + sp + (z, Bz), \tag{29.2.20}$$

$$p = e + tq + up + (z, Cz). \tag{29.2.21}$$

Taken separately, (2.20) and (2.21) each describe conic sections in the q, p plane; and taken jointly they specify the intersection of these two conic sections. The nature of these conic sections is governed by the *discriminants* (determinants) of the quadratic forms (z, Bz) and (z, Cz). For example, if

$$\det B = b_1 b_3 - b_2^2 > 0, \text{ (2.20) describes an ellipse;}$$

$$\det B = 0, \text{ (2.20) describes a parabola;}$$

$$\det B < 0, \text{ (2.20) describes a hyperbola.}$$

According to a theorem of *Bézout*, two polynomials of degrees m and n in two variables intersect in mn points that may be complex, may be at infinity, and may be repeated. For two conic sections, $m = n = 2$; so we expect *four* intersections. For our purposes we assume that the coefficients that appear in (2.20) and (2.21), which are all taken to be real, are such that there is at least one such *real* intersection so that $\mathcal{M}_{gq}$ has a real fixed point. Upon taking γ to be such a real fixed point, (2.17) takes the general form

$$\bar{q}' = \tilde{r}q' + \tilde{s}p' + (z', Bz'),$$

$$\bar{p}' = \tilde{t}q' + \tilde{u}p' + (z', Cz'). \tag{29.2.22}$$

Here we have replaced r, s, t, u by $\tilde{r}, \tilde{s}, \tilde{t}, \tilde{u}$ to take into account the linear terms $2(z', B\gamma)$ and $2(z', C\gamma)$ that "feed down" from the quadratic terms (z, Bz) and (z, Cz) as a result of the displacement (2.14). Since γ is assumed to be real, the quantities $\tilde{r}, \tilde{s}, \tilde{t}, \tilde{u}$ will also be real. Upon dropping the prime and tilde notation, we see that we are interested in studying maps of the simpler form

$$\bar{q} = rq + sp + (z, Bz),$$

$$\bar{p} = tq + up + (z, Cz). \tag{29.2.23}$$

Using the notation (2.8) and (2.9), the map (2.23) can be written more compactly in the form

$$\bar{z} = Rz + V(z). \tag{29.2.24}$$

Now make the linear change of variables

$$q = a_{11}q' + a_{12}p',$$

$$p = a_{21}q' + a_{22}p', \tag{29.2.25}$$

which can be written more compactly as

$$z = Az' \tag{29.2.26}$$

where A is the matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}. \tag{29.2.27}$$

Corresponding to (2.26), we also define $\bar{z}'$ by writing

$$\bar{z} = A\bar{z}'. \tag{29.2.28}$$

With this change of variables the map (2.24) takes the form

$$A\bar{z}' = RAz' + V(Az'), \tag{29.2.29}$$

or

$$\bar{z}' = A^{-1}RAz' + A^{-1}V(Az'). \tag{29.2.30}$$

Under the assumption that the map is orientation preserving (see Exercise 1.4.6), R will have a positive determinant. In this case, define a matrix S by the rule

$$S = R/(\det R)^{1/2}. \tag{29.2.31}$$

According to Exercise 3.1.3, S will be symplectic. Moreover we know, from the discussion of Section 3.4 for the instance of 2×2 symplectic matrices, that there are five possible cases for S. For each of these cases we can select a suitable matrix A such that the *transformed* S given by

$$S^{tr} = A^{-1}SA \tag{29.2.32}$$

has a particularly simple form, which we will call a *normal* form. For example, if the eigenvalues of S lie on the unit circle (Case 3, elliptic), S can be transformed so that S^{tr} is of the form (3.5.58). As a second example, suppose the eigenvalues of S are positive (Case 1, hyperbolic). Then there is a choice of A such that S^{tr} takes the diagonal form

$$S^{tr} = \begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix}. \tag{29.2.33}$$

To continue with the hyperbolic case, let $D(\mu,\nu)$ denote the diagonal matrix

$$D(\mu,\nu) = \begin{pmatrix} \mu & 0 \\ 0 & \nu \end{pmatrix}. \tag{29.2.34}$$

Upon combining (2.31) through (2.33) we see that in the hyperbolic case there is an A (which is real) such that

$$A^{-1}RA = D(\mu,\nu) \tag{29.2.35}$$

where μ and ν are the eigenvalues of R. With this choice of A the map (2.30) takes the form

$$\bar{z}' = D(\mu,\nu)z' + V'(z') \tag{29.2.36}$$

where here

$$V'(z') = A^{-1}V(Az'). \tag{29.2.37}$$

To improve notation, we again drop the primes to rewrite (2.36) in the general form

$$\bar{z} = D(\mu,\nu)z + V(z). \tag{29.2.38}$$

There is one further simplifying transformation that can be made. Make the further linear change of variables

$$z = D(\sigma,\tau)z', \tag{29.2.39}$$

$$\bar{z} = D(\sigma,\tau)\bar{z}'. \tag{29.2.40}$$

With this change of variables the map (2.38) takes the form

$$\begin{aligned}\bar{z}' &= D^{-1}(\sigma,\tau)D(\mu,\nu)D(\sigma,\tau)z' + D^{-1}(\sigma,\tau)V[D(\sigma,\tau)z'] \\ &= D(\mu,\nu)z' + D^{-1}(\sigma,\tau)V[D(\sigma,\tau)z'].\end{aligned} \tag{29.2.41}$$

Here we have used the fact that diagonal matrices commute. We will call this map the *transformed general quadratic* map and denote it by $\mathcal{M}_{gq}^{tr}$.

Let us again drop primes and write out (2.41), the transformed two-variable general quadratic map $\mathcal{M}_{gq}^{tr}$, more explicitly in the form below.

Action of $\mathcal{M}_{gq}^{tr}$:

$$\begin{aligned}&\bar{q} = \mu q + (1/\sigma)[b_1\sigma^2 q^2 + 2b_2\sigma\tau qp + b_3\tau^2 p^2], \\ &\bar{p} = \nu p + (1/\tau)[c_1\sigma^2 q^2 + 2c_2\sigma\tau qp + c_3\tau^2 p^2].\end{aligned} \tag{29.2.42}$$

We now see that σ and τ can be selected in such a way that *any two* of the six coefficients b_1 through b_3 and c_1 through c_3 can be normalized to one. Thus, in the hyperbolic case, there is a six-parameter family of maps labeled by μ and ν and *four* of the six coefficients b_1 through b_3 and c_1 through c_3 with the remaining two coefficients set to 1.

A broad picture now emerges: The general two-dimensional quadratic map $\mathcal{M}_{gq}$ (assuming it has a real fixed point) can be classified according to the nature of its linear part about the fixed point. This linear part can next be brought to a normal form that is generally labelled by two parameters. There is then an associated two parameter family of linear transformations that leaves the normal form unchanged, and these transformations can be used to simplify the nonlinear part V of the map so that it is described by $6-2=4$ parameters. Thus, after a suitable choice of variables, the general two-dimensional quadratic map $\mathcal{M}_{gq}$ is brought to the form $\mathcal{M}_{gq}^{tr}$, and this transformed map is described by 6 parameters: 2 for the linear part in normal form, and 4 for the nonlinear part.

Exercises

29.2.1. Application to harmonic maps and the complex logistic map.

29.2.2. Application to Tinkerbell map.

29.3 Lie Factorization of General Two-Dimensional Quadratic Maps

We next turn to the problem of factorizing $\mathcal{M}_{gq}$ as given by (2.5) in Lie form. We know that $\mathcal{T}$ can always be written in Lie form. See Section 7.7. We also know that $\mathcal{R}$ can be written as a product of at most three Lie transformations if det $R > 0$. See the discussion at the end of Section 4.3 and the beginning of Section 18.1. What remains is to explore the Lie factorization of $\mathcal{N}$. See (2.1). In general it appears that $\mathcal{N}$ will have an infinite number of Lie factors. See Exercise 3.1. In this section we will study under what conditions a quadratic map in two variables can be factored into a product of a *finite* number of Lie transformations. In particular, we will require that $\mathcal{N}$ have a *single-factor* Lie representation.

Let $\mathcal{L}_{\boldsymbol{g}^2}$ be the Lie operator defined by the equation

$$\begin{aligned}\mathcal{L}_{\boldsymbol{g}^2} &= (z,Bz)(\partial/\partial q)+(z,Cz)(\partial/\partial p)\\ &= (b_1q^2+2b_2qp+b_3p^2)(\partial/\partial q)+(c_1q^2+2c_2qp+c_3p^2)(\partial/\partial p). \qquad (29.3.1)\end{aligned}$$

Then we have the result

$$\begin{aligned}\exp(\mathcal{L}_{\boldsymbol{g}^2})q &= q+\mathcal{L}_{\boldsymbol{g}^2}q+(1/2!)(\mathcal{L}_{\boldsymbol{g}^2})^2q+\cdots\\ &= q+(z,Bz)+(1/2!)(\mathcal{L}_{\boldsymbol{g}^2})^2q+\cdots\\ &= \bar{q}+(1/2!)(\mathcal{L}_{\boldsymbol{g}^2})^2q+\cdots,\end{aligned}$$

$$\begin{aligned}\exp(\mathcal{L}_{\boldsymbol{g}^2})p &= p+\mathcal{L}_{\boldsymbol{g}^2}p+(1/2!)(\mathcal{L}_{\boldsymbol{g}^2})^2p+\cdots\\ &= p+(z,Cz)+(1/2!)(\mathcal{L}_{\boldsymbol{g}^2})^2p+\cdots\\ &= \bar{p}+(1/2!)(\mathcal{L}_{\boldsymbol{g}^2})^2p+\cdots. \qquad (29.3.2)\end{aligned}$$

Here we have used (2.1). We conclude that $\mathcal{N}$ will have the single-factor Lie representation

$$\mathcal{N}=\exp(\mathcal{L}_{\boldsymbol{g}^2}) \qquad (29.3.3)$$

if $\mathcal{L}_{\boldsymbol{g}^2}$ satisfies the relations

$$(\mathcal{L}_{\boldsymbol{g}^2})^2q=0, \qquad (29.3.4)$$

$$(\mathcal{L}_{\boldsymbol{g}^2})^2p=0, \qquad (29.3.5)$$

for then each series in (3.2) will terminate after the first two terms. Otherwise, $\mathcal{N}$ will in general have an infinite number of Lie factors. We also note for future reference that $\mathcal{L}_{\boldsymbol{g}^2}$ has the decomposition

$$\mathcal{L}_{\boldsymbol{g}^2}=:h_3:+\mathcal{G}^1 \qquad (29.3.6)$$

where

$$h_3=(1/3)[c_1q^3+(2c_2-b_1)q^2p+(c_3-2b_2)qp^2-b_3p^3], \qquad (29.3.7)$$

$$\mathcal{G}^1=(2/3)[(b_1+c_2)q+(b_2+c_3)p]\Sigma. \qquad (29.3.8)$$

See (21.3.15) and (21.3.16).

Imposition of (3.4) produces the relations

$$b_1^2+b_2c_1=0, \qquad (29.3.9)$$

$$3b_1b_2+2b_2c_2+b_3c_1=0, \qquad (29.3.10)$$

$$2b_2^2+b_1b_3+b_2c_3+2b_3c_2=0, \qquad (29.3.11)$$

$$b_2b_3+b_3c_3=b_3(b_2+c_3)=0; \qquad (29.3.12)$$

and imposition of (3.5) produces the relations

$$b_1c_1+c_1c_2=c_1(b_1+c_2)=0, \qquad (29.3.13)$$

$$b_1c_2 + 2b_2c_1 + c_1c_3 + 2c_2^2 = 0, \tag{29.3.14}$$

$$2b_2c_2 + b_3c_1 + 3c_2c_3 = 0, \tag{29.3.15}$$

$$b_3c_2 + c_3^2 = 0. \tag{29.3.16}$$

Let us focus on those generic solutions of (3.9) through (3.16) for which $b_3 \neq 0$ and $c_1 \neq 0$. In this case, from (3.12) and (3.13), we find the results

$$b_2 + c_3 = 0 \text{ or } b_2 = -c_3, \tag{29.3.17}$$

$$b_1 + c_2 = 0 \text{ or } c_2 = -b_1. \tag{29.3.18}$$

Inspection of (3.7) and (3.8) shows that in this case $\mathcal{G}^1$ vanishes,

$$\mathcal{G}^1 = 0, \tag{29.3.19}$$

and h_3 takes the form

$$h_3 = (1/3)(c_1q^3 - 3b_1q^2p + 3c_3qp^2 - b_3p^3). \tag{29.3.20}$$

Also, the remaining relations in the collection (3.9) through (3.16) take the form

$$b_1^2 - c_1c_3 = 0, \tag{29.3.21}$$

$$-b_1c_3 + b_3c_1 = 0, \tag{29.3.22}$$

$$c_3^2 - b_1b_3 = 0. \tag{29.3.23}$$

Without loss of generality we may make the Ansätze

$$c_1 = \alpha^3 \ , \ b_3 = \beta^3. \tag{29.3.24}$$

We then find that (3.21) through (3.23) have the unique solution

$$b_1 = \alpha^2\beta, \tag{29.3.25}$$

$$c_3 = \alpha\beta^2; \tag{29.3.26}$$

and the remaining quantities have the values

$$c_2 = -\alpha^2\beta, \tag{29.3.27}$$

$$b_2 = -\alpha\beta^2. \tag{29.3.28}$$

As a result of these relations h_3 takes the form

$$h_3 = (1/3)(\alpha q - \beta p)^3. \tag{29.3.29}$$

Note that h_3 is the cube of a first-order polynomial. Consequently, we imediately have the properties

$$: h_3 : q = [h_3, q] = \beta(\alpha q - \beta p)^2, \tag{29.3.30}$$

$$: h_3 :^2 q = (\beta/3)[(\alpha q - \beta p)^3, (\alpha q - \beta p)^2] = 0; \tag{29.3.31}$$

$$: h_3 : p = [h_3, p] = \alpha(\alpha q - \beta p)^2, \tag{29.3.32}$$

$$: h_3 :^2 p = (\alpha/3)[(\alpha q - \beta p)^3, (\alpha q - \beta p)^2] = 0. \tag{29.3.33}$$

Correspondingly, $\mathcal{N}$ is a *symplectic* map with the Lie representation

$$\mathcal{N} = \exp : (1/3)(\alpha q - \beta p)^3 :, \tag{29.3.34}$$

and the Taylor expansion of this map terminates beyond second degree terms as a result of (3.31) and (3.33). This expansion is given explicitly by the relations

$$\bar{q} = q + \beta(\alpha q - \beta p)^2 = q + \alpha^2\beta q^2 - 2\alpha\beta^2 qp + \beta^3 p^2, \tag{29.3.35}$$

$$\bar{p} = p + \alpha(\alpha q - \beta p)^2 = p + \alpha^3 q^2 - 2\alpha^2\beta qp + \alpha\beta^2 p^2. \tag{29.3.36}$$

In this case the full map $\mathcal{M}_{gq}$ takes a form which we will call $\mathcal{M}_{ffq}$ to indicate that it is a *quadratic* map that has a *finite* product *factorization.*

Action of $\mathcal{M}_{ffq}$:

$$\overset{\equiv}{q} = d + rq + sp + (r\alpha^2\beta + s\alpha^3)q^2 - 2(r\alpha\beta^2 + s\alpha^2\beta)qp + (r\beta^3 + s\alpha\beta^2)p^2, \tag{29.3.37}$$

$$\overset{\equiv}{p} = e + tq + up + (t\alpha^2\beta + u\alpha^3)q^2 - 2(t\alpha\beta^2 + u\alpha^2\beta)qp + (t\beta^3 + u\alpha\beta^2)p^2. \tag{29.3.38}$$

There are two properties of $\mathcal{M}_{ffq}$ that are worth noting. First we observe from (3.34) that $\mathcal{N}$ is invertible with the Lie representation

$$\mathcal{N}^{-1} = \exp : -(1/3)(\alpha q - \beta p)^3 :, \tag{29.3.39}$$

and in view of (3.31) and (3.33) the Taylor expansion of this inverse map also terminates beyond terms of degree two. From (2.4) we know that $\mathcal{T}$ is invertible, and (2.3) shows that $\mathcal{R}$ is invertible if R given by (2.8) satisfies $\det R \neq 0$. Therefore, from (2.5) we conclude that $\mathcal{M}_{ffq}$ is also invertible in this case,

$$\mathcal{M}_{ffq}^{-1} = \mathcal{N}^{-1}\mathcal{R}^{-1}\mathcal{T}^{-1}. \tag{29.3.40}$$

Moreover, it is easily verified that the Taylor expansion for $\mathcal{M}_{ffq}^{-1}$ also terminates beyond terms of degree two.

The second observation concerns the Jacobian matrix M_{ffq} of $\mathcal{M}_{ffq}$. The Jacobian matrix of $\mathcal{T}$ is the identity I and the Jacobian matrix of $\mathcal{R}$ is R. Let N be the Jacobian matrix of $\mathcal{N}$. Then from (2.5) and the chain rule we have the matrix relation

$$M_{ffq} = IRN. \tag{29.3.41}$$

Since in this case $\mathcal{N}$ is a symplectic map, N will be a symplectic matrix and therefore have determinant one. It follows that the determinant of M_{ffq} in this case satisfies the relation

$$\det M_{ffq} = \det R. \tag{29.3.42}$$

We see that the imposition of the requirement that $\mathcal{M}_{qg}$ be factorable into a *finite* product of Lie transformations leads generically to the result that the determinant of its Jacobian matrix must be constant, and cannot depend on where in phase space it is evaluated. Such maps are sometimes called *Cremona* maps. However, we shall use this designation to refer to maps that are both symplectic and polynomial (have Taylor series that terminate beyond some finite order). See Section 29.6.

There is one other important observation to be made. We have seen that requiring that $\mathcal{N}$ have a finite number of Lie factors generically forces $\mathcal{N}$ to be symplectic, which in turn forces N to have a constant (unit) determinant. What if we reverse the situation, and require that N have a constant determinant? From (2.1) we find that N has the form

$$N = \begin{pmatrix} (1+2b_1q+2b_2p) & (2b_2q+2b_3p) \\ (2c_1q+2c_2p) & (1+2c_2q+2c_3p) \end{pmatrix}; \tag{29.3.43}$$

and its determinant is of the form

$$\det N = 1 + \text{ linear terms } + \text{ quadratic terms} \tag{29.3.44}$$

with

$$\text{linear terms } = (2b_1+2c_2)q + (2b_2+2c_3)p, \tag{29.3.45}$$

$$\text{quadratic terms } = 4q^2(b_1c_2 - b_2c_1) + 4qp(b_1c_3 - b_3c_1) + 4p^2(b_2c_3 - b_3c_2). \tag{29.3.46}$$

Forcing N to have constant determinant yields the relations

$$b_1 + c_2 = 0 \text{ or } c_2 = -b_1, \tag{29.3.47}$$

$$b_2 + c_3 = 0 \text{ or } b_2 = -c_3; \tag{29.3.48}$$

$$b_1c_2 - b_2c_1 = 0, \tag{29.3.49}$$

$$b_1c_3 - b_3c_1 = 0, \tag{29.3.50}$$

$$b_2c_3 - b_3c_2 = 0. \tag{29.3.51}$$

We see that (3.47) and (3.48) agree with (3.18) and (3.17); and (3.50) agrees with (3.22). Moreover, substituting (3.47) and (3.48) into (3.49) produces the relation

$$-b_1^2 + c_1c_3 = 0, \tag{29.3.52}$$

which agrees with (3.21); and substituting (3.47) and (3.48) into (3.51) produces the relation

$$-c_3^2 + b_1b_3 = 0, \tag{29.3.53}$$

which agrees with (3.23). It follows that the quantities b_1 through b_3 and c_1 through c_3 are again given by the relations (3.24) through (3.28). Thus, requiring $\mathcal{N}$ to have a constant Jacobian determinant is equivalent to requiring that $\mathcal{N}$ have a finite number of Lie factors, and vice versa. And either requirement forces $\mathcal{N}$ to be symplectic.

In the spirit of the previous section, let us classify maps $\mathcal{M}_{ffq}$ of the form given by (3.37) and (3.38). Upon replacement of $\overset{\equiv}{q}$ by $\bar{q}$ and $\overset{\equiv}{p}$ by $\bar{p}$, and after some algebraic rearrangement, these maps can be written as

Action of $\mathcal{M}_{ffq}$:

$$\bar{q} = d + rq + sp + (r\beta + s\alpha)(\alpha q - \beta p)^2,$$

$$\bar{p} = e + tq + up + (t\beta + u\alpha)(\alpha q - \beta p)^2. \tag{29.3.54}$$

We will seek a sequence of affine transformations that simplifies (3.54) as much as possible. This sequence will be quite long; patience on the part of the reader will be required.

Because of the special form of the nonlinear terms on the right side of (3.54), it is advantageous to make a linear change of variables even before dealing with the translation part. Define new variables z' and $\bar{z}'$ by the relations

$$z' = Az, \tag{29.3.55}$$

$$\bar{z}' = A\bar{z}, \tag{29.3.56}$$

where A is the matrix

$$A = \gamma^{-1} \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} \tag{29.3.57}$$

with

$$\gamma = (\alpha^2 + \beta^2)^{1/2}. \tag{29.3.58}$$

Evidently A is symplectic and orthogonal so that it has the inverse

$$A^{-1} = A^T = \gamma^{-1} \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}. \tag{29.3.59}$$

When this change of variables is made, $\mathcal{M}_{ffq}$ is transformed to the map that we will call $\mathcal{M}^*_{ffq}$. It takes the form

Action of $\mathcal{M}^*_{ffq}$:

$$\bar{q}' = d' + r'q' + s'p' + s'\gamma(q')^2, \tag{29.3.60}$$

$$\bar{p}' = e' + t'q' + u'p' + u'\gamma(q')^2, \tag{29.3.61}$$

where

$$d' = (\alpha d - \beta e)/\gamma, \tag{29.3.62}$$

$$e' = (\beta d + \alpha e)/\gamma, \tag{29.3.63}$$

$$r' = [\alpha^2 r - \alpha\beta(s + t) + \beta^2 u]/\gamma^2, \tag{29.3.64}$$

$$s' = [\alpha^2 s - \alpha\beta(-r + u) - \beta^2 t]/\gamma^2, \tag{29.3.65}$$

$$t' = [\alpha^2 t - \alpha\beta(-r + u) - \beta^2 s]/\gamma^2, \tag{29.3.66}$$

$$u' = [\alpha^2 u + \alpha\beta(s + t) + \beta^2 r]/\gamma^2. \tag{29.3.67}$$

Now seek the fixed points of the map $\mathcal{M}^*_{ffq}$ as given by (3.60) and (3.61). Upon dropping primes this map, again call it $\mathcal{M}^*_{ffq}$, takes the form

Action of $\mathcal{M}^*_{ffq}$:

$$\bar{q} = d + rq + sp + s\gamma q^2, \tag{29.3.68}$$

$$\bar{p} = e + tq + up + u\gamma q^2. \tag{29.3.69}$$

The fixed point equations, after completing squares, give the conditions

$$0 = [d - (r-1)^2/(4s\gamma)] + sp + s\gamma[q + (r-1)/(2s\gamma)]^2, \tag{29.3.70}$$

$$0 = [e - t^2/(4u\gamma)] + (u-1)p + u\gamma[q + t/(2u\gamma)]^2. \tag{29.3.71}$$

These two conditions describe two parabolas, both whose principal axes are parallel to the p axis. There are therefore (see Exercise 3.4) two finite fixed points with each fixed point having a multiplicity of one, or there is a single finite fixed point with multiplicity two. It can also happen that there is a single finite fixed point with multiplicity one. In addition there are two or three fixed points at infinity whose presence becomes apparent with the use of homogeneous coordinates.

Suppose the fixed-point equations (3.70) and (3.71) have a *real* solution, call it $\tilde{q}, \tilde{p}$. Again make a displacement change of variables of the form

$$q = q' + \tilde{q}\ ,\ \ p = p' + \tilde{p}; \tag{29.3.72}$$

$$\bar{q} = \bar{q}' + \tilde{q}\ ,\ \ \bar{p} = \bar{p}' + \tilde{p}. \tag{29.3.73}$$

With this change of variables the map $\mathcal{M}^*_{ffq}$ given by (3.68) and (3.69) is transformed to the origin preserving map $\mathcal{M}^{**}_{ffq}$ of the form

Action of $\mathcal{M}^{**}_{ffq}$:

$$\bar{q}' = r'q' + s'p' + s'\gamma(q')^2, \tag{29.3.74}$$

$$\bar{p}' = t'q' + u'p' + u'\gamma(q')^2, \tag{29.3.75}$$

where

$$r' = r + 2s\gamma\tilde{q}, \tag{29.3.76}$$

$$s' = s, \tag{29.3.77}$$

$$t' = t + 2u\gamma\tilde{q}, \tag{29.3.78}$$

$$u' = u. \tag{29.3.79}$$

Again dropping primes, the map now becomes

Action of $\mathcal{M}^{**}_{ffq}$:

$$\bar{q} = rq + sp + s\gamma q^2, \tag{29.3.80}$$

$$\bar{p} = tq + up + u\gamma q^2. \tag{29.3.81}$$

To continue assume, as a possible case, that R [see (2.8)] has real eigenvalues, and one can select an A such that R is diagonalized as in (2.35). When this is done, the map $\mathcal{M}^{**}_{ffq}$ is transformed to become the map $\mathcal{M}^{***}_{ffq}$ which takes the general form

Action of $\mathcal{M}^{***}_{ffq}$:

$$\bar{q} = \mu(q + b_1q^2 + 2b_2qp + b_3p^2), \tag{29.3.82}$$

$$\bar{p} = \nu(p + c_1q^2 + 2c_2qp + c_3p^2). \tag{29.3.83}$$

What can be said about the coefficients b_1 through b_3 and c_1 through c_3? Evidently the map $\mathcal{M}^{***}_{ffq}$ given by (3.82) and (3.83) has the Jacobian matrix M^{***}_{ffq} given by

$$M^{***}_{ffq} = \begin{pmatrix} \mu(1+2b_1q+2b_2p) & \mu(2b_2q+2b_3p) \\ \nu(2c_1q+2c_2p) & \nu(1+2c_2q+2c_3p) \end{pmatrix}. \tag{29.3.84}$$

Also, the changes of variables we have been making consist of making the *same* affine transformations on both z and $\bar{z}$. See, for example, (2.14), (2.15), (2.26), and (2.28). The determinants of their Jacobian matrices are all constant (z independent). Therefore, by the chain rule, these variable changes cannot alter the determinant of the Jacobian matrix of the various forms of $\mathcal{M}_{ffq}$, and there must be the relation

$$\det M^{***}_{ffq} = \det M_{ffq} = \det R. \tag{29.3.85}$$

From (3.84) we find the result

$$\det M^{***}_{ffq} = \mu\nu(1 + \text{ linear terms } + \text{ quadratic terms}), \tag{29.3.86}$$

and comparison with (3.43) and (3.44) shows that the linear terms and quadratic terms are the same as those given by (3.45) and (3.46). According to (3.85) these terms must vanish. Therefore, by a now familiar argument, the coefficients b_1 through b_3 and c_1 through c_3 must be given by relations of the form (3.24) through (3.28). Correspondingly, the map given by (3.82) and (3.83) takes the form

Action of $\mathcal{M}^{***}_{ffq}$:

$$\bar{q} = \mu(q + \alpha^2\beta q^2 - 2\alpha\beta^2 qp + \beta^3 p^2), \tag{29.3.87}$$

$$\bar{p} = \nu(p + \alpha^3 q^2 - 2\alpha^2\beta qp + \beta^2 p^2). \tag{29.3.88}$$

Let us make one last change of variables. Make the transformations

$$q' = \alpha^2\beta q\ ,\ p' = \alpha\beta^2 p; \tag{29.3.89}$$

$$\bar{q}' = \alpha^2\beta\bar{q}\ ,\ \bar{p}' = \alpha\beta^2\bar{p}. \tag{29.3.90}$$

Multiply both sides of (3.87) by $\alpha^2\beta$ and both sides of (3.88) by $\alpha\beta^2$. So doing gives the results

$$\alpha^2\beta\bar{q} = \mu(\alpha^2\beta q + \alpha^4\beta^2 q^2 - 2\alpha^3\beta^3 qp + \alpha^2\beta^4 p^2), \tag{29.3.91}$$

$$\alpha\beta^2\bar{p} = \nu(\alpha\beta^2 p + \alpha^4\beta^2 q^2 - 2\alpha^3\beta^3 qp + \alpha^2\beta^4 p^2). \tag{29.3.92}$$

In view of (3.89) and (3.90) these results can be rewritten in the form

$$\bar{q}' = \mu[q' + (q')^2 - 2q'p' + (p')^2] = \mu[q' + (q' - p')^2], \tag{29.3.93}$$

$$\bar{p}' = \nu[p' + (q')^2 - 2q'p' + (p')^2] = \nu[p' + (q' - p')^2]. \tag{29.3.94}$$

The map $\mathcal{M}^{***}_{ffq}$ has been *transformed* to become the final map $\mathcal{M}^{tr}_{ffq}$. Dropping the primes gives the final relations

Action of $\mathcal{M}^{tr}_{ffq}$:

$$\bar{q} = \mu[q + (q-p)^2]. \tag{29.3.95}$$

$$\bar{p} = \nu[p + (q-p)^2]. \tag{29.3.96}$$

Thanks to heroic effort, we have shown that if a map $\mathcal{M}_{ffq}$ given by relations of the form (3.37) and (3.38) has a real fixed point, and if the linear part of the map about this fixed point has two distinct real eigenvalues μ and ν, then under a suitable affine change of variables this map is equivalent to the transformed map $\mathcal{M}^{tr}_{ffq}$ given by the far simpler relations (3.95) and (3.96). Subsequently, in Sections 18.7 and 18.8, we will find that there is no loss of generality in the assumption we have made about the eigenvalues of the linear part. We will find that the map has a second fixed point, and that the other eigenvalue possibilities are realized by the linear part of the map about the second fixed point when, about the origin, the map has the form given by (3.95) and (3.96).

In summary, we conclude that the condition that the nonlinear part of $\mathcal{M}_{gq}$ have a finite number of Lie factors (or, equivalently, a constant Jacobian determinant) reduces the number of free nonlinear parameters (see the end of Section 18.2) from 4 to 0. Consequently, such maps are labelled completely by the *two* real parameters, μ and ν, that describe their linear (when brought to normal form) parts.

Exercises

29.3.1. Look again at Exercise 1.2.8. It examined the general single-complex-variable analytic quadratic map, which evidently depends on three complex parameters a, b, and c. Hence it depends on six real parameters. In addition, the exercise showed that, under the affine change of variables (1.2.115), this map could be brought to the simpler form (1.2.111) which depends on only one complex parameter μ, and hence on two real parameters. Moreover, Exercise 1.2.7 showed that this simpler form is equivalent to the complex logistic map which, according to (1.37) and (1.38), yields a two-parameter family of quadratic maps of the plane into itself.

In Section 24.2 we found that $\mathcal{M}_{gq}$, the general quadratic map in two variables, depends on twelve real parameters. See (2.13). Among these maps will be the two-parameter subset (1.37) and (1.38). We also found that, under a suitable affine change of variables, $\mathcal{M}_{gq}$ could be brought to the form $\mathcal{M}^{tr}_{gq}$ whose action is described by (2.42), and that $\mathcal{M}^{tr}_{gq}$ depends on six real parameters.

In this section we found that requiring $\mathcal{M}_{gq}$ to have a finite product Lie factorization produced the subset of maps $\mathcal{M}_{ffq}$ given by (3.54). This subset of maps is characterized by eight real parameters, namely d, e, r, s, t, u, α, and β. Also we found that these maps are globally invertible; and we found that under an affine transformation each could be brought to the two-parameter form $\mathcal{M}^{tr}_{ffq}$ described by (3.95) and (3.96). Show that, unlike $\mathcal{M}_{ffq}$, complex logistic maps are not globally invertible.

29.3.2. Find $\mathcal{N}$ for the complex logistic map given by (1.37) and (1.38), and show that it appears to have an infinite number of Lie factors.

29.3.3. Find the decomposition (3.6) using the machinery of Section 21.10.

29.3.4. Verify that the relations (3.9) through (3.16) follow from (3.4) and (3.5). Verify that the choice (3.17) and (3.18) reduces the set of relations (3.9) through (3.16) to the set of relations (3.21) through (3.23). Verify that they have the unique solutions (3.24) through (3.26). Explore what happens when one considers the non-generic possibilities for which a_3 and/or b_1 equal zero.

29.3.5. Exercise on fixed point counting.

29.4 Fixed Points

We have been studying two-variable quadratic maps. There is still more to be done on this subject, and we will do that in Sections 29.7 through 29.10. But before doing so, it is useful to interrupt our discussion to take up the study of fixed points for maps in general, and two-variable maps in particular. We will then be better prepared for our further study of two-variable quadratic maps, and we will also find important results having wide application.

29.4.1 Attack a Map at its Fixed Points

We have seen in Chapter 1 that maps, even the simplest maps, generally exhibit extremely complicated behavior under iteration so that it is difficult to know where to begin in characterizing them. One fruitful starting point for the analysis of a map is to find and classify its fixed points. As Poincaré wrote,

> ⋯ what renders these periodic solutions so precious is that they are, so to speak, the only breach through which we may try to penetrate a stronghold previously reputed to be impregnable.

This section and the next summarize some important aspects of fixed-point analysis.

29.4.2 Fixed Points are Generally Isolated

We begin by observing that the fixed points of a map are, in general, *isolated*. That is, if z^α is a fixed point of a map $\mathcal{M}$, then there is no other fixed point in a neighborhood of z^α providing this neighborhood is sufficiently small. To verify this result, assume the contrary: Suppose there is a second fixed point z^β in any neighborhood of z^α. Then we have the relations

$$\mathcal{M}z^\alpha = z^\alpha, \tag{29.4.1}$$

$$\mathcal{M}z^\beta = z^\beta, \tag{29.4.2}$$

$$z^\beta = z^\alpha + \delta, \tag{29.4.3}$$

where δ is a small vector. Taken together, (4.2) and (4.3) give the result

$$\mathcal{M}(z^\alpha + \delta) = z^\alpha + \delta. \tag{29.4.4}$$

Now let $\mathcal{L}_\alpha$ be the linear part of $\mathcal{M}$ at z^α so that there is the relation

$$\mathcal{M}(z^\alpha + \delta) = z^\alpha + \mathcal{L}_\alpha\delta + O(\delta^2) = z^\alpha + L_\alpha\delta + O(\delta^2). \tag{29.4.5}$$

Here L_α is the matrix that represents $\mathcal{L}_\alpha$. By combining (4.4) and (4.5) we find the result

$$L_\alpha \delta = \delta - O(\delta^2). \tag{29.4.6}$$

Since δ can be arbitrarily small (we have assumed *any* neighborhood), we conclude that L_α must have an eigenvector with eigenvalue +1, and hence

$$\det(L_\alpha - I) = 0. \tag{29.4.7}$$

In general, (4.7) will not be true. We conclude that if L_α does not have +1 as an eigenvalue, then z^α is an isolated fixed point of $\mathcal{M}$. Note that this analysis makes no assumption about the number of variables involved. In particular, it also holds for one-dimensional maps. See, for example, Exercise 1.2.1 where it was found that two fixed points for the logistic map coincided when the eigenvalue μ of the linear part took the value +1, but were isolated for all other values of μ.

If linear analysis is sufficient to demonstrate that a fixed point is isolated, we will call such a fixed point *manifestly* isolated. If (4.7) holds at a fixed point, a beyond linear analysis is required to determine whether or not it is isolated. A fixed point that requires beyond linear analysis to show that it is isolated might be called *expectant* since, as we will see in some examples, what often happens in this case is that as some parameter is varied an expectant fixed point gives birth to additional fixed points. In the analytic case, where complex analysis can be employed, these additional fixed points exist for all parameter values, but may be complex. They merge during the birth process to a common real point and then again separate while now remaining within the real domain. Finally, there are cases where a fixed point lies on a line, or in some higher dimensional domain, all of whose points are fixed points, and therefore such a fixed point is indeed not isolated.

29.4.3 Finding Fixed Points with Contraction Maps

To find a fixed point of $\mathcal{M}$ it is useful to construct a *contraction* map $\mathcal{C}$. Suppose z^{fx} is a *fixed* point of some map $\mathcal{M}$,

$$\mathcal{M} z^{fx} = z^{fx}, \tag{29.4.8}$$

and suppose z^e is an *arbitrary* point in the *vicinity* of z^{fx}. The contraction map will be shown to have the remarkable property

$$\lim_{n\to\infty} \mathcal{C}^n z^e = z^{fx}. \tag{29.4.9}$$

That is, a good guess as to the location of a fixed point is sufficient starting information to contract in on it exactly.

The construction of $\mathcal{C}$ is a generalization of Newton's method to the case of several variables. Suppose z^α is an arbitrary point, and suppose it is sent to the point z^β under the action of $\mathcal{M}$,

$$\mathcal{M} z^\alpha = z^\beta. \tag{29.4.10}$$

Let $\mathcal{L}_\alpha$ be the linear part of $\mathcal{M}$ at z^α. That is, $\mathcal{L}_\alpha$ has the property

$$\mathcal{M}(z^\alpha + \delta) = z^\beta + \mathcal{L}_\alpha \delta + O(\delta^2) \tag{29.4.11}$$

where δ is small. Evidently the action of $\mathcal{L}_\alpha$ can be represented by a matrix L_{z^α},

$$\mathcal{L}_\alpha \delta = L_{z^\alpha}\delta. \tag{29.4.12}$$

(In the differential equation case this matrix is the solution to the *linear* variational equations when $z^d(t^i) = z^\alpha$. See Section 10.12) The map $\mathcal{C}$ is now defined by requiring that its action on the arbitrary point z^α be given by the rule

$$\mathcal{C}z^\alpha = z^\alpha - (I - L_{z^\alpha})^{-1}(z^\alpha - \mathcal{M}z^\alpha). \tag{29.4.13}$$

It is easily verified that $\mathcal{C}$ as defined by (4.13) has the advertised property (4.9). First, suppose that z^{fx} is a fixed pont of $\mathcal{M}$. Then, from (4.13), z^{fx} is also a fixed point of $\mathcal{C}$,

$$\mathcal{C}z^{fx} = z^{fx}. \tag{29.4.14}$$

Next, suppose that z^e is some point in the vicinity of z^{fx}. Then z^e is of the form

$$z^e = z^{fx} + \delta \tag{29.4.15}$$

where δ is small. Upon inserting (4.15) into (4.13), we find after a short calculation the result

$$\begin{aligned} \mathcal{C}z^e &= \mathcal{C}(z^{fx}+\delta) = z^{fx} + \delta - (I - L_{z^{fx}+\delta})^{-1}[(z^{fx}+\delta) - \mathcal{M}(z^{fx}+\delta)] \\ &= z^{fx} + \delta - (I - L_{z^{fx}+\delta})^{-1}[(I - L_{z^{fx}})\delta + O(\delta^2)] = z^{fx} + O(\delta^2). \end{aligned} \tag{29.4.16}$$

Here we have used the relation

$$\mathcal{M}(z^{fx}+\delta) = z^{fx} + L_{z^{fx}}\delta + O(\delta^2) \tag{29.4.17}$$

and the observation that

$$L_{z^{fx}+\delta} = L_{z^{fx}} + O(\delta). \tag{29.4.18}$$

Thus, according to (4.16), although the initial point z^e differs from the desired fixed point z^{fx} by an amount δ, the point $\mathcal{C}z^e$ differs from the point z^{fx} only by an amount of order δ^2. Similarly, the point $\mathcal{C}^2z^e$ difffers from z^{fx} only by an amount of order $(\delta^2)^2$, and $\mathcal{C}^nz^e$ differs from z^{fx} only by an amount of order δ^{2^n}. Consequently, as expected for Newton's method, the convergence of the limit (4.9) to z^{fx} is extremely fast. Note also that Newton's method succeeds even if z^{fx} is *not* an attractor. Indeed, any isolated fixed point of $\mathcal{M}$, whether stable or unstable, is a super stable fixed point of $\mathcal{C}$. See Exercise 1.2.1.

In order to complete the discussion it is necessary to check whether the matrix $(I - L_{z^{fx}})$ has an inverse. Evidently the inverse exists provided the related determinant satisfies the condition

$$\det(L_{z^{fx}} - I) \neq 0. \tag{29.4.19}$$

That is, the matrix $L_{z^{fx}}$ does not have +1 as an eigenvalue. (As we have seen, this requirement is sufficient to guarantee that z^{fx} be isolated.) Therefore, the procedure (4.9) will succeed as long as z^e is sufficiently close to z^{fx} and L_{z^e} does not have +1 as an eigenvalue for all z^e in the vicinity of z^{fx}.

29.4.4 Persistence of Fixed Points

Even more can be said. Suppose the map $\mathcal{M}$ has a fixed point. Do maps near $\mathcal{M}$ (in map space) also have fixed points? Suppose $\mathcal{M}$ has the fixed point z^{fx}, and that $\mathcal{M}'$ is a map near $\mathcal{M}$. Let us speculate that $\mathcal{M}'$ has a fixed point of the form $z^{fx} + \Delta$ where Δ is small,

$$\mathcal{M}'(z^{fx} + \Delta) \overset{?}{=} z^{fx} + \Delta. \tag{29.4.20}$$

Define a map $\mathcal{N}$ by the equation

$$\mathcal{N} = \mathcal{M}^{-1}\mathcal{M}' \tag{29.4.21}$$

so that we have the relation

$$\mathcal{M}' = \mathcal{M}\mathcal{N}. \tag{29.4.22}$$

Since $\mathcal{M}'$ is assumed to be near $\mathcal{M}$, $\mathcal{N}$ is a map near the identity map. Inserting (4.22) in (4.20) gives the hypothesis

$$\mathcal{M}\mathcal{N}(z^{fx} + \Delta) \overset{?}{=} z^{fx} + \Delta. \tag{29.4.23}$$

However, since $\mathcal{N}$ is near the identity map, we must have a result of the form

$$\mathcal{N}(z^{fx} + \Delta) = z^{fx} + \Delta + \tilde{\Delta} \tag{29.4.24}$$

where $\tilde{\Delta}$ is also small. Therefore, we now have the equivalent speculation

$$\mathcal{M}(z^{fx} + \Delta + \tilde{\Delta}) \overset{?}{=} z^{fx} + \Delta. \tag{29.4.25}$$

However, using the relation (4.17), we may rewrite (4.25) in the form

$$z^{fx} + L_{z^{fx}}(\Delta + \tilde{\Delta}) \overset{?}{=} z^{fx} + \Delta, \tag{29.4.26}$$

which is equivalent to the speculation

$$(I - L_{z^{fx}})\Delta \overset{?}{=} L_{z^{fx}}\tilde{\Delta}. \tag{29.4.27}$$

(Here we have omitted higher order terms in Δ and $\tilde{\Delta}$.) We see that (4.27) can be solved for Δ, and therefore our speculation is correct, provided the matrix $(I - L_{z^{fx}})$ is invertible. We conclude that if $\mathcal{M}$ is varied over some path in map space and $\mathcal{M}$ initially has a fixed point, then its fixed point persists and moves over some path in z space, provided that over that path $L_{z^{fx}}$ never has +1 as an eigenvalue.[1]

Suppose the path in map space and the corresponding path in z space are parameterized by some parameter τ. That is, we write $\mathcal{M} = \mathcal{M}(\tau)$, and suppose that for each value of τ the map $\mathcal{M}(\tau)$ has the fixed point $z^{fx}(\tau)$. We will now show that there is a differential equation, whose solution is $z^{fx}(\tau)$, that can be used to "track" z^{fx} as τ is varied.

In the spirit of Section 6.4.2, make the action of $\mathcal{M}(\tau)$ explicit by writing the relations

$$\bar{z}_a(\tau) = u_a(z, \tau). \tag{29.4.28}$$

[1] See Exercise 9.2.6 for an $ISp(2n, \mathbb{R})$ example of this result.

Again define functions $w_a(z,\tau)$ by the rule

$$w_a(z,\tau) = \partial u_a(z,\tau)/\partial\tau. \tag{29.4.29}$$

By definition the fixed point $z^{fx}(\tau)$ obeys the relation

$$z_a^{fx}(\tau) = u_a(z^{fx}(\tau),\tau). \tag{29.4.30}$$

Since (4.30) is presumed to hold for a range of τ values, there is also the relation

$$z_a^{fx}(\tau+d\tau) = u_a(z^{fx}(\tau+d\tau),\tau+d\tau). \tag{29.4.31}$$

The left side of (4.31) has the expansion

$$z_a^{fx}(\tau+d\tau) = z_a^{fx}(\tau) + (dz_a^{fx}/d\tau)d\tau. \tag{29.4.32}$$

The right side of (4.31) has the expansion

$$u_a(z^{fx}(\tau+d\tau),\tau+d\tau) = u_a(z^{fx}(\tau),\tau) + d\tau[\partial u_a/\partial\tau + \sum_b(\partial u_a/\partial z_b)(dz_b^{fx}/d\tau)]. \tag{29.4.33}$$

Equating powers of $d\tau$ in (4.31) through (4.33) and using (4.30) gives the result

$$dz_a^{fx}/d\tau = \partial u_a/\partial\tau + \sum_b(\partial u_a/\partial z_b)(dz_b^{fx}/d\tau). \tag{29.4.34}$$

The second term on the right side of (4.34) contains L, the Jacobian matrix for the map $\mathcal{M}$,

$$L_{ab} = \partial u_a/\partial z_b. \tag{29.4.35}$$

Consequently with the aid of (4.29) and (4.35), and employing vector and matrix notation, the relation (4.34) can also be written in the form

$$(I-L)(dz^{fx}/d\tau) = w. \tag{29.4.36}$$

Finally, again under the assumption the $(L-I)$ is invertible, we obtain the desired result,

$$dz^{fx}/d\tau = (I-L_{z^{fx}})^{-1}w(z^{fx},\tau). \tag{29.4.37}$$

Of course, to solve (4.37) requires an initial condition. That is, we must know $z^{fx}(\tau^i)$ for some initial τ^i. It might be possible to choose τ^i in such a way that $\mathcal{M}(\tau^i)$ has an obvious fixed point, perhaps by construction. Or it might be necessary to do a Newton search or apply some other procedure to find $z^{fx}(\tau^i)$.

Suppose the map in question arises from integrating a differential equation with independent variable t and parameter dependence τ. Then the map will be of the form $\mathcal{M} = \mathcal{M}(t;\tau)$. If we simply wish to find a fixed point for a specified value of τ using Newton's method, then the required Jacobian matrix L could be found by simultaneously integrating the variational equations. See Exercise 4.6 of Section 1.4. We could write a computer program with two nested loops. The inner loop would integrate the equations of motion along with the variational equations in order to compute $\mathcal{M}z$ and L. These are the required ingredients for $\mathcal{C}$. The outer loop would apply $\mathcal{C}$ to achieve a Newton iteration.

If we wish to find a fixed point for a range of τ values, then in this case the Jacobian matrix L and the vector w could be found by integrating "augmented" variational equations where deviations are made in both the original phase-space variables z and the parameter τ. That is, the set of variables and associated variational equations would be enlarged. See Section 10.12.6. We could again write a computer program with two nested loops. Now the inner loop would integrate the equations of motion along with the augmented variational equations in order to compute L and w, and the outer loop would integrate the equations (4.36).

29.4.5 Application to Accelerator Physics

The fixed-point considerations we have described so far have implications for accelerator physics. Suppose $\mathcal{M}$ is the one-turn map for a circular machine. Then a fixed point of $\mathcal{M}$ corresponds to a closed orbit. We know that L is the Jacobian matrix for the map $\mathcal{M}$. (Elsewhere, we have sometimes denoted this Jacobian matrix by the symbols M or R.) In the case that $\mathcal{M}$ is a symplectic map, L will be a symplectic matrix. According to Section 3.4, all the eigenvalues of a symplectic matrix generally differ from +1. Moreover, an eigenvalue taking on the value +1 corresponds to an integer tune. See (3.5.39) and (3.5.40) of Section 3.5. Suppose $\mathcal{M}^{\text{ideal}}$ is the one-turn map for an ideal machine operating at its design energy, and suppose the design tunes do not have integer values. By design, the ideal machine has a closed orbit and therefore $\mathcal{M}^{\text{ideal}}$ has a fixed point. Moreover, this fixed point is isolated since the design tunes are assumed to not have integer values. Now, the difference between an ideal machine operating at its design energy and the imperfect machine realized in actual construction and operating at some nearby energy may be regarded as the result of a variation in $\mathcal{M}^{\text{ideal}}$. Consequently, according to the previous discussion, if the tunes for the closed design orbit in the ideal machine have noninteger values, then the imperfect machine will also have a closed orbit at the design energy (and other nearby energies as well) provided the perturbations in the machine lattice are not so large as to drive some tune to an integer value. In particular, small imperfections in a machine lattice, such as arise from magnet misalignment and misplacement, magnet under or over powering, magnetic fringe fields and general magnetic field inhomogeneities, etc., do not destroy the existence of a closed orbit but merely cause it to be slightly distorted. (See also Exercise 3.4.3.) Therefore, in accelerator physics, integer tunes should be avoided in order to assure the continued existence of a closed orbit under unavoidable perturbations/imperfections.

Exercises

29.4.1. A function f is called *invariant* under the action of a map $\mathcal{M}$ if it has the property

$$f(\mathcal{M}z) = f(z). \tag{29.4.38}$$

[See Section 5.2 and review the discussion surrounding (7.1.12).] For our definition to have significance, we exclude the trivial case where f is simply a constant function (which will always be invariant) and therefore assume that ∇f is generally nonzero. An invariant function is also called an *integral* of $\mathcal{M}$.

Suppose the map $\mathcal{M}$ has the manifestly isolated fixed point z^α, and also has an integral f. You are to prove that then $\nabla f = 0$ at this point. Show from (4.38) that for arbitrary δ there is the relation

$$f[\mathcal{M}(z^\alpha + \delta)] = f(z^\alpha + \delta). \tag{29.4.39}$$

Next show from (4.5) that for small δ there is the relation

$$f[z^\alpha + L_\alpha \delta + O(\delta^2)] = f(z^\alpha + \delta). \tag{29.4.40}$$

Now expand both sides of (4.40) in a Taylor series to get the result

$$f(z^\alpha) + (L_\alpha \delta) \cdot \nabla f = f(z^\alpha) + \delta \cdot \nabla f + O(\delta^2), \tag{29.4.41}$$

from which it follows that

$$[(L_\alpha - I)\delta] \cdot \nabla f = 0. \tag{29.4.42}$$

Suppose that $(L_\alpha - I)$ is invertible. Let ϵ be an arbitrary vector and specify δ by the relation

$$\delta = (L_\alpha - I)^{-1}\epsilon. \tag{29.4.43}$$

Deduce the relation

$$\epsilon \cdot \nabla f = 0. \tag{29.4.44}$$

Since ϵ is assumed to be arbitrary, it must follow that $\nabla f = 0$ at the point z^α. Therefore ∇f must vanish at any manifestly isolated fixed point. Conversely, suppose (as we would generally like to be the case) that ∇f does not vanish at some fixed point z^α. Then, $(L_\alpha - I)$ must not be invertible, $\det(L_\alpha - I)$=0 and hence L_α has +1 as an eigenvalue, and z^α is not manifestly isolated. For example, if $\mathcal{M}$ arises from integrating Hamilton's equations of motion for a time-independent Hamiltonian H, then H is an integral and the L_α for any periodic orbit must have +1 as an eigenvalue. Moreover, since L_α is symplectic in the Hamiltonian case, this eigenvalue must have even multiplicity (generally two, but possibly higher). See Section 3.4.

29.4.2. Consider the map $\mathcal{M}$ given by the relations

$$\bar{q} = q, \tag{29.4.45}$$

$$\bar{p} = p + q. \tag{29.4.46}$$

Show that $\mathcal{M}$ has the origin as a fixed point and that the linear part of $\mathcal{M}$ about the origin is given by the matrix (3.8.28). Verify that this matrix has eigenvalue +1. Show that all points on the p axis are fixed points of $\mathcal{M}$, and thus the origin is not an isolated fixed point. Review Exercise 4.1. Show that the function f given by

$$f(z) = q \tag{29.4.47}$$

is an invariant function under the action of $\mathcal{M}$, and that ∇f does not vanish anywhere despite the existence of all these fixed points.

29.4.3. The logistic map (1.2.5) has the fixed points (1.2.8) and (1.2.9). Verify that these fixed points obey (4.37) when λ is taken to to be the parameter.

29.5 Poincaré Index

For the two-dimensional case the study of the nature of fixed points is facilitated by use of the Poincaré index. It is defined as follows: Use the map $\mathcal{M}$ to produce a vector $v(z)$ at each point z in phase space by the rule

$$v(z) = \mathcal{M}z - z. \tag{29.5.1}$$

Geometrically, $v(z)$ is the vector that extends from z to $\mathcal{M}z$. By construction, $v(z)$ vanishes if and only if z is a fixed point of $\mathcal{M}$. Now consider the two-dimensional case, and let z^α be an isolated fixed point of $\mathcal{M}$. Draw a circle c around z^α small enough that no other fixed points are enclosed and none of the points on the circle itself are fixed points. The vector $v(z)$ for all points z on c can never vanish because, by hypothesis, none of the points on the circle are fixed points. Moreover, since $\mathcal{M}$ is assumed to be continuous, this vector will vary continuously over c. For each point z on c draw, in a separate plane, a vector that is parallel to $v(z)$ and has its tail at the origin. See Figure 5.1. Start at some point on c and traverse c once in the counterclockwise direction to return to this starting point. The "translated" vector $v(z)$ in the separate plane will then vary continuously as well, and will ultimately return to its starting value. The *Poincaré index* of the fixed point z^α is defined to be the number of counterclockwise revolutions that this vector undergoes as c is traversed once in the counterclockwise direction. Evidently the Poincaré index will be a positive or negative *integer* or *zero*. Moreover, its value would be unchanged had we chosen to traverse c in the clockwise direction and correspondingly counted the number of revolutions made by v in the clockwise direction. Finally, if c is continuously deformed to become another closed curve c', the index remains unchanged provided all the intermediate closed curves and the final closed curve c' contain no fixed points.

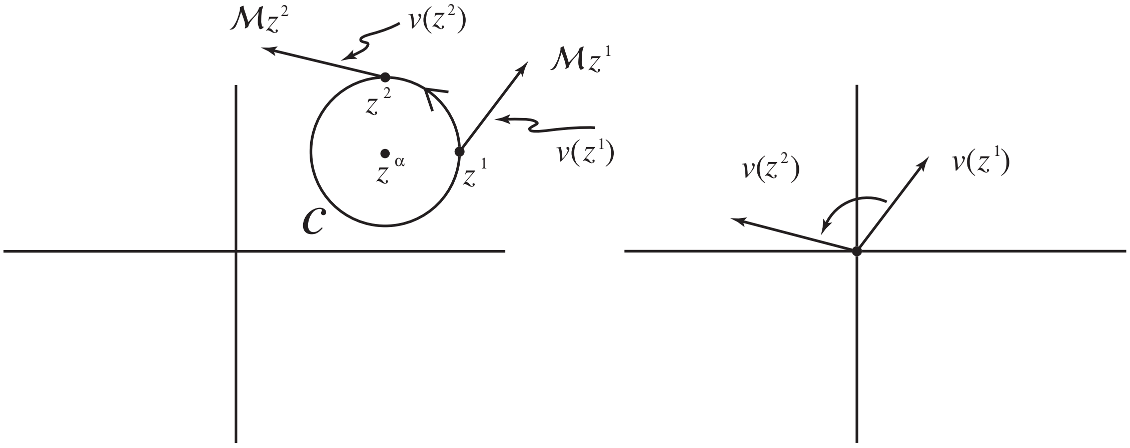


Figure 29.5.1: The isolated fixed point z^α surrounded by a small circle c and the associated vectors $v(z)$ drawn from the common origin (0,0).

If the circle c about the isolated fixed point z^α is small enough, the behavior of $\mathcal{M}$ for $z \in c$ is well approximated by the linear part of $\mathcal{M}$ at z^α; and (since the index is known to be an integer) this approximation can be used to compute the index of z^α. That is, the index of z^α is completely determined by the nature of the matrix L_α.

To see this suppose z^α is a fixed point and $z = z^\alpha + \delta$ is a nearby point. We then have from (4.5) and (5.1) the result

$$v(z) = \mathcal{M}z - z = (L_\alpha - I)\delta + O(\delta^2). \tag{29.5.2}$$

Now let δ be sufficiently small so that terms of order δ^2 can be neglected. It is easily verified that as the points δ traverse a circle about the origin, the points $(L_\alpha - I)\delta$ will, in general, traverse an ellipse about the origin provided $(L_\alpha - I)$ is invertible. (Here we assume that z^α is manifestly isolated.) Moreover, the circle and ellipse will be traversed in the *same* sense if $\det(L_\alpha - I) > 0$, and in the *opposite* sense if $\det(L_\alpha - I) < 0$. See Exercise 5.1. Finally, suppose that N is the normal form of L_α. That is, there is the relation

$$L_\alpha = ANA^{-1} \tag{29.5.3}$$

where A is a matrix that brings L_α to normal form. Then there is also the relation

$$L_\alpha - I = A(N - I)A^{-1}, \tag{29.5.4}$$

and consequently

$$\det(L_\alpha - I) = \det(N - I). \tag{29.5.5}$$

Therefore the index of z^α depends only on the normal form N of L_α, which in turn depends only on the spectrum of L_α. Specifically, the index of z^α is $+1$ if $\det(N - I) > 0$, and -1 if $\det(N - I) < 0$.

As in Section 3.4, let λ_1 and λ_2 be the eigenvalues of L_α. Then (under the assumption that $\mathcal{M}$ is orientation preserving so that $\det L_\alpha > 0$), there are the following possibilities, and for each possibility it is a simple matter to compute the sign of $\det(N - I)$ to determine the index:

1. Both λ_1 and λ_2 are real and positive. For this case the normal forms N for L_α are

$$N = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \text{ if } \lambda_1 \neq \lambda_2 \text{ or } \lambda_1 = \lambda_2 \text{ but } L_\alpha \text{ is diagonalizable,} \tag{29.5.6}$$

$$N = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \text{ if } \lambda_1 = \lambda_2 = \lambda \text{ and } L_\alpha \text{ is not diagonalizable.} \tag{29.5.7}$$

For this case there are 4 subcases:

(a) Both $0 < \lambda_1 < 1$ and $0 < \lambda_2 < 1$. In this subcase z^α is called *attracting*. It is also sometimes called a *node*. The normal forms N for L_α are either (5.6) or (5.7). The index in this subcase is $+1$.

(b) Both $\lambda_1 > 1$ and $\lambda_2 > 1$. In this subcase z^α is called *repelling*. The possible normal forms are again given by (5.6) and (5.7). The index in this subcase is $+1$.

(c) One eigenvalue, say λ_1, satisfies $0 < \lambda_1 < 1$ and the other satisfies $\lambda_2 > 1$. In this subcase z^α is called *hyperbolic*, and the normal form is given by (5.6). It is also sometimes called a *saddle*. The index in this subcase is -1.

   (d) At least one eigenvalue is 0 or is +1. The first possibility is excluded by the requirement $\det L_\alpha > 0$, and the second (which is called *parabolic*) is excluded by the requirement that z^α be manifestly isolated.

2. Both λ_1 and λ_2 are real and negative. For this case there are 4 subcases:

   (a) Both $\lambda_1 < -1$ and $\lambda_2 < -1$. In this subcase z^α is called *inversion repelling*. The possible normal forms are given by (5.6) or (5.7). The index in this subcase is +1.

   (b) Both $-1 < \lambda_1 < 0$ and $-1 < \lambda_2 < 0$. In this subcase z^α is called *inversion attracting*, and the possible normal forms are given by (5.6) or (5.7). The index in this subcase is +1.

   (c) One eigenvalue, say λ_1, satisfies $-1 < \lambda_1 < 0$ and the other satisfies $\lambda_2 < -1$. In this subcase z^α is called *inversion hyperbolic*, and the normal form is given by (5.6). The index in this case is +1.

   (d) One or both eigenvalues is -1. In this subcase z^α is called *inversion parabolic*, and the normal form is given by (5.6) or (5.7). The index in this subcase is +1.

3. Both λ_1 and λ_2 are complex. Since L_α is a real matrix, its eigenvalues will in fact be complex conjugate so that they can be written in the form

$$\lambda_1 = \mu e^{i\phi}, \tag{29.5.8}$$

$$\lambda_2 = \mu e^{-i\phi}, \tag{29.5.9}$$

   with

$$\mu > 0. \tag{29.5.10}$$

   In this case the normal form is given by

$$N = \mu \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix}. \tag{29.5.11}$$

   There are now 3 subcases:

   (a) The quantity μ satisfies $\mu > 1$. In this subcase z^α is called *repelling*. The index is +1.

   (b) The quantity μ satisfies $0 < \mu < 1$. In this subcase z^α is called *attracting*. It is also sometimes called a *node*. The index is +1.

   (c) The quantity μ satisfies $\mu = 1$. (We also must then have $\phi \neq 2n\pi$ because, if not, $N = I$, which in turn implies $L_\alpha = I$ so that L_α has +1 as an eigenvalue contrary to the requirement that z^α be manifestly isolated. See Exercise 5.2.) In this subcase z^α is called *elliptic*. The index is +1.

We conclude that the Poincaré index of a manifestly isolated fixed point z^α is always $+1$ unless z^α is hyperbolic, in which case the index is -1.

Let C be a closed curve that may surround several manifestly isolated fixed points, but does not itself contain any fixed points. That is, no points on C are fixed. We have seen how to define the index of a manifestly isolated fixed point. We will now see that we can also define the index of C. By assumption the vector $v(z)$ given by (5.1) does not vanish on C, and therefore we may count the number of counterclockwise revolutions it makes as C is traversed once in the counterclockwise direction. Call this (integer or zero) number the index of C. As before, the index of C does not change if C is continuously deformed as long as no enclosed fixed points are crossed during the deformation process.

Let n^+ and n^- be the number of fixed points, all surrounded by C, with indices $+1$ and -1, respectively. Then there is the remarkable relation

$$\text{index of } C = n^+ - n^-. \tag{29.5.12}$$

That is, the index of C is the sum of the indices of the fixed points it surrounds. To verify this result, shrink C in a manner analogous to contour integration in such a way that none of the enclosed fixed points are crossed as C is deformed to become C'. Then the index of C' will be the same as that of C. See Figure 5.2. Evidently the contributions to the index of C' made by the anti-parallel portions of C' cancel, and the only nonzero contributions are those made by the small loops around the enclosed fixed points. By the previous discussion, the contribution of each of these is ±1, and these contributions all add. Therefore (5.12) is correct.

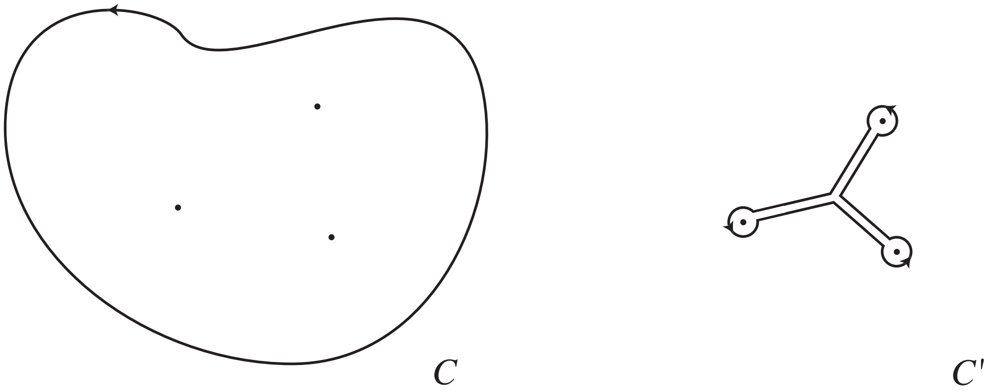


Figure 29.5.2: A closed curve C that surrounds several fixed points, and the curve C' formed by shrinking C.

As an illustration of how the index relation (5.12) can be employed, let us make a detour to consider again the stroboscopic Duffing map of Section 1.4.3 and Chapter 23. Multiply both sides of the Duffing equation of motion (23.1.1) by p and rearrange terms to find the result

$$p(\ddot{q} + q + q^3) = -2\beta p\dot{q} - \epsilon p \sin \omega\tau. \tag{29.5.13}$$

Define a quantity E, which may be viewed as the oscillator *energy*, by the rule

$$E = (1/2)p^2 + (1/2)q^2 + (1/4)q^4. \tag{29.5.14}$$

Then, recalling that $p = \dot{q}$, (5.13) can be rewritten in the form

$$dE/dt = -2\beta p^2 - \epsilon p \sin \omega \tau. \tag{29.5.15}$$

Correspondingly, ΔE, the change in E over one drive period, is given by the integral

$$\Delta E = \int_0^T (-2\beta p^2 - \epsilon p \sin \omega \tau) d\tau. \tag{29.5.16}$$

We next claim that if the initial conditions are such that E is initially sufficiently large (and "largeness" will depend on the values of the parameters β, ϵ, and ω), then ΔE will be negative. That is, there will be a net decrease in the energy over the course of a drive period. There are two reasons to believe this claim. Note that the first term in the integrand of (5.16) is always negative (we assume $\beta > 0$) while the second is generally oscillatory. If E is large, then we expect that p will be large over most of the driving period, and therefore the first term in the integrand will dominate the second since p^2 will generally greatly exceed the magnitude of p. Moreover, if the amplitude of oscillation is large, as it will be if E is large, then the frequency of oscillation will also be large (the frequency of oscillation increases with amplitude in the case of a hard spring) so that the integral over the second term will essentially average to zero.

Let us see how this works out by looking at a numerical example. Figure 5.3 shows two contours $C(E_j)$ in the q, p plane consisting of those points that satisfy (5.14) for the values $E_1 = 25$ and $E_2 = 100$. Also shown are selected (and labeled) points $z = (q, p)$ on the outer contour, their images $\mathcal{M}z$ under the action of the stroboscopic Duffing map $\mathcal{M}$, and the vector joining them. That is, the vectors $v(z)$ given by (5.1) are also shown. (For simplicity we have used the same values for β, ϵ, and ω as those employed in making the basins illustration, Figure 23.4.3.) Evidently all the vectors point inward and terminate on some lesser energy contour thereby indicating that $\Delta E < 0$ for all initial conditions that lie on the outer contour, as desired. We further suppose that $\Delta E < 0$ for all initial conditions that lie on all other contours $C(E_k)$ having $E_k > E_2$. By the above discussion we know this will be true if E_2 is large enough. From this supposition we infer that all fixed points of $\mathcal{M}$, if any, must lie within the inner region bounded by the outer contour. For imagine some fixed point lay outside the contour $C(E_2)$. Then there must be some contour $C(E_\ell)$, with $E_\ell > E_2$, on which this fixed point lies. But, if we use this fixed point as an initial condition, we must have $\Delta E = 0$ since for this point $\mathcal{M}z = z$ and therefore E cannot change. We have arrived at a contradiction; therefore no fixed points lie outside $C(E_2)$. Similarly we conclude that, for each choice of the parameters β, ϵ, and ω (with $\beta > 0$), the fixed points of $\mathcal{M}$ must lie within a bounded region of phase space surrounding the origin.

Even more can be said. Figure 5.4 shows the vectors $v(z)$ labeled and drawn from a common origin as in Figure 5.1. Evidently these vectors make one counterclockwise revolution as the points on $C(E_2)$ make one counterclockwise revolution. We conclude that, for any choice of $\beta > 0$, ϵ, and ω, the index of $C(E_2)$ is $+1$ providing E_2 is large enough. Therefore, by (5.12), the stroboscopic Duffing map must have at least one fixed point for each choice of parameters, and the number of fixed points with positive index must exceed by one the number with negative index. Observe that all the fixed points described in Chapter 23 have this property.

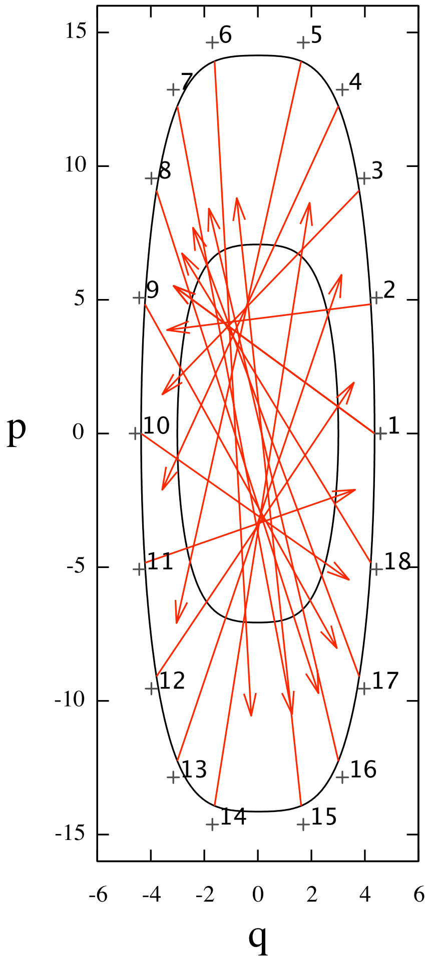


Figure 29.5.3: The contours $C(E_1)$ (inner) and $C(E_2)$ (outer) for the values $E_1 = 25$ and $E_2 = 100$. Also shown are the vectors $v(z)$ for selected (and labeled) points z on $C(E_2)$. Observe that all vectors point inward and terminate on some lesser (inner) energy contour. The Duffing parameters have the values $\omega = 2.25$, $\beta = .1$, and $\epsilon = 1.5$.

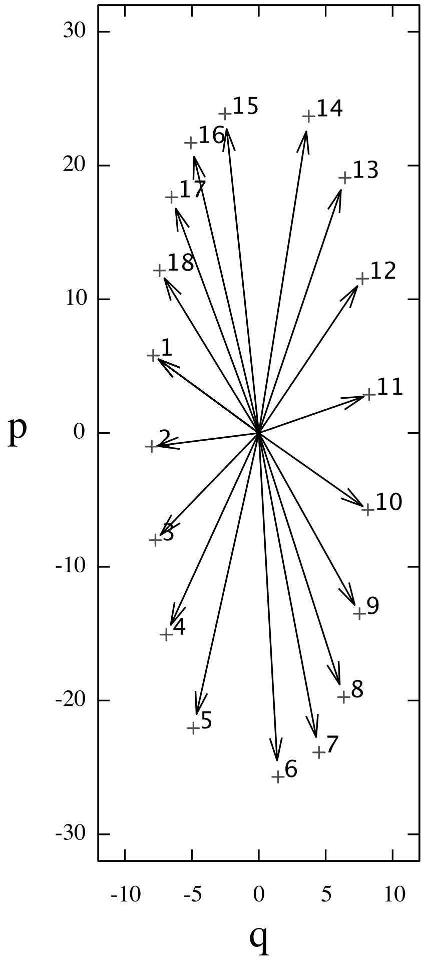


Figure 29.5.4: The vectors $v(z)$ of Figure 5.3 labeled and drawn from the common origin $(0, 0)$. As the points z on $C(E_2)$ make one counterclockwise revolution, the vectors $v(z)$ undergo one counterclockwise revolution, thereby indicating that $C(E_2)$ has index $+1$.

Let us return to our main discussion. From what we have learned it follows that fixed points cannot be born or die (disappear) singly as some parameter τ in a problem is varied. To see this, surround the region where the fixed point is to be born or die by some closed curve C that is free of fixed points (and therefore has a well-defined index) and surrounds no other fixed points. Before the fixed point is born or after it dies, the index of C will be zero. (Reader, verify that the index of a closed curve that can be deformed to a point without encountering any fixed points must be zero.) But since the index is also a continuous function of τ, it must remain zero as τ is varied, and therefore no single fixed point can appear or disappear inside C. Instead, fixed points must be born or die in pairs, and the two fixed points in a pair must be initially infinitesimally nearby and have opposite indices. Moreover, from the discussion leading to (4.7), an eigenvalue of L_α must have the value $+1$ at the birth or death of a pair. This is what occurs at saddle-node bifurcations. As an example, again see the discussion of the Duffing stroboscopic map in Chapter 23.

It can also happen that the index of a fixed point can change as some parameter is varied. If this happens, additional fixed points with compensating indices must also be born or disappear infinitesimally nearby in such a way that the sum of all the indices remains unchanged. This is what occurs at pitchfork bifurcations. Once again see the discussion of the Duffing stroboscopic map in Chapter 23.

The tools used to analyze the fixed points of a map $\mathcal{M}$ can also be applied to analyze the fixed points of $\mathcal{M}^2$. Suppose z_f is a fixed point of $\mathcal{M}$ so that we may write

$$\mathcal{M}(z_f+\delta)=z_f+L_f\delta+O(\delta^2) \tag{29.5.17}$$

where L_f is the linear part of $\mathcal{M}$ about the point z_f. The map $\mathcal{M}^2$ will also have z_f as a fixed point, and we find the relation

$$\mathcal{M}^2(z_f+\delta)=\mathcal{M}(z_f+L_f\delta)+O(\delta^2)=z_f+(L_f)^2\delta+O(\delta^2). \tag{29.5.18}$$

We see that at a fixed point of $\mathcal{M}$ the linear part of $\mathcal{M}^2$ is the square of the linear part of $\mathcal{M}$. Moreover, if λ is an eigenvalue of L_f with associated eigenvector v, then we find that

$$(L_f)^2v=L_f\lambda v=\lambda^2v. \tag{29.5.19}$$

Thus, at a fixed point of $\mathcal{M}$ the eigenvalues of the linear part of $\mathcal{M}^2$ are the squares of the eigenvalues of the linear parts of $\mathcal{M}$.

Now suppose that z_f is initially stable and, as some parameter is varied, suppose also that some eigenvalue of L_f takes the value -1. Then we expect that z_f will become unstable, but will remain isolated. However, the linear part of $\mathcal{M}^2$ will have $+1$ as an eigenvalue, and we may expect bifurcation for $\mathcal{M}^2$ as some parameter is varied. After the bifurcation, $\mathcal{M}^2$ may be expected to have three fixed points. One of them will be the now unstable fixed point of $\mathcal{M}$ itself. The remaining two, call them z_a and z_b, will not be fixed points of $\mathcal{M}$ (since z_f remains isolated), but will have the property

$$\mathcal{M}z_a=z_b, \tag{29.5.20}$$

$$\mathcal{M}z_b=z_a, \tag{29.5.21}$$

from which it follows, of course, that

$$\mathcal{M}^2 z_a = \mathcal{M} z_b = z_a, \tag{29.5.22}$$

$$\mathcal{M}^2 z_b = \mathcal{M} z_a = z_b. \tag{29.5.23}$$

Thus, *period doubling* has occurred. Note that, strictly speaking, a period doubling bifurcation is not a bifurcation of $\mathcal{M}$, for there is only one fixed point of $\mathcal{M}$ both before and after the bifurcation. Rather, it is a bifurcation of $\mathcal{M}^2$, for there is one fixed point of $\mathcal{M}^2$ before the bifurcation (namely z_f), and three fixed points after the bifurcation (z_f, z_a, and z_b). Finally, from index considerations, z_a and z_b will be stable fixed points of $\mathcal{M}^2$. This must be true because, before the bifurcation, z_f had index $+1$ with respect to $\mathcal{M}^2$; and after the bifurcation, z_f must have index -1 with respect to $\mathcal{M}^2$. Thus, to preserve the total index, the fixed points z_a and z_b must have index $+1$ with respect to $\mathcal{M}^2$.

Note that this is what can happen if a storage ring is operated near a half-integer tune. Then, under perturbation, an eigenvalue pair can leave the unit circle through the value -1 so that the fixed point corresponding to the closed orbit becomes inversion hyperbolic. See Figure 3.4.1. The closed orbit persists under perturbation, but becomes unstable. At the same time, a pair of fixed points of $\mathcal{M}^2$ appears, and these fixed points are stable. Correspondingly, there are then two nearby stable orbits that close after *two* turns (but not one turn). Under normal circumstances such a situation is undesirable because the beam is then less well confined. A possible counter-situation might arise if one were thinking about beam extraction.

What can be said about the eigenvalues of the linear part of $\mathcal{M}^2$ at the fixed points z_a, and z_b? Suppose that (5.20) and (5.21) hold. Then we may write the relations

$$\mathcal{M}(z_a + \delta) = z_b + L_a\delta + O(\delta^2), \tag{29.5.24}$$

$$\mathcal{M}(z_b + \delta) = z_a + L_b\delta + O(\delta^2), \tag{29.5.25}$$

from which it follows that

$$\mathcal{M}^2(z_a + \delta) = \mathcal{M}(z_b + L_a\delta) + O(\delta^2) = z_a + L_bL_a\delta + O(\delta^2), \tag{29.5.26}$$

$$\mathcal{M}^2(z_b + \delta) = \mathcal{M}(z_a + L_b\delta) + O(\delta^2) = z_b + L_aL_b\delta + O(\delta^2). \tag{29.5.27}$$

We see that the linear part of $\mathcal{M}^2$ about the fixed point z_a is described by the matrix L_bL_a, and the linear part about the fixed point z_b is described by the matrix L_aL_b. According to Exercise 3.17.15, these matrices have the same eigenvalues. Therefore, the linear parts of $\mathcal{M}^2$ about the corresponding fixed points z_a and z_b have the same eigenvalues.

Exercises

29.5.1. In the two-dimensional case define the matrix B by writing

$$B = L_\alpha - I, \tag{29.5.28}$$

and consider the locus of points Δ produced by the relation

$$\Delta = B\delta \tag{29.5.29}$$

as δ traces out a circle about the origin in the counterclockwise direction. Using (orthogonal) polar decomposition, write B in the form

$$B = PO. \tag{29.5.30}$$

Since P is positive-definite symmetric, there is a proper orthogonal matrix R such that

$$P = RDR^{-1} \tag{29.5.31}$$

where D is a diagonal matrix of the form

$$D = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix} \tag{29.5.32}$$

and the Λ_i are positive. Correspondingly, verify that B can be rewritten as

$$B = RDO' \tag{29.5.33}$$

where O' is the orthogonal matrix given by the relation

$$O' = R^{-1}O. \tag{29.5.34}$$

Now there are two possibilities: First, it could happen that O' is proper orthogonal (has determinant $+1$) in which case we write $O' = R'$ and (5.22) becomes

$$B = RDR'. \tag{29.5.35}$$

Show that for this possibility

$$\det B > 0. \tag{29.5.36}$$

In the second possibility O' has determinant -1. Show that in this possibility O' can be written in the form

$$O' = R'\sigma^3 \tag{29.5.37}$$

where R' is again a proper orthogonal matrix and σ^3 is the improper orthogonal (reflection) matrix given by (5.7.3). Now (5.22) becomes

$$B = RDR'\sigma^3 \tag{29.5.38}$$

and

$$\det B < 0. \tag{29.5.39}$$

In the first possibility we have the relation

$$\Delta = RDR'\delta. \tag{29.5.40}$$

For this possibility verify that as δ traces out a counterclockwise circle, the matrix R' simply rotates this circle by a fixed amount, D squashes and/or stretches it, and R again rotates it by a fixed amount. Verify that the net result is that Δ traces out an ellipse in the counterclockwise direction. Correspondingly, the fixed point z^α has index $+1$, and this fact is to be correlated with the condition (5.25).

In the second possibility we have the relation

$$\Delta = RDR'\sigma^3\delta. \tag{29.5.41}$$

Show that $(\sigma^3\delta)$ traces out a circle in the *clockwise* direction when δ traces out a counterclockwise circle. Referring to (5.30), show that R' now rotates this circle by fixed amount, D squashes and/or stretches it, and R again rotates it by a fixed amount. Verify that the net result is that Δ now traces out an ellipse in the *clockwise* direction. Correspondingly, the fixed point z^α has index -1, and this fact is to be correlated with the condition (5.28).

Verify the index assignments for all the fixed-point cases listed at the beginning of this section.

29.5.2. For N as given by (5.11) show that $\det(N-I) \geq 0$. In the case that $\det(N-I) = 0$, show that $\mu = 1$ and $\phi = 2n\pi$. You may use (5.10).

29.5.3. Let $\mathcal{M}$ be the map defined by

$$\mathcal{M} = \exp : h : \tag{29.5.42}$$

where

$$h = \lambda(p^2 + q^2)^2 \tag{29.5.43}$$

and λ is a parameter. Show that $\mathcal{M}$ has the origin as an isolated, but not manifestly isolated, fixed point. Find the index of this fixed point. Show that $\mathcal{M}$ also has an infinite number of fixed points that are not isolated. Review Exercise 4.1. Show that $f = (p^2 + q^2)$ is an integral of $\mathcal{M}$. At what fixed points does $\nabla f = 0$?

29.5.4. Let $\mathcal{M}$ be the map defined by

$$\mathcal{M} = \exp : h : \tag{29.5.44}$$

where

$$h = p^3 - 3q^2p. \tag{29.5.45}$$

Show that $\mathcal{M}$ has the origin as an isolated, but not manifestly isolated, fixed point. Find the index of this fixed point. Hint: Review Exercise 21.3.13. For the function $w = f(z) = z^2$, verify that if z traces out in a counter-clockwise direction a circular path about the origin of the complex plane, then w goes about the origin twice in the counter-clockwise direction, and $\bar{w}$ goes about the origin twice in the clockwise direction. To do this, write z in the polar form $z = r\exp(i\theta)$. Comment: Let p, q, and r be three Cartesian coordinates. The surface $r = p^3 - 3q^2p$ is called a *monkey saddle.* A cowboy/girl saddle has two up-going parts before and behind the rider, and two down-going parts for the rider's legs. A monkey saddle has three down-going parts, two for the monkey's legs and one for its tail. The monkey saddle also has three up-going parts, with one such up-going part lying between every down-going part. Google on *monkey saddle* to see pictures of them.

29.5.5. Review Section 1.2.2. Show that the complex logistic map $\mathcal{M}$ given by (1.2.29) has the fixed points

$$z_f = 0 \tag{29.5.46}$$

and

$$z_f = (\gamma - 1)/\gamma = 1 - 1/\gamma. \tag{29.5.47}$$

See (1.2.8) and (1.2.9). Exercise 1.1 provides the linear part of $\mathcal{M}$ about the fixed point (5.35). Find the linear part of $\mathcal{M}$ about the second fixed point (5.36). Classify each of these fixed points according to its stability and index, and give each its associated name according to the possibilities 1 through 3 found in this section. According to Exercise 1.2.6 the complex logistic map also has ∞ as a fixed point. Classify this fixed point. Consider a large circular contour C about the origin, large enough to contain both the fixed points (5.35) and (5.36). Compute the index of C. Explain why Douady's γ value, see Figures 1.2.6 and 1.2.8 and Exercise 1.2.12, results in $\mathcal{M}$ having three complex period-three fixed points z^1, z^2, and z^3. What value should γ have for $\mathcal{M}$ to have four complex period-four fixed points? Would this question be any easier to answer for an analysis about the fixed point at the origin using (1.41)? Hint: Given γ, define γ' by the rule

$$\gamma' = 2 - \gamma. \tag{29.5.48}$$

Show that γ and γ' lead to the same value of μ. See Exercise 1.2.7. Call Douady's γ value γ_D,

$$\gamma_D = 2.55268 - .959456i \tag{29.5.49}$$

and define γ'_D by the relation

$$\gamma'_D = 2 - \gamma_D = -.55268 + .959456i. \tag{29.5.50}$$

Figure 5.5 shows the basin structure structure for the complex logistic map $\mathcal{M}$ when $\gamma = \gamma'_D$. Evidently, as expected, the result is again a rabbit (or cactus), but the rabbit now sits more symmetrically on the page. The period-three fixed points are located at

$$z^1 = 0.500003730675024 + (6.968273875812428d - 6)i \;\; \text{(red)}, \tag{29.5.51}$$

$$z^2 = -0.138169999969259 + (0.239864000061970)i \;\; \text{(green)}, \tag{29.5.52}$$

$$z^3 = -0.238618870661709 - (0.264884797354373)i \;\; \text{(yellow)}. \tag{29.5.53}$$

Show that the fixed points of $\mathcal{M}$ itself are located at

$$z_f = 0, \tag{29.5.54}$$

$$z_f = 1 - 1/\gamma'_D = 1.450795 + .7825835i, \tag{29.5.55}$$

$$z_f = \infty. \tag{29.5.56}$$

Verify from Figure 5.5 that the three major lobes on the left, which contain the period-three fixed points z^1, z^2, and z^3, meet at the fixed point $z_f = 0$, and their counterparts on the right meet at the point

$$z' = 1. \tag{29.5.57}$$

Verify analytically that

$$\mathcal{M}z' = z_f = 0, \tag{29.5.58}$$

which explains the symmetry present in Figure 5.5.

Locate γ'_D in Figure 1.2.7. You should find that it lies in the sprout located at the eleven-o'clock position of the left disc. Show, referring to Exercise 1.1, that it has the polar decomposition

$$\rho = 1.1072538, \tag{29.5.59}$$

$$\phi = 119.943^\circ \simeq 2\pi/3. \tag{29.5.60}$$

It follows, see Exercise 1.1, that the effect of $\mathcal{M}$ on points near the origin consists of a counterclockwise rotation about the origin of very nearly 120° followed by scaling (dilation) by a factor of ρ. Consequently, verify that the relations (1.2.33) through (1.2.38) again hold. What value should γ' have in order to achieve $\phi \simeq 90^\circ$ and ρ slightly greater than 1?

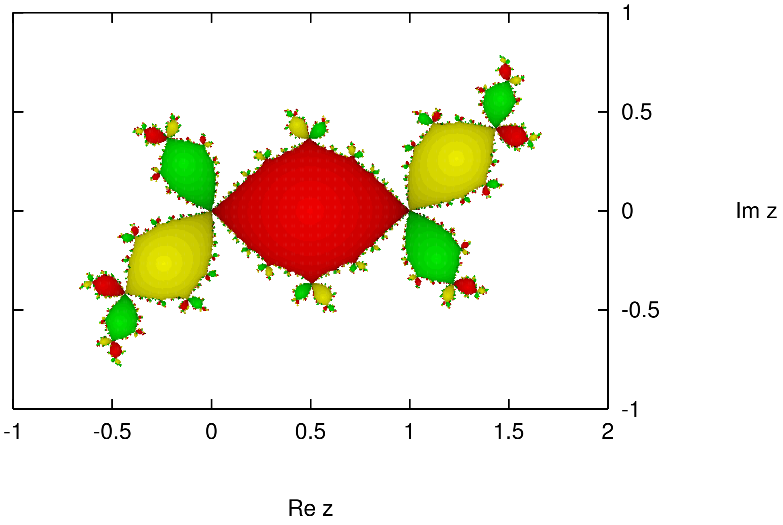


Figure 29.5.5: The basin structure for the complex logistic map when $\gamma = \gamma'_D = -.55268 + .959456i$. The origin is a repelling fixed point, and the three lobes that meet at the origin each contain one of the three attracting period-three fixed points z^1 (red), z^2 (green), and z^3 (yellow). The action of $\mathcal{M}$ on points near the origin is essentially a counterclockwise rotation about the origin by 120°.

29.5.6. Suppose C is a closed curve that is free of fixed points (of some map $\mathcal{M}$ of some two-dimensional space into itself) and suppose that C can be deformed to a point without encountering any fixed points. That is, C does not surround any fixed points. Show that the index of C must be zero.

29.6 Manifolds, and Homoclinic Points and Tangles

Assume for the time being that phase space is 2 dimensional. Suppose z^α is a fixed point of $\mathcal{M}$ and consider the repeated action of $\mathcal{M}$ (and $\mathcal{M}^{-1}$) on points near z^α. In the linear approximation these points are *translates* from the origin to z^α of points obtained by the repeated action of L_α (and ${L_\alpha}^{-1}$) on points near the origin. See Exercise 6.1 and (6.10). Next consider the two-dimensional case and suppose that z^α is a *hyperbolic* fixed point. As Exercise 6.1 goes on to show, in the hyperbolic case the repeated action of $\mathcal{M}$ on points near z^α produces (in the linear approximation) points lying on (transformed) hyperbolas and their asymptotes [or (transformed) generalized hyperbolas and their asymptotes] centered on z^α. Figure 6.1 illustrates these points schematically. The lines $v_<$ and $v_>$ are the translates of the points σv_1 and σv_2 described in Exercise 6.1, and they form the asymptotes of the hyperbolas. According to (6.10), under the repeated action of $\mathcal{M}$ (and in the linear approximation) points on $v_<$ are moved inward toward z^α, and points on $v_<$ are moved outward. See (6.20) and recall that we have assumed $\lambda_1 < 1$ and $\lambda_2 > 1$. Thus, points on $v_<$ are *stable* in the linear approximation, and points on $v_>$ are *unstable*. Points not on the asymptotes move on hyperbolas. They are all unstable since they eventually move away from z^α under the repeated action of $\mathcal{M}$.

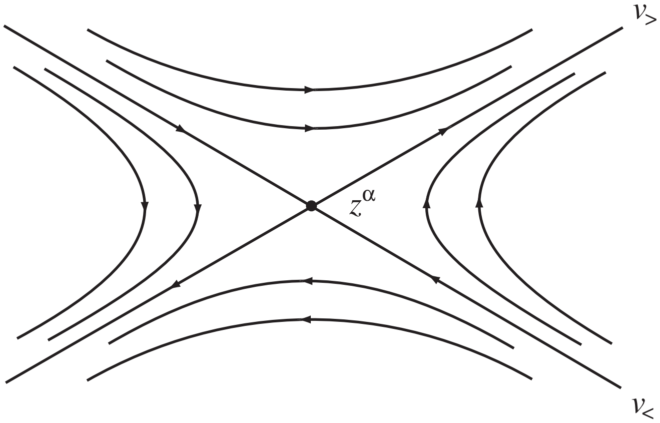


Figure 29.6.1: Schematic illustration of the action of $\mathcal{M}$ on points near z^α in the linear approximation. Points on $v_<$ are moved inward toward z^α, and points on $v_>$ are moved outward. Others are moved on hyperbolas.

When all nonlinearities are taken into account, it can be shown that (in the hyperbolic case) the repeated action of $\mathcal{M}$ on points near z^α is similar. Indeed, according to a celebrated theorem of *Hartman* whose proof lies beyond the scope of our current discussion, in the hyperbolic case there is a conjugating map $\mathcal{A}$ such that in the vicinity of z^α there is a relation of the form

$$\mathcal{M} = \mathcal{A}\mathcal{L}_\alpha\mathcal{A}^{-1} \tag{29.6.1}$$

where $\mathcal{A}^{-1}$ maps z^α to the origin and $\mathcal{L}_\alpha$ is the linear part of $\mathcal{M}$ at z^α. In particular there are *stable* and *unstable manifolds* W_s and W_u that are the analogs of the asymptotes $v_<$ and $v_>$, and nearly coincide with them in the vicinity of z^α where the linear approximation is valid. The stable manifold is defined to be the set of all points that are sent into z^α under the action of $\mathcal{M}^n$ in the limit of large n,

$$W_s = \{z| \lim_{n\to\infty} \mathcal{M}^n z = z^\alpha\}. \tag{29.6.2}$$

The definition of W_u is a bit more subtle. Note that in the linear approximation points on $v_>$ are moved *inward* under the action of $\mathcal{M}^{-1}$. We therefore define the unstable manifold to be the set of all points that are sent into z^α under the action of $\mathcal{M}^{-n}$ in the limit of large n,

$$W_u = \{z| \lim_{n\to\infty} \mathcal{M}^{-n} z = z^\alpha\}. \tag{29.6.3}$$

We observe that thanks to Hartman's theorem W_s and W_u are one dimensional in the neighborhood of z^α. Moreover, if S is a point on W_s, then by the definition (6.2) so are the points $\mathcal{M}^m S$ for any positive or negative values of m. Similarly, if U is a point on W_u, so are the points $\mathcal{M}^m U$. Consequently, if W_u is known in the neighborhood of z^α, it can be extended away from z^α by repeatedly applying $\mathcal{M}$ to the known portion. Similarly, W_s can extended away from its known portion near z^α by repeated application of $\mathcal{M}^{-1}$. Finally, since $\mathcal{M}$ is assumed to be continuous (and dimensionality is conserved by a continuous map), W_s and W_u must be one dimensional globally.

An illuminating example of the behavior of stable and unstable manifolds is provided by the map $\mathcal{M}$ given by (3.95) and (3.96). For simplicity, specify μ and ν by a single parameter Λ by writing the relations

$$\mu = \Lambda,\ \nu = 1/\Lambda. \tag{29.6.4}$$

When this is done, the map becomes symplectic. Next, for purposes of illustration, assign Λ the value $\Lambda = 3$. Inspection of (3.95) and (3.96) shows that in this case the origin in q,p space is a hyperbolic fixed point, and the q and p axes are the unstable and stable asymptotes $v_>$ and $v_<$, respectively. We therefore expect the unstable and stable manifolds W_u and W_s to lie along the q and p axes, respectively, in the neighborhood of the origin.

Further examination of (3.95) and (3.96) shows that $\mathcal{M}$ has a second fixed point given (when $\Lambda = 3$) by

$$\{q,p\} = \{-3/8, 1/8\}, \tag{29.6.5}$$

and that this fixed point is elliptic with a tune $T = .1959\cdots$. See Exercise 6.2.

Figure 6.2 shows the actual unstable and stable manifolds in the vicinity of the origin and somewhat beyond. (See Exercise 6.3 for a description of how they can be calculated.) It also displays the elliptic fixed point and the behavior, under repeated action of $\mathcal{M}$, of points near the elliptic fixed point. Evidently points near the elliptic fixed point do seem to move on ellipses as predicted by the linear analysis of Exercise 6.1. However, we will subsequently learn that their behavior is in fact generally much more complicated. With regard to the unstable and stable manifolds, they do appear to lie along the q,p axes in the neighborhood of the origin. However, they veer away from the axes when farther away from the origin. In particular, the piece of W_u that originated along the negative q axis and the piece of W_s

that originated along the positive p axis *intersect* at the point K.[2] Moreover they do not join smoothly at K, but rather *intersect* with a finite angle of intersection. (They are said to intersect *transversally.*) This point K is called a (transverse) *homoclinic* point, and the angle of intersection is called the *homoclinic angle.* (The word *homoclinic*, a nomenclature due to Poincaré, means “falling into itself”: Since K is on W_u, it must have come from points arbitrarily near z^α under the repeated action of $\mathcal{M}$, and since it is also on W_s, it must also be sent back to points arbitrarily near z^α under the repeated action of $\mathcal{M}$. It can happen that a map has two or more hyperbolic fixed points, and that the unstable manifold of one intersects the stable manifold of another. Such an intersection is called a *heteroclinic* point.) The pieces of W_u and W_s that originated along the positive q and negative p axes, respectively, appear to go off to infinity without intersecting.

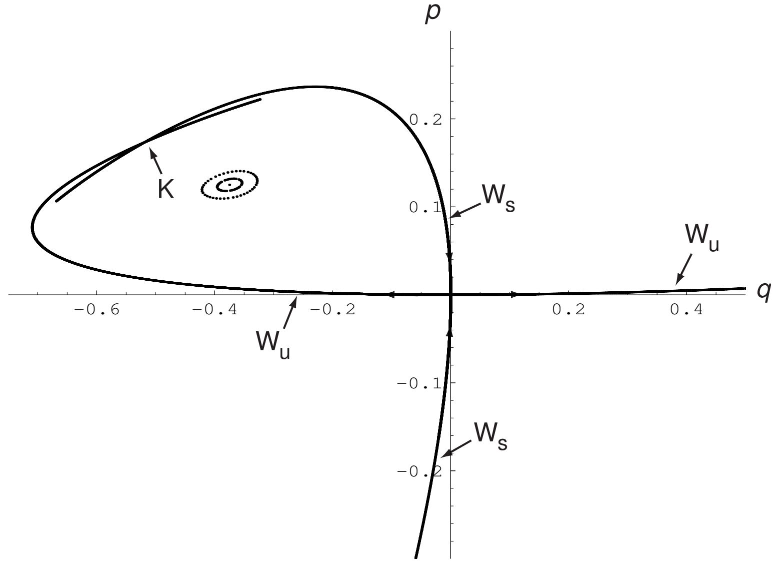


Figure 29.6.2: The (transverse) intersection of the unstable and stable manifolds emanating from a hyperbolic fixed point resulting in a homoclinic point K. Also displayed is the elliptic fixed point and the behavior of points near the elliptic fixed point.

We next claim that the existence of one homoclinic point implies the existence of an infinite number of homoclinic points. To see this, suppose K is a homoclinic point and consider all the points $\mathcal{M}^n K$ for n both positive and negative. Since by assumption K lies on W_u, we know from our earlier discussion that the points $\mathcal{M}^n K$ must also lie on W_u. And since K lies on W_s as well, the points $\mathcal{M}^n K$ must also lie on W_s. Therefore W_u and W_s must intersect at all the points $\mathcal{M}^n K$. Moreover, if the homoclinic angle at K is nonzero,

[2] Suppose an autonomous Hamiltonian system with two degrees of freedom has a periodic orbit. As described in Section 6.9, in the vicinity of this orbit one can set up a Poincaré two-dimensional surface of section return map $\mathcal{M}$, and the periodic orbit will then correspond to a fixed point of $\mathcal{M}$. Suppose this fixed point is hyperbolic. Then, points on its stable manifold correspond to an orbit that asymptotically approaches the periodic orbit as $t \rightarrow +\infty$, and points on its unstable manifold correspond to an orbit that asymptotically approaches the periodic orbit as $t \rightarrow -\infty$. Supose the stable and unstable manifolds intersect. Then these two orbits are part of a common orbit which is said (following Poincaré) to be *doubly asymptotic.*

the homoclinic angles at all the points $\mathcal{M}^n K$ must be finite. See Exercise 6.4.

Figure 6.3 shows the same pieces of the manifolds W_u and W_s that do not go off to infinity as shown on Figure 6.2, but also shows more of them to exhibit some of the points $\mathcal{M}^n K$. We now observe that between any two successive homoclinic points of the form $\mathcal{M}^n K$ there is an *additional* homoclinic point. Why should this be so?

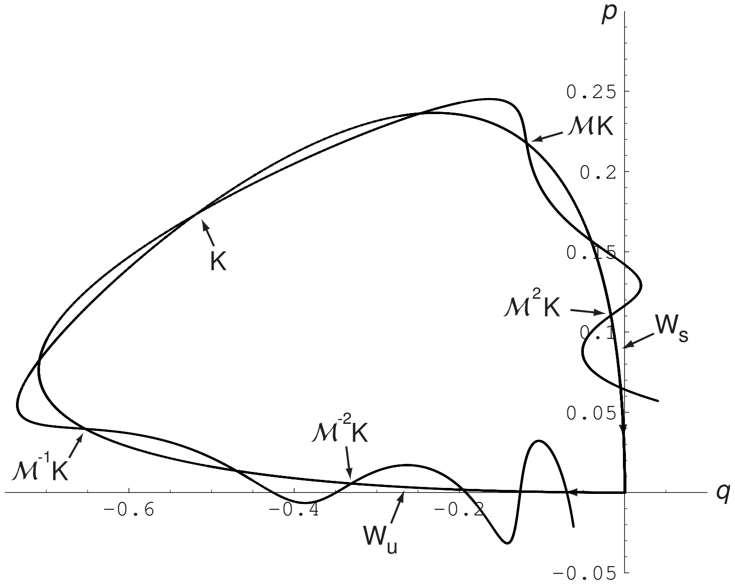


Figure 29.6.3: Successive homoclinic intersections of the unstable and stable manifolds showing the first few points $\mathcal{M}^n K$. The other halves of W_u and W_s, those pieces that go off to infinity, are not shown. Also not shown are the elliptic fixed point and the behavior of points near it.

Consider the closed curve consisting of the part of W_u that extends from z^α to K and the part of W_s that extends from K back to z^α, and let $\mathcal{R}$ be the region enclosed by this curve. Next consider the closed curve consisting of the part of W_u that extends from z^α to $\mathcal{M}K$ and the part of W_s that extends from $\mathcal{M}K$ back to z^α, and let $\bar{\mathcal{R}}$ be the region enclosed by this curve. By construction the boundary of $\mathcal{R}$ is mapped onto the boundary of $\bar{\mathcal{R}}$ under the action of $\mathcal{M}$. It can also be shown that $\mathcal{M}$ maps the interior of $\mathcal{R}$ to the interior of $\bar{\mathcal{R}}$. That is, we may also view $\bar{\mathcal{R}}$ as being the set of points produced by $\mathcal{M}$ acting on all the points in $\mathcal{R}$,

$$\bar{\mathcal{R}} = \mathcal{M}\mathcal{R}. \tag{29.6.6}$$

Let A and $\bar{A}$ be the areas of $\mathcal{R}$ and $\bar{\mathcal{R}}$, respectively,

$$A = \int_{\mathcal{R}} d^2 z, \tag{29.6.7}$$

$$\bar{A} = \int_{\bar{\mathcal{R}}} d^2 \bar{z}. \tag{29.6.8}$$

We claim these two areas are equal. Indeed, by changing variables of integration, we have the result

$$\bar{A} = \int_{\bar{\mathcal{R}}} d^2 \bar{z} = \int_{\mathcal{R}} [\det M(z)] d^2 z = A. \tag{29.6.9}$$

Here we have used the fact that $\mathcal{M}$ is symplectic, and therefore $\det M(z) = 1$. What we have done is recapitulate Liouville's theorem in two dimensions. Recall Subsection 6.6.1.

Now look at Figure 6.4. The left panel displays $\mathcal{R}$; the right displays $\bar{\mathcal{R}}$ and, as a dashed line, the part of W_s that is "removed" under the action of $\mathcal{M}$. We see, because the homoclinic angle is finite, that as a result of the extending action of $\mathcal{M}$ the manifold W_u initially dives under the dashed portion of W_s that is removed. If it did not eventually also curve upward so as to finally lie above the dashed portion of W_s, the result would be $\bar{A} < A$. Therefore W_u must cross W_s between K and $\mathcal{M}K$, thereby producing an intermediate homoclinic point. Moreover, the two small regions (*lobes*) pointed out in the right panel, those that result from the oscillation of W_u about the dashed portion of W_s, must have *equal* areas in order for the relation $\bar{A} = A$ to hold.

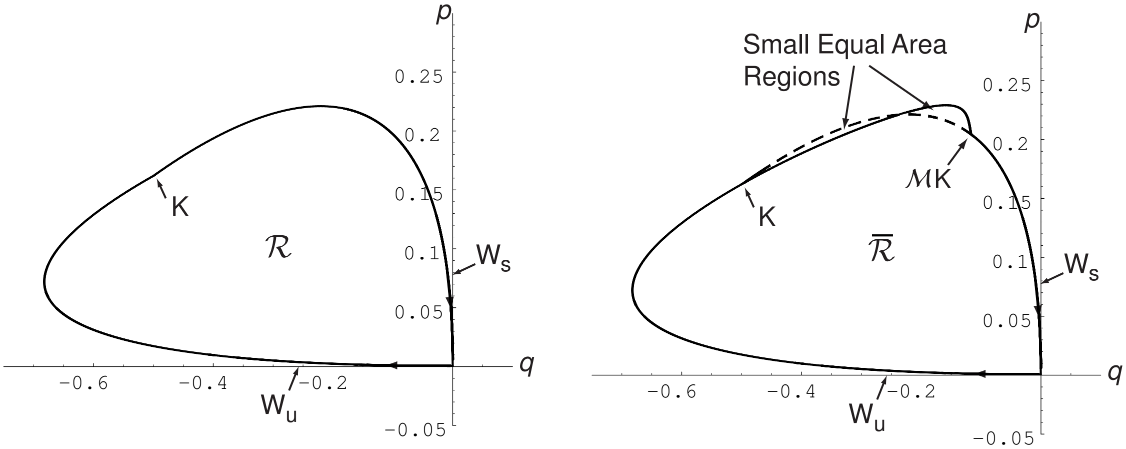


Figure 29.6.4: The regions $\mathcal{R}$ and $\bar{\mathcal{R}}$. Observe that, in the right panel, the unstable manifold "oscillates" about the stable manifold in the interval between K and $\mathcal{M}K$. When $\mathcal{M}$ is symplectic, the two small regions produced by this oscillation must have equal areas.

Let us examine more of the homoclinic points $\mathcal{M}^n K$ by computing more of W_s and W_u. Figure 6.5 shows the result. We see that W_u, as it heads back toward z^α, makes more and more oscillations about W_s and that these oscillations have ever increasing amplitude. Similarly, W_s, as it heads back toward z^α, makes more and more oscillations about W_u with ever increasing amplitude. Why should this be? We have learned that the two lobes in Figure 6.4 must have the same area. A moment's reflection reveals that all the other lobes are images of these lobes under the action of $\mathcal{M}^n$ for suitable positive or negative values of n. See Figure 6.3. Moreover, since $\mathcal{M}$ is symplectic and therefore area preserving, *all* these lobes must have the same area. Finally, in the vicinity of z^α, Exercise 6.1 shows that the spacing of successive homoclinic points must decrease geometrically (exponentially). As a consequence of Hartman's theorem, asymptotically their distances from the origin must be governed by relations of the form (6.15), which in turn imply that their spacings must decrease exponentially. Therefore, in order to maintain constant area, the amplitude (in linear approximation) must grow exponentially. (We remark that a similar phenomena holds in the case of heteroclinic points when the heteroclinic intersection is transverse: The two intersecting manifolds again oscillate about each other with ever finer spacing and ever increasing amplitude as they approach the two associated hyperbolic points.)

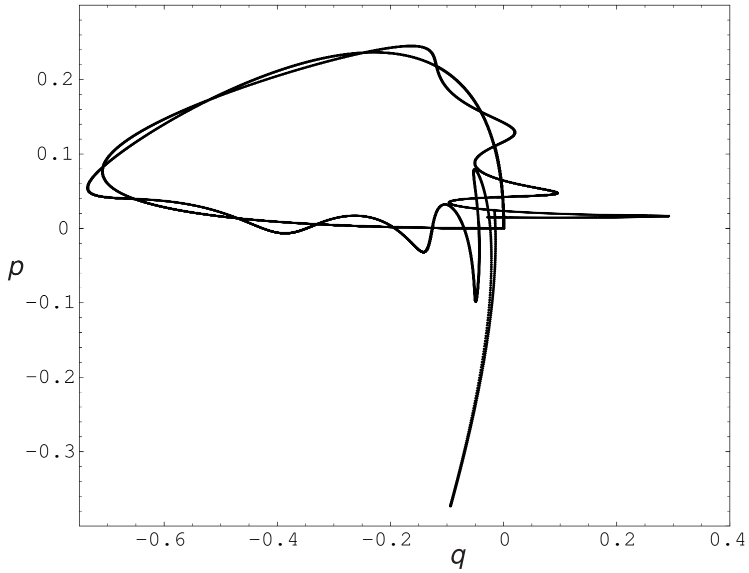


Figure 29.6.5: Successive oscillations of W_u about W_s and of W_s about W_u in the vicinity of the hyperbolic fixed point. The spacing between successive oscillations becomes exponentially finer, and the oscillation amplitude becomes exponentially larger.

The net effect of these oscillations of increasing amplitude is that near the hyperbolic fixed point the oscillations of W_u about W_s must *intersect* the oscillations of W_s about W_u to produce even more homoclinic points. As a result the hyperbolic fixed point is the corner of an ever "denser" cloud of homoclinic points. This property is illustrated in Figure 6.6.

Poincaré first discovered this structure, which we now call a *homoclinic tangle*, without the aid of a computer and computer graphics. About this discovery he wrote:

> *When we try to represent the figure formed by these two curves and their infinitely many intersections, each corresponding to a doubly asymptotic solution, these intersections form a type of trellis, tissue, or grid with infinitely fine mesh. Neither of the two curves must ever cut across itself again, but it must bend back upon itself in a very complex manner in order to cut across all of the meshes in the grid an infinite number of times.*
>
> *The complexity of this figure is striking, and I shall not even try to draw it. Nothing is more suitable for providing us with an idea of the complex nature of the three-body problem, and of all the problems of dynamics in general, where there is no uniform integral and where the Bohlin series are divergent.*

In Exercise 6.5 you will have the pleasure of verifying Poincaré's assertion that W_s cannot intersect itself, nor can W_u intersect itself. And in Exercise 6.6 you will be led to show that the existence of a homoclinic point, which happens generically for nonlinear maps (including those arising from nonlinear systems), precludes the existence of analytic integrals. Indeed, it can be shown that the existence of a homoclinic point implies that there is an infinite

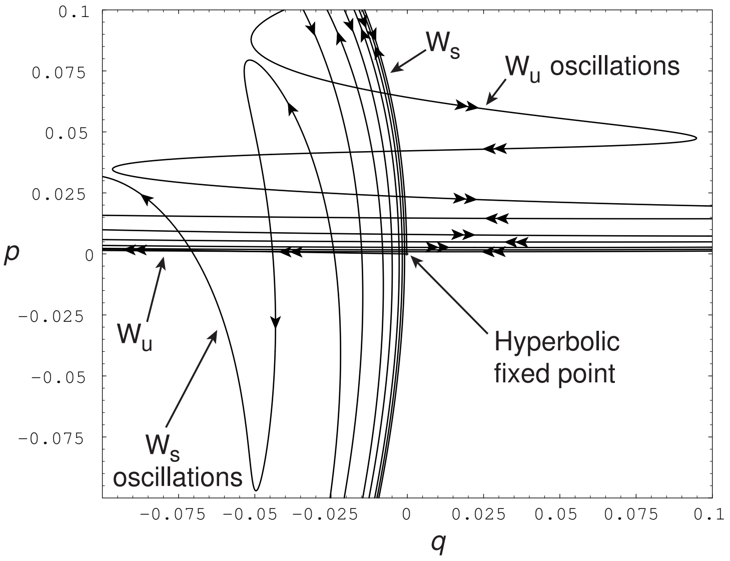


Figure 29.6.6: A continuation of Figure 6.5 near the origin (the hyperbolic fixed point) showing the formation of a grid of intersecting lines. The spacing of the grid becomes finer and finer as it approaches the hyperbolic fixed point. Each grid intersection is a homoclinic point. The result of all these intersections is an ever denser cloud of homoclinic points that has the hyperbolic fixed point as a limit point.

collection of points for which the action of $\mathcal{M}$ is equivalent to that of a Bernoulli shift. (Recall Exercise 1.2.8 for the definition of a Bernoulli shift.) Therefore, most nonlinear problems are not integrable. Finally, it may be remarked that the homoclinic tangle has the same topological structure as that resulting from *Smale's* famous *horseshoe* map.

Strictly speaking, we have only been discussing the two-dimensional case. The higher dimensional cases are still more complicated, and much less detail is known about them. It is known, however, that stable and unstable manifolds and homoclinic (and heteroclinic) behavior also exist in higher dimensions, and that there are additional complex phenomena such as *Arnold diffusion*.

We close this section by noting that, in generically rare cases (but characteristic of maps arising from soluble systems), the unstable and stable manifolds may join smoothly without going into homoclinic oscillations about each other. In such a case their union is called a *separatrix* since (in two dimensions) points inside the separatrix remain so under the repeated action of $\mathcal{M}$, and points outside are eventually mapped to infinity. When the unstable and stable manifolds intersect transversally at a homoclinic point (as is generically the case), this phenomenon is sometimes referred to as separatrix *splitting*.

It can also happen, in systems with sufficient damping, that the stable and unstable manifolds do not intersect at all. As an example, consider the stroboscopic Duffing map with the same parameter values used to make the basins illustration, Figure 23.4.3. Figure 6.7 shows the unstable (*hyperbolic*) fixed point (23.4.3) and the eigenvector v_1, the eigenvector of the linear part of $\mathcal{M}$ about the unstable fixed point, associated with the eigenvalue λ_1 that lies in the interval $0 < \lambda_1 < 1$. Note that the unstable fixed point is on, and v_1 points along, the boundary between the two basins. (Reader, show that this is to be expected.)

Figure 6.8 shows the stable and unstable manifolds associated with the unstable fixed point (23.4.3). They do not intersect, and therefore the stroboscopic Duffing map does not have homoclinic points for these parameter values. Points on the two branches of the unstable manifold spiral into the two stable fixed points (23.4.1) and (23.4.2), the stable manifold separates the two basins, and points on the stable manifold are moved toward the unstable fixed point (23.4.3).

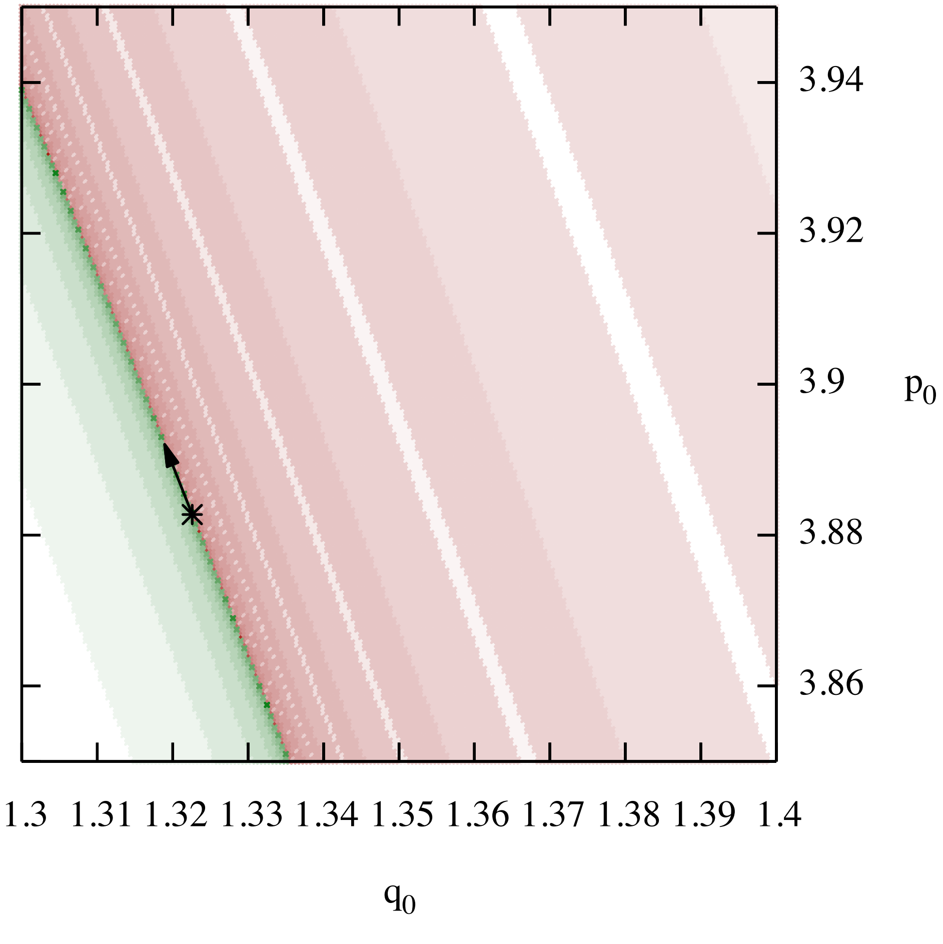


Figure 29.6.7: A blow up of part of Figure 23.4.3 illustrating the basins of attraction, the unstable fixed point (23.4.3), and the eigenvector v_1 (the one with eigenvalue less than 1) for the stroboscopic Duffing map. The unstable fixed point lies on, and the vector v_1 points along, the boundary between the two basins.

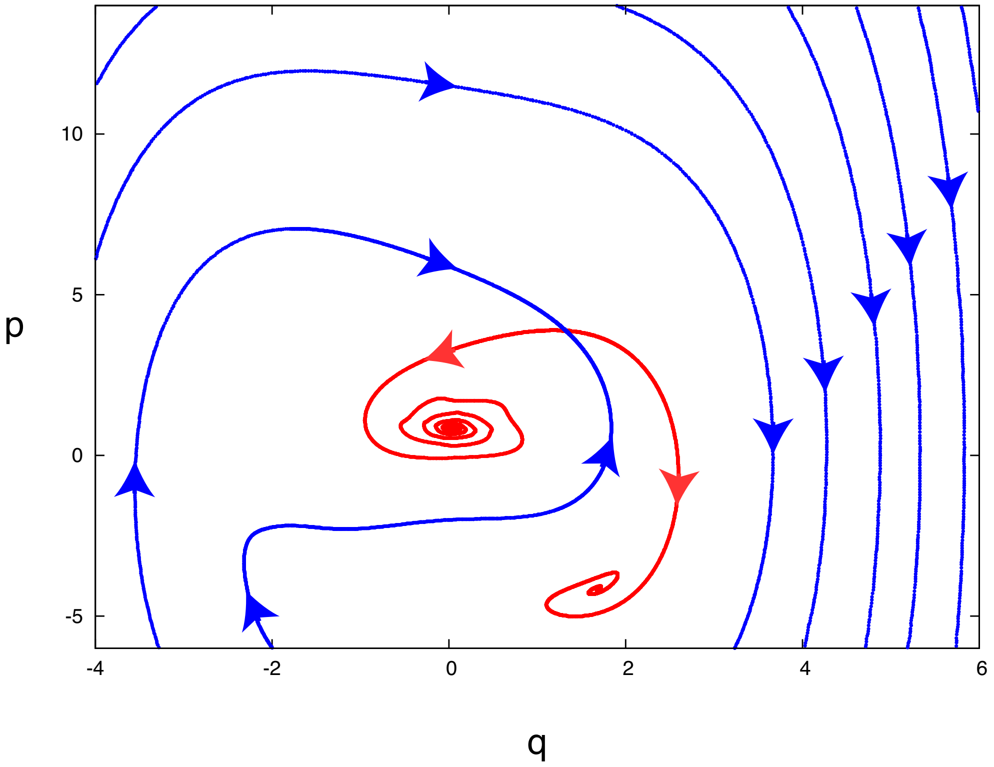


Figure 29.6.8: The stable (blue) and unstable (red) manifolds for the unstable fixed point (23.4.3) of the stroboscopic Duffing map. Note that the unstable manifold spirals into the stable fixed points (23.4.1) and (23.4.2), and the stable manifold lies along the basin boundaries. The stable and unstable manifolds do not intersect, so there are no homoclinic points.

Exercises

29.6.1. Show from (4.5) that there is the relation

$$\mathcal{M}^n(z^\alpha + \delta) = z^\alpha + (L_\alpha)^n \delta + O(\delta^2). \tag{29.6.10}$$

Therefore the behavior of points sufficiently near z^α under repeated action of $\mathcal{M}$ is governed by $(L_\alpha)^n$. From (5.3) there is the relation

$$(L_\alpha)^n = AN^nA^{-1}, \tag{29.6.11}$$

and hence $(L_\alpha)^n$ is in turn governed by N^n.

Now consider the two-dimensional case. For any two-vector $\epsilon^{(0)}$, let $\delta^{(0)} = A\epsilon^{(0)}$ and introduce the notation

$$\epsilon^{(n)} = (N)^n \epsilon^{(0)}, \tag{29.6.12}$$

$$\delta^{(n)} = (L_\alpha)^n \delta^{(0)}. \tag{29.6.13}$$

Deduce the relation

$$\delta^{(n)} = A\epsilon^{(n)}. \tag{29.6.14}$$

For any initial point $\epsilon^{(0)}$, find the successive points $\epsilon^{(n)}$ for all the kinds of fixed points listed at the beginning of this section.

Three cases are of particular interest: Show that in case 3b (see Section 18.5) the successive points $\epsilon^{(n)}$ spiral into the origin; and in the elliptic case, case 3c, the successive points lie on a circle. For the hyperbolic case, case 1c, Show that successive points obey the relation

$$(\epsilon^{(n)})_i = {\lambda_i}^n (\epsilon^{(0)})_i \tag{29.6.15}$$

so that they lie on the "generalized" hyperbola

$$[(\epsilon^{(n)})_1]^a [(\epsilon^{(n)})_2]^{1/a} = [(\epsilon^{(0)})_1]^a [(\epsilon^{(0)})_2]^{1/a} \tag{29.6.16}$$

where a is given by the relation

$$a = [-\log(\lambda_2)/\log(\lambda_1)]^{1/2}. \tag{29.6.17}$$

Let e_1 and e_2 be unit vectors such that

$$Ne_i = \lambda_i e_i. \tag{29.6.18}$$

Show that the vectors σe_i, where σ is any scalar, are the asymptotes of this generalized hyperbola. Verify that the generalized hyperbola becomes an "ordinary" hyperbola in the (symplectic) case $\lambda_1\lambda_2 = 1$.

Finally, verify that the effect of A, which transforms (maps) $\epsilon^{(n)}$ to $\delta^{(n)}$ by (6.14), is simply that of possible reflection followed by possible rotation, possible squashing and/or stretching, and a final possible rotation. Thus, the behavior of the points $\delta^{(n)}$ is similar to that of the $\epsilon^{(n)}$. Hint: As done for B in Exercise 5.1, use polar decomposition for A. When, without loss of generality, may one assume $\det A = 1$?

For the hyperbolic case in particular define vectors v_i by the relation

$$v_i = Ae_i. \tag{29.6.19}$$

Verify that they obey the rule

$$L_\alpha v_i = \lambda_i v_i, \tag{29.6.20}$$

and that the vectors σv_i are the asymptotes of the transformed generalized hyperbola.

29.6.2. Show that the map $\mathcal{M}$ given by (3.95), (3.96), and (4.23) has two fixed points. The first is the origin $\{0, 0\}$. Verify that the second is given by

$$q_f = -\Lambda(\Lambda - 1)/(\Lambda + 1)^2, \tag{29.6.21}$$

$$p_f = (\Lambda - 1)/(\Lambda + 1)^2. \tag{29.6.22}$$

Verify (6.5) for the case $\Lambda = 3$. Show that the linear part of $\mathcal{M}$ about the second fixed point is described by the matrix

$$M_f = \begin{pmatrix} -\Lambda(\Lambda-3)/(\Lambda+1) & 2\Lambda(\Lambda-1)/(\Lambda+1) \\ -2(\Lambda-1)/[\Lambda(\Lambda+1)] & (3\Lambda-1)/[\Lambda(\Lambda+1)] \end{pmatrix}, \tag{29.6.23}$$

and therefore the trace of M_f is given by the relation

$$\operatorname{tr} M_f = -(\Lambda^2 - 4\Lambda + 1)/\Lambda. \tag{29.6.24}$$

Verify, using (3.5.39) through (3.5.41), that when $\Lambda = 3$ the tune T of M_f is given by

$$T = .1959\cdots . \tag{29.6.25}$$

29.6.3. Verify that the map $\mathcal{M}$ given by (3.95), (3.96), and (4.23) has the inverse

$$q = \bar{q}/\Lambda - (\bar{q}/\Lambda - \Lambda\bar{p})^2, \tag{29.6.26}$$

$$p = \Lambda\bar{p} - (\bar{q}/\Lambda - \Lambda\bar{p})^2. \tag{29.6.27}$$

Set $\Lambda = 3$. Consider L equally spaced points along the q axis lying in the small interval $[-\epsilon, 0)$ where ϵ is some small number. Also consider L equally spaced points along the p axis lying in the small interval $(0, \epsilon]$. Let N be some modest positive integer. Apply the maps $\mathcal{M}^n$, for $n = 0$ through N, to the equally spaced set of points along the q axis. Apply the maps $\mathcal{M}^{-n}$, again for $n = 0$ through N, to the equally spaced set of points along the p axis. Use double precision (64 bit) or still higher precision arithmetic. Try, for starters, the case $L = 100$, $\epsilon = .03$, and $N = 5$. Verify that so doing should give (for the case $\Lambda = 3$) some approximation to points lying on the initial portion of some pieces of W_u and W_s, respectively. What pieces can be gotten this way? How can the other pieces be found? For an improved approximation, replace ϵ by ϵ/Λ and N by $N + 1$. Verify that making this replacement repeatedly leads to more and more accurate results for W_u and W_s. Experiment with various values of L, ϵ, and N to get satisfactory graphics.

29.6.4. All homoclinic angles are finite.

29.6.5. W_u and W_s cannot self-intersect.

29.6.6. Existence of homoclinic point implies non-integrability.

29.7 The General Hénon Map

We are now prepared to resume our study of two-variable quadratic maps. Consider the map $\mathcal{M}_{ffq}$ given by (3.37) and (3.38), or by (3.54), for the parameter values

$$d = 1 \ , \ e = 0 \ , \ r = 0 \ , \ s = 1 \ , \ t = b \ , \ u = 0 \ , \ \alpha^3 = -a \ , \ \beta = 0. \tag{29.7.1}$$

This map is called the *general* (not necessarily symplectic) *Hénon* map. We will denote it by $\mathcal{M}_h$. For it the matrix R takes the form

$$R = \begin{pmatrix} 0 & 1 \\ b & 0 \end{pmatrix}; \tag{29.7.2}$$

and, after replacing $\overset{\equiv}{q}$, $\overset{\equiv}{p}$ by the symbols $\bar{q}$, $\bar{p}$ for the sake of improved notation, the map $\mathcal{M}_h$ itself takes the form

Action of $\mathcal{M}_h$:

$$\bar{q} = 1 + p - aq^2, \tag{29.7.3}$$

$$\bar{p} = bq. \tag{29.7.4}$$

The determinant of its Jacobian matrix has the value

$$\det M_h = \det R = -b. \tag{29.7.5}$$

Consequently, the Hénon map is orientation preserving only when $b < 0$, and symplectic only when $b = -1$.

We have discovered that the general Hénon map is an instance of the class of two-dimensional quadratic maps whose nonlinear parts have a finite product Lie factorization (or, equivalently, have a constant Jacobian determinant.) In Section 8.3 we learned that all such maps are equivalent (conjugate), under affine changes of variables, to a two-parameter family of maps. See, for example, (3.95) and (3.96). Note that the Hénon maps also form a two-parameter family. Therefore, as we will eventually confirm, it must also be possible to bring any Hénon map to some standard form. Indeed, the general Hénon map and the class of quadratic two-dimensional maps having a finite Lie product factorization are equivalent.

The general Hénon map has been much studied for the case $b = .3$ and variable a. This case is non-symplectic and non-orientation preserving. For given values of a and b, repeated iteration of $\mathcal{M}_h$ will produce a set in the q, p plane. Figure 7.1 shows the projection of this set onto the q axis as a a function of a. Projection of this set on the p axis gives a similar picture. [See (7.4).] Evidently the Feigenbaum diagram for the general Hénon map is similar to that of the logistic map. See Figures 1.2.2 and 1.2.10. There is a cascade at period doublings followed by what appears to be chaotic behavior. Numerical calculation shows that the sequence of period doublings is again governed by the Feigenbaum constant δ as given by (1.2.13). However, closer inspection of Figure 7.1 reveals additional features in the vicinity of $a = 1.08$ in the form of additional cascades. Figure 7.2 shows an enlargement of one of these features revealing that it is an independent cascade. This is but one indication that the behavior of the Hénon map under iteration is far more complicated than that of the logistic map.

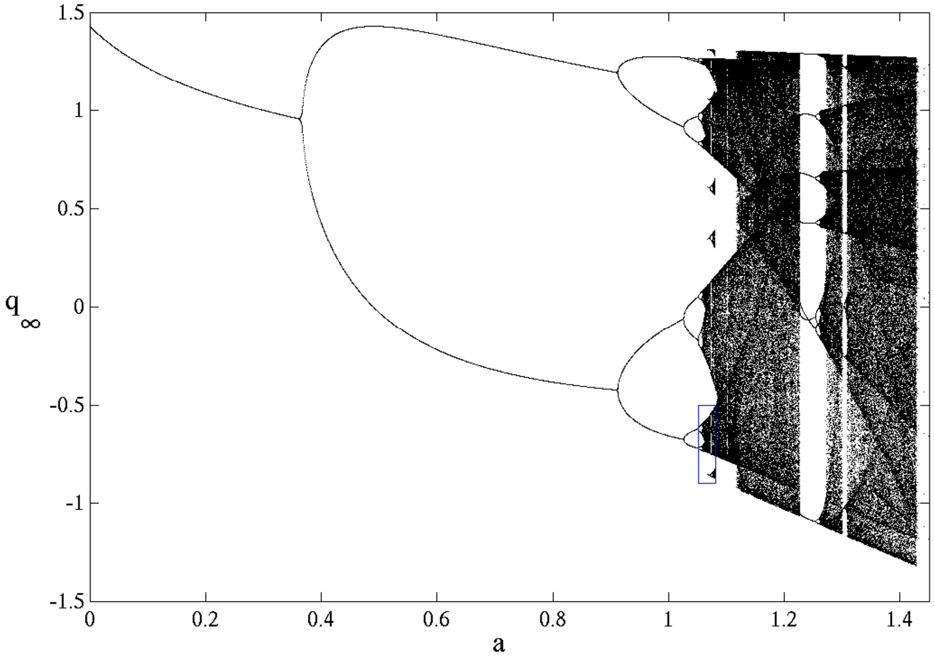


Figure 29.7.1: Feigenbaum diagram showing limiting values q_∞ as a function of a (and b held at $b = +.3$) for a non-orientation preserving case of the general Hénon map.

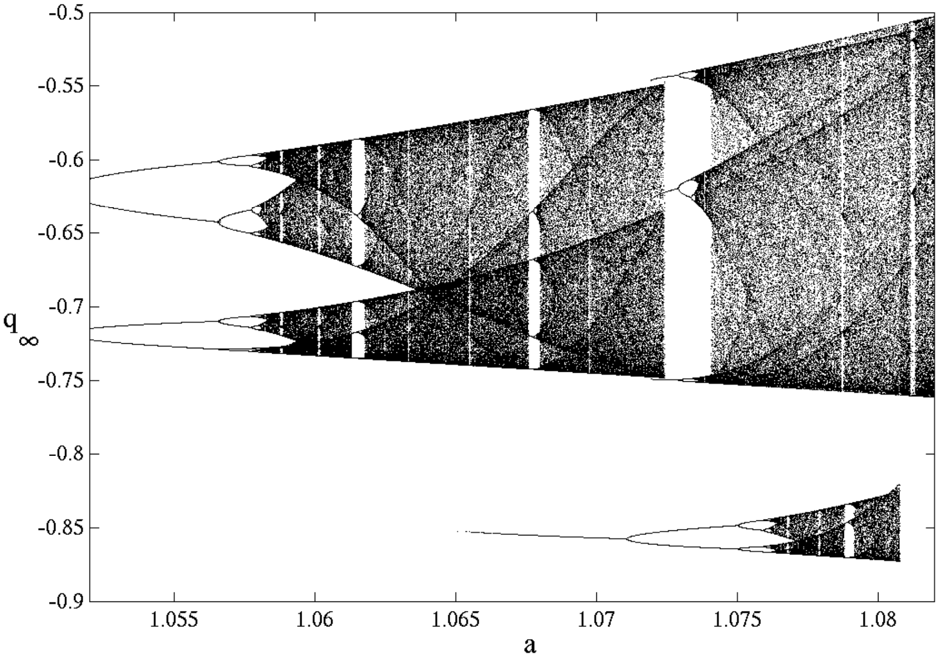


Figure 29.7.2: Enlargement of the boxed region in Figure 7.1. The upper cascade is that readily visible in the box in Figure 7.1. The lower cascade, which seems to appear out of nowhere and then terminate abruptly, corresponds to the small speck near the bottom of the box in Figure 7.1.

Figures 7.3 show various portions of the full set in the q, p plane for the parameter values $b = .3$ and $a = 1.4$. In the first figure there seem to be three curves that form part of an attractor. The next figure shows an enlargement of the "boxed" portion of the first figure. At this magnification it is evident that the top curve in the first figure is "thicker" then the other two. The third figure shows an enlargement of the boxed portion of the second figure. Now it is evident that the thick curve is itself composed of three curves spaced in a way that is similar to the three curves in the first figure. A fourth figure shows a further enlargement. Evidentally the attractor appears to be *fractal*.

There is also the (for us) much more interesting *orientation preserving* case for which $b < 0$, for only then can the general Hénon map model a transfer map arising from a differential equation. [See (7.5) and Exercise 1.4.6.] Indeed, by following a procedure similar to that used in Section 1.2 to produce the symplectic Hénon map, one can use the vector fields appearing in an analogous factorization of the orientation preserving general Hénon map to produce a differential equation whose time-one transfer map is the general Hénon map (with $b < 0$). We will also want the map to be symplectic or area contracting (but not area expanding). Therefore we are primarily interested in the cases $-1 \leq b < 0$. That is, we are mainly interested in maps that describe Hamiltonian or damped systems.

Suppose we give b the value $b = -.3$, and then study the properties of $\mathcal{M}_h$ for various values of a. Repeated iteration of $\mathcal{M}_h$ will again produce a set in the q, p plane. Figure 7.4 shows the projection of this set onto the q axis as a function of a. The graphic is similar to that of Figure 1.2.9: There is an infinite sequence of period doublings as a increases, and this cascade seems to be complete (and then some) by the time a reaches the value $a = 2.11$. However, there are also again additional cascades, this time in the vicinity of $a = 1.9$. In order to provide a complete picture of the attracting set, Figure 7.5 shows, in 3-dimensional perspective, both q_∞ and p_∞ as a function of a.

Figures 7.6 show various portions of the full set in the q, p plane for the parameter values $b = -.3$ and $a = 2.11$. That is, Figures 7.6 show the intersection of Figure 7.5 with the plane $a = 2.11$. In the first figure there again seem to be three curves that form part of an attractor. The remaining three figures show successive enlargements of the "boxed" portion of the previous figure. Again there appears to be a fractal structure, and this evidence suggests that there can also be a strange attractor for the case $b < 0$. However, the fractal structure is more complicated than that of the corresponding Figures 7.3 for the non-orientation preserving case. For the case of Figures 7.3, there is a similarity between every successive picture in the magnification sequence. In Figures 7.6 there is a similarity between every other picture in the magnification sequence.

Why is there a resemblance between the Feigenbaum diagram for the general Hénon map and that of the logistic map? Iterate once the Hénon map, as given by (7.3) and (7.4), to get the result

$$\bar{\bar{q}} = 1 + \bar{p} - a\bar{q}^2, \tag{29.7.6}$$

$$\bar{\bar{p}} = b\bar{q}. \tag{29.7.7}$$

Now imagine that b is small. Then, in view of (7.4), the relations (7.6) and (7.7) can be expanded in the form

$$\bar{\bar{q}} = 1 - a\bar{q}^2 + O(b), \tag{29.7.8}$$

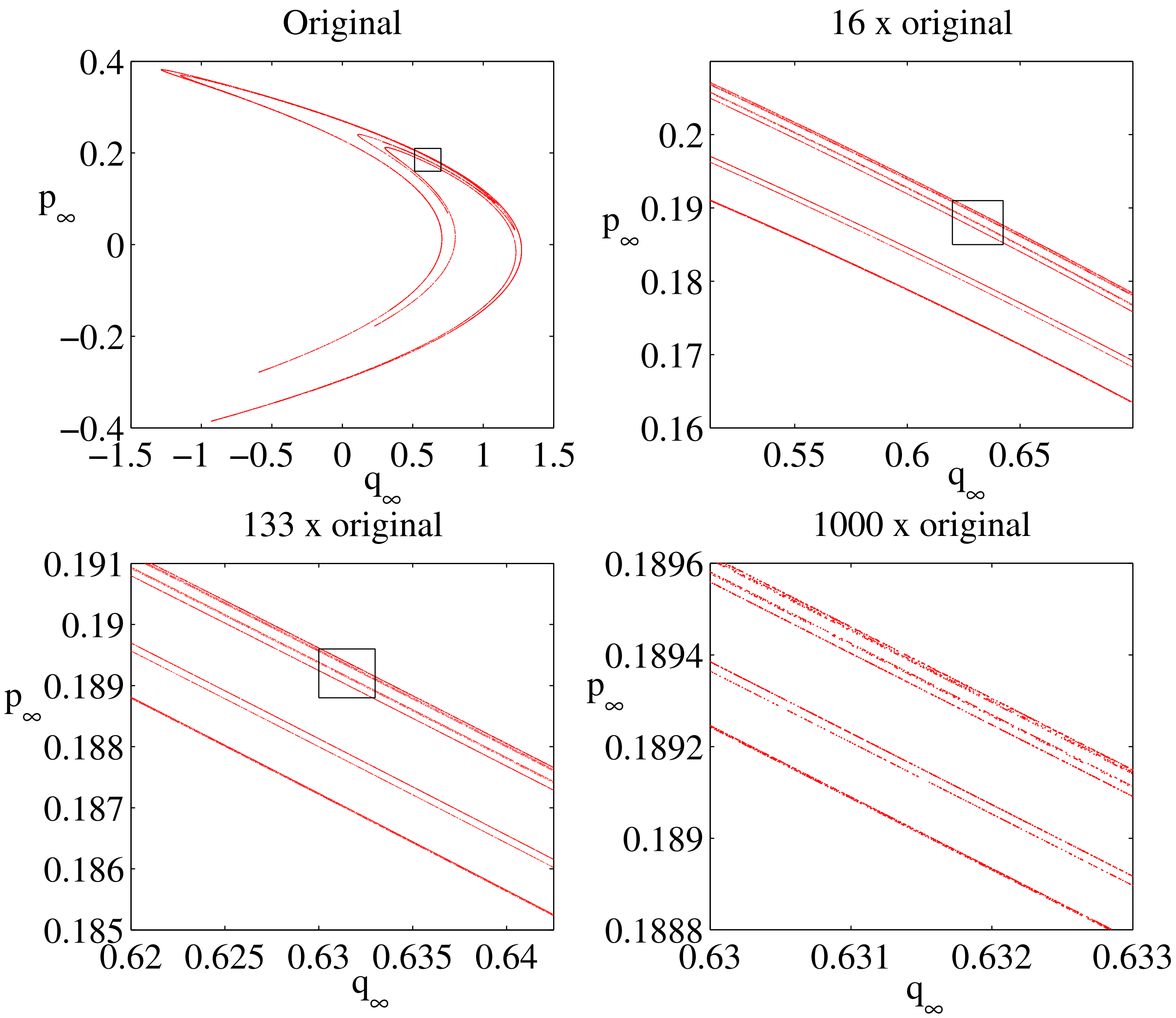


Figure 29.7.3: Successive enlargements of the attracting set q_∞, p_∞ for a non-orientation preserving case of the general Hénon map ($a = 1.4$, $b = +.3$). The attractor appears to be fractal, and therefore strange.

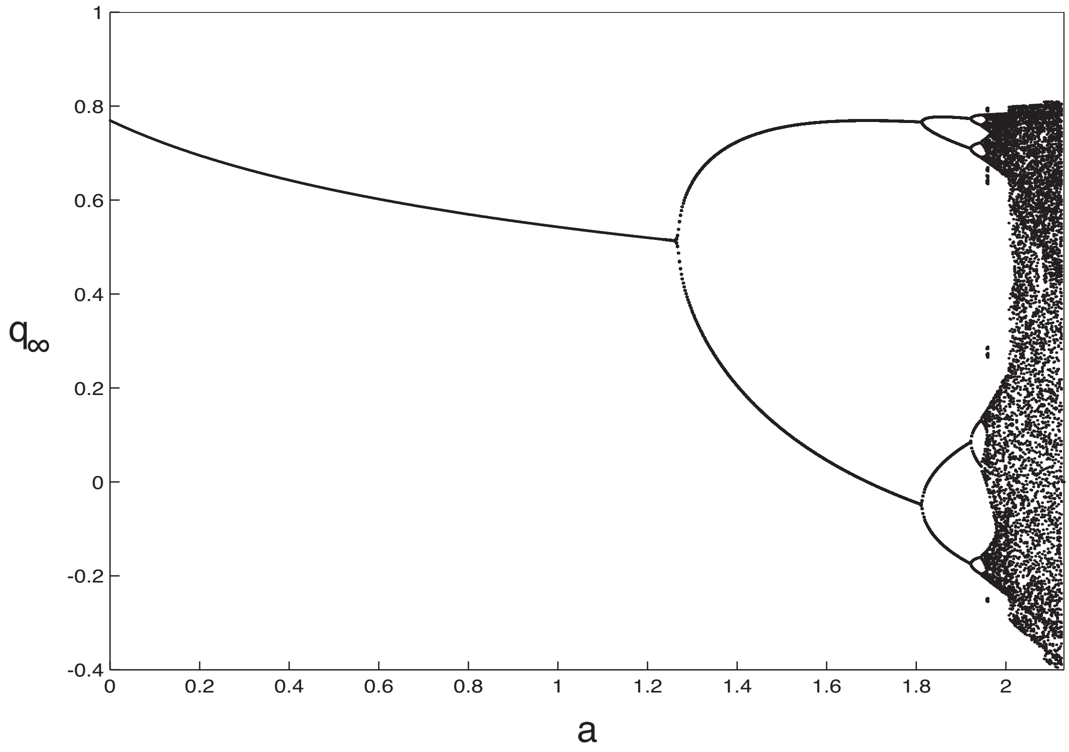


Figure 29.7.4: (Partial) Feigenbaum diagram showing limiting values q_∞ as a function of a (and b held at $b = -.3$) for an orientation preserving case of the general Hénon map.

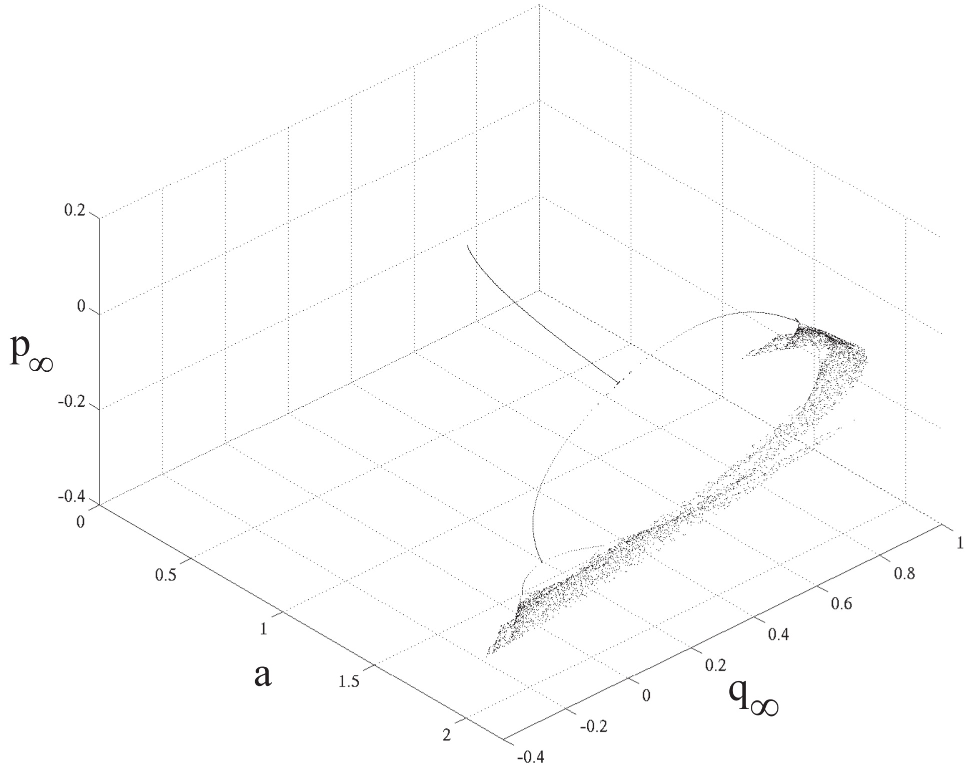


Figure 29.7.5: Full Feigenbaum diagram showing limiting values q_∞ and p_∞ as a function of a (and b held at $b = -.3$) for an orientation preserving case of the general Hénon map.

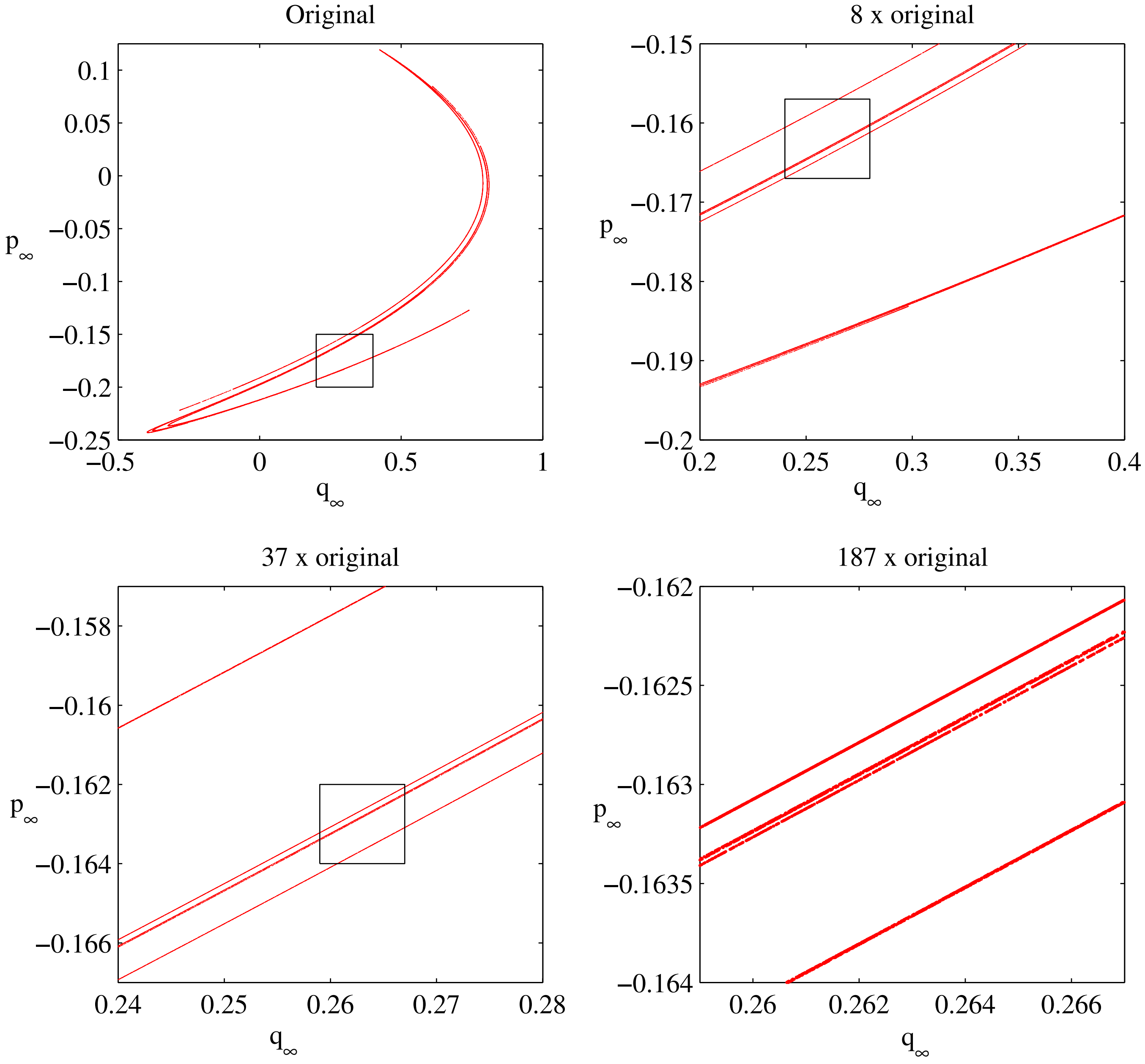


Figure 29.7.6: Successive enlargements of the attracting set q_∞, p_∞ for an orientation preserving case of the general Hénon map ($a = 2.11$, $b = -.3$). The attractor appears to be fractal, and therefore strange.

$$\bar{\bar{p}} = 0 + O(b). \tag{29.7.9}$$

Now make the change of variable

$$\bar{q} = -\bar{w}/a. \tag{29.7.10}$$

When this done (7.8) takes the form

$$\bar{\bar{w}} = -a + \bar{w}^2 + O(b) \tag{29.7.11}$$

which, up to $O(b)$ corrections, is the logistic map in the form (1.2.56) with $\mu = a$. Thus, for sufficiently small b, the Hénon map may be viewed as a *perturbation* of the logistic map. We therefore expect that the general Hénon map will exhibit all the richness of the logistic map and, as Figures 7.1, 7.2, 7.4, and 7.5 hint, even more.

At this point, it is still unclear what produces the strange (fractal) attractors illustrated in Figures 7.3 and 7.6, and where they are located relative to other significant sites in the mapping plane. We will study this question in subsequent sections. For simplicity, we will consider only the orientation preserving case $b < 0$. We will begin with the observation that the general Hénon map has two fixed points, and then factorize the map about each of these points.

Exercises

29.7.1. Here is another way to see that there is a relation between the logistic map and the Hénon map. For the general Hénon map in the form given by (7.3) and (7.4) make the change of variables

$$q = -Q/a \ , \ \bar{q} = -\bar{Q}/a; \tag{29.7.12}$$

$$p = -bP/a \ , \ \bar{p} = -b\bar{P}/a. \tag{29.7.13}$$

Show that in terms of these variables the general Hénon map takes the form

$$\bar{Q} = -a + bP + Q^2, \tag{29.7.14}$$

$$\bar{P} = Q, \tag{29.7.15}$$

which is evidently a perturbation of the logistic map. Indeed, in the limit $b = 0$ the relation (7.14) degenerates to the map

$$\bar{Q} = -a + Q^2, \tag{29.7.16}$$

29.7.2. Use the machinery of Section 18.1 to show that in the orientation preserving case the general Hénon map given by (7.3) and (7.4) has the Lie factorization

$$\mathcal{M}_h = \exp(\mathcal{G}_2)\mathcal{R}\exp[(a/3b^2) : p^3 :]\exp(- : p :) \tag{29.7.17}$$

where

$$\mathcal{G}_2 = \mathcal{G}^0 = (1/2)\log(-b)\Sigma, \tag{29.7.18}$$

$$\mathcal{R} = \exp[-(\pi/4) : q^2 + p^2 :]\exp(1/2)\log(-b) : qp :]. \tag{29.7.19}$$

29.8 Preliminary Study of General Hénon Map

In this section we begin a study of the general Hénon map. We start by locating its fixed points, and find there are two. Next we expand the map about each of these fixed points and analyze their natures. We learn that one of them, without loss of generality, can always be taken to be hyperbolic, or to at least have a diagonal linear part with positive eigenvalues. We then translate this hyperbolic fixed point to the origin, factor the map about it, and show that this factorized map depends on two parameters. Next we explore how the location (relative to the origin) of the second fixed point and its nature depend on these parameters. Finally, we translate the second fixed point to the origin and factor the map about it. In all our work we take particular care to exhibit whatever symmetries exist in the Hénon map.

29.8.1 Location, Expansion About, and Nature of Fixed Points

It is easily verified that the Hénon map $\mathcal{M}_h$ in the form (7.3) and (7.4) has *two* fixed points $q_f^{\pm}, p_f^{\pm}$ given by the relations

$$q_f^+ = \{-(1-b) + [(1-b)^2 + 4a]^{1/2}\}/(2a) = 2/\{(1-b) + [(1-b)^2 + 4a]^{1/2}\},$$

$$q_f^- = \{-(1-b) - [(1-b)^2 + 4a]^{1/2}\}/(2a), \tag{29.8.1}$$

$$p_f^{\pm} = bq_f^{\pm}, \tag{29.8.2}$$

and these fixed points are real when

$$a \geq -(1-b)^2/4. \tag{29.8.3}$$

In our future discussion we will refer to $\{q_f^-, p_f^-\}$ as the *first* fixed point and to $\{q_f^+, p_f^+\}$ as the *second* fixed point.

Figure 8.1 shows $q_f^{\pm}$ as a function of a for the case $b = -.3$, the same b value used in Figure 7.4. [The plot employs values of a for which (8.3) is satisfied.] The behavior of $p_f^{\pm}$ is similar as can be inferred from (8.2). Observe that $\{q_f^-, p_f^-\}$ are singular at $a = 0$. Inspection of (7.3) shows that $\mathcal{M}_h$ becomes *linear* at this value of a, and therefore this case is of less interest.

We will be primarily concerned with maps that are nearly symplectic, the case where $b \simeq -1$. Figure 8.2 shows $q_f^{\pm}$ as a function of a in the case $b = -.9$, an instance where the general Hénon map is more nearly symplectic. From (8.1) and (8.2) it is evident that $\{q_f^-, p_f^-\}$ and $\{q_f^+, p_f^+\}$ *coincide* when a takes on the minimum value allowed by (8.3),

$$a_{\min} = -(1-b)^2/4. \tag{29.8.4}$$

Calculation shows that for this value of a there are the relations

$$q_f^- = q_f^+ = 2/(1-b), \tag{29.8.5}$$

$$p_f^- = p_f^+ = 2b/(1-b). \tag{29.8.6}$$

Look at Figures 8.1 and 8.2, and see Exercise 8.1.

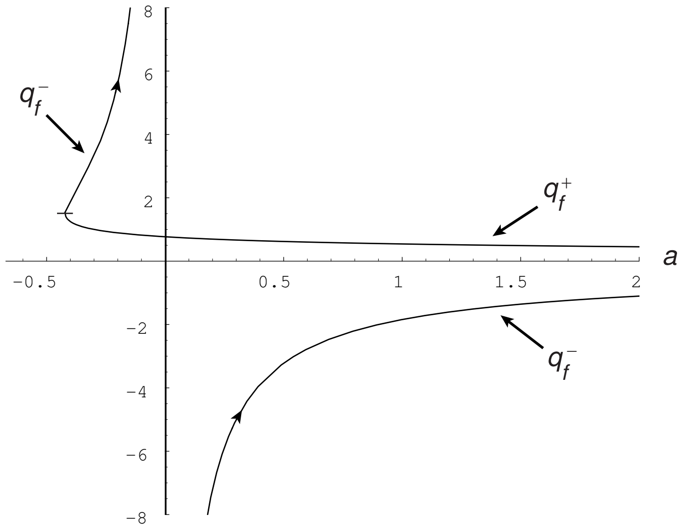


Figure 29.8.1: Values of $q_f^{\pm}$, when $b = -.3$, as a function of a. A horizontal tic mark indicates where q_f^+ and q_f^- meet when $a = a_{\min}$.

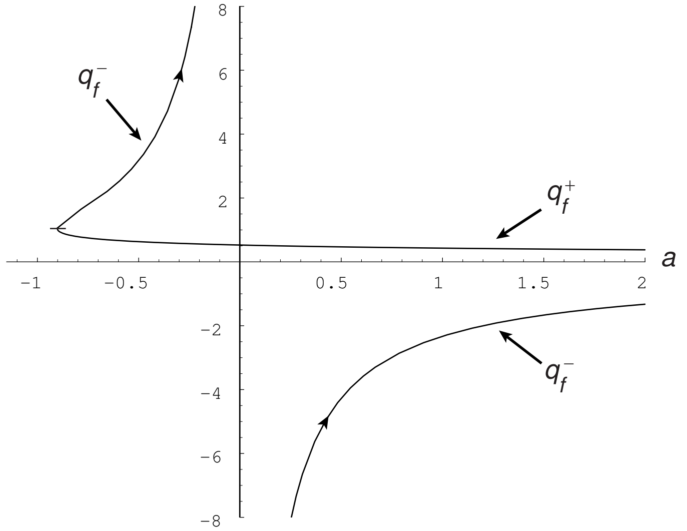


Figure 29.8.2: Values of $q_f^{\pm}$, when $b = -.9$, as a function of a. A horizontal tic mark indicates where q_f^+ and q_f^- meet when $a = a_{\min}$.

We observe that both fixed points *move* as the parameters a and b are varied, and consequently this form of the Hénon map is awkward for further analysis. To overcome this problem, introduce deviation (about a fixed point) variables Q, P by writing

$$q = q_f + Q \ , \ p = p_f + P; \tag{29.8.7}$$

$$\bar{q} = q_f + \bar{Q} \ , \ \bar{p} = p_f + \bar{P}. \tag{29.8.8}$$

So doing brings the general Hénon map $\mathcal{M}_h$ given by (7.3) and (7.4) to the transformed map $\mathcal{M}_h^*$ given by the relation

Action of $\mathcal{M}_h^*$:

$$\bar{Q} = -2aq_fQ + P - aQ^2,$$

$$\bar{P} = bQ, \tag{29.8.9}$$

and we see that the linear part R_h^* of the map about the fixed point (now the origin) is given by the relation

$$R_h^*(q_f, p_f) = \begin{pmatrix} -2aq_f & 1 \\ b & 0 \end{pmatrix}. \tag{29.8.10}$$

Let us find the eigenvalues of R_h^*. They satisfy the equation

$$P(\lambda) = \det(R_h^* - \lambda I) = (-2aq_f - \lambda)(-\lambda) - b = 0, \tag{29.8.11}$$

and are given by the relation

$$\lambda = -aq_f \pm \sqrt{(aq_f)^2 + b}. \tag{29.8.12}$$

The eigenvalues will be real if

$$(aq_f)^2 + b \geq 0, \tag{29.8.13}$$

and complex if

$$(aq_f)^2 + b < 0. \tag{29.8.14}$$

Note that, by (8.1), there is the relation

$$aq_f^{\pm} = \{-(1-b) \pm [(1-b)^2 + 4a]^{1/2}\}/2, \tag{29.8.15}$$

which can be employed in (8.12) through (8.14).

When the eigenvalues are complex, they will be complex conjugate as (8.12) shows. Therefore by (8.10), when the eigenvalues are complex, there will be the relation

$$\lambda\bar{\lambda} = \det R_h^* = -b, \tag{29.8.16}$$

and hence

$$|\lambda| = \sqrt{-b}. \tag{29.8.17}$$

That is, when a varies and b is fixed (and the eigenvalues are complex), the eigenvalues must move on a circle of radius $\sqrt{-b}$ in the complex plane.

Figure 8.3 shows the eigenvalues of R_h^* given by (8.12), for the case $q_f = q_f^-$ and $b = -.3$, as a function of a. Figure 8.4 shows the same thing when $b = -.9$, the more nearly symplectic instance. As expected, both eigenvalues are *real* for $q_f = q_f^-$ because (8.13) is then satisfied, and both are *positive*. Moreover, if $a > a_{\rm min}$, one of them satisfies $\lambda > 1$ and the other satisfies $0 < \lambda < 1$ so that $\{q_f^-, p_f^-\}$ is a hyperbolic fixed point. Finally, when $a = a_{\rm min}$, the eigenvalues take the values

$$\lambda = 1, -b. \tag{29.8.18}$$

See Exercise 8.3. That one of the eigenvalues then has the value +1 should not surprise us. For as a approaches $a_{\rm min}$, we have already seen that $\mathcal{M}_h$ has two fixed points that coincide, and the results of Section 18.4 then apply.

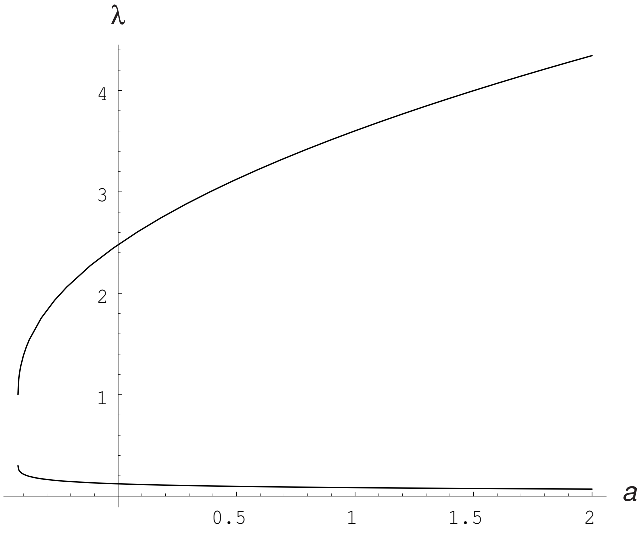


Figure 29.8.3: Eigenvalues λ of R_h^*, when $b = -.3$ and $q_f = q_f^-$, as a function of a.

The eigenvalues can become complex in the case $q_f = q_f^+$ because (8.14) can then be satisfied. As explained earlier, when this happens the eigenvalues are complex conjugate and their magnitude is given by (8.17). Figures 8.5 and 8.6 display the eigenvalues for the case $q_f = q_f^+$ as a function of a when b has the values $b = -.3$ and $b = -.9$, respectively. As the figures show, the eigenvalues are either positive, complex, or negative. When they are complex only their negative magnitude, shown as a dashed line, is plotted when $a > 0$, and only their positive magnitude, also shown as a dashed line, is plotted when $a < 0$. As a increases (and b is held fixed) the eigenvalues can leave the complex plane and become negative. This happens when

$$a = [(1 + \sqrt{-b})^4 - (1 - b)^2]/4. \tag{29.8.19}$$

See Exercise 8.4. Just as they become real and negative, according to (8.16), they both must have the value $-\sqrt{-b}$. As a increases further one of the eigenvalues can eventually take on the value -1. It can be shown that this occurs when

$$a = (3/4)(1 - b)^2, \tag{29.8.20}$$

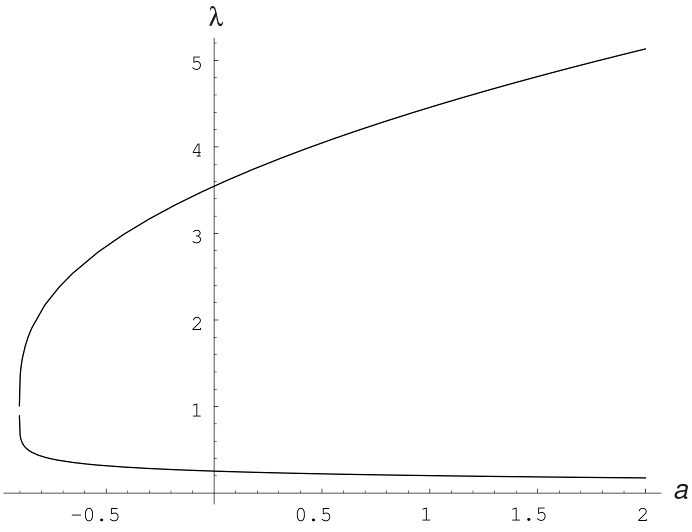


Figure 29.8.4: Eigenvalues λ of R_h^*, when $b = -.9$ and $q_f = q_f^-$, as a function of a.

again see Exercise 8.4, and in Section 18.9 we will learn that this is the condition for the first period doubling. Specifically, for $b = -.3$, we expect period doubling when $a = 1.2675$, which is consistent with Figure 7.4. For $a < 1.2675$ both eigenvalues satisfy $|\lambda| < 1$, and therefore the fixed point $\{q_f^+, p_f^+\}$ will be an attractor so that $q_\infty = q_f^+$. Comparison of Figures 7.4 and 8.5 verifies numerically that this is the case. We also note that for sufficiently small a the eigenvalues can leave the complex plane and become positive. This happens when

$$a = [(1 - \sqrt{-b})^4 - (1 - b)^2]/4. \tag{29.8.21}$$

Yet again see Exercise 8.4. Again according to (8.16), just as they become real and positive, they both must have the value $+\sqrt{-b}$. Finally, when $a = a_{\min}$, the fixed points coincide and therefore the eigenvalues for the case $q_f = q_f^+$ must also satisfy (8.18) at this value of a.

So far we have characterized the the general Hénon map by the parameters a and b. However, inspection of Figures 8.3 through 8.6 shows that (as a and b are varied over their allowed ranges and for a suitable choice of fixed point) the associated eigenvalue pair can take on *any positive* pair of values. Therefore, according to the work of Section 18.3, there is a linear (affine) change of variables that will always bring the general Hénon map to the form given by (3.95) and (3.96), and we may equally well use μ and ν as parameters in place of a and b. Finally, in order to conserve symbols, it is convenient to replace μ, ν by r, u. This replacement makes μ, ν available for a different use later on. With this replacement, the relations (3.95) and (3.96) become

Action of $\mathcal{M}_{ffq}^{tr}$:

$$\begin{aligned} \bar{q} &= r[q + (q-p)^2], \\ \bar{p} &= u[p + (q-p)^2]. \end{aligned} \tag{29.8.22}$$

For this map, consistent with our discussion, we will select the parameter ranges

$$r > 0, \quad u > 0. \tag{29.8.23}$$

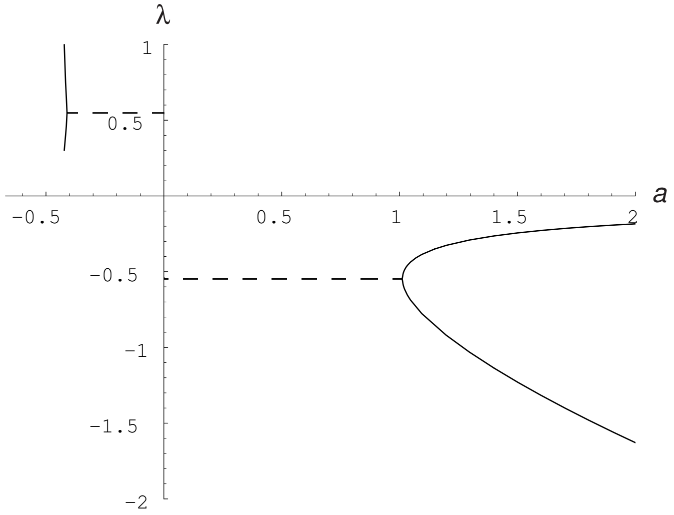


Figure 29.8.5: Eigenvalues λ of R_h^*, when $b = -.3$ and $q_f = q_f^+$, as a function of a. Note the small "line" of positive real eigenvalues for $a < 0$. Its endpoints coincide with the edges of the gap in the curve shown in Figure 8.3.

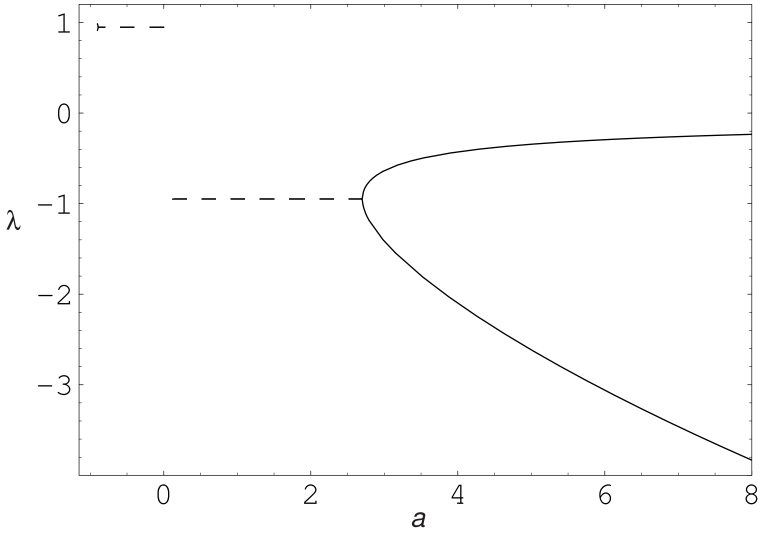


Figure 29.8.6: Eigenvalues λ of R_h^*, when $b = -.9$ and $q_f = q_f^+$, as a function of a. Note the barely visible line of positive real eigenvalues for $a < 0$. Its endpoints coincide with the edges of the tiny gap in the curve shown in Figure 8.4.

By construction $\mathcal{M}_{ffq}^{tr}$ sends the origin into itself. Let L be the Jacobian matrix (linear part) of $\mathcal{M}_{ffq}^{tr}$ about the origin. Evidently L is a diagonal matrix with diagonal entries r and u,

$$L = \begin{pmatrix} r & 0 \\ 0 & u \end{pmatrix}. \tag{29.8.24}$$

It follows that there are the relations

$$\operatorname{tr} L = r + u, \tag{29.8.25}$$

$$\det L = ru. \tag{29.8.26}$$

The map $\mathcal{M}_h^*$ given by (8.9) also sends the origin into itself, and according to (8.10) its linear part R_h^* about the origin satisfies the relations

$$\operatorname{tr} R_h^* = -2aq_f, \tag{29.8.27}$$

$$\det R_h^* = -b. \tag{29.8.28}$$

Since the maps $\mathcal{M}_h^*$ and $\mathcal{M}_{ffq}^{tr}$ are related by a linear change of variables, the matrices R_h^* and L must be related by a similarity transformation. Consequently there must be the equalities

$$\operatorname{tr} R_h^* = \operatorname{tr} L, \tag{29.8.29}$$

$$\det R_h^* = \det L. \tag{29.8.30}$$

It follows that there are the relations

$$r + u = -2aq_f, \tag{29.8.31}$$

$$ru = -b. \tag{29.8.32}$$

Note that (8.23), when employed in (8.31), implies that

$$-b > 0. \tag{29.8.33}$$

Consequently $\mathcal{M}_{ffq}^{tr}$ will be orientation preserving, and the origin will be an attracting, repelling, or hyperbolic fixed point depending on the values given to r and u.

Inspection of Figures 8.5 and 8.6, and reference to (8.18), show that, when r, u lie in the interval $(-b, 1)$, we should set $q_f = q_f^+$ so that (8.31) takes the form

$$r + u = -2aq_f^+ = (1-b) - [(1-b)^2 + 4a]^{1/2}. \tag{29.8.34}$$

Here we have used (8.15). Solving (8.34) for a gives the result

$$a = [(r+u)^2 - 2(r+u)(1+ru)]/4. \tag{29.8.35}$$

For other values of r, u we should set $q_f = q_f^-$. See Figures 8.3 and 8.4. Then (8.31) takes the form

$$r + u = -2aq_f^- = (1-b) + [(1-b)^2 + 4a]^{1/2}. \tag{29.8.36}$$

Solving (8.36) for a again gives the result (8.35). Therefore the only thing we have to remember is that $\mathcal{M}^{tr}_{ffq}$ as given by (8.22) is a form of the general Hénon map $\mathcal{M}^*_h$ about the *second* fixed point $\{q^+_f, p^+_f\}$ when r, u lie *within* the interval $(-b, 1)$, and (8.22) is a form of the general Hénon map $\mathcal{M}^*_h$ about the *first* fixed point $\{q^-_f, p^-_f\}$ when r, u lie *outside* the interval $(-b, 1)$. For most of our future discussion we will be interested in cases where b is very near -1 so that the length of the interval $(-b, 1)$ is very small. Consequently, r, u will generally lie outside the interval $(-b, 1)$. Correspondingly the origin will generally be a hyperbolic fixed point of $\mathcal{M}^{tr}_{ffq}$. At any rate, we will henceforth refer to the origin as the *first* fixed point of $\mathcal{M}^{tr}_{ffq}$, and its other fixed point (still to be found) as its *second* fixed point.

29.8.2 Lie Factorization About the First (Hyperbolic) Fixed Point

To study $\mathcal{M}^{tr}_{ffq}$ it is instructive to factorize it and then make yet more linear changes of variables. Define linear *scaling* and *damping* maps $\mathcal{S}$ and $\mathcal{D}$ by the relations

$$\mathcal{S} = \exp : \left(\log \sqrt{u/r}\right) qp :, \tag{29.8.37}$$

$$\mathcal{D} = \exp\left[\left(\log \sqrt{ru}\right) \Sigma\right]. \tag{29.8.38}$$

Note that $\mathcal{S}$ and $\mathcal{D}$ commute because of (21.3.9). It is easily verified that their actions on the variables q, p are given by the relations

$$\mathcal{S}q = (\sqrt{r/u})q, \tag{29.8.39}$$

$$\mathcal{S}p = (\sqrt{u/r})p, \tag{29.8.40}$$

$$\mathcal{D}q = (\sqrt{ru})q, \tag{29.8.41}$$

$$\mathcal{D}p = (\sqrt{ru})p. \tag{29.8.42}$$

Consequently, there are the results

$$\mathcal{S}\mathcal{D}q = rq, \tag{29.8.43}$$

$$\mathcal{S}\mathcal{D}p = up. \tag{29.8.44}$$

Next let $\mathcal{N}$ be the nonlinear map

$$\mathcal{N} = \exp : (q-p)^3/3 : . \tag{29.8.45}$$

Simple computation shows that it has the properties

$$\mathcal{N}q = q + (q-p)^2, \tag{29.8.46}$$

$$\mathcal{N}p = p + (q-p)^2. \tag{29.8.47}$$

It follows that the map $\mathcal{M}^{tr}_{ffq}$ given by (8.22) can be written in the factorized form

$$\mathcal{M}^{tr}_{ffq} = [\exp : (q-p)^3/3 :]\mathcal{S}\mathcal{D}. \tag{29.8.48}$$

Finally, let $\mathcal{O}(\theta)$ be the linear transformation (rotation)

$$\mathcal{O}(\theta) = \exp : (-\theta/2)(p^2 + q^2) : . \tag{29.8.49}$$

It is easily verified that $\mathcal{O}(5\pi/4)$ has the properties

$$\mathcal{O}(5\pi/4)q = -(q + p)/\sqrt{2}, \tag{29.8.50}$$

$$\mathcal{O}(5\pi/4)p = (q - p)/\sqrt{2}, \tag{29.8.51}$$

$$\mathcal{O}(5\pi/4)(q - p) = -\sqrt{2}q, \tag{29.8.52}$$

$$\mathcal{O}(5\pi/4)(qp) = (p^2 - q^2)/2. \tag{29.8.53}$$

See (1.2.37) and (1.2.38) and Exercise 5.4.5. Note also that $\mathcal{O}(\theta)$ and $\mathcal{D}$ commute.

Now use $\mathcal{S}$, $\mathcal{D}$, and $\mathcal{O}$ to make linear changes of variables, thereby bringing $\mathcal{M}^{tr}_{ffq}$ to a new form which we will call $\mathcal{M}_{-}$, by writing

$$\mathcal{M}_{-} = \mathcal{O}(5\pi/4)\mathcal{D}^{1/2}\mathcal{S}^{1/2}\mathcal{M}^{tr}_{ffq}\mathcal{S}^{-1/2}\mathcal{D}^{-1/2}\mathcal{O}^{-1}(5\pi/4). \tag{29.8.54}$$

Here $\mathcal{S}^{\pm 1/2}$ and $\mathcal{D}^{\pm 1/2}$ are defined by

$$\mathcal{S}^{\pm 1/2} = \exp : (\pm 1/2)(\log \sqrt{u/r})qp :, \tag{29.8.55}$$

$$\mathcal{D}^{\pm 1/2} = \exp[(\pm 1/2)(\log \sqrt{ru})\Sigma]. \tag{29.8.56}$$

From (8.48) and (8.50) through (8.56) we find that $\mathcal{M}_{-}$ has the pleasing factorization

$$\mathcal{M}_{-} = \mathcal{D}^{1/2}\mathcal{H}\exp[(-\sqrt{8}/3) : q^3 :]\mathcal{H}\mathcal{D}^{1/2} \tag{29.8.57}$$

where $\mathcal{H}$ is the linear *hyperbolic* map

$$\mathcal{H} = \exp : (1/4)(\log \sqrt{u/r})(p^2 - q^2) : . \tag{29.8.58}$$

(See Exercise 8.8.) According to our previous discussion, for the most part the origin will be a hyperbolic fixed point of $\mathcal{M}^{tr}_{ffq}$. Correspondingly, since the transformations $\mathcal{O}$, $\mathcal{D}$, and $\mathcal{S}$ involved in (8.54) are linear, the origin will also a hyperbolic fixed point of $\mathcal{M}_{-}$. According to item 1b of Section 18.5, hyperbolic fixed points have index -1. That is why we have used the symbols $\mathcal{M}_{-}$ to denote the map given by (8.54) and (8.57)

To proceed further, we need explicit formulas for the action of $\mathcal{M}_{-}$. Since we want to treat r and u on a similar footing, it is convenient to employ a variable ν defined by writing

$$r = \sqrt{ru}\exp(\nu), \tag{29.8.59}$$

$$u = \sqrt{ru}\exp(-\nu). \tag{29.8.60}$$

Then there are the relations

$$u/r = \exp(-2\nu),\ \sqrt{u/r} = \exp(-\nu), \tag{29.8.61}$$

and $\mathcal{H}$ takes the form

$$\mathcal{H} = \exp : (-\nu/4)(p^2 - q^2) : . \tag{29.8.62}$$

The action of $\mathcal{H}$ may be found from Exercise 5.4.6,

$$\mathcal{H}q = cq + sp\ ,\ \mathcal{H}p = sq + cp, \tag{29.8.63}$$

with

$$c = \cosh(\nu/2)\ ,\ s = \sinh(\nu/2); \tag{29.8.64}$$

and the action of $\mathcal{D}^{1/2}$ follows from (8.41) and (8.42),

$$\mathcal{D}^{1/2}q = (ru)^{1/4}q, \tag{29.8.65}$$

$$\mathcal{D}^{1/2}p = (ru)^{1/4}p. \tag{29.8.66}$$

From these relations we deduce that $\mathcal{M}_-$ takes the form

Action of $\mathcal{M}_-$:

$$\bar{q} = \mathcal{M}_-q = (ru)^{1/2}[q(c^2+s^2) + p(2cs) - s\sqrt{8}(ru)^{1/4}(cq+sp)^2], \tag{29.8.67}$$

$$\bar{p} = \mathcal{M}_-p = (ru)^{1/2}[q(2cs) + p(c^2+s^2) - c\sqrt{8}(ru)^{1/4}(cq+sp)^2]. \tag{29.8.68}$$

To make contact with the parameters a, b of the previous subsection, (8.35) can be solved for a in terms of b and ν using (8.32), (8.59), and (8.60) to give the result

$$a = -(1-b)(\sqrt{-b})\cosh\nu - b\cosh^2\nu. \tag{29.8.69}$$

Figure 8.7 displays a as a function of ν for various values of b. It can be shown that a as given by (8.69) always satisfies the inequality (8.3) provided ν is real. See Exercise 8.10.

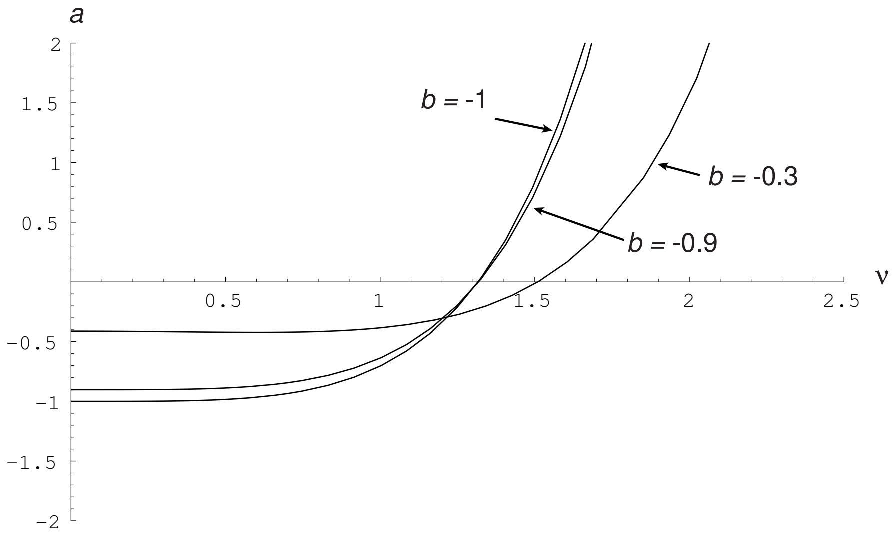


Figure 29.8.7: The parameter a as a function of ν for various values of b.

29.8.3 Location and Nature of Second Fixed Point

We know from the previous discussion, and as is evident from (8.67) and (8.68), that the origin is a fixed point of $\mathcal{M}_-$. What can be said about the *second* fixed point of $\mathcal{M}_-$ as given in the form (8.67) and (8.68)? Its position will depend on the parameters r, u; consequently it will be called the *mobile* fixed point, while the origin will be referred to as the *stationary* fixed point. Somewhat involved calculation shows that the fixed-point equations

$$\bar{q} = q,\ \bar{p} = p \tag{29.8.70}$$

have the second solution

$$q_f = \left(\frac{1}{\sqrt{2}}\right)\left[\frac{\sqrt{u}}{1-u} - \frac{\sqrt{r}}{1-r}\right]\left[\frac{(1-r)(1-u)}{(r-u)}\right]^2, \tag{29.8.71}$$

$$p_f = -\left(\frac{1}{\sqrt{2}}\right)\left[\frac{\sqrt{u}}{1-u} + \frac{\sqrt{r}}{1-r}\right]\left[\frac{(1-r)(1-u)}{r-u}\right]^2. \tag{29.8.72}$$

See Exercise 8.11. We observe that q_f is odd under the interchange of r and u, and p_f is even. According to (8.59) and (8.60), interchanging r and u is equivalent to replacing ν by $-\nu$. Therefore, q_f is an odd function of ν, and p_f is even in ν.

Note also that there is the algebraic identity

$$\frac{\sqrt{u}}{1-u} + \frac{\sqrt{r}}{1-r} = \frac{(\sqrt{u}+\sqrt{r})(1-\sqrt{ru})}{(1-u)(1-r)}. \tag{29.8.73}$$

In the symplectic case there is the relation $ru = 1$, from which it follows, using (8.72) and (8.73), that

$$p_f = 0,\ \text{symplectic case.} \tag{29.8.74}$$

[That is why we chose to bring the general Hénon map to the form given by (8.54) and (8.57).] Moreover, the relation (8.74) will be nearly satisfied if $\mathcal{M}_-$ is nearly symplectic. The expression for q_f also simplifies in the symplectic case. Insertion of (8.74) into (8.67) gives for q_f the result

$$q_f = (1/\sqrt{2})(s/c^2),\ \text{symplectic case.} \tag{29.8.75}$$

Figure 8.8 displays q_f as a function of ν in the symplectic case. Note that in the symplectic case q_f is bounded by the extrema $\pm 1/\sqrt{8}$. We also observe that when $\nu = 0$, the stationary fixed point (the one at the origin) and the mobile fixed point coincide.

To study the location of the mobile fixed point in general, and for future use, it is convenient to introduce a quantity τ defined by the relation

$$\tau = (1-r)(1-u)/(r-u). \tag{29.8.76}$$

Note that τ occurs as a factor in (8.71) and (8.72). From (8.59) through (8.61) we see that it depends on ν and ru in the fashion

$$\tau = [(ru)^{1/2} + (ru)^{-1/2} - 2\cosh\nu]/[2\sinh\nu]. \tag{29.8.77}$$

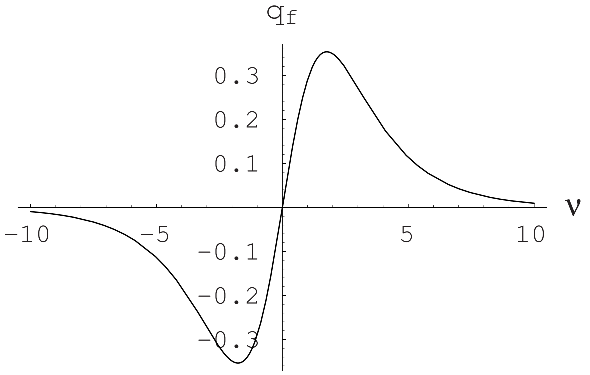


Figure 29.8.8: The quantity q_f for the mobile fixed point as a function of ν in the symplectic case $b = -1$.

It is easily checked that the function f defined by the relation

$$f(x) = x + 1/x \tag{29.8.78}$$

satisfies the inequality

$$f \geq 2 \text{ for } x > 0 \tag{29.8.79}$$

and

$$f = 2 \text{ only when } x = 1. \tag{29.8.80}$$

It follows that τ is singular at $\nu = 0$ except for the symplectic case $ru = 1$, in which case it is regular at $\nu = 0$ and in fact vanishes there. See Figure 8.9.

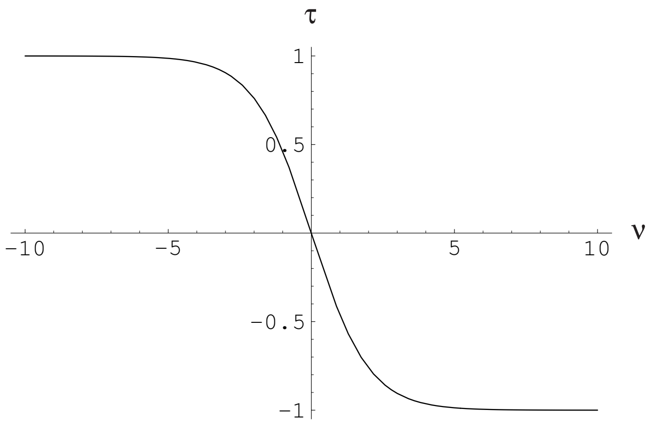


Figure 29.8.9: The quantity τ as a function of ν in the symplectic case $b = -1$.

Let us now look at a numerical example of a nonsymplectic (but orientation preserving) case. Suppose $ru = -b = .3$. Now τ is singular as a function of ν as illustrated in Figure

8.10. Correspondingly, as shown in Figures 8.11 and 8.12, q_f and p_f can also move to infinity. (See Exercise 8.9.) Figure 8.13 shows q_f and p_f simultaneously, and illustrates that now the relation $p_f = 0$ is no longer exactly maintained. Instead the mobile fixed point traces out loops. We also observe that when ν has values such that $\tau = 0$ (see Figure 8.10), then $q_f = p_f = 0$ and the stationary fixed point and the mobile fixed point again coincide. Look at Figures 8.11 through 8.13. As is consistent with the results of Sections 18.4 and 18.5, this happens when $r = 1$ or $u = 1$. See (8.71), (8.72), (8.76), and Exercise 8.10.

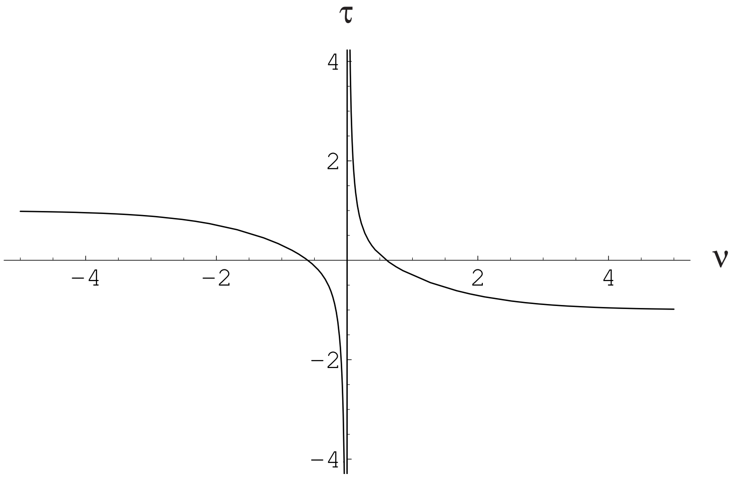


Figure 29.8.10: The quantity τ as a function of ν in the nonsymplectic case $b = -.3$.

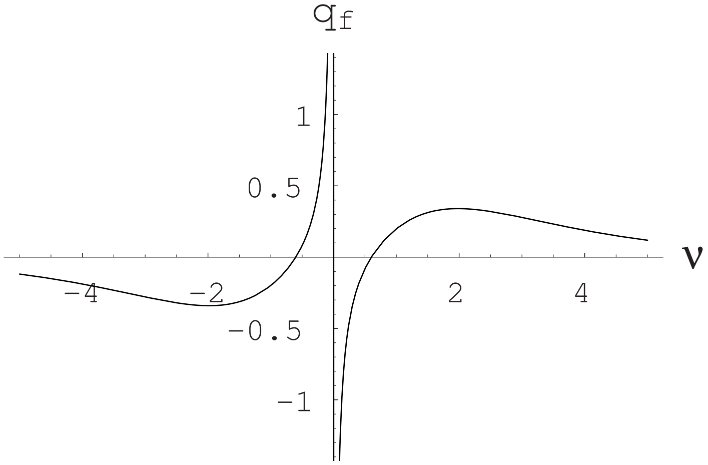


Figure 29.8.11: The quantity q_f for the mobile fixed point as a function of ν in the nonsymplectic case $b = -.3$.

We close this subsection by analyzing the nature of the second (mobile) fixed point. Rather than working with the map $\mathcal{M}_-$ [as given by (8.54), (8.57), (8.67), and (8.68)], it is algebraically simpler to work with the related map $\tilde{\mathcal{M}}$ given by

$$\begin{aligned}\tilde{\mathcal{M}} &= \mathcal{O}^{-1}(5\pi/4)\mathcal{M}_{-}\mathcal{O}(5\pi/4) = \mathcal{D}^{1/2}\mathcal{S}^{1/2}\mathcal{M}^{tr}_{ffq}\mathcal{S}^{-1/2}\mathcal{D}^{-1/2} \\ &= \mathcal{D}^{1/2}\mathcal{S}^{1/2}[\exp:(q-p)^3/3:]\mathcal{S}^{1/2}\mathcal{D}^{1/2}. \qquad (29.8.81)\end{aligned}$$

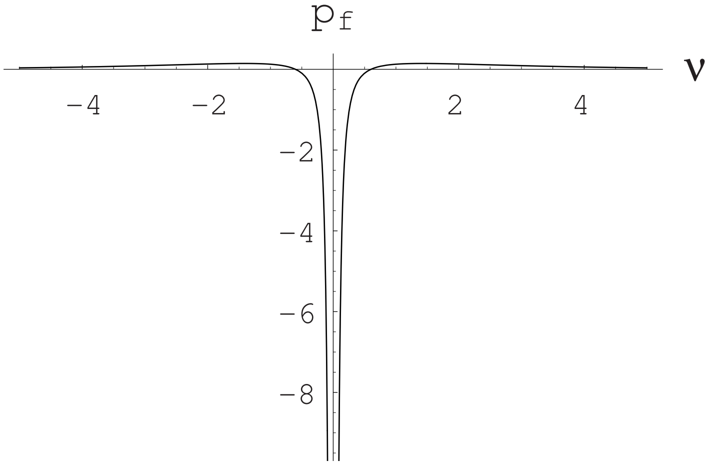


Figure 29.8.12: The quantity p_f for the mobile fixed point as a function of ν in the nonsymplectic case $b = -.3$.

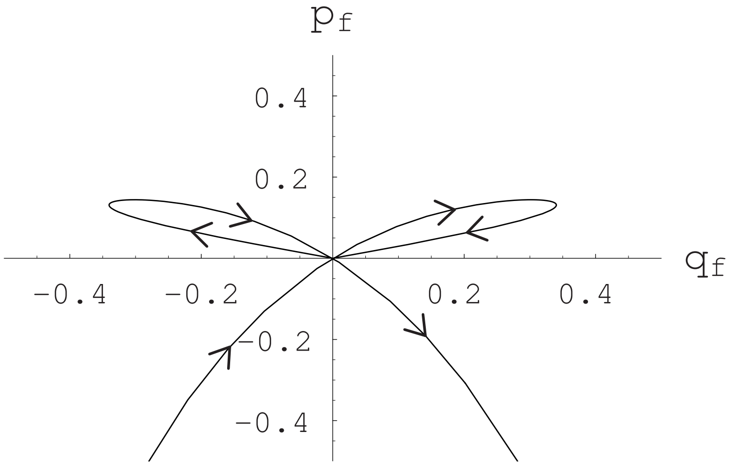


Figure 29.8.13: Location of the mobile fixed point for the nonsymplectic case $b = -.3$ and ν varying over the range $[-20, 20]$.

From (8.39) through (8.47) we find that this map takes the form

Action of $\tilde{\mathcal{M}}$:

$$\bar{q} = \tilde{\mathcal{M}}q = rq + \sqrt{r}(\sqrt{r}q - \sqrt{u}p)^2,$$

$$\bar{p} = \tilde{\mathcal{M}}p = up + \sqrt{u}(\sqrt{r}q - \sqrt{u}p)^2, \qquad (29.8.82)$$

and has the second fixed point

$$\tilde{q}_f = \left(\frac{\sqrt{r}}{1-r}\right)\left[\frac{(1-r)(1-u)}{(r-u)}\right]^2, \tag{29.8.83}$$

$$\tilde{p}_f = \left(\frac{\sqrt{u}}{1-u}\right)\left[\frac{(1-r)(1-u)}{(r-u)}\right]^2. \tag{29.8.84}$$

Since $\mathcal{O}(5\pi/4)$ is simply the r and u independent linear transformation described by the rotation matrix

$$O(5\pi/4) = \begin{pmatrix} -1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}, \tag{29.8.85}$$

the Jacobian matrices (linear parts) of $\mathcal{M}_-$ and $\tilde{\mathcal{M}}$ will be similar (with the matrix O providing the requisite similarity transformation) and therefore will have the *same* spectrum.

At the fixed point $\tilde{q}_f, \tilde{p}_f$ we find the Jacobian matrix of $\tilde{\mathcal{M}}$ as given by (8.82) takes the form

$$\tilde{M}_f = \tilde{M}(\tilde{q}_f, \tilde{p}_f) = L + 2[(1-r)(1-u)/(r-u)]\begin{pmatrix} r & -\sqrt{ru} \\ \sqrt{ru} & -u \end{pmatrix}, \tag{29.8.86}$$

where L denotes the diagonal matrix (8.24). [Note that the factor τ, as given by (8.76), also appears in (8.86).] The nature of this fixed point can be determined by finding the spectrum of $\tilde{M}_f$. We know from (3.42) and (8.28), or find by direct calculation, that

$$\det \tilde{M}_f = ru. \tag{29.8.87}$$

Also, direct calculation gives the result

$$2\sigma = \text{ tr } \tilde{M}_f = 2(1+ru) - (r+u). \tag{29.8.88}$$

Suppose λ_1, λ_2 are the eigenvalues of $\tilde{M}_f$. Then from standard matrix relations we again have the results

$$\lambda_1\lambda_2 = ru \ , \ \lambda_1 + \lambda_2 = 2\sigma. \tag{29.8.89}$$

Define quantities μ_1, μ_2 by the rules

$$\mu_1 = \lambda_1/\sqrt{ru} \ , \ \mu_2 = \lambda_2/\sqrt{ru}. \tag{29.8.90}$$

They evidently satisfy the relations

$$\mu_1\mu_2 = 1 \ , \ \mu_1 + \mu_2 = 2\sigma/\sqrt{ru}. \tag{29.8.91}$$

Equations (8.91) have the immediate solution

$$\mu_\pm = \sigma/\sqrt{ru} \pm \sqrt{\sigma^2/(ru) - 1}. \tag{29.8.92}$$

We see that the μ behave like the eigenvalues of a 2 × 2 symplectic matrix as in Figure 3.4.1. Correspondingly, the λ are given by the relations

$$\lambda_\pm = (\sqrt{ru})\mu_\pm = \sigma \pm \sqrt{\sigma^2 - ru}. \tag{29.8.93}$$

They behave in a similar fashion, but are governed by a circle of radius $\sqrt{ru} = \sqrt{-b}$. According to (8.93), they can potentially come off or on this circle when

$$\sigma^2 = ru, \tag{29.8.94}$$

from which condition we deduce, using (8.88), the relation

$$4r^2u^2 - 4ru(r+u) + r^2 + 6ru + u^2 - 4(r+u) + 4 = 0. \tag{29.8.95}$$

Finally, it is easily verified that -1 becomes an eigenvalue of $\tilde{M}_f$ when r and u satisfy the relation

$$3(1+ru) = (r+u). \tag{29.8.96}$$

See Exercise 8.15.

In terms of the variables ν and ru, the quantity σ is given by the relation

$$\sigma = (ru)^{1/2}[(ru)^{1/2} + (ru)^{-1/2} - \cosh\nu]. \tag{29.8.97}$$

It is also useful to examine the matrix $\tilde{M}_f'$ defined by

$$\tilde{M}_f' = (ru)^{-1/2}\tilde{M}_f. \tag{29.8.98}$$

It is symplectic, and indeed is the symplectic factor in the symplectic polar decomposition of $\tilde{M}_f$. In the case that its eigenvalues lie on the unit circle, it can be brought to the normal form (3.5.61) by a symplectic similarity transformation. See (3.5.53). Moreover, since the trace is preserved by a similarity transformation, we find from (3.5.61), (8.88), (8.97), and (8.98) the result that its phase advance ϕ satisfies the relation

$$\cos\phi = [(ru)^{1/2} + (ru)^{-1/2} - \cosh\nu]. \tag{29.8.99}$$

Now let us look at two numerical examples. In the first example $\tilde{M}_f$ will be symplectic, and thus we set $ru = -b = 1$. Figures 8.14 and 8.15 show $\sigma/\sqrt{ru}$ and the spectrum of $\tilde{M}_f$ as ν is varied. Finally, Figure 8.16 shows the phase advance ϕ for $\tilde{M}_f$ as a function of ν. Note that (8.99) only gives the cosine of the phase advance, and therefore only determines the phase advance up to a sign. The sign of ϕ was determined by bringing $\tilde{M}_f$ to normal form numerically and then examining the 1,2 entry of the normal form. See (3.5.41).

As a second numerical example we consider the nonsymplectic (but orientation preserving) case $ru = -b = .3$. Figures 8.17 and 8.18 show $\sigma/\sqrt{ru}$ and the spectrum of $\tilde{M}_f$ as ν is varied. Figure 8.19 shows the phase advance ϕ for $\tilde{M}_f'$ as a function of ν. The eigenvalues leave the unit circle for $-.839 \leq \nu \leq .839$ and $|\nu| > .886$, and therefore the phase advance is undefined in these ranges. See Exercises 8.18 and 8.19. As is evident from (8.97) and Figure 8.17, σ takes on its maximum value for $\nu = 0$. Consequently, the maximum value of λ_+ as given by (8.93) depends only on the value of the product ru. Figure (8.20) displays this maximum value, $\lambda_+^{\max}$, as a function of ru over the range $0 \leq ru \leq 1$. As the graphic suggests, and computation confirms, $\lambda_+^{\max}$ decreases monotonically from $\lambda_+^{\max} = 2$ for $ru = 0$ to $\lambda_+^{\max} = 1$ for $ru = 1$.

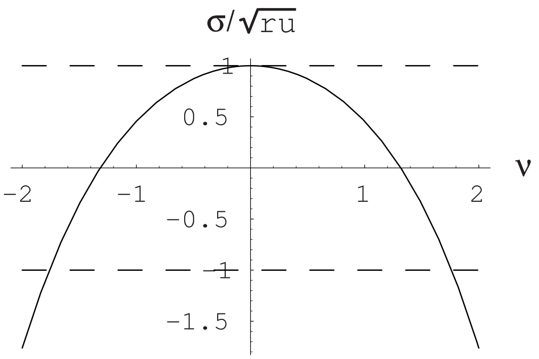


Figure 29.8.14: Value of $\sigma/\sqrt{ru}$ as a function ν in the symplectic case $b = -1$.

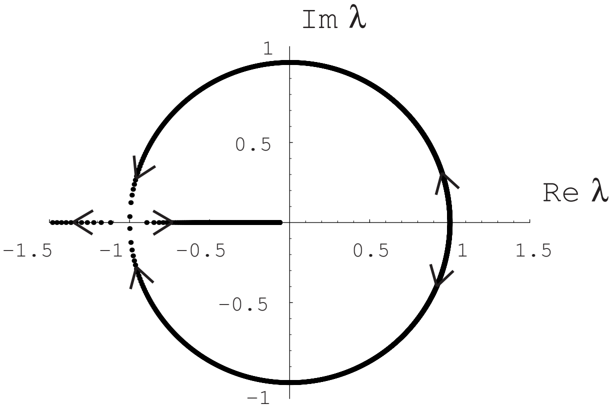


Figure 29.8.15: Spectrum of $\tilde{M}_f$ (and of the linear part of $\mathcal{M}_-$ about its second fixed point) for ν varying over the range $[0, 3]$ in the symplectic case $b = -1$. An identical picture is produced for ν varying in the range $[-3, 0]$ as is evident from (8.97) and Figure 8.14. The eigenvalues leave the unit circle at $\nu = \pm 1.763$ as is also evident from (8.97) and Figure 8.14. See Exercise 8.18.

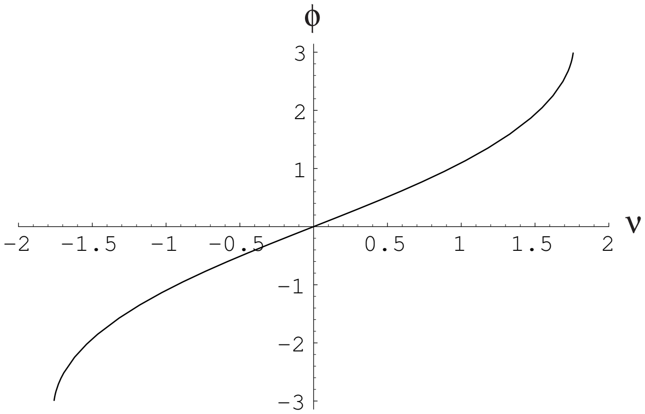


Figure 29.8.16: Phase advance of $\tilde{M}_f$ (and of the linear part of $\mathcal{M}_-$ about its second fixed point) as a function of ν in the symplectic case $b = -1$. Only the range $\nu \in [-1.763, 1.763]$ is shown because the eigenvalues leave the unit circle outside this range.

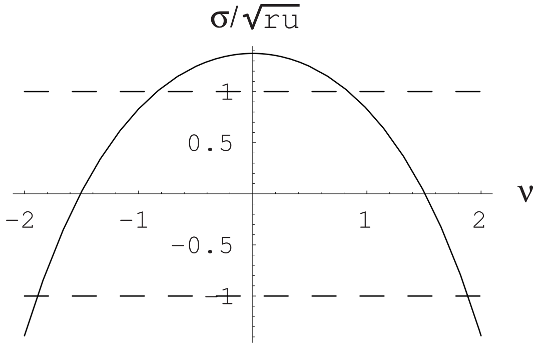


Figure 29.8.17: Value of $\sigma/\sqrt{ru}$ as a function of ν in the nonsymplectic case $b = -.3$.

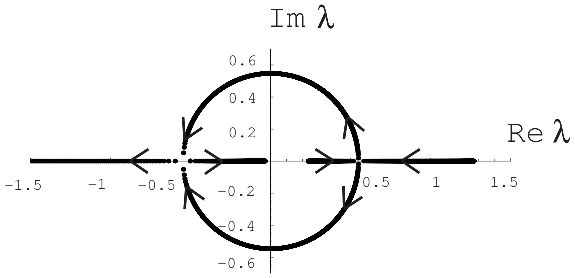


Figure 29.8.18: Spectrum of $\tilde{M}_f$ (and of the linear part of $\mathcal{M}_-$ about its second fixed point) for ν varying over the range $[0, 3]$ in the nonsymplectic case $b = -.3$.

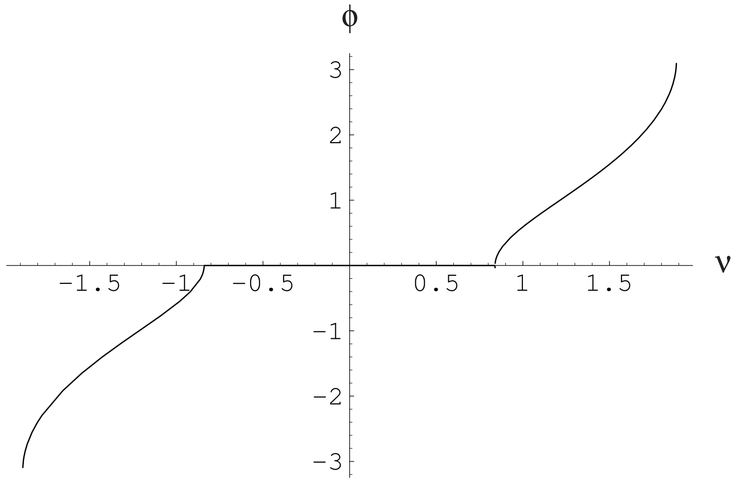


Figure 29.8.19: Phase advance of $\tilde{M}_f'$ as a function of ν in the nonsymplectic case $b = -.3$. Only the ranges $\nu \in [-1.886, -.839]$ and $\nu \in [.839, 1.886]$ are shown because the eigenvalues of $\tilde{M}_f'$ leave the unit circle for ν outside these ranges. See Figure 8.17 and Exercise 8.15.

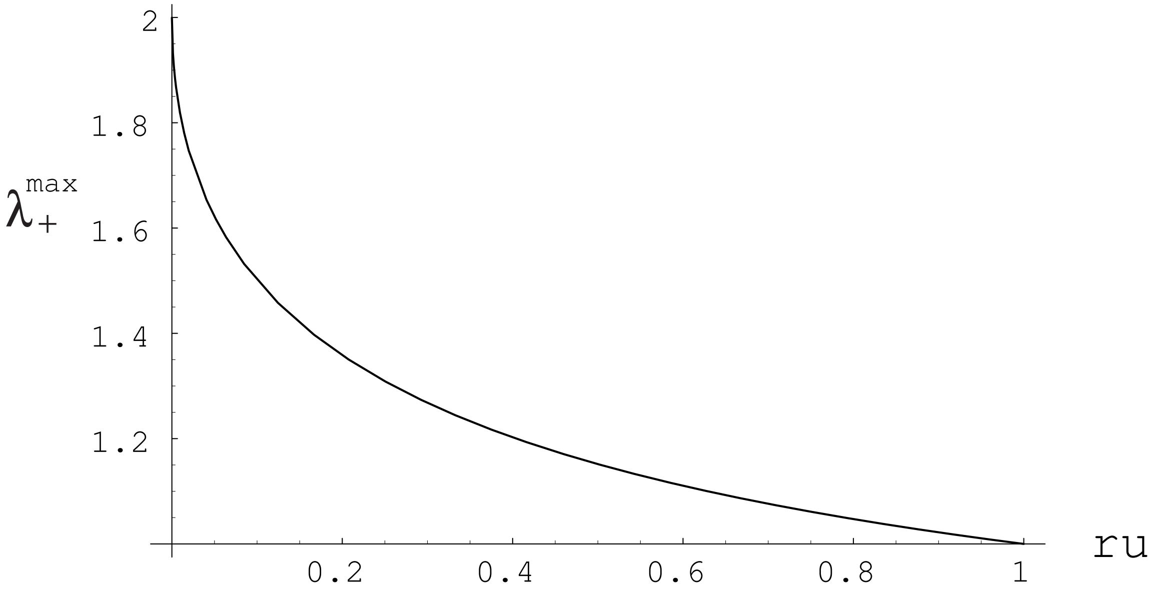


Figure 29.8.20: Maximum value of λ_+ as a function of ru.

29.8.4 Expansion and Lie Factorization About Second Fixed Point

To make contact with earlier work, and for future use, it is also useful to have an expansion and factorization of the Hénon map $\mathcal{M}_-$ about its second fixed point (in which case the second fixed point is placed at the origin and the first fixed point, which was the origin, becomes mobile). We will first expand and factorize $\tilde{\mathcal{M}}$, see (8.81) and (8.82), about its second fixed point $\tilde{q}_f, \tilde{p}_f$; and we will then use the transformation $\mathcal{O}(5\pi/4)$ to obtain an expansion and factorization of $\mathcal{M}_-$ about its second fixed point q_f, p_f.

To expand $\tilde{\mathcal{M}}$ as given by (8.82) about the fixed point $\tilde{q}_f, \tilde{p}_f$, make the change of variables

$$q = Q + \tilde{q}_f,\ p = P + \tilde{p}_f,\ \overline{q} = \overline{Q} + \tilde{q}_f,\ \overline{p} = \overline{P} + \tilde{p}_f, \tag{29.8.100}$$

which amounts to a translation. We will use the notation $\tilde{\mathcal{M}}_f$ to denote the map $\tilde{\mathcal{M}}$ expanded about the fixed point $\tilde{q}_f, \tilde{p}_f$. Inserting (8.100) into (8.82) yields the result

Action of $\tilde{\mathcal{M}}_f$:

$$\begin{aligned}\overline{Q} &= rQ + 2r[(1-r)(1-u)/(r-u)]Q - 2\sqrt{ru}[(1-r)(1-u)/(r-u)]P \\ &\quad + \sqrt{r}(\sqrt{r}Q - \sqrt{u}P)^2,\end{aligned} \tag{29.8.101}$$

$$\begin{aligned}\overline{P} &= 2\sqrt{ru}[(1-r)(1-u)/(r-u)]Q + uP - 2u[(1-r)(1-u)/(r-u)]P \\ &\quad + \sqrt{u}(\sqrt{r}Q - \sqrt{u}P)^2.\end{aligned} \tag{29.8.102}$$

Note that the linear part of $\tilde{\mathcal{M}}_f$ gives a matrix that agrees with (8.86).

We will now seek to Lie factorize $\tilde{\mathcal{M}}_f$. The first step is to show that $\tilde{\mathcal{M}}_f$ can be written as the product of 3 maps. Let $F(\tau)$ be the symplectic matrix

$$F(\tau) = \begin{pmatrix} 1+\tau & -\tau \\ \tau & 1-\tau \end{pmatrix}. \tag{29.8.103}$$

With L given by (8.24), define matrices ${}^1M$ and ${}^2M$ by the rules

$${}^1M = L^{1/2}F = \begin{pmatrix} \sqrt{r}(1+\tau) & -\sqrt{r}\tau \\ \sqrt{u}\tau & \sqrt{u}(1-\tau) \end{pmatrix}, \tag{29.8.104}$$

$${}^2M = FL^{1/2} = \begin{pmatrix} \sqrt{r}(1+\tau) & -\sqrt{u}\tau \\ \sqrt{r}\tau & \sqrt{u}(1-\tau) \end{pmatrix}. \tag{29.8.105}$$

Then it is easily verified that there is the relation

$${}^f\tilde{M} = ({}^1M)\ ({}^2M) \tag{29.8.106}$$

provided τ has the value

$$\tau = [(1-r)(1-u)/(r-u)] = (\sqrt{r}\tilde{q}_f - \sqrt{u}\tilde{p}_f). \tag{29.8.107}$$

Here we have used the symbol ${}^f\tilde{M}$ to replace the symbol $\tilde{M}_f$ since we will soon be adding subscripts to denote specific matrix elements. Note also that (8.107) agrees with (8.76) as the notation is intended to indicate.

Next consider three maps:

$$\bar{Q} = {}^2M_{11}Q + {}^2M_{12}P,\ \bar{P} = {}^2M_{21}Q + {}^2M_{22}P; \tag{29.8.108}$$

$$\bar{\bar{Q}} = \bar{Q} + (\bar{Q} - \bar{P})^2,\ \bar{\bar{P}} = \bar{P} + (\bar{Q} - \bar{P})^2; \tag{29.8.109}$$

$$\bar{\bar{\bar{Q}}} = {}^1M_{11}\ \bar{\bar{Q}} + {}^1M_{12}\ \bar{\bar{P}},\ \bar{\bar{\bar{P}}} = {}^1M_{21}\ \bar{\bar{Q}} + {}^1M_{22}\ \bar{\bar{P}}\ . \tag{29.8.110}$$

Concatenating the first two yields the result

$$\bar{\bar{Q}} = {}^2M_{11}Q + {}^2M_{12}P + ({}^2M_{11}Q + {}^2M_{12}P - {}^2M_{21}Q - {}^2M_{22}P)^2, \tag{29.8.111}$$

$$\bar{\bar{P}} = {}^2M_{21}Q + {}^2M_{22}P + ({}^2M_{11}Q + {}^2M_{12}P - {}^2M_{21}Q - {}^2M_{22}P)^2. \tag{29.8.112}$$

Exploiting the specific form of ${}^2M$ simplifies the common parenthetical term in (8.111) and (8.112) to give the result

$$({}^2M_{11}Q + {}^2M_{12}P - {}^2M_{21}Q - {}^2M_{22}P) = (\sqrt{r}Q - \sqrt{u}P). \tag{29.8.113}$$

Thus the product of the first two maps can also be rewritten in the form

$$\bar{\bar{Q}} = {}^2M_{11}Q + {}^2M_{12}P + (\sqrt{r}Q - \sqrt{u}P)^2, \tag{29.8.114}$$

$$\bar{\bar{P}} = {}^2M_{21}Q + {}^2M_{22}P + (\sqrt{r}Q - \sqrt{u}P)^2. \tag{29.8.115}$$

Now concatenate in the third map by solving (8.110) for $\bar{\bar{\bar{Q}}}$,$\bar{\bar{\bar{P}}}$ to find the result

$$\begin{aligned}\bar{\bar{\bar{Q}}} &= [({}^1M_{11})({}^2M_{11}) + ({}^1M_{12})({}^2M_{21})]Q + [({}^1M_{11})({}^2M_{12}) + ({}^1M_{12})({}^2M_{22})]P \\ &\quad + ({}^1M_{11} + {}^1M_{12})(\sqrt{r}Q - \sqrt{u}P)^2,\end{aligned} \tag{29.8.116}$$

$$\begin{aligned}\bar{\bar{\bar{P}}} &= [({}^1M_{21})({}^2M_{11}) + ({}^1M_{22})({}^2M_{21})]Q + [({}^1M_{21})({}^2M_{12}) + ({}^1M_{22})({}^2M_{22})]P \\ &\quad + ({}^1M_{21} + {}^1M_{22})(\sqrt{r}Q - \sqrt{u}P)^2.\end{aligned} \tag{29.8.117}$$

In view of (8.106) and the explicit form (8.104) of ${}^1M$, this result can be rewritten in the form

$$\bar{\bar{\bar{Q}}} = {}^f\tilde{M}_{11}Q + {}^f\tilde{M}_{12}P + \sqrt{r}(\sqrt{r}Q - \sqrt{u}P)^2, \tag{29.8.118}$$

$$\bar{\bar{\bar{P}}} = {}^f\tilde{M}_{21}Q + {}^f\tilde{M}_{22}P + \sqrt{u}(\sqrt{r}Q - \sqrt{u}P)^2, \tag{29.8.119}$$

which is identical to the map (8.101), (8.102) upon replacing $\bar{\bar{\bar{Q}}}$,$\bar{\bar{\bar{P}}}$ by $\bar{Q}$,$\bar{P}$. Thus $\tilde{\mathcal{M}}_f$ can be written as the product of the 3 maps (8.108) through (8.110)

Lie factorization of $\tilde{\mathcal{M}}_f$ is now straightforward. It is easily verified that the Lie transformation for $F(\tau)$ is given (in q,p variables) by the relation

$$\mathcal{F}(\tau) = \exp : (\tau/2)(q-p)^2 : . \tag{29.8.120}$$

Also, the nonlinear relation (8.109) is produced by the Lie transformation $\exp :(q-p)^3/3:$. Finally, the Lie transformation $\mathcal{L}^{1/2}$ corresponding to $L^{1/2}$ is given by

$$\mathcal{L}^{1/2} = \exp[(1/2)(\log\sqrt{ru}\Sigma]\exp :(1/2)(\log\sqrt{u/r})qp := \mathcal{D}^{1/2}\mathcal{S}^{1/2}. \tag{29.8.121}$$

See (8.37) through (8.44). Consequently, the map $\tilde{\mathcal{M}}_f$ (again in q,p variables) has the factorization

$$\tilde{\mathcal{M}}_f = \mathcal{L}^{1/2}\exp :(\tau/2)(q-p)^2 : \exp :(q-p)^3/3 : \exp :(\tau/2)(q-p)^2 : \mathcal{L}^{1/2}. \tag{29.8.122}$$

Note that in deducing (8.122) it is necessary to recall that Lie transformations act in the opposite order of matrices. See Section 8.3. Note also that the factorization (8.122), like (8.81), has the pleasing feature that the outer linear maps (which depend on r,u) appear in a symmetric way, and the central factor again has a fixed simple form.

As indicated at the beginning of this section, our real aim is to expand $\mathcal{M}_-$ about q_f, p_f. Since the fixed point q_f, p_f is generally elliptic or inversion hyperbolic, it generally has index $+1$. (See Section 18.5). We will therefore call this desired map $\mathcal{M}_+$. Then, in view of (8.81), there is the relation

$$\mathcal{M}_+ = \mathcal{O}(5\pi/4)\tilde{\mathcal{M}}_f\mathcal{O}^{-1}(5\pi/4). \tag{29.8.123}$$

From (8.52) there is the result

$$\mathcal{O}(5\pi/4)\exp :(\tau/2)(q-p)^2 : \mathcal{O}^{-1}(5\pi/4) = \exp :\tau q^2 : . \tag{29.8.124}$$

Also, there are the relations (8.147) and (8.148). See Exercise 8.8. It follows that $\mathcal{M}_+$ has the factorization

$$\mathcal{M}_+ = \mathcal{D}^{1/2}\mathcal{H}\mathcal{B}\mathcal{N}\mathcal{B}\mathcal{H}\mathcal{D}^{1/2} \tag{29.8.125}$$

where $\mathcal{B}$ and $\mathcal{N}$ are defined by

$$\mathcal{B} = \exp :\tau q^2 :, \tag{29.8.126}$$

$$\mathcal{N} = \exp[(-\sqrt{8}/3) : q^3 :]. \tag{29.8.127}$$

Here we have again used the result that $\mathcal{D}$ commutes with $\mathcal{S}$ and $\mathcal{O}(\theta)$.

In the work to follow we will study the properties of the general Hénon map either in the form $\mathcal{M}_-$ given by (8.57) or the form $\mathcal{M}_+$ given by (8.123) or (8.125). However, before doing so, we will demonstrate that in the symplectic case the map $\mathcal{M}_+$ can be brought to the form (1.2.39) provided the eigenvalues of the linear part of $\mathcal{M}_+$ (about the origin) are complex. Starting from (8.125) we write

$$\mathcal{D}^{-1/2}\mathcal{M}_+\mathcal{D}^{-1/2} = \mathcal{H}\mathcal{B}\mathcal{N}\mathcal{B}\mathcal{H}, \tag{29.8.128}$$

$$\mathcal{B}\mathcal{H}\mathcal{D}^{-1/2}\mathcal{M}_+\mathcal{D}^{-1/2}\mathcal{H}^{-1}\mathcal{B}^{-1} = \mathcal{B}\mathcal{H}\mathcal{H}\mathcal{B}\mathcal{N} = \mathcal{C}\mathcal{N} \tag{29.8.129}$$

where $\mathcal{C}$ is the symplectic map

$$\mathcal{C} = \mathcal{B}\mathcal{H}\mathcal{H}\mathcal{B}. \tag{29.8.130}$$

It can be shown that $\mathcal{C}$ has a square root (that is also symplectic) if the eigenvalues of $\tilde{M}_f'$ are complex. Moreover, the eigenvalues of $\mathcal{C}$ and $\mathcal{C}^{1/2}$ will be on the unit circle. See Exercise 8.23. Consequently, in this case, (8.129) may be rewritten in the form

$$\mathcal{C}^{-1/2}\mathcal{B}\mathcal{H}\mathcal{D}^{-1/2}\mathcal{M}_+\mathcal{D}^{-1/2}\mathcal{H}^{-1}\mathcal{B}^{-1}\mathcal{C}^{1/2} = \mathcal{C}^{1/2}\mathcal{N}\mathcal{C}^{1/2}, \tag{29.8.131}$$

or

$$\mathcal{C}^{-1/2}\mathcal{B}\mathcal{H}\mathcal{M}_{+}\mathcal{H}^{-1}\mathcal{B}^{-1}\mathcal{C}^{1/2} = \mathcal{D}^{1/2}\mathcal{C}^{1/2}\mathcal{N}\mathcal{C}^{1/2}\mathcal{D}^{1/2}. \tag{29.8.132}$$

Here we have used the fact that $\mathcal{D}$ also commutes with $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{H}$.

Next, for any $\mathcal{C}$ of the form (8.130), there is a κ such that

$$\exp : \kappa qp : \mathcal{C}^{1/2} \exp : -\kappa qp : = \exp[-(\phi/4) : p^2 + q^2 :] = \mathcal{O}(\phi/2), \tag{29.8.133}$$

where ϕ is the phase advance of M'_f. See Exercise 8.24. Also, we have the relation

$$\hat{\mathcal{N}} \overset{\text{def}}{=} \exp : \kappa qp : \mathcal{N} \exp : -\kappa qp : = \exp : -[\exp(-\kappa)](\sqrt{8}/3)q^3 : . \tag{29.8.134}$$

Let $\mathcal{D}_\lambda$ denote the map defined by the relation

$$\mathcal{D}_\lambda = \exp(\lambda\Sigma). \tag{29.8.135}$$

Then there is a choice of λ such that

$$\mathcal{D}_\lambda \hat{\mathcal{N}} \mathcal{D}_\lambda^{-1} = \exp : -q^3 : . \tag{29.8.136}$$

See Exercise 8.25. Finally, consider the transformation $\mathcal{O}(\pi)$. According to (5.4.19) there are the relations

$$\mathcal{O}(\pi)q = -q \ , \ \mathcal{O}(\pi)p = -p, \tag{29.8.137}$$

from which it follows that

$$\mathcal{O}(\pi)[\exp : -q^3 :]\mathcal{O}^{-1}(\pi) = \exp(: q^3 :). \tag{29.8.138}$$

Now put all this information together. Doing so gives the final result

$$\mathcal{A}\mathcal{M}_{+}\mathcal{A}^{-1} = \mathcal{D}^{1/2}\mathcal{O}(\phi/2)[\exp(: q^3 :)]\mathcal{O}(\phi/2)\mathcal{D}^{1/2} \tag{29.8.139}$$

where $\mathcal{A}$ is the linear map

$$\mathcal{A} = \mathcal{D}_\lambda \mathcal{O}(\pi)[\exp : \kappa qp :]\mathcal{C}^{-1/2}\mathcal{B}\mathcal{H}. \tag{29.8.140}$$

See Exercise 8.26. We observe that, apart from the damping map $\mathcal{D}^{1/2}$, the right sides of (1.2.39) and (8.139) are identical. Thus, under the assumption made about the spectrum of $\tilde{M}'_f$ and in the absence of damping, the map $\mathcal{M}_{+}$ given by (8.125) and the map $\mathcal{M}(\theta)$ given by (1.2.39) are physically equivalent.

Exercises

29.8.1. Verify (8.1) and (8.2). Show that the fixed points $\{q_f^{\pm}, p_f^{\pm}\}$ are real when $a > a_{\min}$. See (8.4). Show that, as illustrated in Figures 8.1 and 8.4, the quantity q_f^{+} has a finite limit as $a \to 0$, and $q_f^{-} \to \infty$ as $a \to 0$. Verify (8.5) and (8.6). Find the asymptotes of $q_f^{\pm}$ as $a \to \infty$.

29.8.2. Verify (8.9).

29.8.3. Show that (8.13) is satisfied for $q_f = q_f^-$ if (8.3) holds, and hence the spectrum in this case is real. Show that at the fixed points $q_f^\pm, p_f^\pm$ given by (8.1) and (8.2) the Jacobian matrices of $\mathcal{M}_h^*$ are given by the relations

$$R_\pm = R_h^*(q_f^\pm, p_f^\pm) = \begin{pmatrix} -2aq_f^\pm & 1 \\ b & 0 \end{pmatrix}. \tag{29.8.141}$$

Show that $R_\pm$ has the trace

$$2\sigma_\pm = \mathrm{tr} R_\pm = (1-b) \mp [(1-b)^2 + 4a]^{1/2}. \tag{29.8.142}$$

Show that the spectrum of R_- is positive, and verify (8.18). Evaluate (8.18) for the cases $b = -.3, -.9$ and compare your results with Figures 8.3 and 8.4.

29.8.4. Review Exercise 8.3. Show that the spectrum of R_+ can be complex, and that the transition from real to complex occurs at the a values given by (8.19) and (8.21). Show that at the value of a given by (8.19) the eigenvalues leave the complex plane and have the value

$$\lambda = -\sqrt{-b}. \tag{29.8.143}$$

Compute, using (8.19), a and λ for the cases $b = -.3$ and $-.9$, and compare these values with those shown in Figures 8.5 and 8.6. Show that at the value of a given by (8.21) the eigenvalues also leave the complex plane and have the value

$$\lambda = +\sqrt{-b}. \tag{29.8.144}$$

Again compute numerical results and compare these results with those shown in Figures 8.5 and 8.6. Show that R_+ has -1 as an eigenvalue when

$$(-1+b) = 2\sigma_+, \tag{29.8.145}$$

and derive the condition (8.20). Assuming that period doubling occurs when R_+ has -1 as an eigenvalue, see Section 18.9, show that (8.20) predicts period doubling for $b = -.3$ when $a = 1.2675$, which is consistent with Figure 7.4. Consider also the cases $b = 0$ and $b = +.3$. Compute the corresponding value of a for each and relate your results to Figures 1.2.10 and 7.1.

29.8.5. Solve (8.34) and (8.36) for a to verify (8.35).

29.8.6. Verify the relations (8.39) through (8.44).

29.8.7. Verify the factorization (8.48).

29.8.8. Verify (8.57) and (8.58) using relations of the form

$$\mathcal{O}\mathcal{S}^{1/2}\mathcal{O}^{-1} = \exp:(1/2)(\log\sqrt{u/r})\mathcal{O}qp: \tag{29.8.146}$$

to show that

$$\mathcal{O}(5\pi/4)\mathcal{S}^{1/2}\mathcal{O}^{-1}(5\pi/4) = \mathcal{H}, \tag{29.8.147}$$

$$\mathcal{O}(5\pi/4)\exp:(q-p)^3/3:\mathcal{O}^{-1}(5\pi/4) = \exp[(-\sqrt{8}/3):q^3:]. \tag{29.8.148}$$

29.8.9. Verify (8.67) and (8.68).

29.8.10. Show from (8.1) and (8.2) that the fixed points q_f^+, p_f^+ and q_f^-, p_f^- coincide when

$$(1-b)^2 + 4a = 0. \tag{29.8.149}$$

Show, using (8.31) and (8.32), that (8.149) implies $r = 1$ or $u = 1$. Verify that (8.69) can be rewritten in the form

$$a = -(1-b)^2/4 + (-b)[\cosh\nu - (1-b)/(2\sqrt{-b})]^2, \tag{29.8.150}$$

thereby demonstrating that a as given by (8.69) always satisfies the inequality (8.3) provided ν is real. Note also that equality, (8.149), is achieved for real ν.

29.8.11. Verify that the map (8.81) has the action (8.82), and has the second (besides the origin) fixed point (8.83) and (8.84). Now use (8.85) to verify that $\mathcal{M}_-$ has the second fixed point (8.71) and (8.72).

29.8.12. Show that the first factor in q_f as given by (8.71) vanishes when $\nu = 0$. Nevertheless, show that the second factor, which is τ^2, is sufficiently divergent at $\nu = 0$ (in the nonsymplectic case) so that q_f is also divergent at $\nu = 0$ in the nonsymplectic case. Verify (8.77).

29.8.13. Verify (8.73) through (8.75).

29.8.14. Verify (8.86).

29.8.15. Verify (8.88).

29.8.16. Verify (8.96) and show that it is equivalent to (8.20).

29.8.17. Verify (8.97)

29.8.18. Verify (8.95). Determine the condition for ϕ, as defined by (8.99), to be real in both the symplectic and nonsymplectic cases. Consider the symplectic case by writing $r = \Lambda$, $u = 1/\Lambda$. Show, using (8.88), that then

$$2\sigma = 4 - \Lambda - 1/\Lambda, \tag{29.8.151}$$

and compare your result with (6.24). Show that in the symplectic case the eigenvalues of $\tilde{M}_f$ leave the unit circle through the point -1 when $\Lambda = 3 \pm \sqrt{8}$.

29.8.19. Correlate the data shown in Figures 8.5 and 8.18.

29.8.20. Problem on Poincaré index: show that index on large circle is zero for Hénon map.

29.8.21. Verify (8.106) using (8.86), (8.103) through (8.105), and (8.107).

29.8.22. Verify that $\mathcal{F}(\tau)$ as given by (8.120) corresponds to $F(\tau)$ as given by (8.103).

29.8.23.

29.8.24.

29.8.25.

29.8.26. Verify (8.139) and (8.140) using (8.131) through (8.134), (8.136), and (8.138). Hint: Recall that $\mathcal{D}$ and $\mathcal{D}_\lambda$ commute with $\mathcal{O}$ and $[\exp : \kappa qp :]$. In the course of your verification use relations of the form

$$\begin{aligned}&[\exp : \kappa qp :]\mathcal{C}^{1/2}\mathcal{N}\mathcal{C}^{1/2}[\exp : -\kappa qp :] =\\ &[\exp : \kappa qp :]\mathcal{C}^{1/2}[\exp : -\kappa qp :][\exp : \kappa qp :]\mathcal{N}[\exp : -\kappa qp :] \times\\ &[\exp : \kappa qp :]\mathcal{C}^{1/2}[\exp : -\kappa qp :] = \mathcal{O}(\phi/2)\hat{\mathcal{N}}\mathcal{O}(\phi/2).\end{aligned} \tag{29.8.152}$$

29.9 Period Doubling and Strange Attractors

With a preliminary study of the general Hénon map $\mathcal{M}_h$ behind us, let us explore the results of iterating $\mathcal{M}_h$. We will first study the behavior of $\mathcal{M}_-$, the form of the general Hénon map with its hyperbolic fixed point translated to the origin as given by (8.57), (8.67), and (8.68). We will examine the stable and unstable manifolds, and the nature of the second fixed point, for various representative values of r and u. Then, to explore the behavior near the the second fixed point in more detail, we will employ $\mathcal{M}_+$, the form of the general Hénon map with its second fixed point translated to the origin as gven by (8.123) or (8.125)

29.9.1 Behavior about Hyperbolic Fixed Point

In analogy with (6.4), let us begin with the case where

$$r = \Lambda, u = 1/\Lambda, \text{with } \Lambda = 3, \tag{29.9.1}$$

and for which

$$b = -ru = -1,\ a = -5/9 = -.555\cdots,\ \nu = 1.098\cdots. \tag{29.9.2}$$

In this case $\mathcal{M}_-$ is symplectic and, according to (8.71) and (8.72), the second fixed point has the location

$$\{q_f, p_f\} = \{\sqrt{3/32}, 0\} = \{.306\cdots, 0\}. \tag{29.9.3}$$

Figure 9.1 shows the stable and unstable manifolds for this case as well as the behavior of points near the second fixed point. The second fixed point is elliptic, in accord with Figure 8.14, and has a tune $T = .1959\cdots$. See Exercise 9.1. Evidently Figure 9.1 is similar to Figure 6.2. The difference consists of a rescaling of the axes and a counterclockwise rotation by $5\pi/4$ radians to achieve symmetry about the q axis. See (8.54) and (8.57). Note that the homoclinic point K now lies on the q axis, and that the whole figure is symmetric about the q axis.

Discussion of period doubling.

29.9.2 Behavior about Second Fixed Point

Discussion of strange attractor.

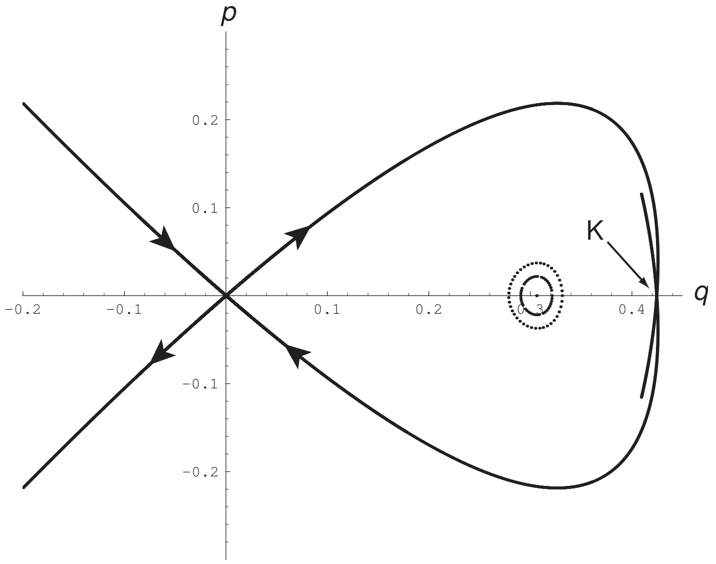


Figure 29.9.1: Stable and unstable manifolds for $\mathcal{M}_{-}$ and behavior of points near the second fixed point for the case $\Lambda = 3$.

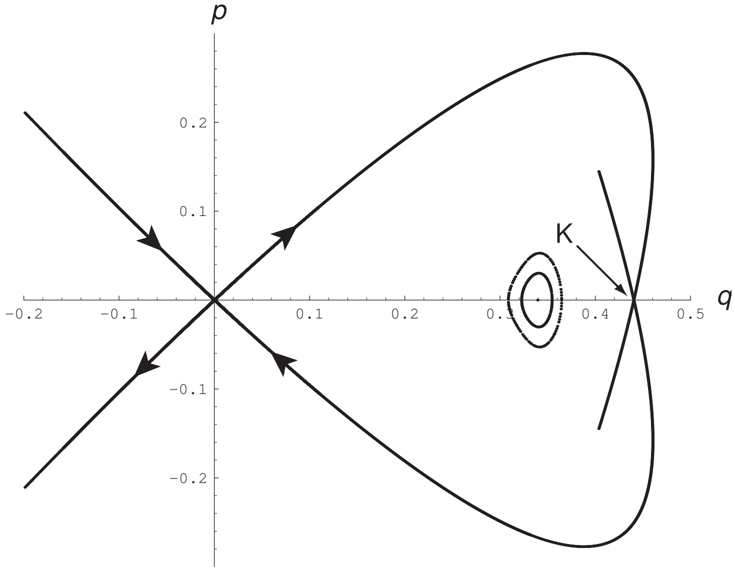


Figure 29.9.2: Stable and unstable manifolds for $\mathcal{M}_{-}$ and behavior of points near the second fixed point for the case $\Lambda = 4$.

Exercises

29.9.1.

29.10 Attempts at Integrals

29.11 Quadratic Maps in Higher Dimensions

This section will discuss Moser's generalization of the quadratic symplectic map to more than 2 dimensions.

29.12 Truncated Taylor Approximations to Stroboscopic Duffing Map

We learned in Section 1.3 that the solutions of ordinary differential equations, under quite general analyticity conditions on their right sides, are analytic in the initial conditions and in whatever parameters that occur. See Theorem 3.3 in Section 1.3. In terms of maps, this means that the maps produced by integrating analytic differential equations will be analytic in the initial conditions and parameters. We may therefore consider approximating such maps by truncated Taylor series in the initial conditions and parameters. Moreover, Section 10.10 described methods for obtaining these truncated Taylor maps.

This section will illustrate, as an example, how the stroboscopic Duffing map can be approximated by truncated Taylor maps, including parameter dependence. At first glance it might appear that the approximation of the stroboscopic Duffing map by a polynomial map, for that is what a truncated Taylor map amounts to, is a foolish enterprise. We know from Chapter 25 that the stroboscopic Duffing map exhibits very complicated properties that almost defy description. Could any polynomial map have similar properties? On the other hand, we have also seen from the logistic and Hénon map examples that polynomial maps can also exhibit complicated behavior. So perhaps there is hope. In fact, we will find that the main features of the stroboscopic Duffing map found in Chapter 25 are reproduced by truncated Taylor maps.

In our actual calculations we will use the quantities Q, σ, and t given by (10.10.103) through (10.10.105). However, for ease of comparison with previous material, when making graphics we will present results in terms of the quantities q, p, and ω of Chapter 25. Finally, for details of how the truncated Taylor maps were computed, see Appendix S.

29.12.1 Saddle-Node Bifurcations

We first explore the possible duplication of saddle-node bifurcations by polynomial maps. This turns out to be the most demanding task and, as we will see, is barely possible for the Duffing map. The reason for the difficulty is that the fixed points move over a considerable region of phase space in the course of a saddle-node bifurcation followed by an inverse saddle-node bifurcation, and a Taylor map may not converge over such a wide domain. See, for example, Figures 25.4.1 and 25.4.2.

To minimize phase-space excursions, let us work with a small value of ϵ, but a value still large enough that saddle-node bifurcations still occur. Figure 12.1 shows the bifurcation diagram for the case $\epsilon = 0.30$. It is this behavior that we wish to reproduce with a Taylor map.

There is another feature of the exact map that is also of interest. Suppose, as in Subsection 10.10.7, we work in terms of the variables z_1 and z_2. Also, let L be the linear part of the stroboscopic map in these variables. Then we may compute the eigenvalues λ of $L(2\pi)$ for each of the fixed points as σ (and therefore also ω) varies. Since $L(2\pi)$ is 2×2, there will be two eigenvalues. These will be the eigenvalues of the linear part of the stroboscopic map (in the variables z_1 and z_2) about any fixed point . Figures 12.2 and 12.3 show (in two perspective views) these eigenvalues plotted as a function of ω over the interval $\omega \in [\omega_{\text{low}}, \omega_{\text{high}}]$ with $\omega_{\text{low}} = 1.27$ and $\omega_{\text{high}} = 1.4$. Notice that we have arranged to have the ordering $\omega_{\text{low}} < \omega_1 < \omega_2 < \omega_{\text{high}}$. The eigenvalues are color coded the same as their corresponding fixed points in Figure 12.1. We would also like to explore to what extent this feature of the exact map can be reproduced by a Taylor map.

The behavior of the eigenvalues shown in these figures can be understood based on the following facts:

1. From (10.10.128) we have the relations

$$\det L(2\pi) = \exp(-4\pi\beta\sigma) = \exp(-4\pi\beta/\omega). \tag{29.12.1}$$

2. For $\omega \in [\omega_{\text{low}}, \omega_{\text{high}}]$, and given that $\beta = .1$, there is the result

$$\det L(2\pi) \in [.37, .41]. \tag{29.12.2}$$

3. Since $L(2\pi)$ is a real matrix, its eigenvalues must be real or complex conjugate. If they are complex conjugate, they must satisfy the relation

$$|\lambda|^2 \in [.37, .41]. \tag{29.12.3}$$

   If they are real, call them λ_1 and λ_2, they must satisfy the relation

$$\lambda_1\lambda_2 \in [.37, .41]. \tag{29.12.4}$$

Armed with these facts, first consider the red curves associated with the eigenvalues of $L(2\pi)$ about the unstable fixed points. These fixed points exist only for ω in the range $\omega \in [\omega_1, \omega_2]$, and hence the red curves appear only in this range. Also, since these fixed points are unstable, for each ω value one of the eigenvalues must satisfy $|\lambda| \geq 1$. In view of (12.3), this fact rules out the possibility of complex conjugate pairs. Thus the eigenvalues for each unstable fixed point must be real, and by (12.4) they must have the same sign. Finally, in the vicinity of the endpoints ω_1 and ω_1, we know that there are two fixed points that are very nearby because that is where two fixed points are born or are annihilated.

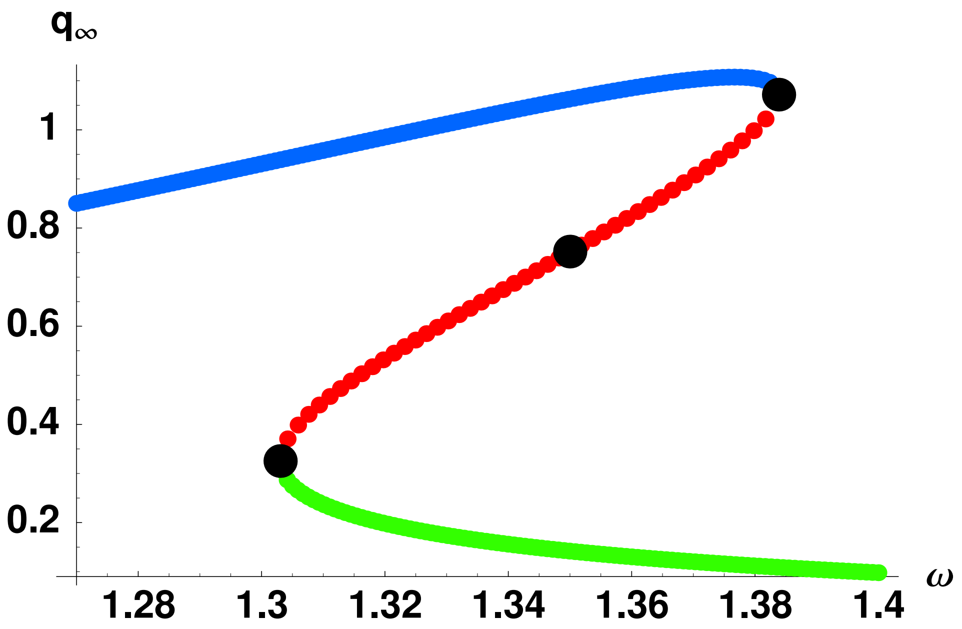


Figure 29.12.1: Bifurcation diagram showing limiting values q_∞ as a function of ω (when $\beta = 0.1$ and $\epsilon = 0.3$) for the stroboscopic Duffing map. The trail of the fixed point that is unique and stable for small values of ω is shown in blue. A pair of fixed points, one stable and one unstable, is born at $\omega = \omega_1 = 1.30305 \cdots$. The trail of the stable fixed point is shown in green and the trail of the unstable fixed point is shown in red. The black dot at the left end of the red trail is the value of $\omega = \omega_1$ at which the pair is born. The blue stable fixed point and the red unstable fixed point annihilate at $\omega = \omega_2 = 1.38386 \cdots$. This point is indicated by the black dot at the right end of the red trail. For larger ω values only the green fixed point remains. The black dot near the center of the red trail marks the expansion point to be used in preparing Figures 12.4 through 12.6.

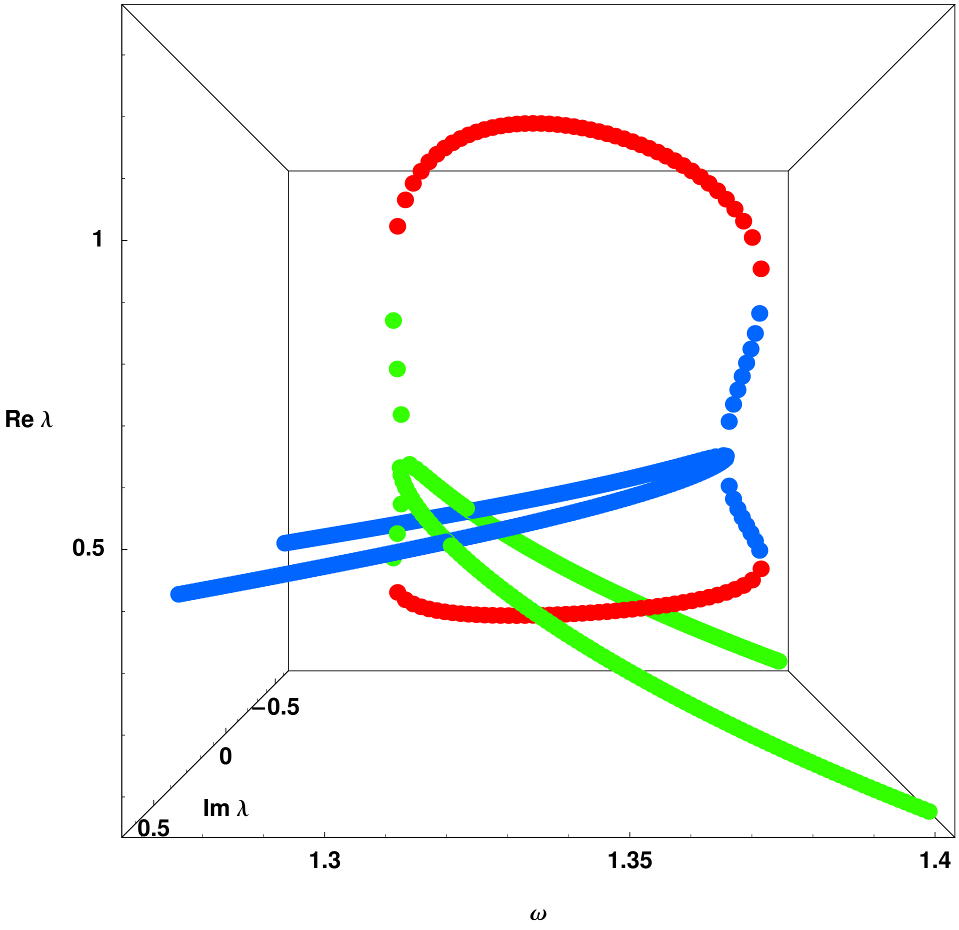


Figure 29.12.2: Eigenvalues of $L(2\pi)$, the linear part of the stroboscopic map (in the variables z_1 and z_2), about the fixed points shown in Figure 12.1. The color coding is that same as in Figure 12.1.

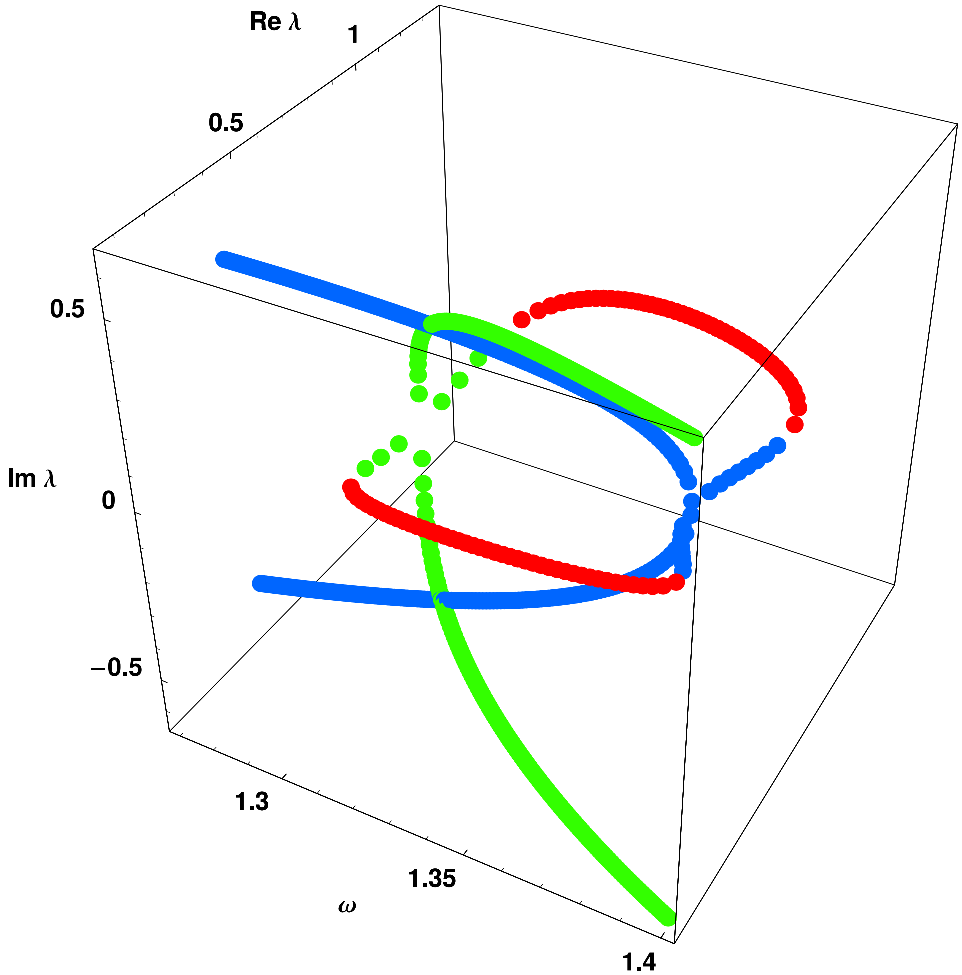


Figure 29.12.3: Eigenvalues of $L(2\pi)$ shown from a different perspective.

Consequently, from the work of Section 4, we know that one eigenvalue, call it λ_1, must be near $+1$. Since the eigenvalues must be real, they cannot change sign without passing through zero, which is forbidden by (12.4). Therefore we conclude that both eigenvalues must be positive. Thus, for the red curves, there must be the relations

$$\lambda_1 = 1 \text{ when } \omega = \omega_1 \text{ or } \omega_2, \tag{29.12.5}$$

$$\lambda_1 > 1, \text{ when } \omega \in (\omega_1, \omega_2), \tag{29.12.6}$$

$$0 < \lambda_2 < 1 \text{ when } \omega \in [\omega_1, \omega_2]. \tag{29.12.7}$$

Examination of the red curves shows that these relations are indeed satisfied.

Consider next the blue curves. They end at $\omega = \omega_2$ because that is the ω value at which the blue and red fixed points mutually annihilate. See Figure 12.1. For ω slightly less than ω_2, we know that the eigenvalues associated with the red fixed points are real and positive. Since the blue fixed points are near the red fixed points for these ω values, it follows that the eigenvalues associated with the blue fixed points must also be real and positive for these ω values. They must also be less than 1 because the blue fixed points are stable. Then, as ω is decreased, the smaller eigenvalue grows and the larger eigenvalue decreases until they become equal. For still smaller ω they leave the real axis and become complex conjugates. We know this must eventually happen because for small enough ω the Duffing oscillator is out of resonance with the drive, and its behavior is essentially that of an undriven oscillator. In that case the stable fixed point is nearly the origin, and points launched near this fixed point spiral into it due to the free (and damped) oscillations of the oscillator. The eigenvalues must have imaginary parts to produce this free oscillation. All this behavior is consistent with the facts listed above.

Finally, consider the green curves. The discussion of their behavior is similar to that of the blue curves. They begin at $\omega = \omega_1$ because that is the ω value at which the red and and green fixed points are born. See Figure 12.1. For ω slightly larger than ω_1, we know that the eigenvalues associated with the red fixed points are real and positive. Since the green fixed points are near the red fixed points for these ω values, it follows that the eigenvalues associated with the green fixed points must also be real and positive for these ω values. They must also be less than 1 because the green fixed points are stable. Then, as ω is increased, the smaller eigenvalue grows and the larger eigenvalue decreases until they become equal. For still larger ω they leave the real axis and become complex conjugates. We know this must eventually happen because for large enough ω the Duffing oscillator is again out of resonance with the drive, and its behavior is essentially that of an undriven oscillator. In that case the stable fixed point is nearly the origin, and points launched near this fixed point spiral into it due to the free (and damped) oscillations of the oscillator. The eigenvalues must have imaginary parts to produce this free oscillation. Again, all this behavior is consistent with the facts listed above.

We will now try to duplicate with a Taylor map the results shown in Figures 12.1 through 12.3. Let ω_{mid} be a value of ω that is approximately *midway* between ω_1 and ω_2. A convenient value is $\omega_{\text{mid}} = 1.35$. Let $(q^{\text{rf}}(\omega_{\text{mid}});\ p^{\text{rf}}(\omega_{\text{mid}}))$ be the unstable (red) fixed point corresponding to this value of ω. It is shown in Figure 12.1. Suppose the stroboscopic Duffing map is Taylor expanded about this fixed point through eighth order in both phase-space coordinates and parameter value. (Use the equations described in Section 10.10.6 and

Section 10.10.7 and the parameter σ, but subsequently plot results in terms of the original coordinates q, p and the parameter ω.) We will call this map $\mathcal{M}_8$. With this polynomial map in hand, it is easy to extract its linear part (Jacobian matrix) L. Now carry out the following steps:

1. Slowly *increase* ω from its initial value $\omega_{\rm mid}$ and find and plot in red the trail of the unstable fixed point for the polynomial map as ω is increased. Since L is available, the location of this fixed point for each value of ω can be found using Newton's method.

2. At the same time, since L is known, calculate its determinant and eigenvalues. Examine the determinant to see how much it differs from the exact value given by (12.1). Plot the eigenvalues in red. See if one of them is approaching the value $+1$. Continue increasing ω until one eigenvalue, call it λ_1, takes on the value $+1$. Record the ω value at which this occurs, and call it $\omega_2^{\rm approx}$.

3. When L has eigenvalue $+1$, we know that a pair of fixed points should be born or annihilated. Starting with $\omega = \omega_2^{\rm approx}$, slowly *decrease* ω while now looking for two (initially nearby) fixed points. One of them should be on the trail of red (unstable) fixed points, and the other should be on the trail of the blue (stable) fixed points. Plot the blue points.

4. At the same time, find L about each blue fixed point. Then find its determinant and eigenvalues. Plot these eigenvalues in blue.

5. Continue to decrease ω until the determinant computed as described in item 4 above deviates by from from its exact value by some significant amount thereby indicating that the polynomial map is becoming unreliable. In this example we have required that this deviation be less than .1 which, in view of (12.2), amounts to an error of approximately 25% or less.

6. Carry out analogous calculations to find points on the trail of the green (stable) fixed point and to find the eigenvalues associated with these fixed points. That is, again starting with $\omega_{\rm mid}$ and the the associated fixed point $(q^{\rm rf}(\omega_{\rm mid});\ p^{\rm rf}(\omega_{\rm mid}))$, now slowly *decrease* the value of ω while again plotting red points and computing red eigenvalues. Continue until again $\lambda_1 = 1$. Record the ω value at which this occurs, and call it $\omega_1^{\rm approx}$. Now slowly increase ω, again finding pairs of fixed points, one red and one green. For each green fixed point, find the linear part of the polynomial map, compute its determinant, and find and plot its eigenvalues in green.

What is the outcome of carrying out all these steps? First, Figures 12.4 through 12.6 show the analogs of Figures 12.1 through 12.3, but now computed using the Taylor map. Evidently, there is good qualitative agreement. In particular, the saddle-node bifurcations are qualitatively reproduced by the Taylor map, at least in the neighborhood of these bifurcations. Second, we find that $\omega_1^{\rm approx} = 1.30325\cdots$ and $\omega_2^{\rm approx} = 1.38225\cdots$. These values agree well with the exact values given in the caption of Figure 12.1. Also, for the red fixed points in Figure 12.4 corresponding to the end values $\omega_1^{\rm approx}$ and $\omega_2^{\rm approx}$, we find that the determinant of the linear part of the Taylor map deviates from that of the linear part of the the exact stroboscopic Duffing map by approximately .001, and is still smaller for the interior red points. Correspondingly, the trail of unstable fixed points is well reproduced. Finally, for the smallest ω value on the blue trail and the largest ω value on the green trail, the error in the determinant becomes as large as .1, which is why the computation of the green trail is terminated slightly to the right of ω_1^{approx} and the computation of the blue trail is terminated slightly to the left of ω_2^{approx}.

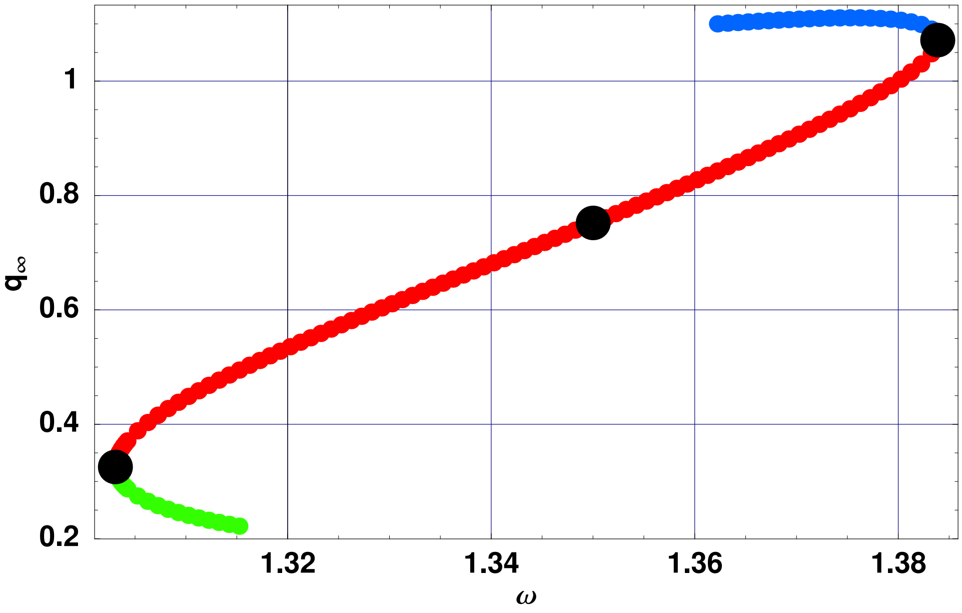


Figure 29.12.4: The analog of Figure 12.1 computed using $\mathcal{M}_8$, an eighth-order approximation to $\mathcal{M}$ including parameter dependence. The black point near the center of the red trail is the point about which the Taylor expansion of the stroboscopic Duffing map is constructed. The black dots near the ends of the red trail are the exact values of the fixed points for the exact values of ω_1 and ω_2.

Having demonstrated that a single map reproduces the saddle-node bifurcations (albeit only over a rather small domain), it is worth exploring if more can be achieved with the use of two or more Taylor maps expanded about different points. The goal would be to cover more of parameter and phase space with overlapping domains associated with multiple maps. We will find that more can indeed be achieved; but, of course, more work is also required.

For simplicity, we will restrict our discussion to the use or two maps. In this case a promising possibility is to choose as expansion points the fixed point in Figure 12.1 where the the red and green trails merge at $\omega = \omega_1$ and the fixed point where the red and blue trails merge at $\omega = \omega_2$. We can then use the Taylor map associated with ω_1 to compute the red and green trails and their associated eigenvalues, and we can use the Taylor map associated with ω_2 to compute the red and blue trails and their associated eigenvalues. As before, we terminate a trail when the determinant of the linear part of the the polynomial-based map differs from its exact value by .1. We expect that it should now be possible to extend the

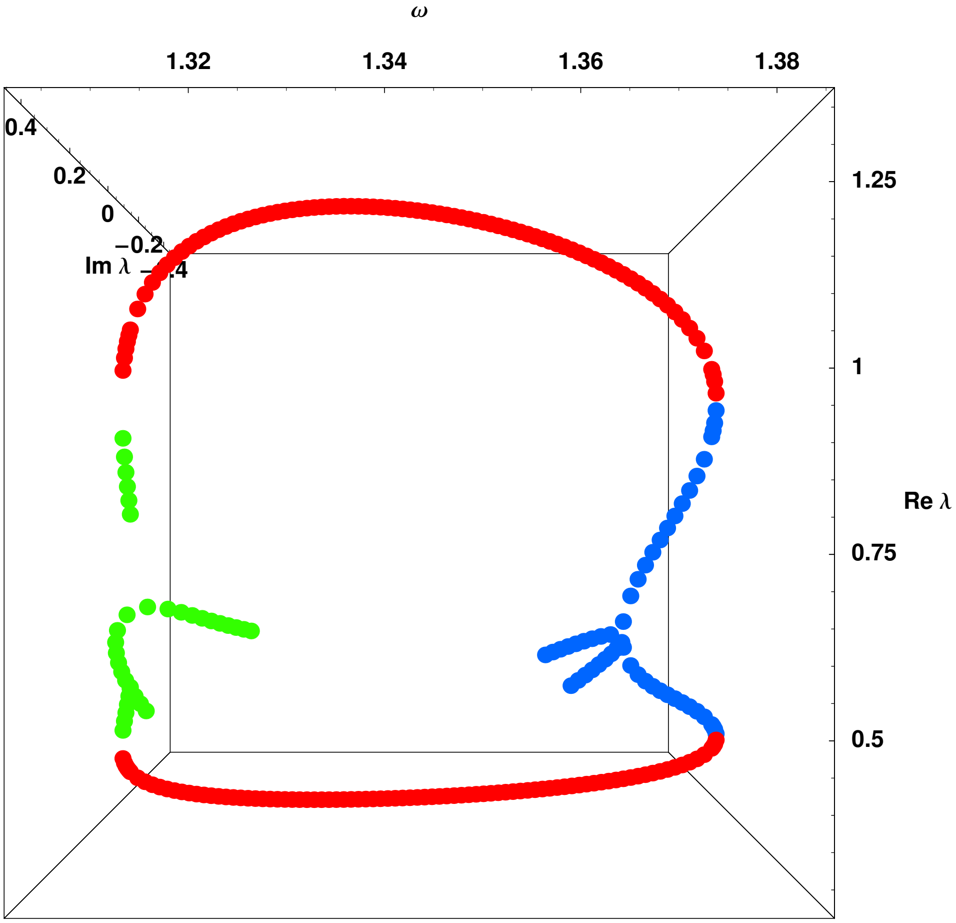


Figure 29.12.5: The analog of Figure 12.2 computed using $\mathcal{M}_8$, an eighth-order approximation to $\mathcal{M}$ including parameter dependence.

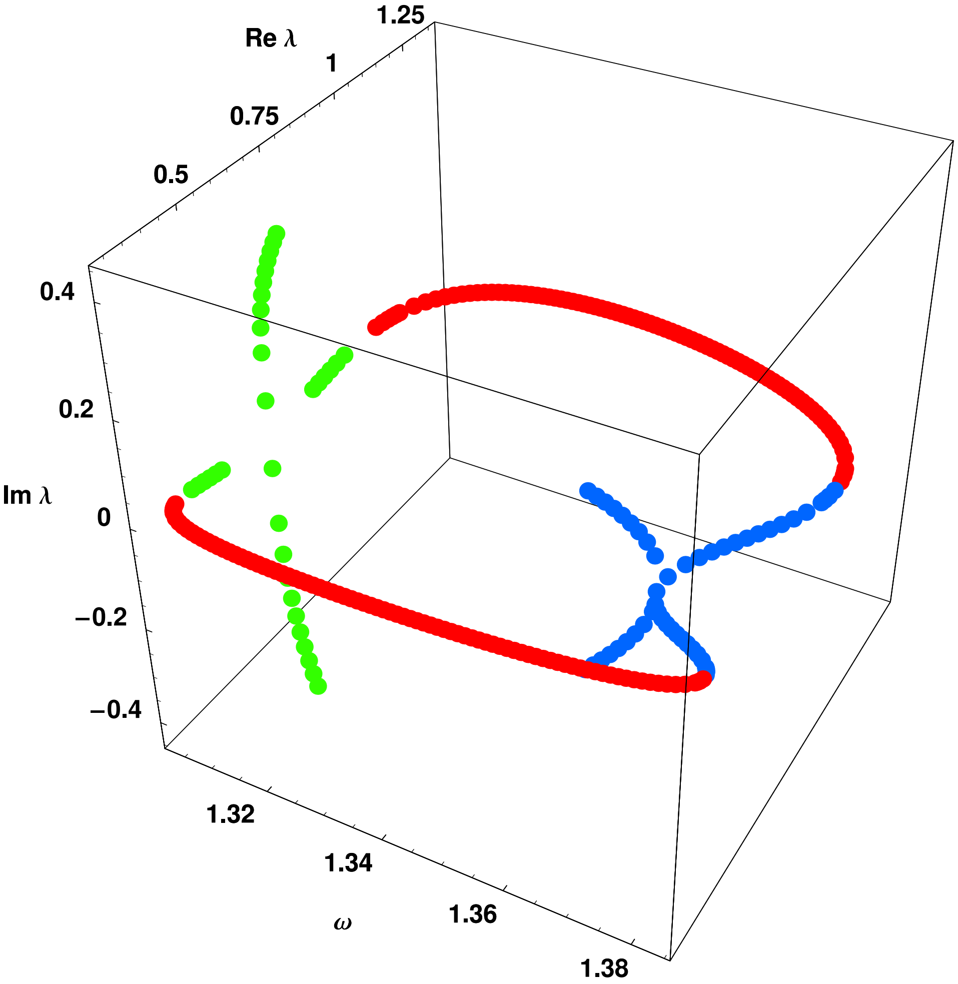


Figure 29.12.6: Data of Figure 12.5 shown from a different perspective.

blue trail farther to the left and the green trail farther to the right. Note that "red" values get computed twice, once for each map, and we can compare them to see how well they agree.

Figures 12.7 through 12.9 show the result of this procedure. Observe that in Figure 12.7 the blue and green trails are indeed extended, and the red points appear to lie almost on a single trail because both Taylor maps give results that nearly agree in this region of parameter and phase space. The red trail computed using the map expanded about ω_1 and its associated fixed point is terminated just to the left of $\omega = \omega_2$, at which point there is an error in the determinant of .009. The green trail is continued on to ω slightly larger than 1.45, at which point the error in the determinant is .1 The red trail computed using the map expanded about ω_2 and its associated fixed point is terminated near $\omega = 1.329$ where the error in the determinant becomes .1. The error in the determinant for the blue branch reaches .1 slightly to the right of $\omega = 1.28$. Observe also that in Figures 12.8 and 12.9 there is some hint of difference in the red-point eigenvalues found using the two different maps, thereby indicating that the computation of the linear-part of a map based on a Taylor expansion can be more demanding than the computation of a fixed point.

We conclude, for the stroboscopic Duffing map, that saddle-node bifurcations can be reproduced by a single polynomial map over a somewhat limited region of parameter and phase space, and that this region can be enlarged by the use of multiple maps.

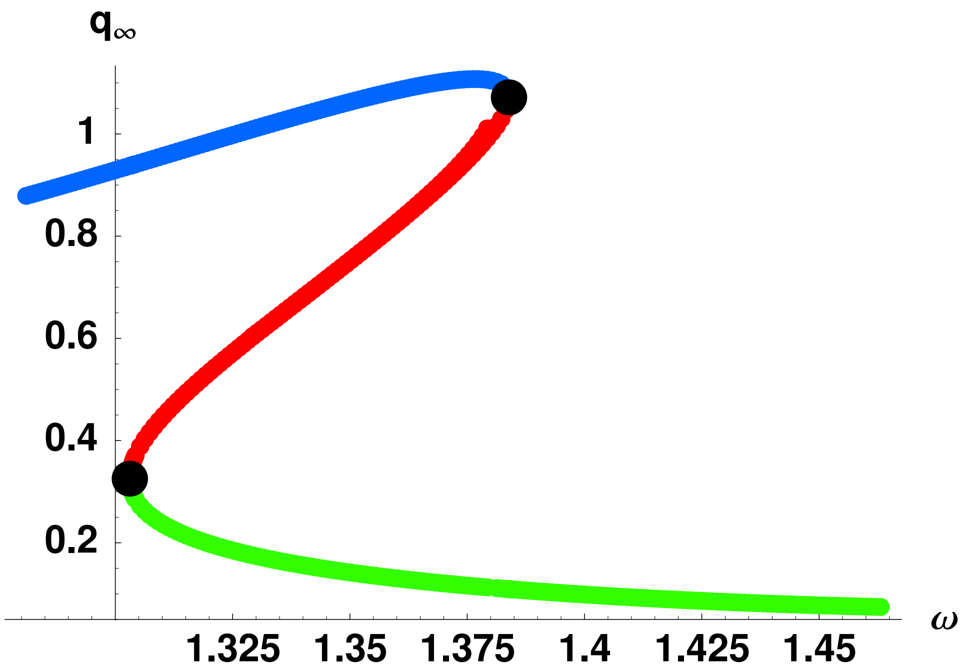


Figure 29.12.7: The analog of Figure 12.1 computed using two eighth-order polynomial maps including parameter dependence. The black dots at the ends of the red trail, located at ω_1 and ω_2, are the points about which the Taylor expansions of the stroboscopic Duffing map are constructed.

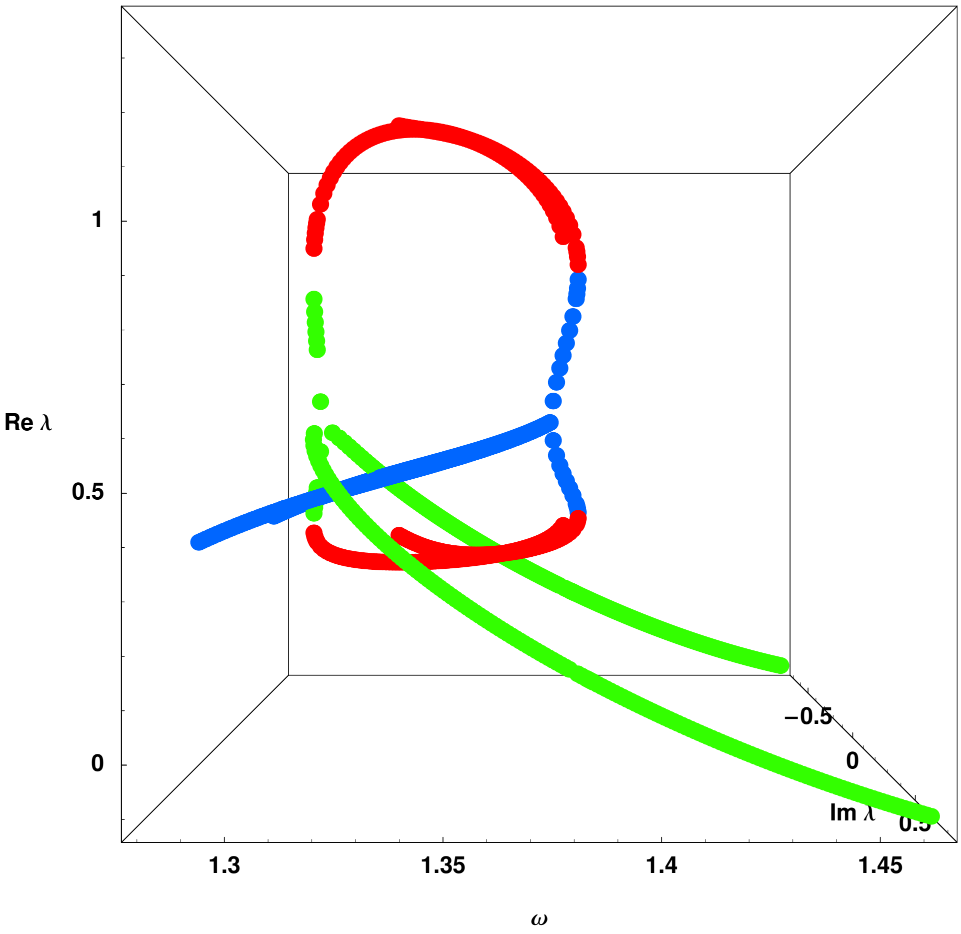


Figure 29.12.8: The analog of Figure 12.2 computed using two eighth-order polynomial maps including parameter dependence.

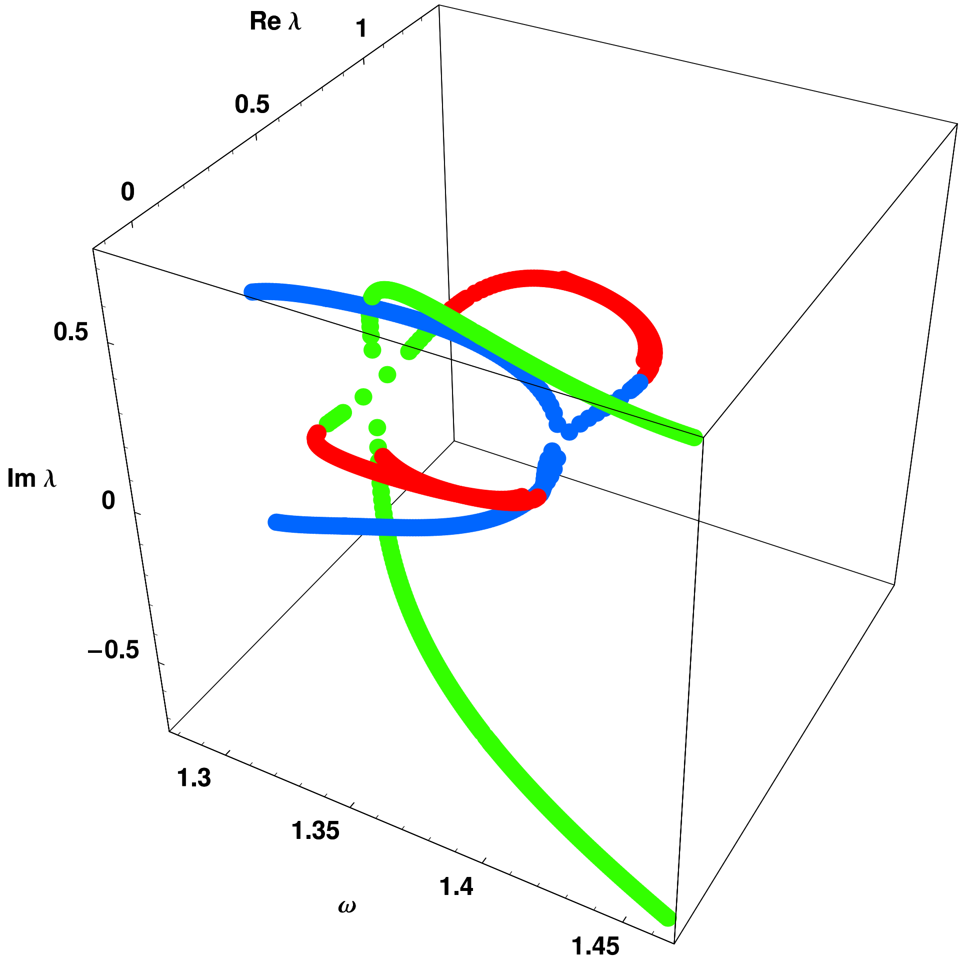


Figure 29.12.9: Data of Figure 12.8 shown from a different perspective.

Exercises

29.12.1. Look at the green and blue curves in Figures 12.2 and 12.3. For small ω only the blue curves exist, and for large ω only the green curves exist. Somewhere within the interval $[\omega_1, \omega_2]$ there appears to be an ω value for which there are blue-green intersections. For example, in Figure 12.3 the bottom blue and green curves appear to intersect and, at the same ω value, the top blue and green curves also appear to intersect. Prove mathematically that these intersections actually do occur and are not an artifact of the thickness of the lines employed in making the figures.

29.12.2. Exercise on equivariance causing eigenvalues to be the same.

29.12.2 Pitchfork Bifurcations

We have seen that Duffing saddle-node bifurcations can be reproduced by truncated Taylor maps over a limited region of parameter and phase space. Now we will see that pitchfork bifurcations can also be reproduced. Here the problem of domain size is somewhat less pressing because the phase-space excursions associated with a pitchfork bifurcation are generally smaller than those for a saddle-node bifurcation. We will find, for example, that the case $\epsilon = 2.5$ can be handled with the use of a single truncated Taylor map.[3]

Figure 12.10 shows the bifurcation diagram for the Duffing stroboscopic map in the case $\epsilon = 2.5$ (and $\beta = 0.1$) for ω in the vicinity of the first bubble. In this figure the fixed points are shown over the interval $\omega \in [\omega_{\text{low}}, \omega_{\text{high}}]$ with $\omega_{\text{low}} = .76$ and $\omega_{\text{high}} = 1.075$. The trail of the stable fixed point, before the pitchfork bifurcation that occurs at $\omega = \omega_1 = .87076\cdots$ and after the pitchfork merger that occurs at $\omega = \omega_2 = .96639\cdots$, is shown in black. The trails of the two stable fixed points that exist after the pitchfork bifurcation and before the pitchfork merger are shown in blue and green. The trail of the associated unstable fixed point is shown in red.

Also shown in Figures 12.11 and 12.12 are the eigenvalues of the linear parts of the map about the various fixed points. The color coding is the same as in Figure 12.10. The relation (12.1) in this case yields, for $\omega \in [\omega_{\text{low}}, \omega_{\text{high}}]$, the alternatives of complex conjugate eigenvalues with

$$|\lambda|^2 \in [.2361, .2724], \tag{29.12.8}$$

or both eigenvalues real with

$$\lambda_1\lambda_2 \in [.2361, .2724]. \tag{29.12.9}$$

For $\omega \approx \omega_{\text{low}}$ the eigenvalues are complex and (12.8) holds. Then, as ω increases, one of the eigenvalues must approach $+1$ because of the imminent bifurcation. Along the way, as the figure illustrates, the eigenvalues must become real in order to not violate (12.8), and (12.9) then holds, in which case both eigenvalues are positive. At bifurcation, when $\omega = \omega_1$, one of the eigenvalues becomes equal to $+1$.

[3] However, larger ϵ values, such as the value $\epsilon = 5.5$ used in Figures 25.5.2 through 25.5.4, do require the use of multiple maps.

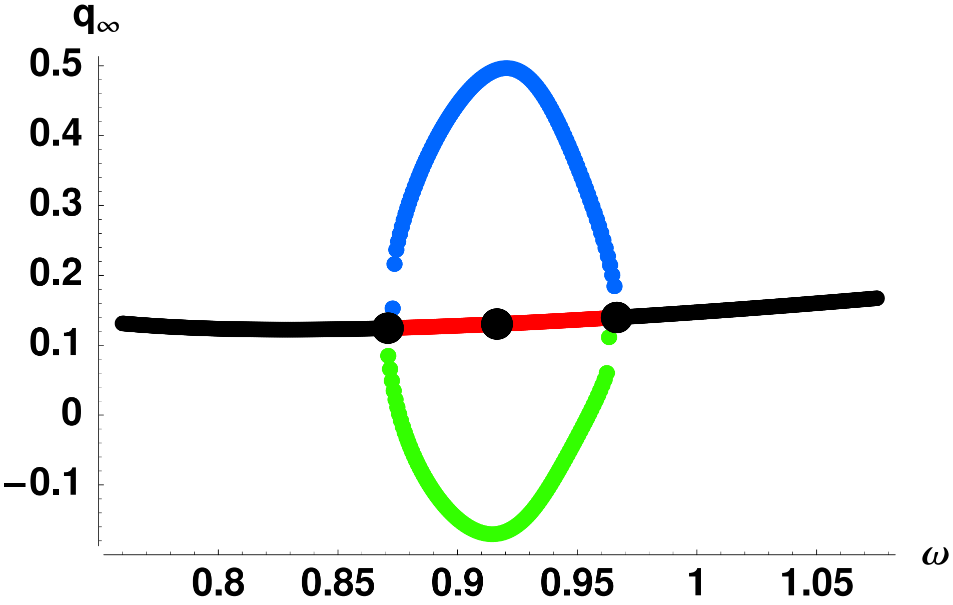


Figure 29.12.10: Bifurcation diagram for the Duffing stroboscopic map in the case $\epsilon = 2.5$ (and $\beta = 0.1$) for ω in the vicinity of the first bubble. The trail of the stable fixed point before the pitchfork bifurcation and after the pitchfork merger is shown in black. The trails of the two stable fixed points that exist after the pitchfork bifurcation and before the pitchfork merger are shown in blue and green. The trail of the associated unstable fixed point is shown in red. The black dot at the left end of the red trail is located at $\omega = \omega_1$. and the black dot at the right end of the red trail is located at $\omega = \omega_2$. The black dot near the middle of the red trail indicates the value $\omega = \omega_{\text{mid}}$ to be used as an expansion point.

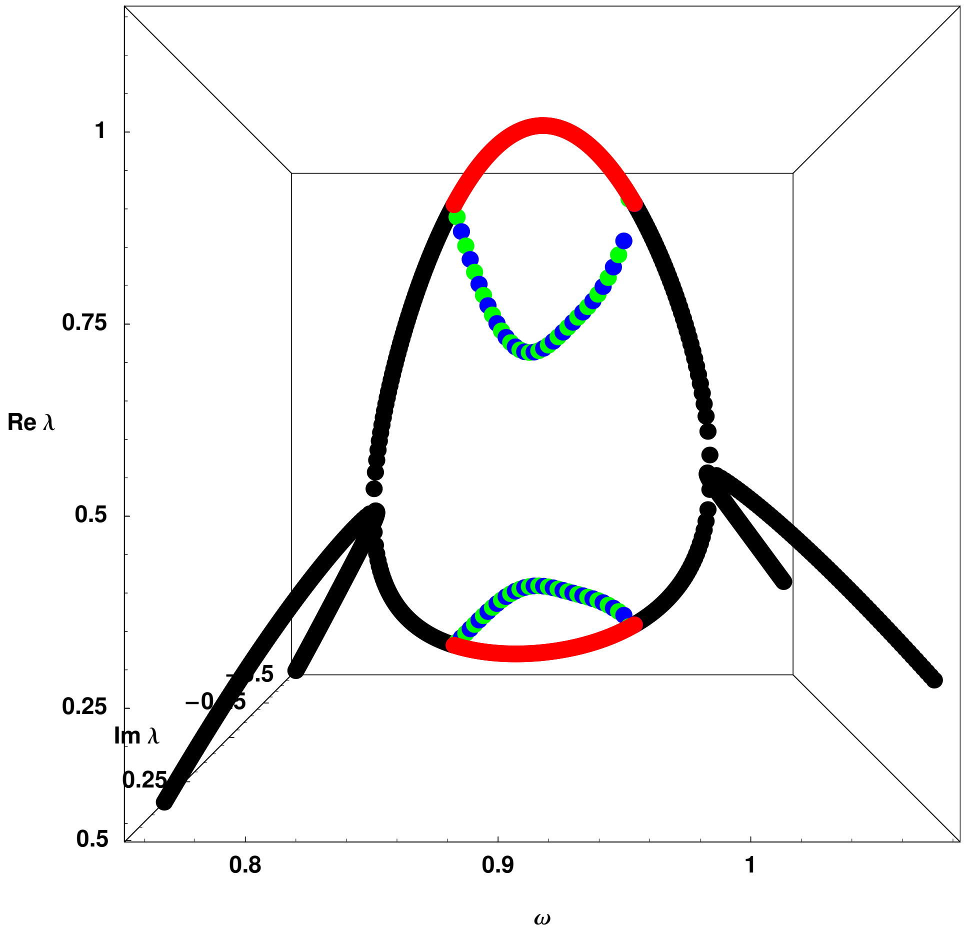


Figure 29.12.11: Eigenvalues of $L(2\pi)$, the linear part of the stroboscopic map (in the variables z_1 and z_2), about the fixed points shown in Figure 12.10. The color coding is that same as in Figure 12.10. Note two of the curves are colored blue-green because, as explained in the text, there is overlap because of equivariance symmetry.

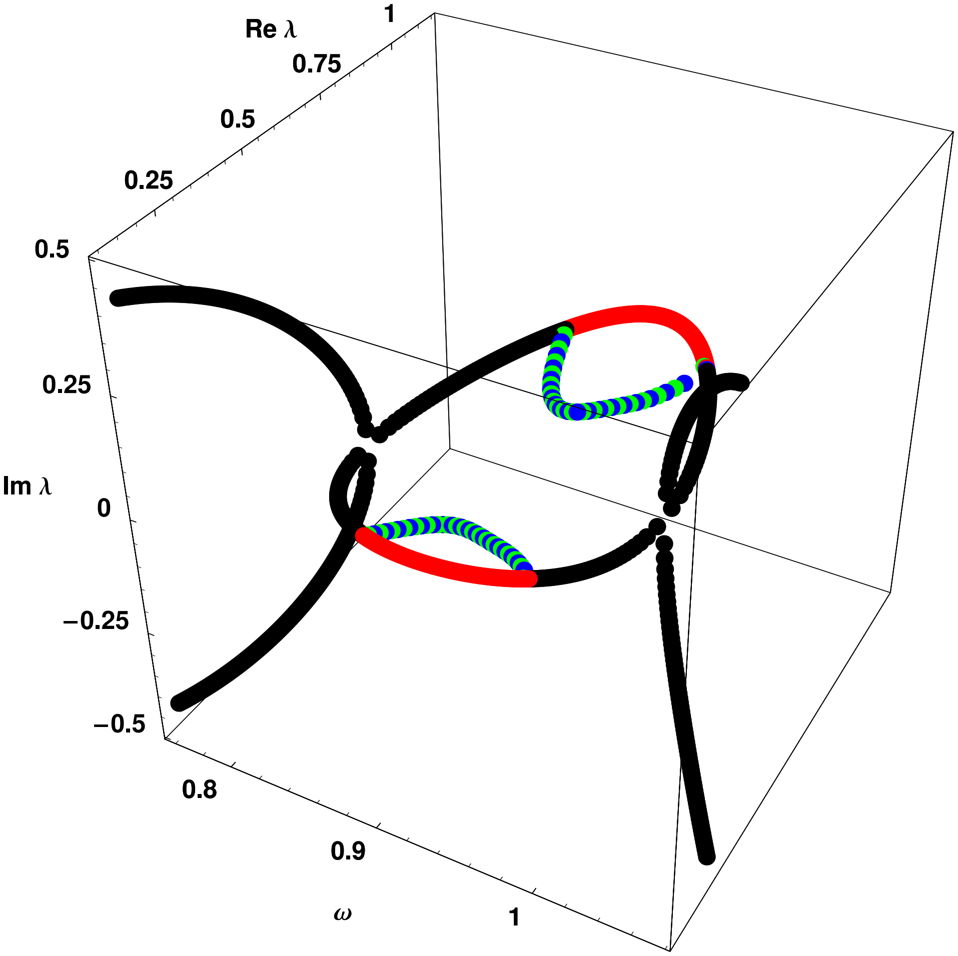


Figure 29.12.12: Data of Figure 12.11 shown from a different perspective.

After bifurcation at ω_1, and within the interval (ω_1, ω_2), we might expect there would be six curves: two red, two blue, and two green. This was the case for the saddle-node bifurcation. Recall Figures 12.2 and 12.3. Inspection of Figures 12.11 and 12.12 shows there is indeed a pair of red curves corresponding to the unstable fixed point trail. However, the blue and green curves overlap in pairs. For this reason, each curve arising from the stable blue and green trails is displayed both in blue and green. It can be shown that this overlap is a consequence of the equivariance relation (25.4.6) that maps the two stable periodic orbits (corresponding to the blue and green stable fixed points) into each other. As a result, for each ω value in (ω_1, ω_2), the linear parts of the map about the blue and green fixed points have the *same* eigenvalues. See Exercise 12.2.

At the inverse bifurcation, when $\omega = \omega_2$, one of the eigenvalues again becomes equal to $+1$. Just beyond the inverse bifurcation there are two real eigenvalues corresponding to the black trail and (12.9) again holds. Eventually these eigenvalues again become complex so that (12.8) then again holds.

It is the behavior shown in Figures 12.10 through 12.12 that we would like to reproduce with a single truncated Taylor map. Figures 12.13 through 12.15 verify that this is indeed possible. Figure 12.13 shows the trails of the fixed points for a single eighth-order map computed by expanding about the black point near the middle of the red trail shown in Figure 12.10 with $\omega_{\text{mid}} = .916349$. Figures 12.14 and 12.15 show the eigenvalues of the linear parts of the eight-order map evaluated at the fixed points on the trails in Figure 12.13. It is found that, over the ω interval displayed, the error in the determinant of the linear part of the map is less than .022. This worst error occurs when $\omega = \omega_{\text{low}}$. To the eye, the truncated Taylor map results appear to be identical to the exact results. A single truncated Taylor map has indeed replicated the pitchfork bifurcation behavior of the exact map.

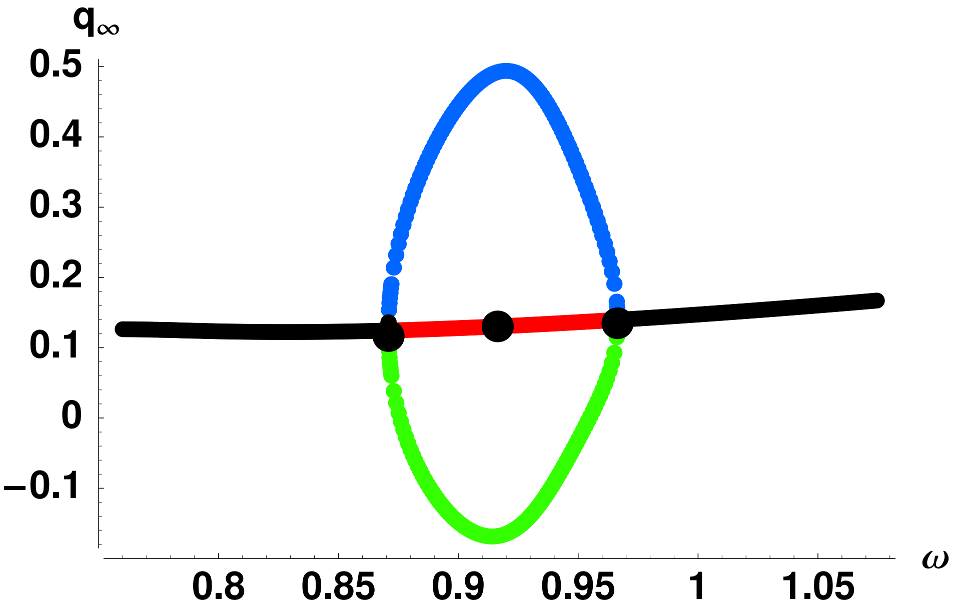


Figure 29.12.13: The analog of Figure 12.10 computed using $\mathcal{M}_8$, an eighth-order approximation to $\mathcal{M}$ including parameter dependence. The black point near the center of the red trail is the point $\omega_{\rm mid} = .916349$ about which the Taylor expansion of the stroboscopic Duffing map was constructed. The other two black points are located at the exact values of ω_1 and ω_2.

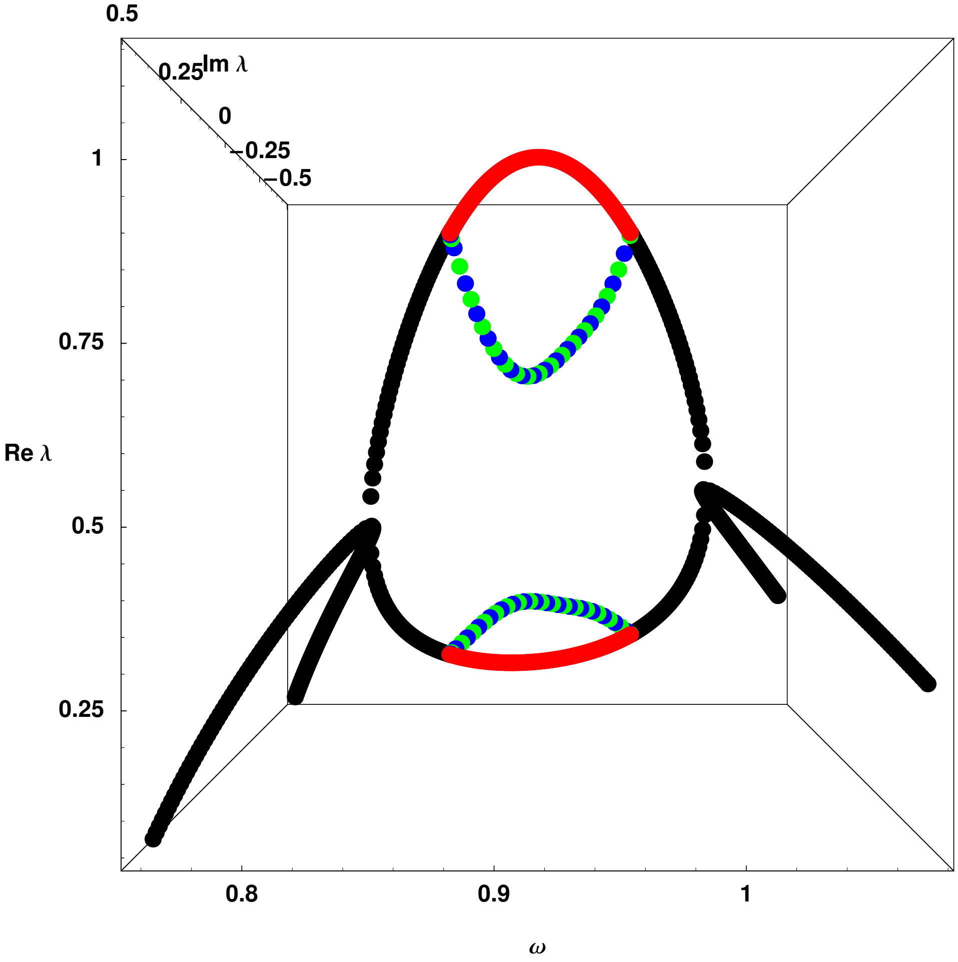


Figure 29.12.14: The analog of Figure 12.11 computed using $\mathcal{M}_8$, an eighth-order approximation to $\mathcal{M}$ including parameter dependence.

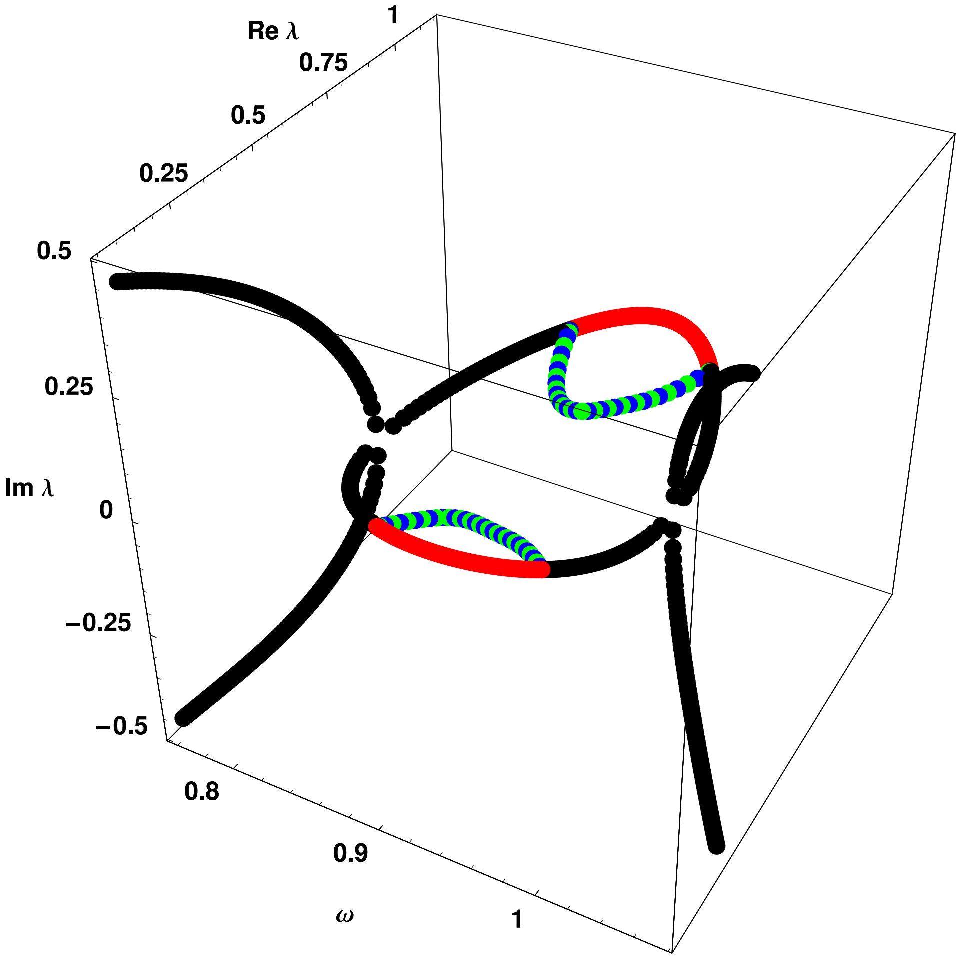


Figure 29.12.15: Data of Figure 12.14 shown from a different perspective.

29.12.3 Infinite Period-Doubling Cascade and Strange Attractor

The main purpose of this subsection is to demonstrate that a single truncated Taylor map can reproduce a Duffing map period doubling cascade and associated strange attractor. However, before doing so, it is instructive to examine the period doubling mechanism. We also verify an earlier conjecture.

Period Doubling Mechanism

Look again at Figure 25.8.6, which shows part of the upper cascade in Figure 25.8.5. Part of it is reproduced below in Figure 12.16 to display the first period doubling bifurcation in more detail. Also, to avoid confusion, the part of the trail of the stable fixed point associated with the second saddle-node bifurcation, and all other high-order fixed points and their bifurcations, are suppressed. Review the discussion of period doubling at the end of Section 9. For ω less than the bifurcation value the stroboscopic Duffing map $\mathcal{M}$ has a single stable fixed point z_f, and its trail is shown in black. After the bifurcation there is still a single fixed point z_f, but now it is unstable and its trail is shown in red. Next consider the map $\mathcal{M}^2$. For each ω value less than the bifurcation value it too has the corresponding point on the black trail as a fixed point, and this fixed point is stable. And for each ω value greater than the bifurcation value it too has the corresponding point on the red trail as a fixed point, and this fixed point is unstable. But, for each ω value greater than the bifurcation value, $\mathcal{M}^2$ has two additional fixed points, and these fixed points are shown in blue and green. These fixed points of $\mathcal{M}^2$ are *not* fixed points of $\mathcal{M}$. Let $z_{\rm blue}$ be a blue fixed point of $\mathcal{M}^2$ corresponding to some value of ω, and let $z_{\rm green}$ be the green fixed point of $\mathcal{M}^2$ for the same value of ω. Then we know there are the relations

$$\mathcal{M}z_{\rm blue} = z_{\rm green}, \tag{29.12.10}$$

$$\mathcal{M}z_{\rm green} = z_{\rm blue}. \tag{29.12.11}$$

Also we know that the linear parts of $\mathcal{M}^2$ about the corresponding fixed points $z_{\rm blue}$ and $z_{\rm green}$ must have the same eigenvalues. Finally, we know that the blue and green fixed points must be (initially) stable. See the discussion at the end of Section 9.

We expect that at the period doubling bifurcation one of the eigenvalues of the linear part of $\mathcal{M}$ will take on the value -1. This is indeed the case. Figures 12.17 and 12.18 show these eigenvalues before, at, and after period doubling. They are colored black while z_f is stable, and red after z_f becomes unstable. Note that for the smaller values of ω the eigenvalues are complex. Then, as ω is increased, they leave the complex plane to become real and negative, but still have magnitude less than 1. Then, as ω is further increased, one of them takes on the value -1, at which point z_f becomes unstable. It remains unstable as ω is increased still further.

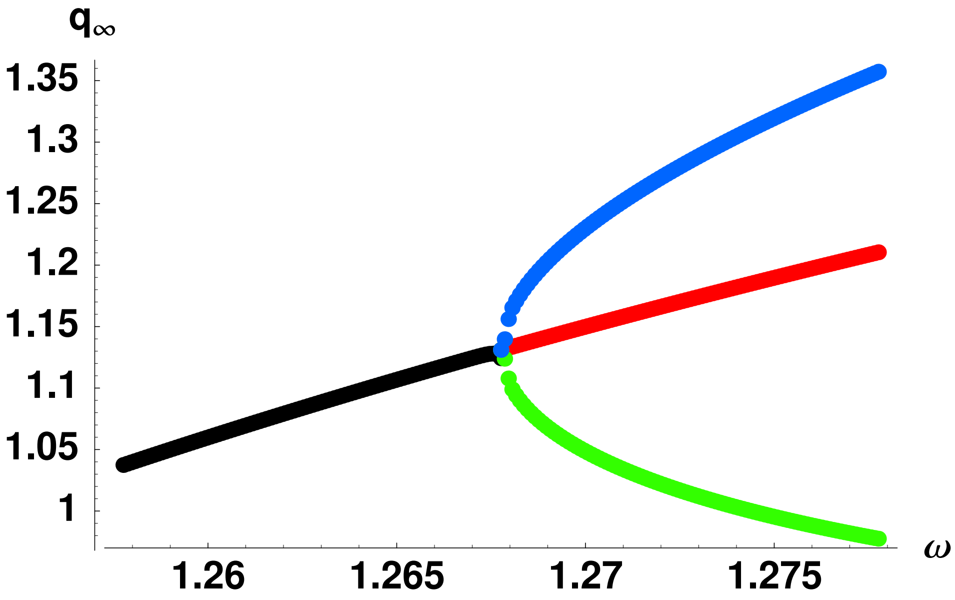


Figure 29.12.16: Detail of part of the period doubling bifurcation shown in Figure 25.8.6. The map $\mathcal{M}$ has one fixed point z_f before period doubling, it is stable, and its trail is shown in black. After period doubling $\mathcal{M}$ still has one fixed point z_f, it is unstable, and its trail is shown in red. These fixed points are, of course, also fixed points of $\mathcal{M}^2$. After period doubling, $\mathcal{M}^2$ has two additional fixed points whose trails are shown in blue and green. These period-two fixed points are not fixed points of $\mathcal{M}$. Instead, they are sent into each other under the action of $\mathcal{M}$.

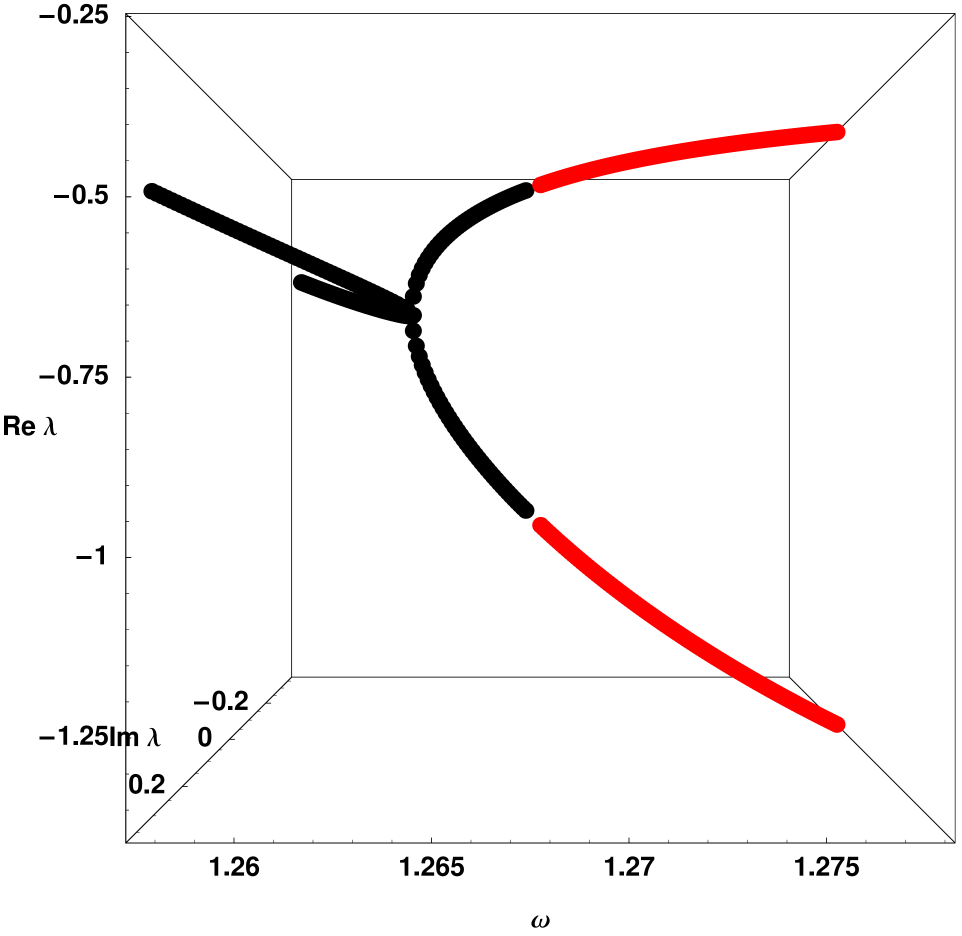


Figure 29.12.17: Eigenvalues of the linear part of $\mathcal{M}$ in the vicinity of its period-doubling bifurcation.

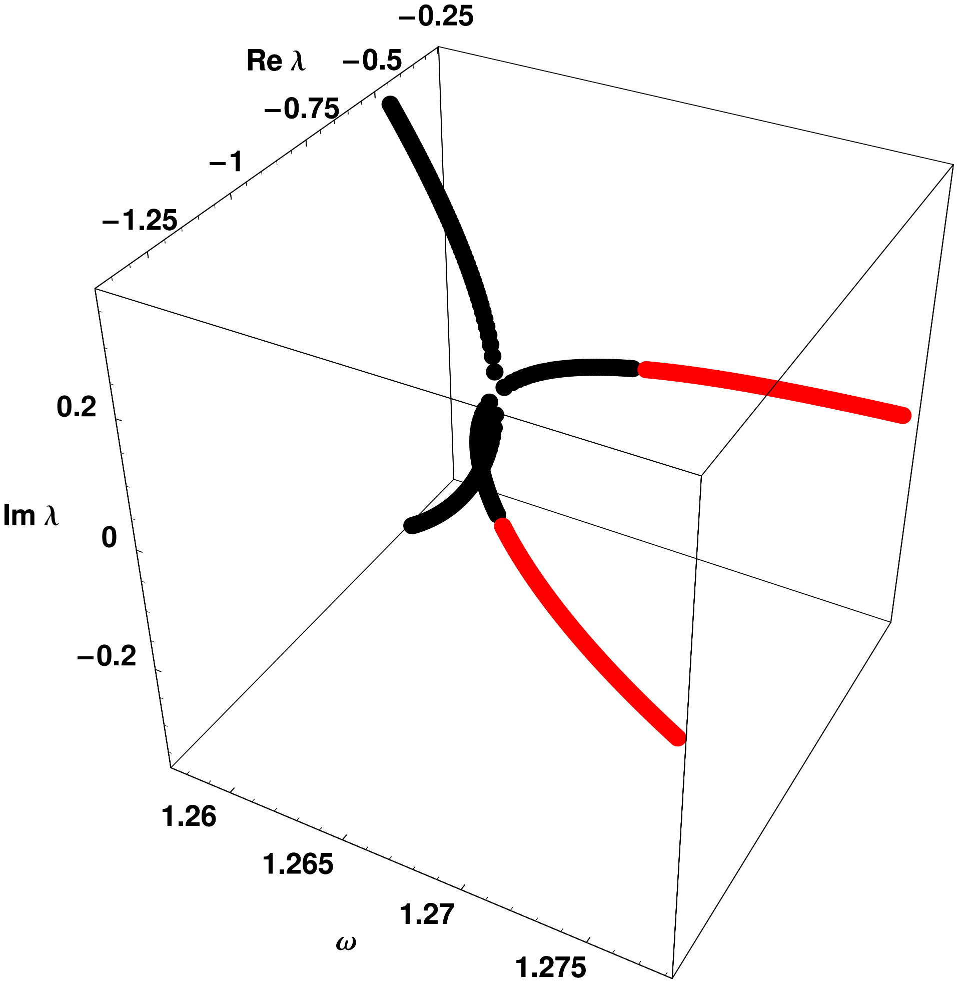


Figure 29.12.18: Different perspective of the eigenvalues of the linear part of $\mathcal{M}$ in the vicinity of its period-doubling bifurcation.

We should also examine the eigenvalues of the linear parts $\mathcal{M}^2$ about its fixed points. They are shown in Figures 12.19 and 12.20. The curves associated with z_f are shown in black before the period doubling bifurcation, and in red afterward. Observe, that as expected, that these eigenvalues are the squares of the eigenvalues of the linear parts of $\mathcal{M}$. The curves associated with $z_{\rm blue}$ and $z_{\rm green}$ are displayed in both blue and green because of their agreement. By continuity, the eigenvalues of the linear parts of $\mathcal{M}^2$ for the period-two fixed points $z_{\rm blue}$ and $z_{\rm green}$ are initially real and positive and, right after birth, have magnitude less than one. Thus $z_{\rm blue}$ and $z_{\rm green}$ are stable. However, as ω is further increased, their eigenvalues eventually become complex while still having magnitude less than one. Moreover, as ω is increased still further, these eigenvalues "circle" through the complex plane until they reach the negative real axis and then again become real. Finally, one of them reaches the value -1. At the corresponding ω value period doubling again occurs so that now $\mathcal{M}^2$ has two period-two fixed points and, correspondingly, $\mathcal{M}$ has four period-four fixed points. This is the mechanism by which continual period doubling can occur to produce a period doubling cascade.

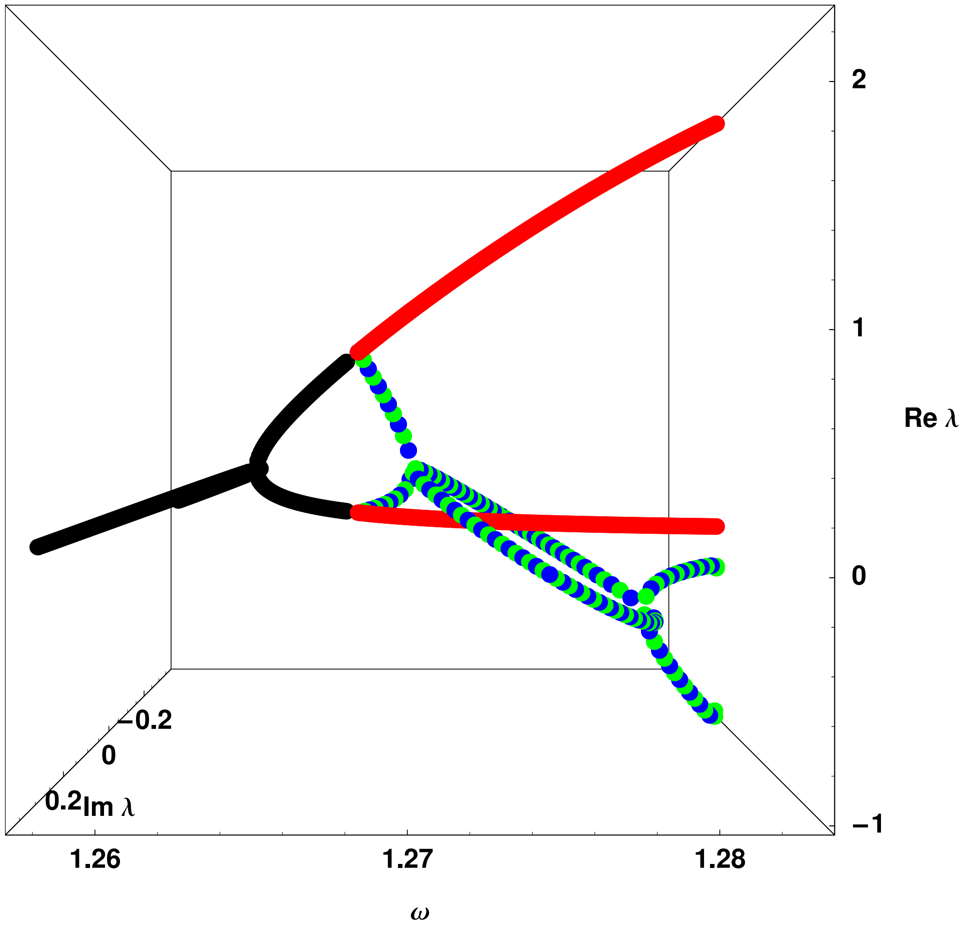


Figure 29.12.19: Eigenvalues of the linear part of $\mathcal{M}^2$ in the vicinity of the period doubling bifurcation of $\mathcal{M}$.

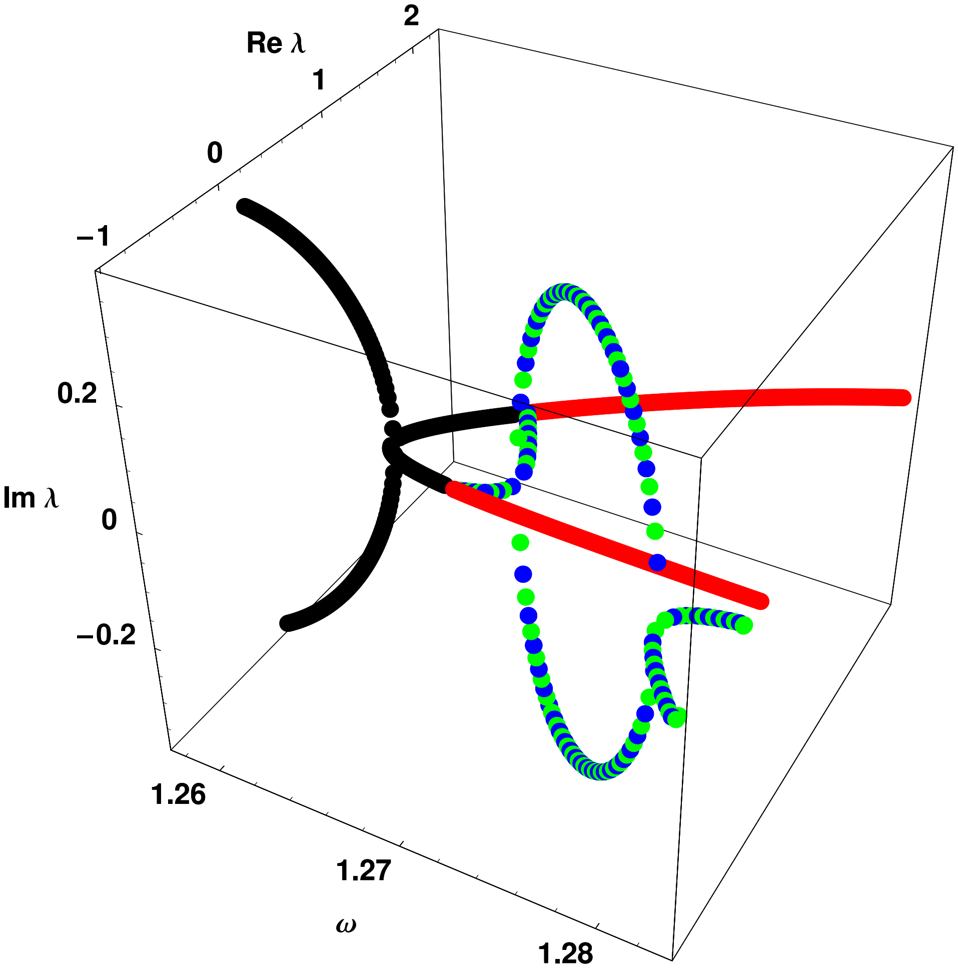


Figure 29.12.20: Different perspective of the eigenvalues of the linear part of $\mathcal{M}^2$ in the vicinity of the period doubling bifurcation of $\mathcal{M}$.

Verification of a Conjecture

Recall the conjecture, made in Exercise 1.2.3, that to find the leading behavior in the case of period doubling it sufficient to know the map through third order. To test this conjecture, let $\mathcal{M}_3$ be the third-order (including parameter dependence) truncated Taylor map expansion of the Dufing stroboscopic map $\mathcal{M}$ about the period-doubling bifurcation point shown in Figures 25.8.6 and 12.16. Figure 12.21 shows its bifurcation diagram. Evidently there is a close resemblance between Figures 12.16 and 12.21. As a further test, we can examine the eigenvalues of $(\mathcal{M}_3)^2$. They are shown in Figures 12.22 and 12.23. Comparison of Figures 12.22 and 12.23 with Figures 12.18 and 12.19 shows there is good quantitative agreement in the vicinity of the bifurcation point, and similar qualitative behavior farther away from the bifurcation point. Indeed, over the ω range displayed, the determinant of the linear part of $(\mathcal{M}_3)^2$ differs from the determinant of the linear part of $\mathcal{M}^2$ by at most .06. Moreover, the *second* period doubling (period quadrupling) for the map $\mathcal{M}_3$ occurs at $\omega = 1.28094\cdots$ while that for the exact map $\mathcal{M}$ occurs at $\omega = 1.28307\cdots$.

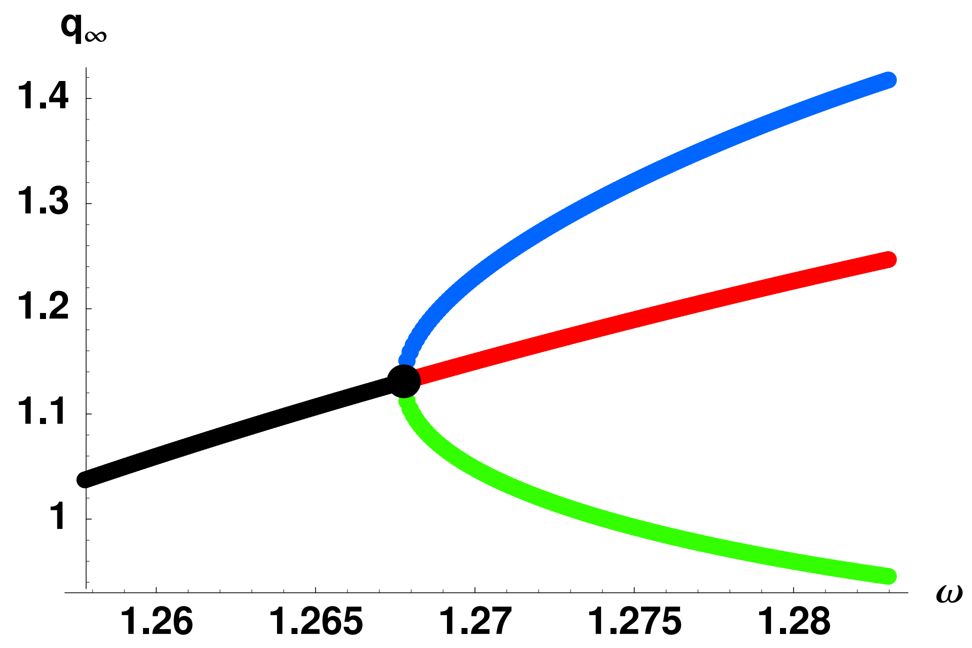


Figure 29.12.21: Bifurcation diagram for the map $\mathcal{M}_3$, the third-order polynomial approximation to $\mathcal{M}$ (including parameter dependence) expanded about the period-doubling bifurcation point shown in black. The polynomial map has one fixed point z_f before period doubling. It is stable and its trail is shown in black. After period doubling $\mathcal{M}_3$ still has one fixed point z_f. It is unstable and its trail is shown in red. These fixed points are, of course, also fixed points of $(\mathcal{M}_3)^2$. After period doubling, $(\mathcal{M}_3)^2$ has two additional fixed points whose trails are shown in blue and green. These period-two fixed points are not fixed points of $\mathcal{M}_3$. Instead, they are sent into each other under the action of $\mathcal{M}_3$.

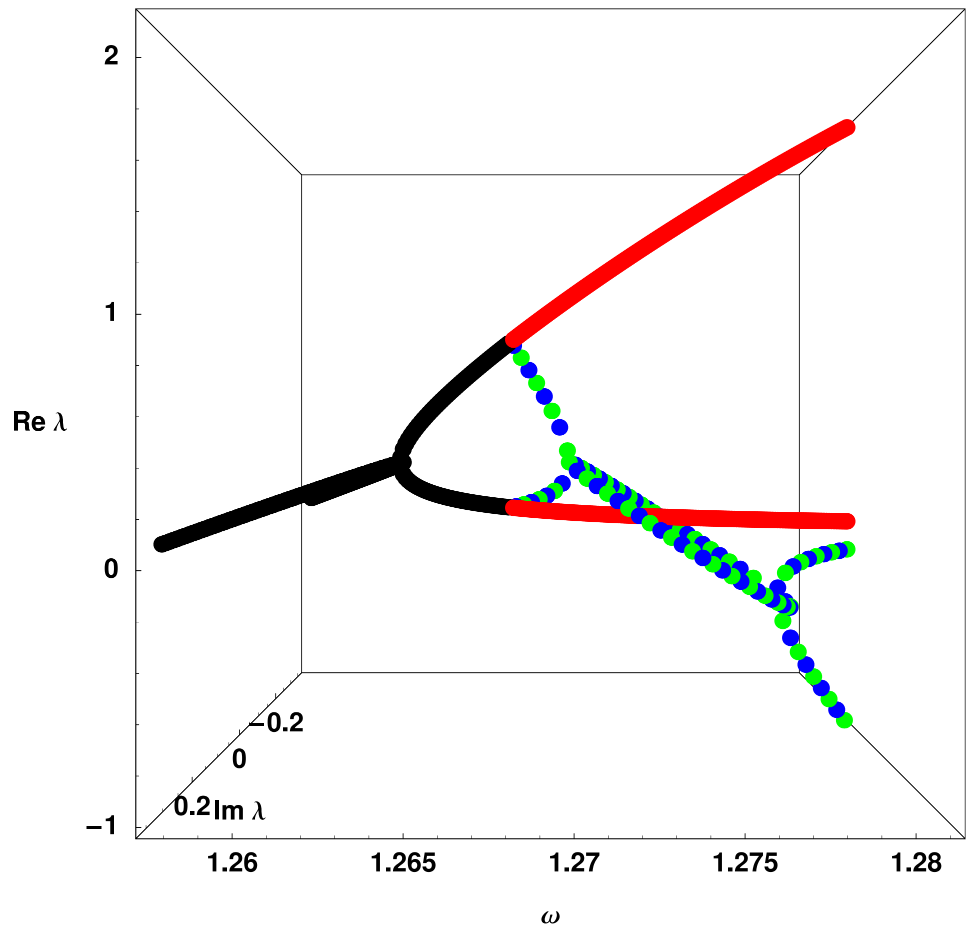


Figure 29.12.22: Eigenvalues of the linear part of $(\mathcal{M}_3)^2$ in the vicinity of the period doubling bifurcation of $\mathcal{M}_3$.

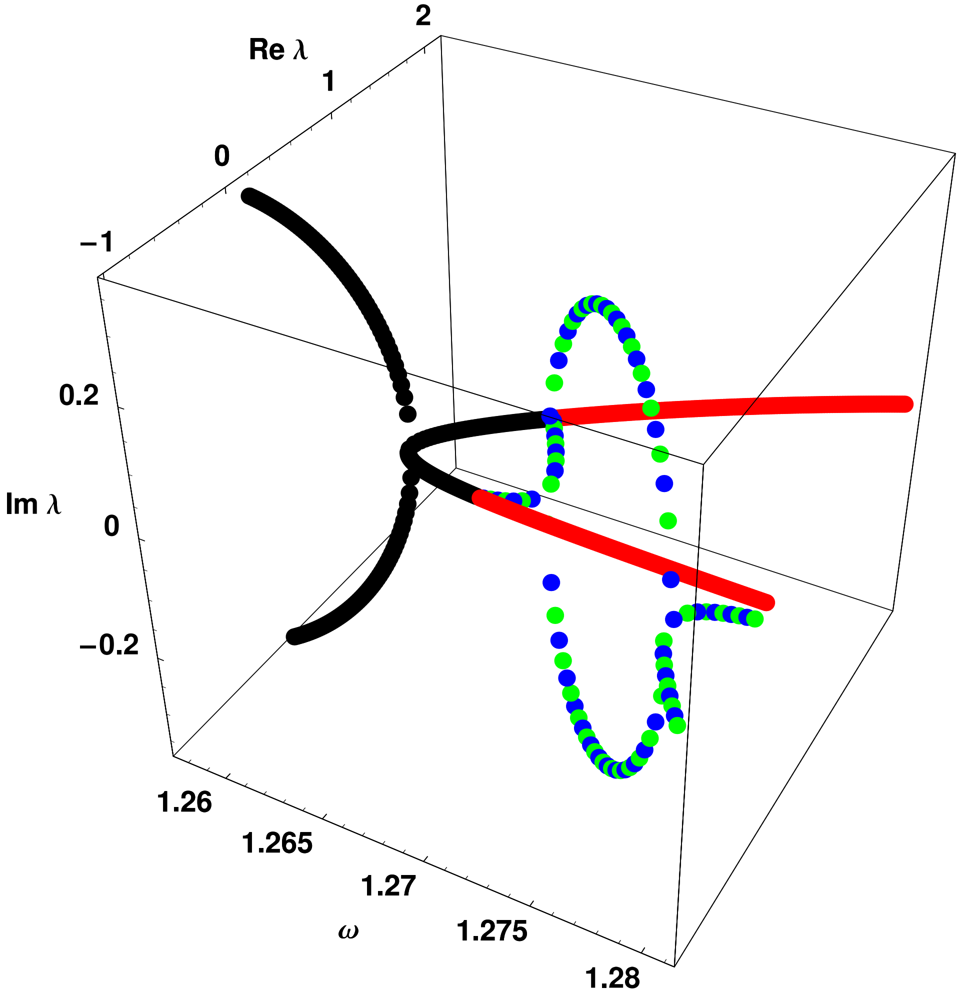


Figure 29.12.23: Different perspective of the eigenvalues of the linear part of $(\mathcal{M}_3)^2$ in the vicinity of the period doubling bifurcation of $\mathcal{M}_3$.

Infinite Period-Doubling Cascade

We now return to the main purpose of this subsection, namely to demonstrate that a single truncated Taylor map can reproduce a Duffing map infinite period-doubling cascade and associated strange attractor. In particular we will try to replicate, using a polynomial map with parameter dependence, the behavior illustrated in Figures 25.8.6, 25.9.1, and 25.9.2.

Figure 12.24 shows a partial Feigenbaum diagram for the map we will call $\mathcal{M}_8$. This is the 8^{th}-order polynomial approximation to the map $\mathcal{M}$. The black dot, situated at $\omega = 1.285\cdots$, is the fixed point z_f of $\mathcal{M}$ for this ω value, and it is used as the expansion point. It lies on a continuation of what is the red trail in Figure 12.21 and has the coordinates

$$z_f = (1.26082\cdots\ ;\ 2.05452\cdots) \text{ and } \omega = 1.285. \tag{29.12.12}$$

Figure 12.25 shows the associated full Feigenbaum diagram. Observe that the points on the full Feigenbaum diagram appear to be very nearly confined to a surface. Therefore, although we are dealing with a map in two dimensions, its behavior is very similar to a map in one dimension. Correspondingly, many aspects of Figure 12.24 are very similar to those of Figure 1.2.4, the Feigenbaum diagram for the logistic map, the simplest map in one dimension.

Comparison of Figures 25.8.6 and 12.24 reveals a striking resemblance.[4] Remarkably, an 8^{th}-order polynomial map approximation fully reproduces the complete period doubling cascade exhibited by the stroboscopic Duffing map. There is good quantitative agreement in the vicinity of the expansion point, and good qualitative agreement over the full cascade. Note that even some of the cascades associated with higher-period fixed points are captured.

[4] Similar color schemes are employed in both figures. They have no dynamical significance save to aid the eye in following successive bifurcations.

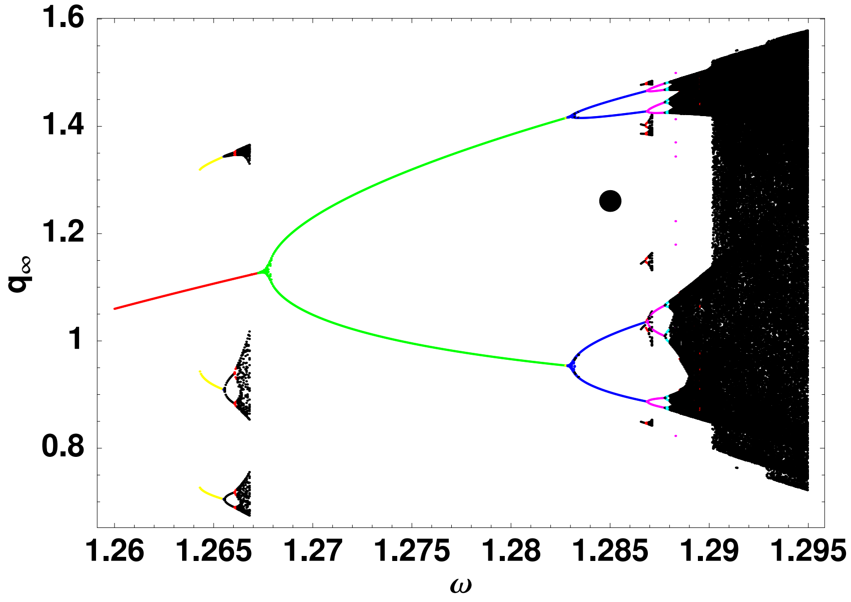


Figure 29.12.24: Partial Feigenbaum diagram for the map $\mathcal{M}_8$. The black dot marks the point about which $\mathcal{M}$ is expanded to yield $\mathcal{M}_8$

.

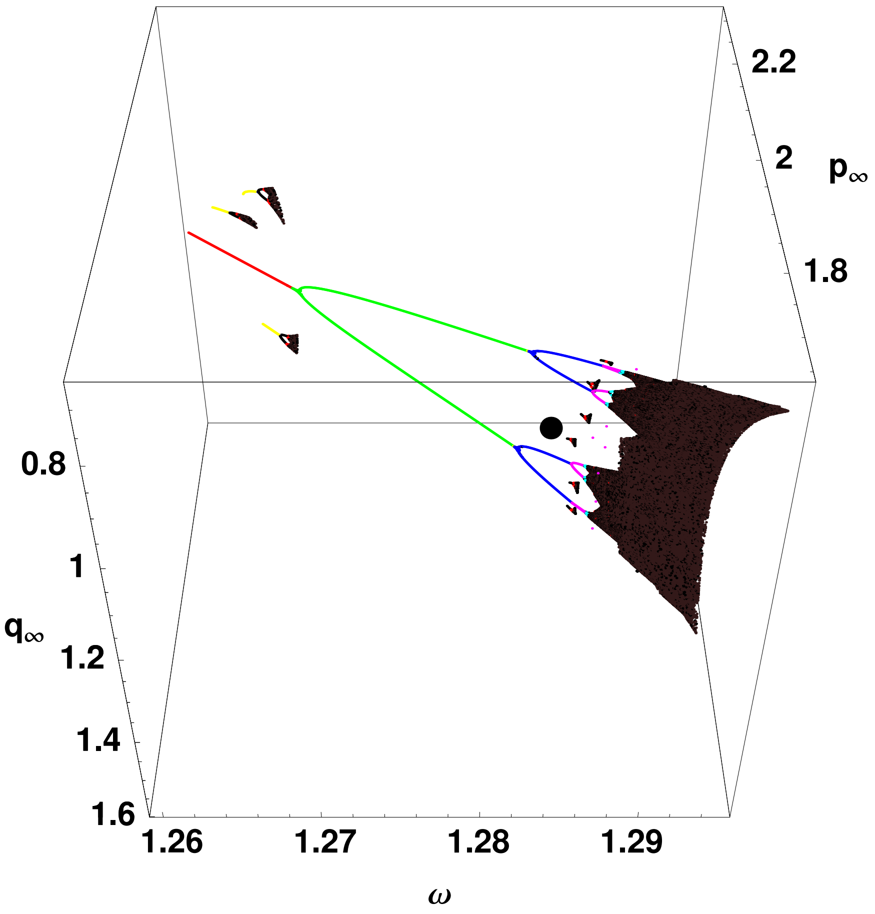


Figure 29.12.25: Full Feigenbaum diagram for the map $\mathcal{M}_8$. The black dot again marks the expansion point.

Strange Attractor

What about the strange attractor that is expected to appear after the completion of the infinite period doubling cascade? Figures 12.26 through 12.29 show the strange attractor for $\mathcal{M}_8$, and magnifications of selected portions, when $\omega = 1.2902$. Comparison of Figures 12.26 and 12.27 with Figures 25.9.1 and 25.9.2 illustrates that the strange attractor is remarkably well reproduced by the polynomial map. Moreover, Figures 12.28 and 12.29 show further successive magnifications, thereby illustrating the continued fractal structure. The exact stroboscopic Duffing map counterparts of these figures would be difficult to produce because of the extensive numerical integration required. However, these magnifications are readily obtained for the polynomial map since it is easily iterated.

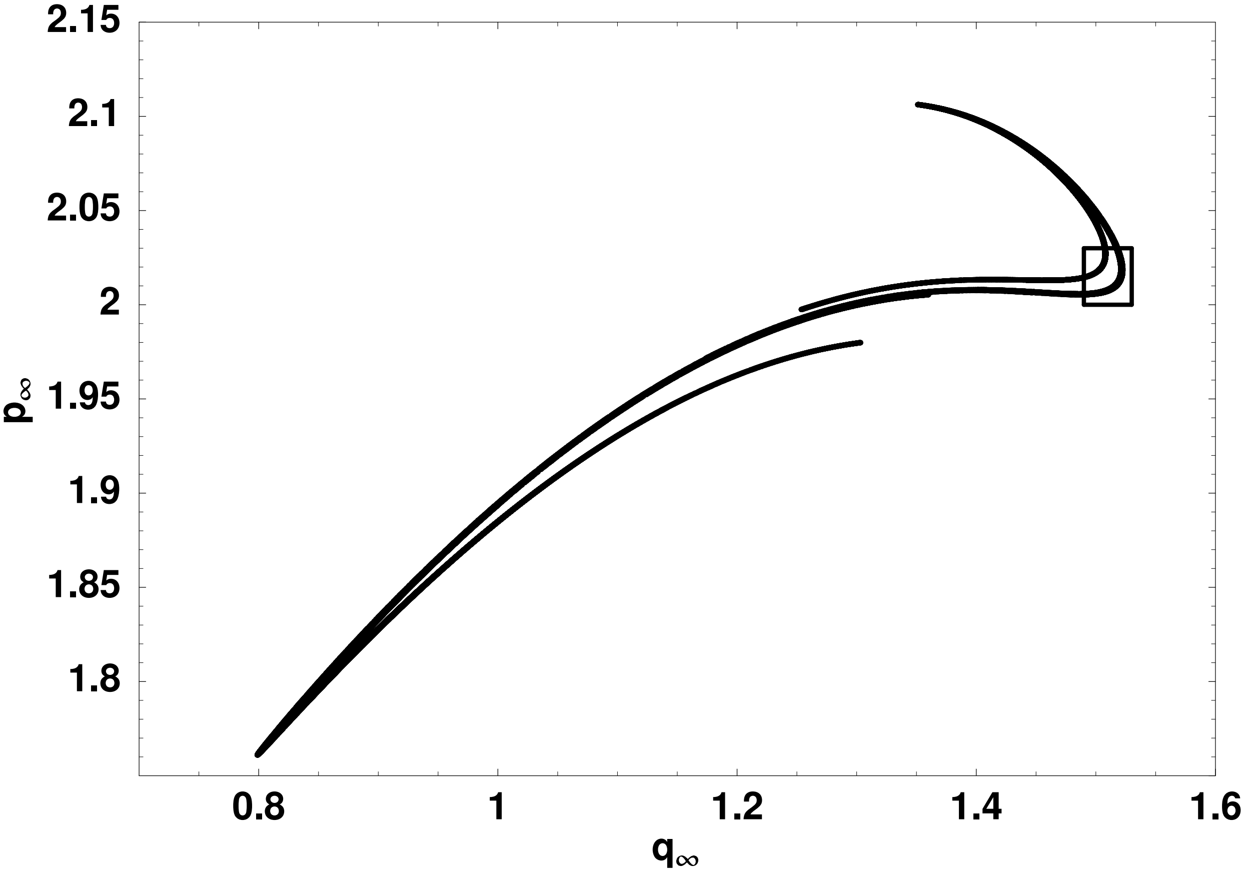


Figure 29.12.26: Limiting values of q_∞, p_∞ for the map $\mathcal{M}_8$ when $\omega = 1.2902$. They appear to lie on a strange attractor.

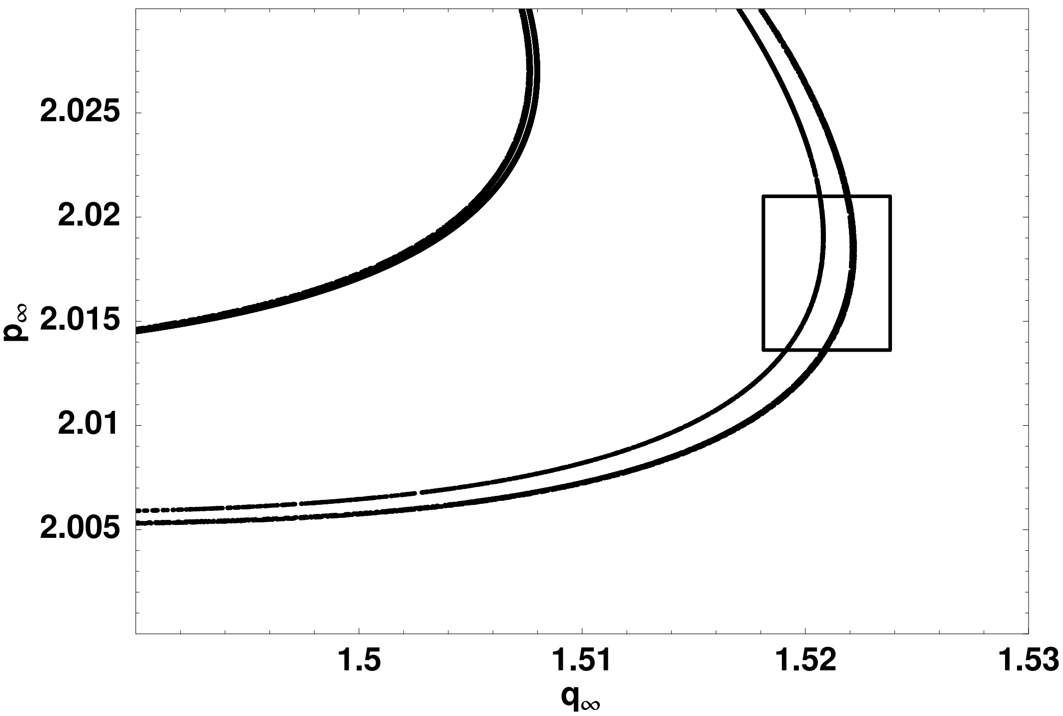


Figure 29.12.27: Enlargement of boxed portion of Figure 12.26 illustrating the beginning of self-similar fractal structure.

29.12.4 Undoing a Cascade by Successive Mergings

According to Figure 25.8.4, the period doubling cascade in Figure 25.8.6 undoes itself by successive mergings as ω becomes sufficiently large. Can this behavior also be reproduced, at least qualitatively, by a polynomial map? The answer is *yes* provided the expansion point is suitably chosen. Figure 12.30 shows the partial Feigenbaum diagram for $\mathcal{M}_8$ when the black dot, located at

$$z_f = (1.59406\cdots\ ;\ 0.565464\cdots), \tag{29.12.13}$$

is used as the expansion point. It is an unstable period-one fixed point when $\omega = 1.4$, and lies on the continuation to smaller ω values of the red fixed-point trail shown on the right side of the figure. Figure 12.31 shows the associated full Feigenbaum diagram. We remark that we were unable to reproduce the full cascade followed by successive merging if the expansion point was chosen to lie on the *left* side of the cascade. Presumably this is because the phase-space and ω excursions associated with the full cascade followed by successive mergings are quite large, and therefore the expansion point has to be sufficiently "centered" in order for the associated Taylor series to have an adequate domain of convergence.[5]

Note that within the q, p range displayed in Figures 12.30 and 12.31 there appears to be for $\mathcal{M}_8$ a gap around $\omega \approx 1.33$ for which there are no stable fixed points nor any attracting set. To the left and right of this gap there are chaotic regions, but there seems to be nothing

[5] In all the examples presented, the expansion point has been chosen to be a fixed point of $\mathcal{M}$ for some value of ω. This is not necessary, and was merely done so that the Taylor map would have the simplifying property of having no constant terms.

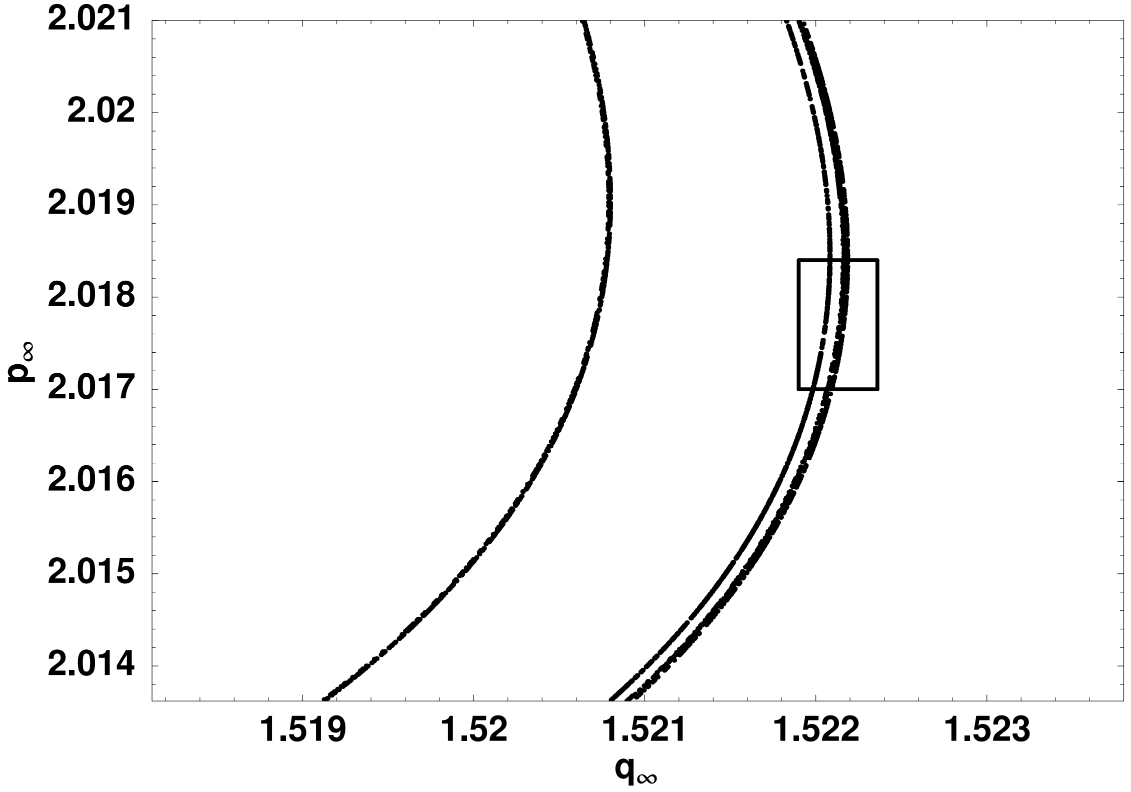


Figure 29.12.28: Enlargement of boxed portion of Figure 12.27 illustrating the continuation of self-similar fractal structure.

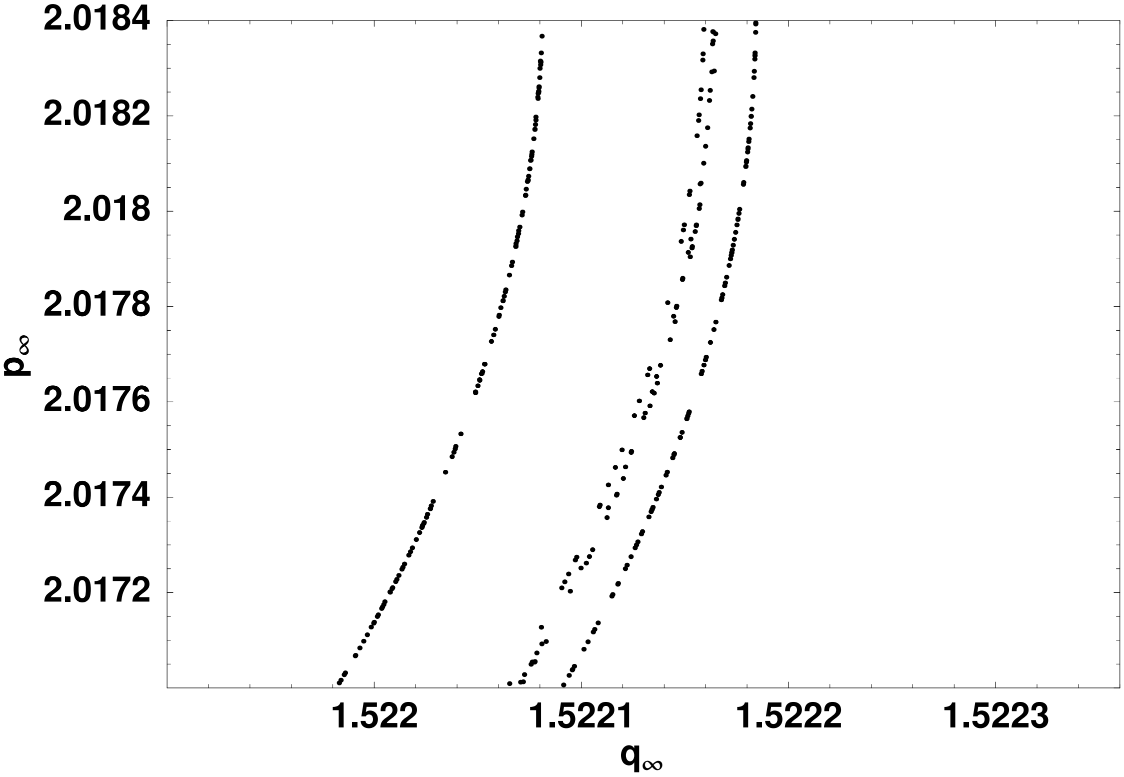


Figure 29.12.29: Enlargement of boxed portion of Figure 12.28 illustrating the further continuation of self-similar fractal structure.

in the gap. Let us examine the behavior of the exact stroboscopic Duffing map $\mathcal{M}$ in this ω range. Figure 25.8.4 provides a partial Feigenbaum diagram and Figure 12.32 provides a full diagram for $\mathcal{M}$.

For ω within the gap and sufficiently large ($\omega \geq 1.335$), Figure 12.32 displays that there are three yellow trails. They correspond to a single stable period-three orbit. That is, there are three stable fixed points of $\mathcal{M}^3$, and these three points are cyclically permuted among themselves under the action of $\mathcal{M}$. The behavior is completely analogous to that of the period-three orbit for the complex logistic map Douady rabbit described by the relations (1.2.33) through (1.2.38). See Section 1.2.2.

Now suppose ω is decreased. Then it happens that *each* of these fixed points of $\mathcal{M}^3$ undergoes a *pitchfork* bifurcation to produce a *triplet* of fixed points of $\mathcal{M}^3$. Thus, there are now three triplets of fixed points of $\mathcal{M}^3$. (Recall that the period does not change at either a saddle-node or pitchfork bifurcaton.) In accord with the pattern for a pitchfork bifurcation, within each triplet two of the fixed points of $\mathcal{M}^3$ are stable, and one is unstable. Consequently, only two points of each triplet will be visible in a Feigenbaum diagram, and there will appear to be three *pairs* of fixed points of $\mathcal{M}^3$ for ω in this range. Correspondingly, there will be *two* period-three stable orbits of $\mathcal{M}$ in this ω range. That is also what Figure 12.32 displays. For ω within the gap and sufficiently large, there are three yellow trails. And, as ω is decreased, each yellow trail splits into two yellow trails so that there are then six yellow trails corresponding to two period-three stable orbits of $\mathcal{M}$. We conclude that the gap is essentially a period-three window.[6]

Finally, we observe that all the stable fixed points in this window have q_∞, p_∞ values that lie outside the q, p range displayed in Figures 12.30 and 12.31, and outside the range for which the Taylor map $\mathcal{M}_8$ well approximates the exact map $\mathcal{M}$. We conclude that $\mathcal{M}_8$ correctly describes the state of affairs for the q, p, ω range depicted in Figure 12.31.

[6] Closer inspection, on a scale finer than that shown in Figure 12.32, reveals that at $\omega = 1.315\cdots$ a *period doubling* occurs so that there are then twelve trails corresponding to two stable period-six orbits. At $\omega = 1.312\cdots$ there is a second period doubling so that now there are twenty-four trails corresponding to two stable period-twelve orbits. The net result is that there appears, on a very fine scale, to be an infinite cascade of period doublings that ultimately merges with the chaotic region to the left of the gap.

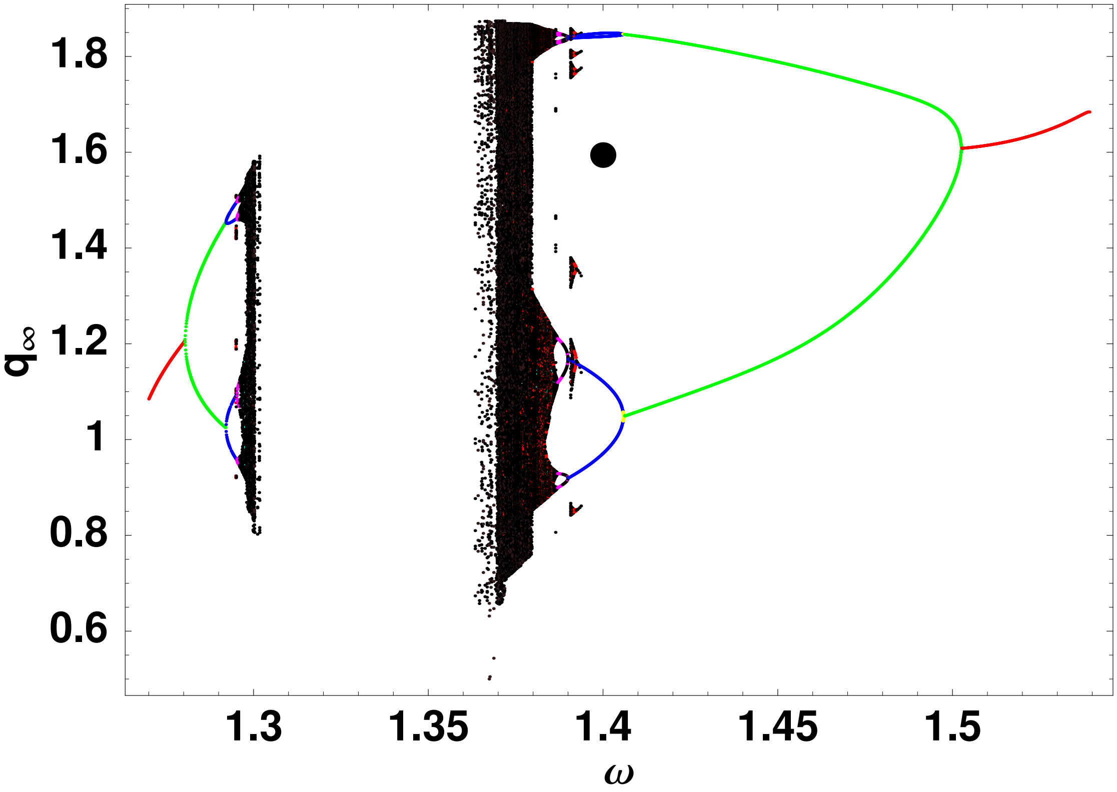


Figure 29.12.30: Partial Feigenbaum diagram for the map $\mathcal{M}_8$ showing a full cascade followed by successive mergings. The black dot marks the point about which $\mathcal{M}$ is now expanded to yield $\mathcal{M}_8$.

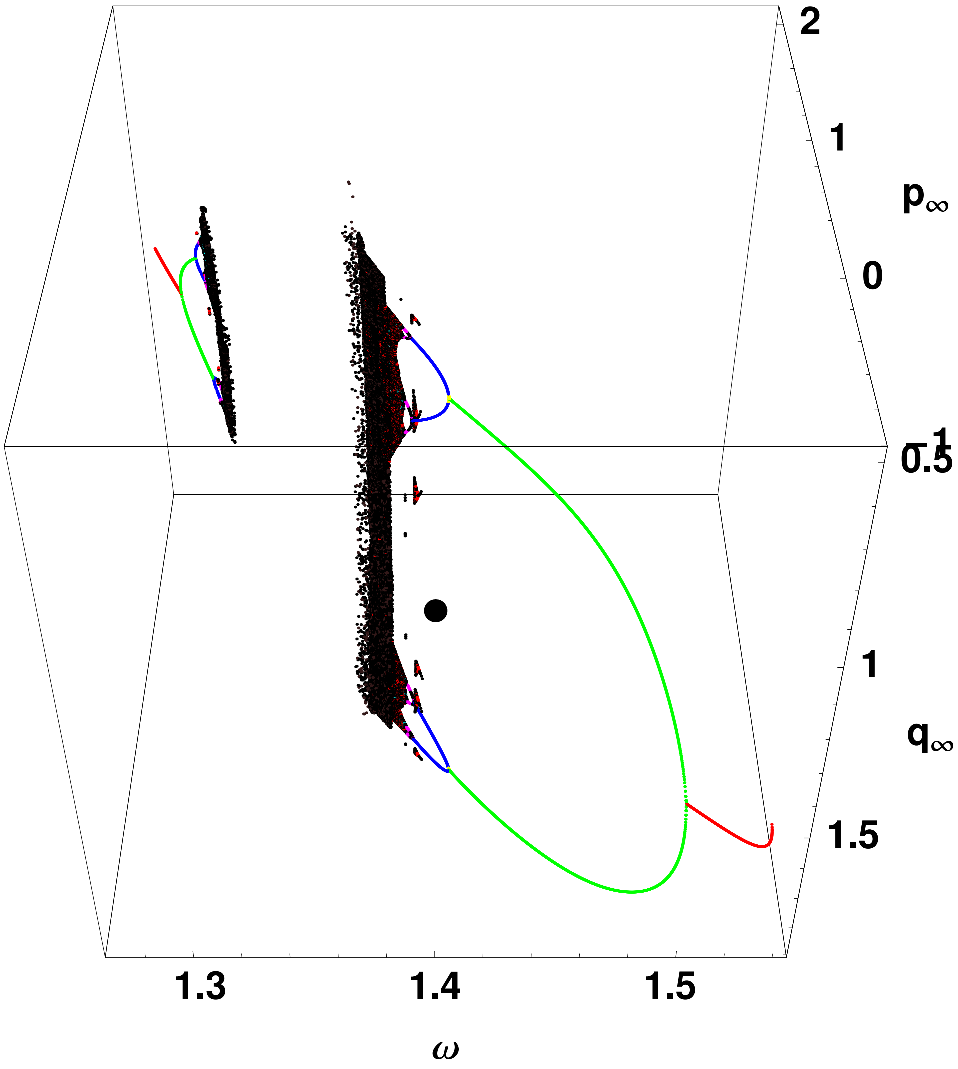


Figure 29.12.31: Full Feigenbaum diagram for the map $\mathcal{M}_8$ showing a full cascade followed by successive mergings. The black dot again marks the point about which $\mathcal{M}$ is expanded to yield $\mathcal{M}_8$.

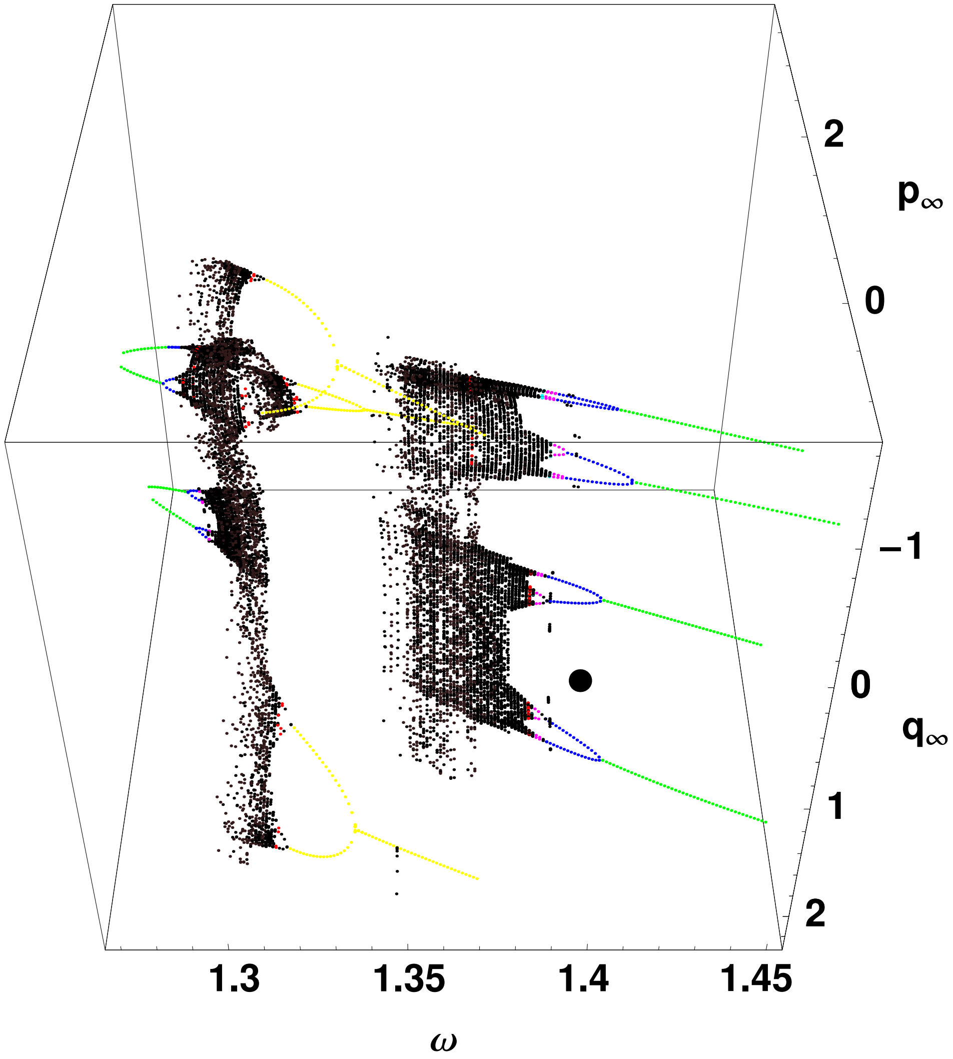


Figure 29.12.32: Full Feigenbaum diagram for the exact map $\mathcal{M}$. See Figure 25.8.4 for a related partial Feigenbaum diagram. The black dot again marks the expansion point used in Figures 12.30 and 12.31. There appears to be a gap around $\omega \approx 1.33$ separating two chaotic regions. Within the right side of the gap (to the right of $\omega \approx 1.335$) there are three yellow trails corresponding to a period-three stable orbit. As ω is decreased, there are pitchfork bifurcations so that each yellow trail splits into two yellow trails. There are then six yellow trails corresponding to two period-three stable orbits. Thus the gap, on the scale shown, is essentially a period-three window.

29.12.5 Convergence of Taylor Maps: Performance of Lower-Order Polynomial Approximations

We close this section with illustrations of the performances of $\mathcal{M}_3$ and $\mathcal{M}_5$, third and fifth-order polynomial approximations (including parameter dependence) to the exact map $\mathcal{M}$. All expansions are made about the point (12.12). Comparison of these performances gives some feeling for the convergence properties of the Taylor approximation to $\mathcal{M}$.

Performance of $\mathcal{M}_3$

Figure 12.33 shows the $\mathcal{M}_3$ counterpart to Figure 12.24 produced using $\mathcal{M}_8$. Evidently the qualitative features of the period doubling cascade are the same. Also, we have found that there is not qualitative agreement if $\mathcal{M}_2$ is used. We conjecture that generically third-order information is necessary and sufficient to obtain qualitative agreement for a period doubling cascade arising from what once was a period-one fixed point.

Note also that $\mathcal{M}_3$ does not reproduce the three features near $\omega = 1.265$ seen in Figure 25.8.6 for the exact $\mathcal{M}$ and in Figure 12.24 for $\mathcal{M}_8$. We have found that these features first appear for $\mathcal{M}_n$ when $n = 5$. They belong to what was initially a period-three fixed point for $\mathcal{M}$.

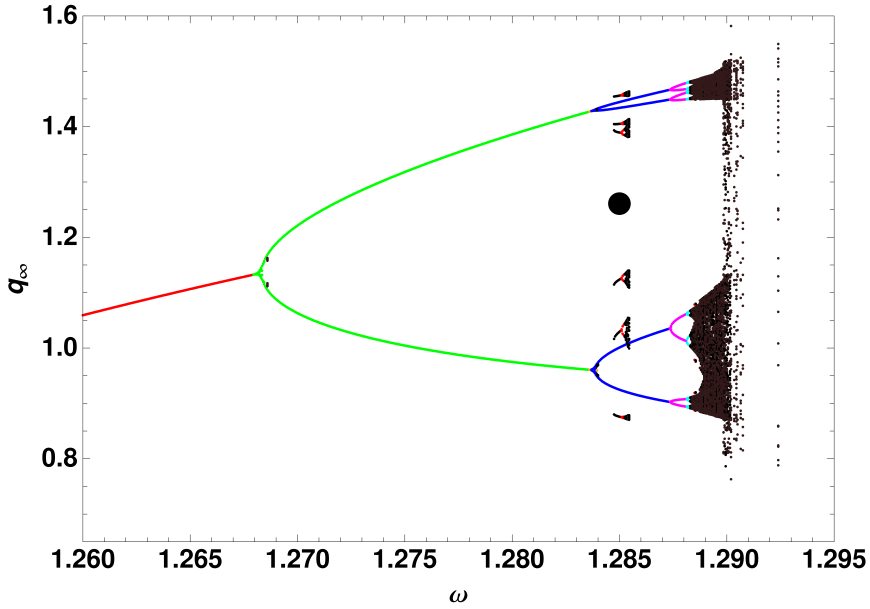


Figure 29.12.33: Partial Feigenbaum diagram for the map $\mathcal{M}_3$. The black dot marks the point about which $\mathcal{M}$ is expanded to yield $\mathcal{M}_3$

.

Figures 12.34 and 12.35 show the $\mathcal{M}_3$ counterparts to Figures 12.26 and 12.27 produced using $\mathcal{M}_8$. Evidently there is qualitative agreement. The attractors in Figures 12.34 and 12.26 look similar. And, when enlarged, both show evidence of fractal structure. Compare Figures 12.35 and 12.27.

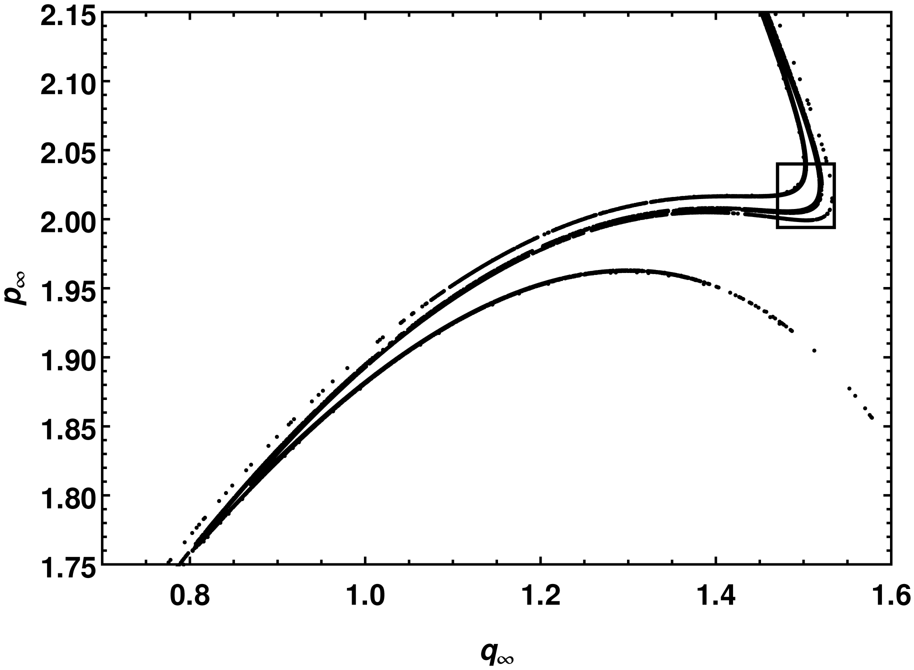


Figure 29.12.34: Limiting values of q_∞, p_∞ for the map $\mathcal{M}_3$ when $\omega = 1.2902$. They appear to lie on a strange attractor.

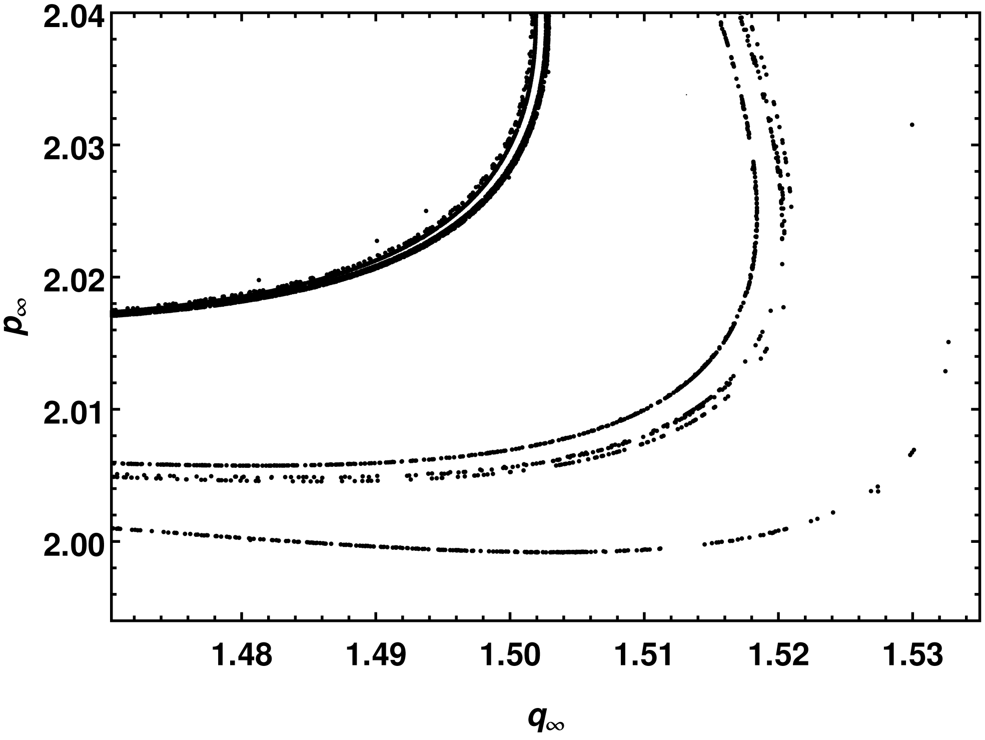


Figure 29.12.35: Enlargement of boxed portion of Figure 12.34 illustrating the beginning of self-similar fractal structure.

Performance of $\mathcal{M}_5$

Figure 12.36 shows the $\mathcal{M}_5$ counterpart to Figure 12.24 produced using $\mathcal{M}_8$. Now there is improved quantitative agreement as well as qualitative agreement. Also, there are now three features near $\omega = 1.265$ that resemble those seen in Figures 25.8.6 and 12.24.

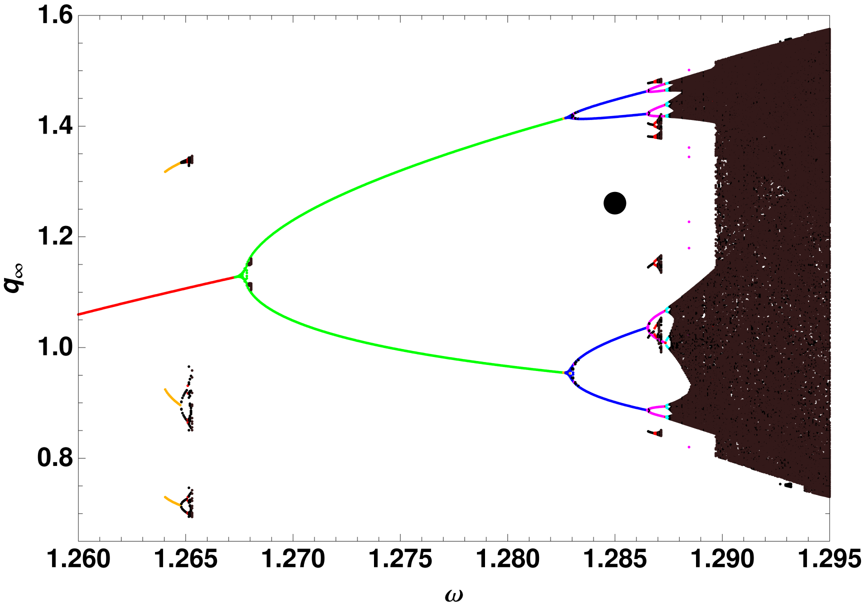


Figure 29.12.36: Partial Feigenbaum diagram for the map $\mathcal{M}_5$. The black dot marks the point about which $\mathcal{M}$ is expanded to yield $\mathcal{M}_5$

.

Figures 12.37 and 12.38 show the $\mathcal{M}_5$ counterparts to Figures 12.26 and 12.27 produced using $\mathcal{M}_8$. Again there is improved quantitative agreement. We surmise that, for the region of phase space and ω range displayed, convergence appears to be well underway.

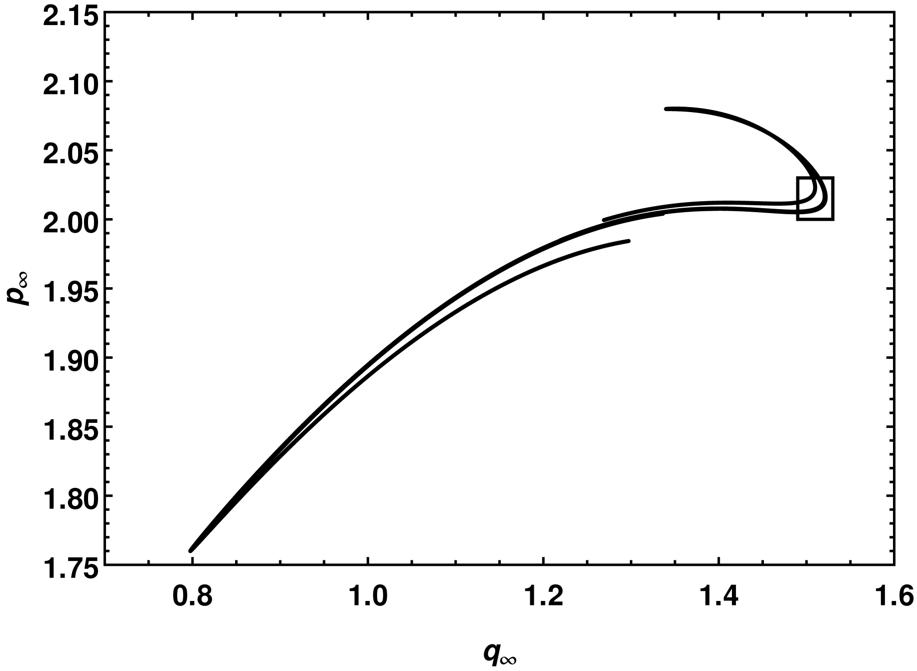


Figure 29.12.37: Limiting values of q_∞, p_∞ for the map $\mathcal{M}_5$ when $\omega = 1.2902$. They appear to lie on a strange attractor.

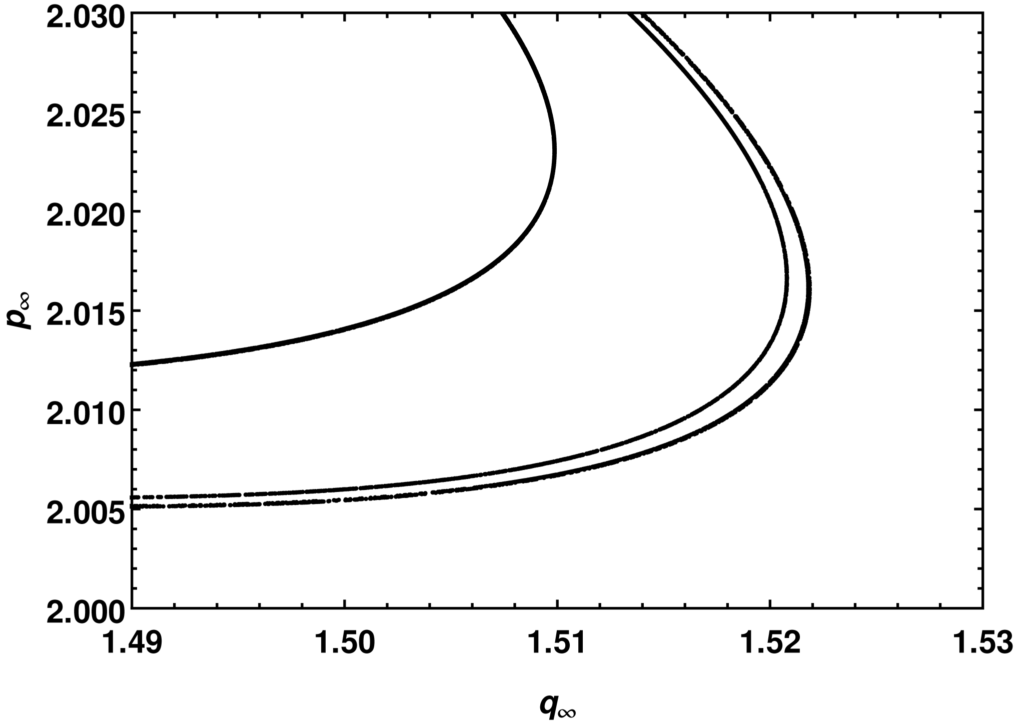


Figure 29.12.38: Enlargement of boxed portion of Figure 12.37 illustrating the beginning of self-similar fractal structure.

29.12.6 Concluding Summary and Discussion

Poincaré analyticity (and its generalization to include parameter dependence) implies that transfer maps $\mathcal{M}$ arising from ordinary differential equations can be expanded as Taylor series in the initial conditions and also in whatever parameters may be present. Section 10.10 described the complete variational equations, and described how the determination of these expansions is equivalent to solving the complete variational equations. Chapter 25 provided an overview of the properties of the stroboscopic transfer map $\mathcal{M}$ for the Duffing equation. The present section described examples of how n^{th} degree approximations $\mathcal{M}_n$ to $\mathcal{M}$ (including parameter dependence) could reproduce various features of the exact $\mathcal{M}$. In particular it illustrated, remarkably, that $\mathcal{M}_8$ produced an infinite period doubling cascade and apparent strange attractor that closely resembled those of the exact map. It also illustrated how the accuracy of $\mathcal{M}_n$ improves with increasing n.

We have seen that there are situations in which a truncated Taylor map well reproduces results obtained by the integration of differential equations. This is comforting since the behavior of polynomial maps, because such maps can easily be evaluated repeatedly, is often studied in detail with the hope that the behavior of such maps is illustrative of what can be expected for maps in general, including the maps that arise from integrating differential equations.

In view of this success, one might wonder if there are situations in which the use of truncated Taylor maps could replace or at least complement direct numerical integration. There is, of course, the question of convergence for Taylor series, and the convergence domain is related to the (generally unknown) singularity structure of the solution to the differential equation in the complex domain. See Section 35.3. However, if satisfactory approximation can be illustrated by the comparison of numerical integration results with truncated Taylor results for representative solutions in some domain, then the use of truncated Taylor maps to find additional results may be faster than continued numerical integration.

For example, in the case of the Duffing equation, although the determination of the relevant $h_a^r(t)$ of Section 10.10 requires the simultaneous numerical integration of a large number of differential equations, these equations need be integrated over only one drive period. Once the truncated Taylor series stroboscopic map has been found, its evaluation for any phase-space point and any parameter value is essentially free. All that is required is the evaluation of two n-degree polynomials (one for ζ_1^f and one for ζ_2^f, the deviation variables associated with q^f and p^f, respectively) in three variables (ζ_1^i, ζ_2^i, and ζ_3^i). (Again see Section 10.10 for notation.) By contrast, the direct construction of a Feigenbaum diagram requires the integration of the Duffing equation for a large number of drive periods and a large number of parameter values. And, determination of the strange attractor associated with the Duffing equation requires the integration of the Duffing equation over thousands of drive periods.

Suppose T_2 is the time required to integrate two equations over a drive period. In our example, it is the time required to integrate the Duffing pair of differential equations (1.4.32) over one drive period. Suppose T_{N_e} is the time required to integrate N_e equations over one drive period. Let $L(m, n)$ be the number of monomials of degree 0 through n in m variables.

It is given by the binomial coefficient

$$L(m,n) = \binom{m+n}{n}. \tag{29.12.14}$$

See Section 7.10. When working with m variables through terms of degree n, the number N_e of differential equations to be integrated to determine the relevant functions $h_a^r(t)$ is given by the relation

$$N_e = mL(m,n), \tag{29.12.15}$$

which amounts to

$$N_e = 3L(3,8) = 3 \times 165 = 495 \tag{29.12.16}$$

in the case of $\mathcal{M}_8$ for the Duffing equation including parameter dependence. We have found in our numerical studies that there is the approximate scaling relation

$$T_{N_e} \simeq (N_e/2)T_2 \tag{29.12.17}$$

for $n \leq 9$. That is, the computation time scales with the number of equations to be integrated. We conclude that in this example the use of $\mathcal{M}_8$ becomes advantageous once the number of drive periods times the number of parameter values exceeds $495/2 \simeq 250$.

With regard to providing complementary information, it is common practice to integrate the first degree variational equations in order to establish the *linear* stability of solutions. Integration of the higher degree variational equations, including possible parameter dependence, provides information about *nonlinear* behavior/stability. As examples, such information is required for the control of orbits in accelerators and the understanding and control of aberrations in optical systems.

In conclusion, there are applications for which use of the higher degree variational equations is advantageous, and the whole subject of the usefulness of truncated Taylor maps merits continued study.

29.12.7 Acknowledgment

Dobrin Kaltchev made major contributions to the work of this section.

29.13 Analytic Properties of Fixed Points and Eigenvalues

As described in the beginning of Section 12, integrating analytic differential equations can be expected to yield analytic maps. For these maps we can compute fixed points and the eigenvalues of the linear parts of these maps about their fixed points. What can be said about the parameter dependence of these fixed points and eigenvalues?

Consider first the behavior of eigenvalues. They are roots of the characteristic polynomial (3.4.1) when M is the linear part of the map. The coefficients of this polynomial depend on the matrix elements of M in an analytic way. See Exercise 3.7.14. Moreover, since M is determined by integrating the variational equations, we may expect these matrix elements to depend analytically on any parameter. Thus, we may expect that the coefficients of the characteristic polynomial will have analytic parameter dependence.

However, it does not follow that the roots will necessarily have analytic parameter dependence. Think, for example, of the roots of a quadratic equation. The solution of such an equation involves square roots of quantities formed from its coefficients. In such a case we expect that there could be branch points where the arguments of these square roots vanish. Generally we may expect continuity, but not analyticity.

Observe, for example, the behavior of the eigenvalues shown in Figures 12.2 and 12.3. For ω values sufficiently large the green eigenvalues have both real and imaginary parts. But as ω is decreased, there comes an ω value below which the green eigenvalues have only real parts. This means that the green eigenvalues are not analytic functions of ω at this threshold ω value. Similarly, for ω values sufficiently small, the blue eigenvalues have both real and imaginary parts. But as ω is increased, there comes an ω value above which the blue eigenvalues have only real parts. This means that the blue eigenvalues are not analytic functions of ω at this threshold ω value. See Exercise 13.1.

Consider next the case of fixed points. They too may be regarded as the roots of some equations with coefficients that have analytic parameter dependence. We may expect that the fixed points will depend analytically on the parameter as long as no eigenvalue of the linear part of the map about this fixed point has eigenvalue +1. See (4.37). However, as we have learned, fixed points are generally born or annihilated when an eigenvalue becomes equal to +1. Actually, for an analytic map, they are not created or destroyed, but only become invisible by becoming complex. Thus we may again expect the existence of branch points and an associated lack of analyticity at these parameter values. Moreover, if period doubling is to occur, we expect that the linear part of $\mathcal{M}^2$ will have +1 as an eigenvalue; so we also expect lack of analyticity for period-two fixed points as they are born. For a simple one-dimensional example, see Exercise 1.2.2. Finally, at parameter values for which the location of a fixed point fails to be analytic, we may also anticipate failure of analyticity for the associated eigenvalues of the linear part of the map about this fixed point.

Exercises

29.13.1. Suppose $f(z)$ is an analytic function of the complex variable $z = x + iy$. Let z_0 be a point on the real axis and suppose that $f(z)$ is real when z is real and $x < z_0$. Suppose also that f is analytic at z_0. Show that then $f(z)$ is also real for z real and $x > z_0$. Thus, if $f(z)$ is to have an imaginary part for z real and $x > z_0$, then f cannot be analytic at z_0.

29.13.2. Suppose S is a symmetric 2×2 matrix whose entries depend analytically on a parameter λ, and suppose these entries are real when λ is real. Show that in this case the eigenvalues of S are also analytic functions of λ.

Bibliography

Map Factorization

[1] S-N. Chow, C. Li, and D. Wang, *Normal Forms and Bifurcation of Planar Vector Fields*, Cambridge University Press (1994).

[2] S. Steinberg, "Factored product expansions of solutions of nonlinear differential equations", *SIAM Journal on Mathematical Analysis*, 15(1):108-115, January 1984.

[3] F. Takens, "Forced oscillations and bifurcations: Applications of global anaysis I", in *Comun. Math.*, vol. 3, Inst. Rijksuniv. Utrecht (1974).

[4] Exponential Form not Generic.

Maps and Bifurcation Theory

See also the Map and Universality references given at the end of Chapter 1.

[5] M. Hénon, "A Two-dimensional Mapping with a Strange Attractor", *Commmun. Math. Phys.* **50**, 69-77 (1976).

[6] D. Ruelle, "What Is a Strange Attractor?", *Notices of the American Mathematical Society* **53**, p. 764, August 2006.

[7] J.H. Curry,"On the Hénon transformation", *Commun. Math. Phy.*, **68**, 129-140 (1979).

[8] M. Benedicks and L. Carleson, "The dynamics of the Hénon map", *Ann. Math.*, **133**, 73-169, (1991).

[9] E. Zeraoulia and J. Sprott, *2-D Quadratic Maps and 3-D ODE Systems: A Rigorous Approach*, World Scientific (2010).

[10] J. Meiss, *Differential Dynamical Systems*, SIAM (2007).

[11] R.S. MacKay, *Renormalization in Area Preserving Maps*, Princeton University Ph.D. Thesis (1982).

[12] H.-O. Peitgen, H. Jürgens, and D. Saupe, *Chaos and Fractals: New Frontiers of Science*, p. 675, (Springer-Verlag 1992).

[13] E. Forest, *Beam Dynamics, A New Attitude and Framework*, p. 191, Harwood Publishers (1998).

[14] J.M. Greene, R.S. MacKay, F. Vivaldi, and M.J. Feigenbaum, “Universal Behaviour of Area-Preserving Maps”, *Physica* **3D**, 468 (1981).

[15] T.C. Bountis, “Period-Doubling Bifurcations and Universality in Conservative Systems”, *Physica* **3D**, 577 (1981).

[16] V.I. Arnold, Ed., *Dynamical Systems V, Bifurcation Theory and Catastrophe Theory*, Springer-Verlag (1994).

[17] V.I. Arnold, Ed., *Dynamical Systems VI, Singularity Theory I*, Springer-Verlag (1993).

[18] V.I. Arnold, *Catastrophe Theory*, Springer-Verlag (1986).

[19] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcatons of Vector Fields*, Springer-Verlag (1983).

[20] R. Rand and D. Armbruster, *Perturbation Methods, Bifurcation Theory and Computer Algebra*, Springer-Verlag (1987).

[21] D. Ruelle., *Elements of Differentiable Dynamics and Bifurcation Theory*, Academic Press (1989).

[22] J. Murdock, *Normal Forms and Unfoldings for Local Dynamical Systems*, Springer-Verlag (2003).

[23] J. Marsden and T. Ratiu, *Introduction to Mechanics and Symmetry*, Second Edition, Springer-Verlag (1999).

[24] J. Hale and H. Kocak, *Dynamics and Bifurcations*, Springer-Verlag (1991).

[25] M. Kubicek and M. Marek, *Computational Methods in Bifurcation Theory and Dissipative Structures*, Springer-Verlag (1983).

[26] J. M. T. Thompson and H. B. Stewart, *Nonlinear Dynamics and Chaos*, Second Edition, John Wiley (2002).

[27] M. Hirsch, C. Pugh, and M. Shub *Invariant Manifolds*, Springer (1977).

[28] Z. Nitecki, *Differentiable Dynamics: An Introduction to the Orbit Structure of Diffeomorphisms*, MIT Press (1971).

[29] J. Palis and W. de Melo, *Geometric Theory of Dynamical Systems*, Springer (1982).

[30] S. Wiggins, *Introduction to Applied Nonlinear Dynamics and Chaos*, Springer (1990).

[31] S. Wiggins, *Global Bifurcations and Chaos: Analytical Methods*, Springer (1988).

[32] D. Schlomiuk, Edit., *Bifurcations and Periodic Orbits of Vector Fields*, Kluwer Academic Publishers (1993).

[33] T. Bridges and J. Furter, *Singularity Theory and Equivariant Symplectic Maps*, Springer-Verlag (1993).

[34] P. Duren, *Harmonic Mappings in the Plane*, Cambridge University Press (2004).

[35] P. Glendinning, *Stability, Instability, and Chaos: an introduction to the theory of nonlinear differential equations*, Cambridge University Press (2004).

[36] Bézout's Theorem

[37] Hartman's Theorem

[38] Smale's Horseshoe

[39] Arnold Diffusion

Contraction Maps and Newton's Method

[40] M.S. Berger, *Nonlinearity and Functional Analysis, Lectures on Nonlinear Problems in Mathematical Analysis*, Academic Press (1977).

[41] L. Collatz, *Functional Analysis and Numerical Mathematics*, Academic Press (1966).

[42] R. Courant and F. John, *Introduction to Calculus and Analysis I*, Springer (1989).

[43] A.N. Kolmogorov and S.V. Fomin, *Elements of the Theory of Functions and Functional Analysis*, Graylock Press (1957).

[44] W. Rudin, *Principles of Mathematical Analysis*, Third Edition, McGraw-Hill (1976).

Taylor Maps from Augmented Complete Variational Equations

[45] D. Kaltchev and A. Dragt, "Poincaré Analyticity and the Complete Variational Equations", *Physica D: Nonlinear Phenomena* **242** (2013). See also

http://arxiv.org/abs/1102.3394.

Non-Integrability of Hamiltonian Systems

[46] A. Dragt and J. Finn, "Insolubility of Trapped Particle Motion in a Magnetic Dipole Field", *Journal of Geophysical Research* **81**, 2327 (1976).

[47] related papers

[48] J. Morales Ruiz, *Differential Galois Theory and Non-Integrability of Hamiltonian Systems*, Birkhäuser (1999).

[49] M. Audin, *Hamiltonian Systems and Their Integrability*, American Mathematical Society and Société Mathématique de France (2001).

Chapter 30

Normal Forms for Symplectic Maps and Their Applications

30.1 Equivalence Relations

Def. 1.1: Let X be some (possibly abstract) set, and let $\sim$ be some relation (something that can be true or false) among pairs of elements in X. The relation $\sim$ is said to be an *equivalence* relation if it satisfies three properties:

i. $x \sim x$ for all x in X (reflexive property).

ii. $x_1 \sim x_2$ implies $x_2 \sim x_1$ for all x_1, x_2 in X (symmetric property) .

iii. $x_1 \sim x_2$ and $x_2 \sim x_3$ implies $x_1 \sim x_3$ for all x_1, x_2, x_3 in X (transitive property).

Def. 1.2: The set of all elements in X that are equivalent (under some given equivalence relation $\sim$) to a given x in X is called the *equivalence class* of x, and is denoted by the symbol $\{x\}$.

Thrm. 1.1: We have the logical relation

$$x_1 \sim x_2 \Leftrightarrow \{x_1\} = \{x_2\}. \tag{30.1.1}$$

Thrm. 1.2: Given an equivalence relation $\sim$ on some set X, show that each x in X belongs to one and only one equivalence class. Thus, under an equivalence relation, a set decomposes in a natural way into disjoint subsets: the equivalence classes produced by the equivalence relation.

Def. 1.3: Let X be some set and x some element in X. Then, given an equivalence relation $\sim$, x belongs to the equivalence class $\{x\}$. A *normal form* x_n for x is an element of $\{x\}$ that has some desired attribute such as “simplicity”. See Figure 1.1.

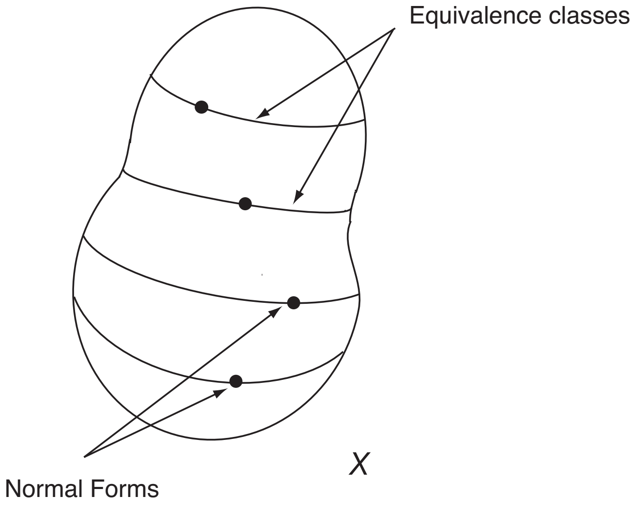


Figure 30.1.1: Decomposition of a set X into disjoint equivalence classes, with a normal form element representative for each equivalence class.

30.2 Symplectic Conjugacy of Symplectic Maps

Def. 2.1: Suppose $\mathcal{M}_1$ and $\mathcal{M}_2$ are two symplectic maps. These maps are said to be (symplectically) *conjugate* if there exists a third (symplectic) map $\mathcal{A}$ such that

$$\mathcal{M}_2 = \mathcal{A}\mathcal{M}_1\mathcal{A}^{-1}. \tag{30.2.1}$$

Def. 2.2: The map $\mathcal{A}$ is called the *conjugating* map.

Thrm. 2.1: Conjugacy and symplectic conjugacy are equivalence relations and therefore determine equivalence classes called *conjugacy classes.* Two maps $\mathcal{M}_1$ and $\mathcal{M}_2$ are equivalent (belong to the same conjugacy class) if a conjugating map $\mathcal{A}$ can be found such that (2.1) holds.

30.3 Normal Forms for Maps

Def. 3.1: A map *normal form* is a representative of an equivalence class (in this case a conjugacy class) selected for its maximal simplicity: Given any map $\mathcal{M}_1$, consider maps $\mathcal{N}_1$ of the form

$$\mathcal{N}_1 = \mathcal{A}_1\mathcal{M}_1\mathcal{A}_1^{-1}. \tag{30.3.1}$$

Select the map $\mathcal{A}_1$, and thereby also the map $\mathcal{N}_1$, in such a way that $\mathcal{N}_1$ is as *simple* as possible. The map $\mathcal{N}_1$ is called the normal form of $\mathcal{M}_1$, and the conjugating map $\mathcal{A}_1$ is called the *normalizing* map. Note that, by construction, we have the relations

$$\mathcal{N}_1 \sim \mathcal{M}_1 \text{ and } \{\mathcal{N}_1\} = \{\mathcal{M}_1\} \tag{30.3.2}$$

so that $\mathcal{N}_1$ is indeed a representative of the conjugacy class of $\mathcal{M}_1$

Thrm. 3.1: Given suitable specifications concerning the set of allowed conjugating maps, there is a *unique* normal form element for each conjugacy class. In other words, the normal form is unique in the sense that if $\mathcal{M}_2$ and $\mathcal{M}_1$ belong to the same conjugacy class,

$$\mathcal{M}_2 \sim \mathcal{M}_1 \text{ and } \{\mathcal{M}_2\} = \{\mathcal{M}_1\}, \tag{30.3.3}$$

then they have the same normal form,

$$\mathcal{N}_2 = \mathcal{N}_1. \tag{30.3.4}$$

Conversely, if two maps $\mathcal{M}_2$ and $\mathcal{M}_1$ have the same normal form, i.e. (3.4) holds, then they are conjugate and (3.4) holds. Thus, we have the logical relation

$$\mathcal{N}_2 = \mathcal{N}_1 \Leftrightarrow \mathcal{M}_2 \sim \mathcal{M}_1 \text{ and } \{\mathcal{M}_2\} = \{\mathcal{M}_1\}. \tag{30.3.5}$$

Proof: Stating that $\mathcal{M}_1$ and $\mathcal{M}_2$ have the normal forms $\mathcal{N}_1$ and $\mathcal{N}_2$ means that there exist normalizing maps $\mathcal{A}_1$ and $\mathcal{A}_2$ such that

$$\mathcal{N}_1 = \mathcal{A}_1\mathcal{M}_1\mathcal{A}_1^{-1}, \tag{30.3.6}$$

$$\mathcal{N}_2 = \mathcal{A}_2\mathcal{M}_2\mathcal{A}_2^{-1}, \tag{30.3.7}$$

and both $\mathcal{N}_1$ and $\mathcal{N}_2$ have maximal simplicity. Now suppose that the left equality in (3.5) holds. This supposition, when combined with (3.6) and (3.7), gives the relation

$$\mathcal{A}_1\mathcal{M}_1\mathcal{A}_1^{-1} = \mathcal{A}_2\mathcal{M}_2\mathcal{A}_2^{-1}, \tag{30.3.8}$$

which can be rewritten in the form

$$\mathcal{M}_1 = (\mathcal{A}_1^{-1}\mathcal{A}_2)\mathcal{M}_2(\mathcal{A}_1^{-1}\mathcal{A}_2)^{-1}. \tag{30.3.9}$$

We conclude that $\mathcal{M}_2$ and $\mathcal{M}_1$ are conjugate,

$$\mathcal{M}_2 \sim \mathcal{M}_1 \text{ and } \{\mathcal{M}_2\} = \{\mathcal{M}_1\}. \tag{30.3.10}$$

Conversely, suppose that $\mathcal{M}_2$ and $\mathcal{M}_1$ are in the same conjugacy class. Then there exists a conjugating map $\mathcal{A}$ such that (2.1) holds. From the relations (2.1), (3.6), and (3.7) we deduce the results

$$\mathcal{N}_1 = \mathcal{A}_1\mathcal{M}_1\mathcal{A}_1^{-1} = \mathcal{A}_1\mathcal{A}^{-1}\mathcal{A}\mathcal{M}_1\mathcal{A}^{-1}\mathcal{A}\mathcal{A}_1^{-1} = (\mathcal{A}_1\mathcal{A}^{-1})\mathcal{M}_2(\mathcal{A}_1\mathcal{A}^{-1})^{-1}, \tag{30.3.11}$$

$$\mathcal{N}_2 = \mathcal{A}_2\mathcal{M}_2\mathcal{A}_2^{-1} = \mathcal{A}_2\mathcal{A}\mathcal{A}^{-1}\mathcal{M}_2\mathcal{A}\mathcal{A}^{-1}\mathcal{A}_2^{-1} = (\mathcal{A}_2\mathcal{A})\mathcal{M}_1(\mathcal{A}_2\mathcal{A})^{-1}. \tag{30.3.12}$$

We see that $\mathcal{M}_2$ has the normal form $\mathcal{N}_1$ when $(\mathcal{A}_1\mathcal{A}^{-1})$ is used as a normalizing map, and $\mathcal{M}_1$ has the normal form $\mathcal{N}_2$ when $(\mathcal{A}_2\mathcal{A})$ is used as a normalizing map. Next, conjugate both sides of (3.6) with the map $(\mathcal{A}_2\mathcal{A}\mathcal{A}_1^{-1})$. Doing so gives the result

$$\begin{aligned}(\mathcal{A}_2\mathcal{A}\mathcal{A}_1^{-1})\mathcal{N}_1(\mathcal{A}_2\mathcal{A}\mathcal{A}_1^{-1})^{-1} &= (\mathcal{A}_2\mathcal{A}\mathcal{A}_1^{-1})\mathcal{A}_1\mathcal{M}_1\mathcal{A}_1^{-1}(\mathcal{A}_2\mathcal{A}\mathcal{A}_1^{-1})^{-1}\\ &= \mathcal{A}_2\mathcal{A}\mathcal{M}_1\mathcal{A}^{-1}\mathcal{A}_2^{-1}\\ &= \mathcal{A}_2\mathcal{M}_2\mathcal{A}_2^{-1} = \mathcal{N}_2.\end{aligned} \tag{30.3.13}$$

Similarly, conjugating both sides of (3.7) with the map $(\mathcal{A}_1\mathcal{A}^{-1}\mathcal{A}_2^{-1})$ gives the result

$$\begin{aligned}(\mathcal{A}_1\mathcal{A}^{-1}\mathcal{A}_2^{-1})\mathcal{N}_2(\mathcal{A}_1\mathcal{A}^{-1}\mathcal{A}_2^{-1})^{-1} &= (\mathcal{A}_1\mathcal{A}^{-1}\mathcal{A}_2^{-1})\mathcal{A}_2\mathcal{M}_2\mathcal{A}_2^{-1}(\mathcal{A}_1\mathcal{A}^{-1}\mathcal{A}_2^{-1})^{-1}\\ &= \mathcal{A}_1\mathcal{A}^{-1}\mathcal{M}_2\mathcal{A}\mathcal{A}_1^{-1}\\ &= \mathcal{A}_1\mathcal{M}_1\mathcal{A}_1^{-1} = \mathcal{N}_1. \qquad (30.3.14)\end{aligned}$$

We have learned that $\mathcal{M}_2$, under the assumption that it is conjugate to $\mathcal{M}_1$, can be normalized to *both* the normal forms $\mathcal{N}_1$ and $\mathcal{N}_2$. Similarly, under the same assumption, $\mathcal{M}_1$ can be normalized to both the normal forms $\mathcal{N}_1$ and $\mathcal{N}_2$. By assumption, the $\mathcal{A}_1$ used in (3.6) and the $\mathcal{A}_2$ used in (3.7) are supposed to make $\mathcal{N}_1$ and $\mathcal{N}_2$ as simple as possible. Moreover, $\mathcal{N}_1$ and $\mathcal{N}_2$ are equally simple. For if $\mathcal{N}_2$ were simpler than $\mathcal{N}_1$, then (3.12) and (3.13) show that the map $(\mathcal{A}_2\mathcal{A})$ normalizes $\mathcal{M}_1$ to the simpler form $\mathcal{N}_2$, which is contrary to the assumption that $\mathcal{A}_1$ has been properly choosen in (3.6) to make $\mathcal{N}_1$ as simple as possible. Similarly, (3.11) and (3.14) show that $\mathcal{N}_1$ cannot be simpler than $\mathcal{N}_2$. We conclude that they must be the same,

$$\mathcal{N}_1 = \mathcal{N}_2. \qquad (30.3.15)$$

Rmk. 3.1: We have seen that the normal form $\mathcal{N}_1$ of a map $\mathcal{M}_1$ is unique. However we remark that, without further requirements, the normalizing map $\mathcal{A}_1$ is not unique. Suppose $\mathcal{B}_1$ is any invertible map that commutes with $\mathcal{N}_1$,

$$\mathcal{N}_1\mathcal{B}_1 = \mathcal{B}_1\mathcal{N}_1. \qquad (30.3.16)$$

Use $\mathcal{B}_1$ to conjugate both sides of (3.6). Doing so and making use of (3.16) gives the result

$$\mathcal{N}_1 = \mathcal{B}_1\mathcal{N}_1\mathcal{B}_1^{-1} = \mathcal{B}_1\mathcal{A}_1\mathcal{M}_1(\mathcal{B}_1\mathcal{A}_1)^{-1}. \qquad (30.3.17)$$

We see that the map $(\mathcal{B}_1\mathcal{A}_1)$ also is a normalizing map that normalizes $\mathcal{M}_1$ to $\mathcal{N}_1$.

Thrm. 3.2: Conversely, suppose $\mathcal{A}_1$ and $\tilde{\mathcal{A}}_1$ are both normalizing maps for $\mathcal{M}_1$,

$$\mathcal{N}_1 = \mathcal{A}_1\mathcal{M}_1\mathcal{A}_1^{-1} = \tilde{\mathcal{A}}_1\mathcal{M}_1\tilde{\mathcal{A}}_1^{-1}. \qquad (30.3.18)$$

Then $\tilde{\mathcal{A}}_1$ and $\mathcal{A}_1$ are related by the equation

$$\tilde{\mathcal{A}}_1 = \mathcal{B}_1\mathcal{A}_1 \qquad (30.3.19)$$

with

$$\mathcal{B}_1 = \tilde{\mathcal{A}}_1\mathcal{A}_1^{-1}, \qquad (30.3.20)$$

where $\mathcal{B}_1$ commutes with $\mathcal{N}_1$.

30.4 Sample Normal Forms

We now describe what normal forms can be achieved in various cases when we are working in the setting of a 6-dimensional phase space with the variables x, y, τ, p_x, p_y, p_τ. As illustrated in Figure 4.1, there are four broad possibilities for general maps $\mathcal{M}$: dynamic (τ-dependent) maps with or without translation factors (characterized by the presence or absence of f_1 terms); static (τ-independent) maps with or without translation factors. These cases are listed below, and will be discussed in subsequent sections.

i. Dynamic map with an f_1 translation factor

ii. Dynamic map without an f_1 translation factor

iii. Static map with an f_1 translation factor

iv. Static map without an f_1 translation factor

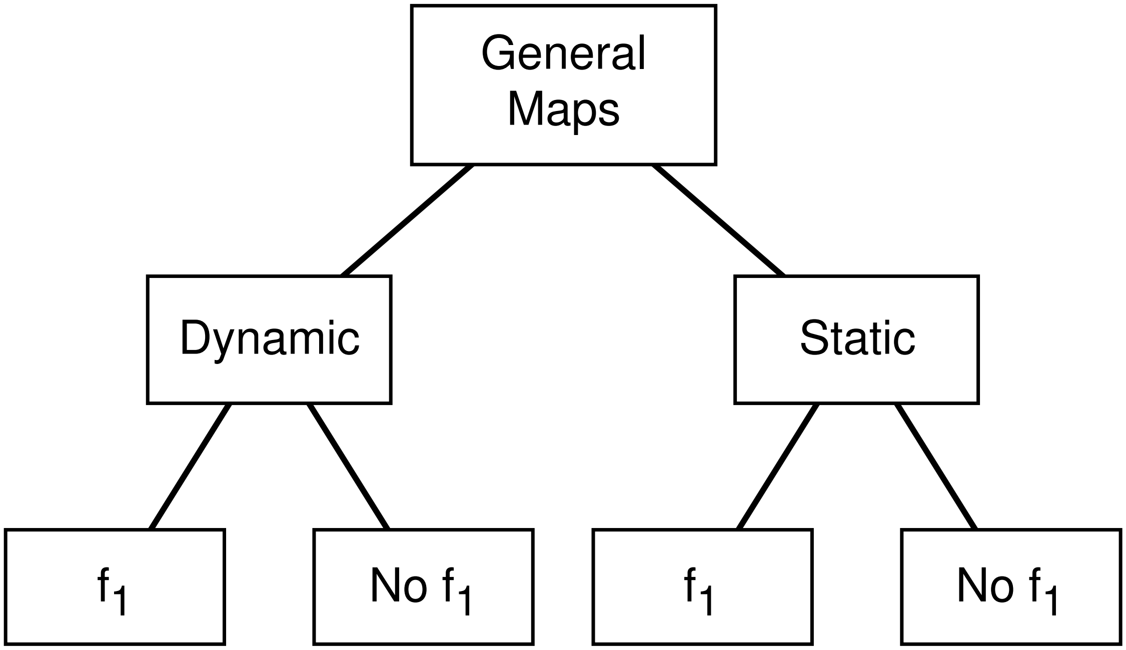


Figure 30.4.1: Four broad possibilities for general maps.

30.5 Dynamic Maps Without Translation Factor

Strangely enough, the easiest case to discuss is dynamic maps without f_1 translation factors. We will therefore treat it first.

30.6 Dynamic Maps With Translation Factor

30.7 Static Maps Without Translation Factor

30.7.1 Preparatory Steps

For the purposes of this section it is convenient to order the phase-space variables as

$$z = (x, p_x; y, p_y; \tau, p_\tau). \tag{30.7.1}$$

According to Section 3.2, J then takes the form (3.2.10) which, in the 6×6 case, is

$$J = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix}. \tag{30.7.2}$$

Suppose $\mathcal{M}$ is an origin preserving symplectic map so that it has the Lie factorization

$$\mathcal{M} = \mathcal{R} \exp(: f_3 :) \exp(: f_4 :) \cdots . \tag{30.7.3}$$

Next assume the $\mathcal{M}$ and also has the property

$$\mathcal{M} p_\tau = p_\tau. \tag{30.7.4}$$

For reasons to become evident shortly, we will call such an $\mathcal{M}$ a *static* map. Upon using the representation (5.2) in (5.3) and equating terms of like degree, it follows that there are the relations

$$\mathcal{R} p_\tau = p_\tau. \tag{30.7.5}$$

$$0 =: f_m : p_\tau = [f_m, p_\tau] = \partial f_m / \partial \tau \text{ for } m \geq 3, \tag{30.7.6}$$

The relation (5.6) says that the generators f_m are τ independent (hence static) for $m \geq 3$, and we will see that (5.5) and the symplectic condition imply that the matrix R associated with $\mathcal{R}$ has a very special form.

To explore the properties of R, let us begin by writing it out in full using standard matrix notation,

$$R = \begin{pmatrix} R_{11} & R_{12} & R_{13} & R_{14} & R_{15} & R_{16} \\ R_{21} & R_{22} & R_{23} & R_{24} & R_{25} & R_{26} \\ R_{31} & R_{32} & R_{33} & R_{34} & R_{35} & R_{36} \\ R_{41} & R_{42} & R_{43} & R_{44} & R_{45} & R_{46} \\ R_{51} & R_{52} & R_{53} & R_{54} & R_{55} & R_{56} \\ R_{61} & R_{62} & R_{63} & R_{64} & R_{65} & R_{66} \end{pmatrix}. \tag{30.7.7}$$

Now require that $\mathcal{R}$ be a symplectic map that satisfies (5.5). From (3.1.10) we know that the symplectic condition requires the relation

$$RJR^T = J. \tag{30.7.8}$$

When written in terms of components, this relation takes the form

$$\sum_{bc} R_{ab} J_{bc} R_{dc} = J_{ad}. \tag{30.7.9}$$

As a result of these two requirements (5.5) and (5.8) we will see that many matrix elements of R are 0 or 1, and others are related.

Define the quantities Δ_1 through Δ_4 by the rules

$$\Delta_1 = R_{52}, \tag{30.7.10}$$

$$\Delta_2 = -R_{51}, \tag{30.7.11}$$

$$\Delta_3 = R_{54}, \tag{30.7.12}$$

$$\Delta_4 = -R_{53}, \tag{30.7.13}$$

and view them as the components of a vector Δ. Also define a matrix $\hat{R}$ by the rule

$$\hat{R} = \begin{pmatrix} R_{11} & R_{12} & R_{13} & R_{14} & 0 & 0 \\ R_{21} & R_{22} & R_{23} & R_{24} & 0 & 0 \\ R_{31} & R_{32} & R_{33} & R_{34} & 0 & 0 \\ R_{41} & R_{42} & R_{43} & R_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{30.7.14}$$

and write (5.14) in the more compact form

$$\hat{R} = \begin{pmatrix} \check{R} & 0 \\ 0 & I \end{pmatrix} \tag{30.7.15}$$

where $\check{R}$ is the 4×4 matrix

$$\check{R} = \begin{pmatrix} R_{11} & R_{12} & R_{13} & R_{14} \\ R_{21} & R_{22} & R_{23} & R_{24} \\ R_{31} & R_{32} & R_{33} & R_{34} \\ R_{41} & R_{42} & R_{43} & R_{44} \end{pmatrix}. \tag{30.7.16}$$

Finally, let $\check{J}$ be the the 4×4 version of J,

$$\check{J} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}. \tag{30.7.17}$$

Then, it is the case that R must be of the more specific form

$$R = \begin{pmatrix} R_{11} & R_{12} & R_{13} & R_{14} & 0 & (\check{R}\Delta)_1 \\ R_{21} & R_{22} & R_{23} & R_{24} & 0 & (\check{R}\Delta)_2 \\ R_{31} & R_{32} & R_{33} & R_{34} & 0 & (\check{R}\Delta)_3 \\ R_{41} & R_{42} & R_{43} & R_{44} & 0 & (\check{R}\Delta)_4 \\ -\Delta_2 & \Delta_1 & -\Delta_4 & \Delta_3 & 1 & R_{56} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{30.7.18}$$

Also, the entries in the upper left 4×4 block of R, the entries in $\check{R}$, must obey the "reduced" symplectic relations

$$\check{R}\check{J}\check{R}^T = \check{J}. \tag{30.7.19}$$

Observe that the entries R_{16}, R_{26}, R_{36}, and R_{46} in R describe dispersive effects. That is, they describe how the transverse coordinates and momenta depend on p_τ. By contrast, the entries R_{51}, R_{52}, R_{53}, and R_{54} in R describe how the time of flight depends on the transverse initial conditions. From (5.18) we see that dispersive effects and time of flight effects are related by the symplectic condition! They are opposite sides of the same coin. This is an example of what we call *symplectic reciprocity*: seemingly unrelated quantities are in fact related by the symplectic condition.

We will prove this result in stages: We recall that the matrix R associated with $\mathcal{R}$ is given by the relation

$$\mathcal{R}z_a = \sum_b R_{ab} z_b. \tag{30.7.20}$$

As a result of (5.20), the condition (5.5) requires that R have the more specific form

$$R = \begin{pmatrix} R_{11} & R_{12} & R_{13} & R_{14} & R_{15} & R_{16} \\ R_{21} & R_{22} & R_{23} & R_{24} & R_{25} & R_{26} \\ R_{31} & R_{32} & R_{33} & R_{34} & R_{35} & R_{36} \\ R_{41} & R_{42} & R_{43} & R_{44} & R_{45} & R_{46} \\ R_{51} & R_{52} & R_{53} & R_{54} & R_{55} & R_{56} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{30.7.21}$$

Next, impose the symplectic condition (5.8) for R of the form (5.21). Set $a = 6$ in the relation (5.9) to get the result

$$\sum_{bc} R_{6b} J_{bc} R_{dc} = J_{6d}. \tag{30.7.22}$$

But, from (5.21), we know that

$$R_{6b} = \delta_{6b}. \tag{30.7.23}$$

Therefore the sum (5.22) becomes

$$\sum_c J_{6c} R_{dc} = J_{6d}. \tag{30.7.24}$$

Also, we see from (5.2) that

$$J_{6c} = -\delta_{5c}. \tag{30.7.25}$$

Therefore the sum (5.24) reduces to the result

$$-R_{d5} = J_{6d} \tag{30.7.26}$$

from which we conclude that

$$R_{d5} = 0 \text{ for } d = 1 \text{ to } 4, \tag{30.7.27}$$

$$R_{55} = 1. \tag{30.7.28}$$

Consequently, R must have the yet more specific form

$$R = \begin{pmatrix} R_{11} & R_{12} & R_{13} & R_{14} & 0 & R_{16} \\ R_{21} & R_{22} & R_{23} & R_{24} & 0 & R_{26} \\ R_{31} & R_{32} & R_{33} & R_{34} & 0 & R_{36} \\ R_{41} & R_{42} & R_{43} & R_{44} & 0 & R_{46} \\ R_{51} & R_{52} & R_{53} & R_{54} & 1 & R_{56} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{30.7.29}$$

Define the quantity ξ by the rule

$$\xi = R_{56}, \tag{30.7.30}$$

and associate with ξ and Δ the matrices $C(\xi)$ and $D(\Delta)$ by the rules

$$C(\xi) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \xi \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & R_{56} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{30.7.31}$$

$$D(\Delta) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \Delta_1 \\ 0 & 1 & 0 & 0 & 0 & \Delta_2 \\ 0 & 0 & 1 & 0 & 0 & \Delta_3 \\ 0 & 0 & 0 & 1 & 0 & \Delta_4 \\ -\Delta_2 & \Delta_1 & -\Delta_4 & \Delta_3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{30.7.32}$$

It is easily verified that the matrices $C(\xi)$ and $D(\Delta)$ are symplectic and have the inverses

$$C^{-1}(\xi) = C(-\xi), \tag{30.7.33}$$

$$D^{-1}(\Delta) = D(-\Delta). \tag{30.7.34}$$

Indeed, $C(\xi)$ and $D(\Delta)$ are the matrices associated with the linear symplectic maps $\mathcal{C}(\xi)$ and $\mathcal{D}(\Delta)$ given by the relations

$$\mathcal{C} = \exp(:-\xi p_\tau^2/2:), \tag{30.7.35}$$

$$\mathcal{D} = \exp(: p_\tau g_1 :), \tag{30.7.36}$$

where

$$g_1(\Delta) = \Delta_2 x - \Delta_1 p_x + \Delta_4 y - \Delta_3 p_y. \tag{30.7.37}$$

Note, for future use, that the matrix C commutes with both the matrices D and $\hat{R}$.

We now assert that R has the factorization

$$R = \hat{R}CD \tag{30.7.38}$$

or, equivalently,

$$\hat{R} = RD^{-1}C^{-1}. \tag{30.7.39}$$

The proof of this assertion involves matrix multiplication and invoking the symplectic condition for R. Define a matrix $\hat{R}'$ by the rule

$$\hat{R}' = RD^{-1}C^{-1}. \tag{30.7.40}$$

Carrying out the indicated multiplications gives the result

$$\hat{R}' = \begin{pmatrix} R_{11} & R_{12} & R_{13} & R_{14} & 0 & \epsilon_1 \\ R_{21} & R_{22} & R_{23} & R_{24} & 0 & \epsilon_2 \\ R_{31} & R_{32} & R_{33} & R_{34} & 0 & \epsilon_3 \\ R_{41} & R_{42} & R_{43} & R_{44} & 0 & \epsilon_4 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{30.7.41}$$

where

$$\begin{aligned} \epsilon_1 &= R_{16} - [R_{11}R_{52} + R_{12}(-R_{51}) + R_{13}R_{54} + R_{14}(-R_{53})] \\ &= R_{16} - (\check{R}\Delta)_1, \end{aligned} \tag{30.7.42}$$

$$\begin{aligned} \epsilon_2 &= R_{26} - [R_{21}R_{32} + R_{22}(-R_{52}) + R_{23}R_{54} + R_{24}(-R_{53})] \\ &= R_{26} - (\check{R}\Delta)_2, \end{aligned} \tag{30.7.43}$$

$$\begin{aligned} \epsilon_3 &= R_{36} - [R_{31}R_{52} + R_{32}(-R_{51}) + R_{33}R_{54} + R_{34}(-R_{53})] \\ &= R_{36} - (\check{R}\Delta)_3, \end{aligned} \tag{30.7.44}$$

$$\begin{aligned} \epsilon_4 &= R_{46} - [R_{41}R_{52} + R_{42}(-R_{51}) + R_{43}R_{54} + R_{44}(-R_{53})] \\ &= R_{46} - (\check{R}\Delta)_4. \end{aligned} \tag{30.7.45}$$

Next, because R, C^{-1}, and D^{-1} are symplectic matrices, $\hat{R}'$ must a symplectic matrix. In analogy with (5.9), the symplectic condition for $\hat{R}'$ can be written in the form

$$\sum_{bc} \hat{R}'_{ab} J_{bc} \hat{R}'_{dc} = J_{ad}. \tag{30.7.46}$$

Now put $a = 5$ in (5.46) and make use of the special forms of J and $\hat{R}'$ as given by (5.2) and (5.41). Doing so gives the result

$$\sum_{bc} \hat{R}'_{5b} J_{bc} \hat{R}'_{dc} = J_{5d}, \tag{30.7.47}$$

which yields the relations

$$\hat{R}'_{d6} = J_{5d}. \tag{30.7.48}$$

From (5.2), (5.41), and (5.48), we conclude that

$$\epsilon_d = 0 \text{ for } d = 1 \text{ to } 4. \tag{30.7.49}$$

Therefore there is the relation

$$\hat{R}' = \hat{R}, \tag{30.7.50}$$

and (3.39) is correct.

Moreover, in view of (5.42) through (5.45) and (5.49), we have found the relations

$$R_{a6} = (\check{R}\Delta)_a \text{ for } a = 1 \text{ to } 4 \tag{30.7.51}$$

with Δ given by (5.10) through (5.13). Thus, (5.18) is correct. To reiterate, what we have learned from the symplectic condition is that in (5.29) the matrix elements R_{a6} for $a = 1$ to 4 are not independent of the matrix elements R_{5a} for $a = 1$ to 4, but instead are related by the conditions (5.10) through (5.13) and (5.51). Finally, because of (5.15), the remaining relations demanded by (5.46) yield the matrix relation (5.19).

We close this section by noting that the matrix relation (5.38) implies a related factorization for the map $\mathcal{R}$. Let $\hat{\mathcal{R}}$ be the symplectic map associated with $\hat{R}$. Then (5.38) is equivalent to the map factorization relation

$$\mathcal{R} = \mathcal{D}\mathcal{C}\hat{\mathcal{R}} = \mathcal{C}\mathcal{D}\hat{\mathcal{R}}. \tag{30.7.52}$$

Here we have used the fact that C and D commute.

Exercises

30.7.1. Verify that the maps $\mathcal{C}$ and $\mathcal{D}$ given by (5.35) and (5.36) are equivalent to the matrices C and D given by (5.31) and (5.32).

30.7.2. Starting with (5.40), verify that carrying out the indicated multiplications yields the results (5.41) through (5.45).

30.8 Static Maps With Translation Factor

30.9 Tunes, Phase Advances and Slips, Momentum Compaction, Chromaticities, and Anharmonicities

- Ring Analysis and Phase Advances
- Equivalent Locations
- Matched Insertions
- Floquet Theory
- Tune Footprints

30.10 Courant-Snyder Invariants and Lattice Functions

30.11 Analysis of Tracking Data

Bibliography

[1] K.R. Meyer, "Normal forms for Hamiltonian systems", *Celestial Mechanics* **9**, 517-522 (1974).

[2] G. Benettin, I. Galgani, A. Giorgilli, J.-M. Strelcyn, "A Proof of Kolmogorov's Theorem on Invariant Tori Using Canonical Transformations Defined by the Lie Method", *Il Nuovo Cimento* **79 B**, 201 (1984).

[3] A. Bruno, *Local Methods in Nonlinear Differential Equations*, Springer-Verlag (1989).

[4] A. Bruno, *The Restricted 3-Body Problem: Plane Periodic Orbits*, de Gruyter (1994).

[5] J. A. Murdock, *Normal Forms and Unfoldings for Local Dynamical Systems*, Springer (2003).

[6] B. Braaksma, G. Immink, M. van der Put, Edit., *The Stokes Phenomenon and Hilbert's 16th Problem*, World Scientific (1996).

[7] Y. Ilyashenko and S. Yakovenko, *Lectures on Analytic Differential Equations*, American Mathematical Society (2008).

[8] H. Zoladek, *The Monodromy Group*, Birkhäser Verlag (2006).

[9] G. Cicogna and G. Gaeta, *Symmetry and Perturbation Theory in Nonlinear Dynamics*, Springer (2010).

[10] J. Cresson, "Mould Calculus and Normalization of Vector Fields", http://web.univ-pau.fr/~jcresson/MouldCalculus.pdf, (2006). See also the Web site http://web.univ-pau.fr/~jcresson/ for links to many other related and interesting publications.

[11] E. Forest, *From Tracking Code to Analysis: Generalized Courant-Snyder Theory for Any Accelerator Model*, Springer Japan (2016).

[12] A. Haro, M. Canadell, J-L. Figueras, A. Luque, and J-M. Mondelo, *The Parameterization Method for Invariant Manifolds: From Rigorous Results to Effective Computations*, Applied Mathematical Sciences Volume 195, Springer (2016).

Chapter 31

Lattice Functions

Chapter 32

Solved and Unsolved Polynomial Orbit Problems: Invariant Theory

32.1 Introduction

As in (8.5.2), let $\mathcal{R}$ be any linear symplectic map [a map corresponding to an $Sp(2n)$ transformation] written in the general form

$$\mathcal{R} = \exp(: f_2^c :) \exp(: f_2^a :). \tag{32.1.1}$$

Suppose $\mathcal{R}$ acts on any homogeneous polynomial g_m in $\mathcal{P}_m$. Then, in view of (21.5.6), the result is a *transformed* polynomial $g_m^{\rm tr}$ that is also in $\mathcal{P}_m$,

$$g_m^{\rm tr}(z) = \mathcal{R} g_m(z) = g_m(\mathcal{R} z) = g_m(Rz). \tag{32.1.2}$$

Here we have also used (8.4.15). Indeed, we know from the work of Chapter 21 that in the two-variable case the $\mathcal{P}_m$ carry the irreducible representation $\Gamma(m)$ of $Sp(2)$; in the four-variable case the $\mathcal{P}_m$ carry the irreducible representation $\Gamma(m,0)$ of $Sp(4)$; in the six-variable case the $\mathcal{P}_m$ carry the irreducible representation $\Gamma(m,0,0)$ of $Sp(6)$; etc.

The set of polynomials $g_m^{\rm tr}$ that can be obtained from any given g_m and arbitrary $\mathcal{R}$ of the form (1.1) is called the *orbit* of g_m under the action of $Sp(2n)$. Now suppose that h_m is any other polynomial in $\mathcal{P}_m$. We will say that h_m is *equivalent* to g_m if there is some $\mathcal{R}$ of the form (1.1) that sends g_m to h_m,

$$h_m \sim g_m \Leftrightarrow h_m = \mathcal{R} g_m \text{ for some } \mathcal{R}. \tag{32.1.3}$$

It is easy to check that (1.3) is indeed an equivalence relation, and we may say that two polynomials in $\mathcal{P}_m$ are equivalent if they lie on the same orbit.

Finally, suppose we are given some polynomial g_m. Then the equivalence class of g_m, which we will denote by $\{g_m\}$, consists of all the $g_m^{\rm tr}$ given by (1.2) for all choices of $\mathcal{R}$. (Thus, the equivalence class $\{g_m\}$ is the orbit of g_m.) Among the $g_m^{\rm tr}$ produced in this fashion there will be one that has some particularly desirable form or property. Various possibilities come to mind: For example, we may attempt to drive to zero as many coefficients in $g_m^{\rm tr}$ as possible by a particular choice of $\mathcal{R}$. Or, if g_m happens to be on the orbit of some monomial, we might like to discover which monomial and determine its coefficient. Or, we might like to find a g_m^{tr} on the oribit of g_m that has the smallest length in the sense of minimizing the scalar product $\langle g_m^{\text{tr}}, g_m^{\text{tr}}\rangle$ as defined in Section 7.3. A or the g_m^{tr} that has some such desirable form or property, or perhaps some other property yet to be discovered, will be called the *normal form* of g_m and will be denoted by the symbols g_m^N. (We remark that in some literature a normal form is called a *canonical* form, and homogeneous polynomials or ratios of homogeneous polynomials are called *quantics*.) Put another way, a normal form of g_m is a particularly simple or pleasing point on the orbit of g_m. Exactly what a normal form for g_m should be is partly a matter of investigation, and partly a matter of choice. Given a g_m, one must first examine all the members of the equivalence class $\{g_m\}$. Then, with their properties clearly in mind, one selects a particularly pleasing g_m^{tr} and calls it g_m^N. Ideally one would like to have an algorithm that takes g_m as an input and provides as outputs g_m^N and the *normalizing* $\mathcal{R}$ that transforms g_m into g_m^N.

Three facts are now obvious practically as a matter of definition. First, $Sp(2n)$ acts transitively on each equivalence class. Second, we may label the equivalence class of g_m by specifying g_m^N. That is. we have the relation

$$\{g_m\} = \{g_m^N\}. \tag{32.1.4}$$

Third, suppose two polynomials g_m and h_m are known or can be shown to have the same normal form,

$$g_m^N = h_m^N. \tag{32.1.5}$$

Then, they are in the same equivalence class and there is an $\mathcal{R}$ that sends one into the other as in (1.3).

There is another terminology that is sometimes used for the situation we have been describing. In this terminology each equivalence class (orbit) is called a *leaf*, and the decomposition of $\mathcal{P}_m$ into equivalence classes is called a *foliation*.

Evidently a general homogeneous polynomial g_m is specified by giving its coefficients. It can be shown (and we will see examples) that there exist polynomial functions of these coefficients that remain unchanged under the transformation (1.2). These functions are called *invariants*. Thus if h_m and g_m are equivalent as in (1.3), each invariant function must have the same value for the coefficients of h_m and the coefficients of g_m.

Why, apart from curiosity, should one care about orbits of g_m in $\mathcal{P}_m$, normal forms g_m^N, and invariants? We will see in Chapter 33 that a knowledge of normal forms for g_2 and invariants for g_m is useful for characterising beams. In Chapter 34 we will see that normal forms for g_3, g_4, $\cdots$ might, if we knew them, be useful in the approximate but exactly symplectic numerical evaluation of the effect of a general map $\mathcal{M}$ on a general phase-space point z as in (7.6.2).

Exercises

32.1.1. Show that (1.3) defines an equivalence relation. See Exercise (5.12.7).

32.2 Solved Polynomial Orbit Problems

In this section we will describe briefly some of what is known about the orbits of g_m under the action of $Sp(2n)$ or, to be more precise, $Sp(2n, \mathbb{R})$. Our results will be fairly complete for the cases $m = 1$ and $m = 2$, and therefore these cases can be characterized as being *solved*. The cases $m > 2$ are much more difficult, and will be characterized largely in terms of what is not known. They are treated in the next section.

First, to dispell a possible false expectation, recall the action of the rotation group on ordinary 3-dimensional Euclidean space. If x, y, z are the usual Cartesian coordinates in Euclidean 3-space, we know that the group $SO(3)$ of rotations about the origin *preserves* the polynomial $(x^2 + y^2 + z^2)$ and any function of this polynomial. Does something analogous happen for the action of $Sp(2n)$ on phase space? The answer is *no.* Suppose that some g_m is preserved,

$$g_m^{\rm tr} = g_m. \tag{32.2.1}$$

Then, from (1.2), we find the result

$$g_m(Rz) = g_m(z) \tag{32.2.2}$$

for all R in $Sp(2n)$. But, from Sections 3.6.5 and 7.2, we know that $Sp(2n)$ acts transitively on phase space. Therefore, any g_m that satisfies (2.2) for all R must have the same value everywhere in phase space, and the only such polynomial is g_0.

There is another instructive way to reach the same conclusion. From (1.1), (1.2), and (2.2) one sees that to be preserved g_m must satisfy the relation

$$\exp(:\epsilon f_2:)g_m = g_m \text{ for all } f_2. \tag{32.2.3}$$

The infinitesimal version of (2.3) is the relation

$$:f_2:g_m = 0 \text{ for all } f_2. \tag{32.2.4}$$

But, say for $sp(6)$, we know that any g_m belongs to the *irreducible* representation $\Gamma(m, 0, 0)$. See Section 1.8. Therefore, the only way that (2.4) can be satisfied is to have $m = 0$.

32.2.1 First-Order Polynomials

We have seen that there is no nontrivial preserved g_m. Thus, $Sp(2n)$ must have some genuine action on each $\mathcal{P}_m$. Let us begin with the case of $\mathcal{P}_1$. Any g_1 in $\mathcal{P}_1$ can be written in the form

$$g_1(a; z) = \sum_j a_j z_j = (a, z). \tag{32.2.5}$$

In this case use of (1.2) gives the result

$$g_1^{\rm tr}(a; z) = (a, Rz) = (R^T a, z) = g_1(R^T a; z) = g_1(a^{\rm tr}; z). \tag{32.2.6}$$

Here we have introduced the notation

$$a^{\rm tr} = R^T a. \tag{32.2.7}$$

We know that R^T is symplectic if R is symplectic; and we again recall that $Sp(2n)$ acts transitively. It follows that if a is any nonzero $2n$-vector, there is a symplectic R such that a^{tr} is any desired vector. Therefore, $\mathcal{P}_1$ decomposes into two equivalence classes: the identically zero polynomial and all the rest. If we ignore the trivial case of the identically zero polynomial, we may say that $\mathcal{P}_1$ consists of only one equivalence class and correspondingly, a single orbit. A convenient normal form is the monomial q_1 with unit coefficient,

$$g_1^N = q_1. \tag{32.2.8}$$

32.2.2 Second-Order Polynomials

Next consider $\mathcal{P}_2$. Here the situation is more complicated. Any g_2 in $\mathcal{P}_2$ can be written in the form

$$g_2(S;z) = \sum_{jk} S_{jk} z_j z_k = (z, Sz) \tag{32.2.9}$$

where S is any symmetric matrix. In this case use of (1.2) gives the result

$$g_2^{\text{tr}}(S;z) = (Rz, SRz) = (z, R^T SRz) = g_2(S^{\text{tr}}; z) \tag{32.2.10}$$

where

$$S^{\text{tr}} = R^T SR. \tag{32.2.11}$$

It is easily checked that S^{tr} is symmetric if S is.

The analysis of the relation (2.11) is facilitated by a trick. Let B denote the *Hamiltonian* matrix gotten from S by the rule

$$B = JS. \tag{32.2.12}$$

Since J is invertible, one can always find S given B, and vice versa. See Section 3.7. Next we define B^{tr} by the rule

$$B^{\text{tr}} = JS^{\text{tr}}. \tag{32.2.13}$$

With the aid of these definitions the relation (2.11) takes the form

$$B^{\text{tr}} = JS^{\text{tr}} = JR^T SR = JR^T J^{-1} JSR = R^{-1} BR. \tag{32.2.14}$$

Here we have used (3.1.9). With the aid of J we have turned a symplectic congruency relation (2.11) into a symplectic conjugacy (similarity) relation (2.14). What we learn from (2.14) is that the problem of finding orbits in $\mathcal{P}_2$ is equivalent to finding orbits in the space of $2n \times 2n$ real Hamiltonian matrices under the action of real symplectic similarity transformations. We know that eigenvalues are unchanged by similarity transformations, and therefore expect that eigenvalues and functions constructed from eigenvalues will play an important role.

Suppose we were allowed to make arbitrary (including complex and nonsymplectic) similarity transformations. Then we know that B, if it has distinct eigenvalues, can be diagonalized. And if the eigenvalues are not distinct, B might still be diagonalizable or, in the worst case, it could still be brought to Jordan normal form. We might define the diagonal or Jordan form for B to be B^{tr}, and then try to form S^{tr} and $g_2^N = g_2^{\text{tr}}$ accordingly. However, we are only allowed to use real symplectic similarity transformations, and we must see to what extent something analogous can be done using only such transformations.

32.2.2.1 Two-Dimensional Phase-Space Case

We will come to the general $2n \times 2n$ case eventually. As a warm-up exercise, consider first the 2×2 case for 2-dimensional phase space.. Then g_2 has the general form

$$g_2 = \beta p^2 + 2\alpha pq + \gamma q^2 \tag{32.2.15}$$

where α, β, γ are arbitrary constants. Correspondingly, the matrices S and B take the forms

$$S = \begin{pmatrix} \gamma & \alpha \\ \alpha & \beta \end{pmatrix}, \tag{32.2.16}$$

$$B = \begin{pmatrix} \alpha & \beta \\ -\gamma & -\alpha \end{pmatrix}. \tag{32.2.17}$$

Evidently the transformation (2.14) cannot change the determinant of B which we will call δ,

$$\delta = \det B = \beta\gamma - \alpha^2. \tag{32.2.18}$$

That is, δ is an *invariant* constructed from the coefficients of g_2. [Note that δ is just the negative of the discriminant. See (8.7.30).] However, as will be seen, we can change α, β, γ while maintaining the condition (2.18). Note that the matrix B has the characteristic polynomial

$$P(\lambda) = \det(B - \lambda I) = \lambda^2 + \beta\gamma - \alpha^2 = \lambda^2 + \delta, \tag{32.2.19}$$

and therefore has the eigenvalues

$$\lambda_\pm = \pm(\alpha^2 - \beta\gamma)^{1/2} = \pm(-\delta)^{1/2}. \tag{32.2.20}$$

It is convenient to consider separately the six cases listed below:

i. $\beta > 0$

ii. $\beta < 0$

iii. $\gamma > 0$

iv. $\gamma < 0$

v. $\beta = \gamma = 0$

vi. $\alpha = \beta = \gamma = 0$

In the next few paragraphs we will treat them one by one.

Case *i*, $\beta > 0$

Let us begin with case i by supposing $\beta > 0$. Let M be the symplectic matrix defined by the equation

$$M = \begin{pmatrix} \sqrt{\beta} & 0 \\ -\alpha/\sqrt{\beta} & 1/\sqrt{\beta} \end{pmatrix}. \tag{32.2.21}$$

Then use of (2.11) with $R = M$ gives a transformed S that we will call S',

$$S' = M^T S M = \begin{pmatrix} \delta & 0 \\ 0 & 1 \end{pmatrix}. \tag{32.2.22}$$

Next suppose that $\delta > 0$. In this case we conclude from (2.18) that $\gamma > 0$ and from (2.20) that the eigenvalues $\lambda_\pm$ are pure imaginary. Let N be the symplectic matrix

$$N = \begin{pmatrix} \delta^{-1/4} & 0 \\ 0 & \delta^{1/4} \end{pmatrix}. \tag{32.2.23}$$

Use of N to transform S' to $S^{\rm tr}$ gives the result

$$S^{\rm tr} = N^T S' N = \begin{pmatrix} \delta^{-1/4} & 0 \\ 0 & \delta^{1/4} \end{pmatrix} \begin{pmatrix} \delta & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \delta^{-1/4} & 0 \\ 0 & \delta^{1/4} \end{pmatrix} = \begin{pmatrix} \delta^{1/2} & 0 \\ 0 & \delta^{1/2} \end{pmatrix}. \tag{32.2.24}$$

Correspondingly $g_2^{\rm tr}$ is given by the relation

$$g_2^{\rm tr} = (z, S^{\rm tr} z) = \delta^{1/2}(p^2 + q^2). \tag{32.2.25}$$

Suppose instead that $\delta = 0$. Now we conclude from (2.18) that $\gamma \geq 0$ and from (2.20) that the eigenvalues $\lambda_\pm$ both vanish. In this case S' as given by (2.22) can be used directly to give the result

$$g_2^{\rm tr} = (z, S' z) = p^2. \tag{32.2.26}$$

It is easily verified that p^2 and q^2 are equivalent,

$$p^2 \sim q^2. \tag{32.2.27}$$

See Exercise 2.1. Therefore, if desired, we can find and employ an R such that

$$g_2^{\rm tr} = q^2. \tag{32.2.28}$$

Finally suppose that $\delta < 0$ (in which case $\beta\gamma < \alpha^2$ and the eigenvalues $\lambda_\pm$ are real). Let N be the symplectic matrix

$$N = \begin{pmatrix} (-\delta)^{-1/4} & 0 \\ 0 & (-\delta)^{1/4} \end{pmatrix}. \tag{32.2.29}$$

Use of N to transform S' to $S^{\rm tr}$ gives the result

$$\begin{aligned} S^{\rm tr} = N^T S' N &= \begin{pmatrix} (-\delta)^{-1/4} & 0 \\ 0 & (-\delta)^{1/4} \end{pmatrix} \begin{pmatrix} \delta & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} (-\delta)^{-1/4} & 0 \\ 0 & (-\delta)^{1/4} \end{pmatrix} \\ &= (-\delta)^{1/2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \end{aligned} \tag{32.2.30}$$

Correspondingly $g_2^{\rm tr}$ is given by the relation

$$g_2^{\rm tr} = (z, S^{\rm tr} z) = (-\delta)^{1/2}(q^2 - p^2). \tag{32.2.31}$$

Case *ii*, $\beta < 0$

To consider case *ii*, suppose $\beta < 0$. Let M be the symplectic matrix defined by the equation

$$M = \begin{pmatrix} \sqrt{-\beta} & 0 \\ \alpha/\sqrt{-\beta} & 1/\sqrt{-\beta} \end{pmatrix}. \tag{32.2.32}$$

Then use of (2.11) with $R = M$ gives a transformed S which we again call S',

$$S' = M^T S M = \begin{pmatrix} -\delta & 0 \\ 0 & -1 \end{pmatrix}. \tag{32.2.33}$$

Next suppose $\delta > 0$. In this case we conclude from (2.18) that $\gamma < 0$ and from (2.20) that the eigenvalues $\lambda_\pm$ are pure imaginary. Again let N be the symplectic matrix (2.23). Then we find for $S^{\rm tr}$ the result

$$S^{\rm tr} = N^T S' N = -\delta^{1/2} I. \tag{32.2.34}$$

Correspondingly $g_2^{\rm tr}$ is given by the relation

$$g_2^{\rm tr} = -\delta^{1/2}(p^2 + q^2). \tag{32.2.35}$$

Suppose instead that $\delta = 0$. Now we conclude from (2.18) that $\gamma \le 0$ and from (2.20) that the eigenvalues $\lambda_\pm$ both vanish. In this case use of S' directly gives the result

$$g_2^{\rm tr} = -p^2. \tag{32.2.36}$$

Alternatively, in view of (2.27), we can find an R such that

$$g_2^{\rm tr} = -q^2. \tag{32.2.37}$$

Finally suppose $\delta < 0$. Then use of (2.33) and N given by (2.29), and calling the result S'', give the relation

$$S'' = -\begin{pmatrix} (-\delta)^{1/2} & 0 \\ 0 & -(-\delta)^{1/2} \end{pmatrix} = -(-\delta)^{1/2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{32.2.38}$$

Also, it is easily verified that in this case use of the symplectic matrix J gives the relation

$$S^{\rm tr} = J^T S'' J = -S'' = (-\delta)^{1/2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{32.2.39}$$

It follows that we may take for $g_2^{\rm tr}$ the polynomial

$$g_2^{\rm tr} = (-\delta)^{1/2}(q^2 - p^2). \tag{32.2.40}$$

Cases *iii* and *iv*, $\gamma > 0$ or $\gamma < 0$

We have covered cases *i* and *ii*. Exercise 2.2 shows that cases *iii* and *iv* give results identical to those for cases *i* and *ii*, respectively. In particular, we still find the results (2.25), (2.26), (2.28), (2.31), (2.35), (2.36), (2.37), and (2.40) [which is identical to (2.31)].

Case *v*, $\beta = \gamma = 0$

For case v we immediately have the result

$$S = \begin{pmatrix} 0 & \alpha \\ \alpha & 0 \end{pmatrix} = (-\delta)^{1/2} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{32.2.41}$$

and g_2 takes the form

$$g_2 = 2(-\delta)^{1/2} qp. \tag{32.2.42}$$

Let O be the symplectic matrix

$$O = (1/\sqrt{2}) \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}. \tag{32.2.43}$$

Use of O to transform S gives the result

$$S^{\rm tr} = O^T S O = (-\delta)^{1/2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{32.2.44}$$

Correspondingly, we find for $g_2^{\rm tr}$ the identical result (2.40). Note that comparison of (2.41) and (2.44) reveals that there are the equivalence relations

$$+(-\delta)^{1/2}(q^2 - p^2) \sim -(-\delta)^{1/2}(q^2 - p^2) \sim -2(-\delta)^{-1/2} qp \sim +2(-\delta)^{1/2} qp. \tag{32.2.45}$$

Case *vi*, $\alpha = \beta = \gamma = 0$

Finally, case vi gives the zero polynomial.

Normal Forms

Evidently, we may take the various g_2^{tr} discovered for cases i through vi to be normal forms. We see that, for the most part, the normal form is labeled by the value of the invariant δ with additional qualifications for the sign of β or γ in the cases $\delta \geq 0$. Therefore, as mentioned before, a necessary condition for $h_2 \sim g_2$ is that they have the *same* invariant. All these results are summarized in Figure 2.1 below.

$\delta < 0$	$\delta = 0$	$\delta > 0$
$(-\delta)^{1/2}(q^2 - p^2)$	q^2 or p^2 if β or $\gamma > 0$	$(\delta)^{1/2}(p^2 + q^2)$ if β or $\gamma > 0$
or	$-\,q^2$ or $-\,p^2$ if β or $\gamma < 0$	$-\,(\delta)^{1/2}(p^2 + q^2)$ if β or $\gamma < 0$
$2(-\delta)^{1/2}pq$		

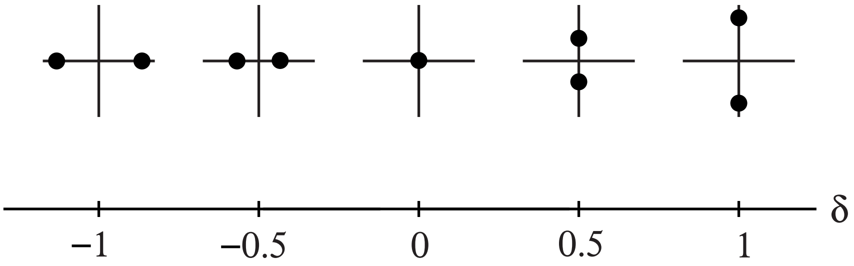


Figure 32.2.1: Normal forms g_2^N and eigenvalue spectrum of associated Hamiltonian matrices in the case of 2-dimensional phase space. The normal forms given in the three columns above are for the cases $\delta < 0$, $\delta = 0$, and $\delta > 0$, respectively.

Geometrical Description

It is also instructive to examine the surfaces $\beta\gamma - \alpha^2 = \delta$ for various values of δ. We will see that each such surface is an orbit. To do so, it is convenient to perform a 45° rotation in the β, γ plane, and to scale α, by introducing new variables ξ, η, ζ by the definitions

$$\beta = (1/\sqrt{2})(\xi - \eta), \tag{32.2.46}$$

$$\gamma = (1/\sqrt{2})(\xi + \eta), \tag{32.2.47}$$

$$\alpha = (1/\sqrt{2})\zeta. \tag{32.2.48}$$

In terms of these variables the relation (2.18) becomes

$$\xi^2 - \eta^2 - \zeta^2 = 2\delta. \tag{32.2.49}$$

Also, use of the scalar product of Section 7.3 and (2.15) gives the result

$$\langle g_2, g_2 \rangle = 2\beta^2 + 4\alpha^2 + 2\gamma^2 = 2(\xi^2 + \eta^2 + \zeta^2). \tag{32.2.50}$$

Thus, polynomials of any given norm are *spheres* in ξ, η, ζ space.

In the case that $\delta \leq 0$, the relation (2.49) can be rewritten in the form

$$\eta^2 + \zeta^2 - \xi^2 = -2\delta. \tag{32.2.51}$$

For $\delta < 0$ this equation yields hyperboloids of one sheet with symmetry axis ξ. See Figure 2.2a below. Evidently for $\delta < 0$ the points on a given hyperboloid that are closest to the origin lie on the plane $\xi = 0$, in which case (2.51) becomes the circle

$$\eta^2 + \zeta^2 = -2\delta. \tag{32.2.52}$$

According to (2.50) all polynomials on this circle have the same squared norm, namely -4δ. The normal form (2.31) lies on this circle and has the coordinates

$$\xi = 0,\ \eta = (-2\delta)^{1/2},\ \zeta = 0, \tag{32.2.53}$$

or, equivalently,

$$-\beta = \gamma = (-\delta)^{1/2},\ \alpha = 0. \tag{32.2.54}$$

Here use has been made of (2.46) through (2.48). Evidently all points on the hyperboloid corresponding to a given value of $\delta < 0$ lie on the same orbit. That is, there is one equivalence class for each value of $\delta < 0$.

The case $\delta = 0$ produces two cones with a common vertex at the origin. Again see Figure 2.2a. Shortly we will discuss it more.

For $\delta > 0$ the equation (2.49) yields a hyperboloid of two sheets with symmetry axis ξ. See Figure 2.2b. On the upper sheet $\xi > 0$, and on the lower sheet $\xi < 0$. Also from (2.49) we conclude that

$$\xi^2 - \eta^2 = \zeta^2 + 2\delta > 0 \text{ when } \delta > 0. \tag{32.2.55}$$

It follows from (2.46) and (2.47) that $\beta, \gamma > 0$ on the upper sheet and $\beta, \gamma < 0$ on the lower sheet. Also, on the upper sheet and for a given value of $\delta > 0$, there is a single point closest to the origin; from (2.49) it has the coordinates

$$\eta = \zeta = 0,\ \xi = (2\delta)^{1/2}, \tag{32.2.56}$$

or, equivalently,

$$\beta = \gamma = \delta^{1/2},\ \alpha = 0. \tag{32.2.57}$$

This point corresponds to the normal form given by (2.25). Similarly, on the corresponding lower sheet, there is also a single point closest to the origin with the coordinates

$$\eta = \zeta = 0,\ \xi = -(2\delta)^{1/2}, \tag{32.2.58}$$

or, equivalently,

$$\beta = \gamma = -\delta^{1/2},\ \alpha = 0. \tag{32.2.59}$$

This point corresponds to the normal form given by (2.35). Evidently, for a fixed value of $\delta > 0$, all points on the upper sheet lie on the same orbit, and those on the lower sheet lie on a second distinct orbit. Consequently, there are two equivalence classes for each positive value of δ. The upper sheet gives the case $\beta, \gamma > 0$, and the lower sheet gives the case $\beta, \gamma < 0$.

There remains the case $\delta = 0$ for which, as already mentioned, the relation (2.49) yields two cones. These cones have no point in common save the origin which corresponds to the single-element equivalence class $g_2 = 0$. Moreover, points on the upper and lower cones belong to separate equivalence classes. It is easy to check that β or $\gamma > 0$ on the upper cone and β or $\gamma < 0$ on the lower cone. (Here the origin is to be excluded.) The monomial $+q^2$ given by (2.28) provides a normal form for polynomials corresponding to points on the upper cone. Its coordinates are given by the relations

$$\xi = \eta = 1/\sqrt{2},\ \zeta = 0, \tag{32.2.60}$$

or, equivalently,

$$\alpha = \beta = 0,\ \gamma = 1. \tag{32.2.61}$$

The monomial $-q^2$ given by (2.37) provides a normal form for polynomials corresponding to points on the lower cone. Its coordinates are given by the relations

$$\xi = \eta = -1/\sqrt{2},\ \zeta = 0, \tag{32.2.62}$$

or, equivalently,

$$\alpha = \beta = 0,\ \gamma = -1. \tag{32.2.63}$$

It follows that each cone is a separate orbit.

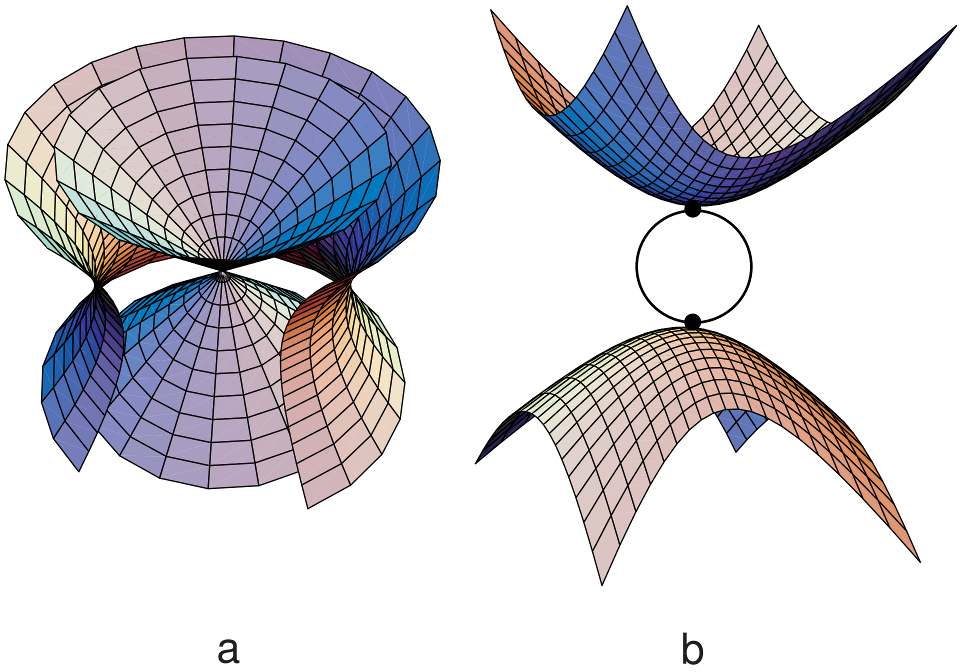


Figure 32.2.2: Equivalence classes (orbits/leaves) for the space $\mathcal{P}_2$ of second-order polynomials in two variables. They are displayed in terms of the variables ξ, η, ζ. In this figure the ξ axis points upward, the η axis points out of the page, and the ζ axis points to the right. Case a, for which $\delta < 0$, shows a typical one-sheeted hyperboloid. The point on the equator given by (2.53) and (2.54) corresponds to the normal form $(-\delta)^{1/2}(q^2 - p^2)$ for that value of δ. Also shown on this diagram are the two cones for $\delta = 0$. The point on the front of the top cone given by ((2.60) and (2.61) corresponds to the normal form $+q^2$ and the point on the rear of the bottom cone given (2.62) and (2.63) by corresponds to the normal form $-q^2$. The origin where the cones meet is the single-element equivalence class $g_2 = 0$. Case b, for which $\delta > 0$, shows a typical two-sheeted hyperboloid. Also shown is the sphere (2.50) that just kisses the hyperboloid. The two kissing points (the points on the upper and lower sheets that are closest to the origin) correspond to the normal forms $\pm\delta^{1/2}(p^2 + q^2)$. For simplicity, cases a and b are shown separately. They should actually be superimposed along with many other such hyperboloids to show all the one-sheeted and two-sheeted hyperboloids for all values of δ.

Observations

Note that generically each equivalence class is two dimensional. However, the origin is zero dimensional. There are four other observations to be drawn from the simple 2-dimensional phase-space example we have been studying.

First, we know that exponentiating Hamiltonian matrices produces symplectic matrices. Therefore the normal form problem for quadratic polynommials is related to the normal form problem for symplectic matrices. For example, exponentiating Hamiltonian matrices corresponding to the quadratic polynomials associated with points on a one-sheeted hyperboloid such as that shown in Figure 2.2a for $\delta < 0$ produces symplectic matrices whose spectrum corresponds to that shown in case 1 of Figure 3.4.1. Also, exponentiating Hamiltonian matrices corresponding to the quadratic polynomials associated with points on either sheet of the two-sheeted hyperboloid such as that shown in Figure 2.2b for $\delta > 0$ produces symplectic matrices whose spectrum corresponds to that shown in case 3 of Figure 3.4.1. In particular, exponentiating the Hamiltonian matrix corresponding to a normal form g_2^N for $\delta > 0$ produces symplectic matrices of the form (3.5.58). Finally, exponentiating Hamiltonian matrices corresponding to the quadratic polynomials associated with points on either cone as shown in Figure 2.2a for $\delta = 0$ produces symplectic matrices whose spectrum corresponds to that shown in case 4 of Figure 3.4.1. Because of the relation between quadratic polynomial and symplectic matrix normal forms, the normal form references cited at the end of Chapter 3 are also relevant to the polynomial case. However, as learned in Section 8.7.2, not every symplectic matrix can be written in single exponential form. Therefore, the classification of symplectic matrices is more complicated than the classification of quadratic polynomials (Hamiltonian matrices).

The second observation is that all quadratic polynomials associated with points on the upper-sheet of the two-sheeted hyperboloid for $\delta > 0$, see Figure 2.2b, are *positive definite.* That is, such polynomials obey $g_2(z) > 0$ for any nonzero z. From this perspective, (2.25) is the normal form for positive-definite quadratic polynomials. Correspondingly, all quadratic polynomials associated with points on the lower sheet of the two-sheeted hyperboloid are *negative definite.* They obey $g_2(z) < 0$ for any nonzero z. Their normal form is given by (2.35).

The third observation is that the normal form problem for quadratic polynomials is identical to that of classifying quadratic Hamiltonians: Given two quadratic Hamiltonians, is there a linear canonical transformation that will send one into the other? Given a quadratic Hamiltonian, how "simple" can it be made by applying a suitable linear canonical transformation? Given a quadratic Hamiltonian, what will be the nature of the motion it generates? For the case of a 2-dimensional phase space we have learned that all quadratic Hamiltonians with $\delta < 0$ can be brought to the form (2.31), and they generate *exponentially unbounded* motion. All nonzero Hamiltonians with $\delta = 0$ can be brought to one of the forms (2.26), (2.36), and they generate *linearly unbounded* motion. All Hamiltonians with $\delta > 0$ can be brought to one of the forms (2.25), (2.35), and they generate *bounded* motion. See Exercise 2.4.

The fourth observation is that we may view the g_2 as elements of the Lie algebra $sp(2)$. From this perspective, we have been studying what elements in the Lie algebra $sp(2)$ can be transformed into each other under the action of the group $Sp(2)$. In the case of a $2n$-dimensional phase space we may view the g_2 as elements of the Lie algebra $sp(2n)$, and studying the action of $Sp(2n)$ on the g_2 is equivalent to studying the action of $Sp(2n)$ on $sp(2n)$. See Exercise 2.6 for a preliminary effort in this direction, mostly devoted to $sp(4)$.

32.2.2.2 Case of Four-Dimensional Phase Space

Having mastered the case of 2-dimensional phase space, we consider the next more complicated case, namely 4-dimensional phase space. Now S takes the general form

$$S = \begin{pmatrix} a & b & c & d \\ b & e & r & s \\ c & r & t & u \\ d & s & u & v \end{pmatrix}. \tag{32.2.64}$$

Invariants

The Hamiltonian matrix $B = JS$ has the characteristic polynomial

$$P(\lambda) = \det(B - \lambda I) = \lambda^4 + C\lambda^2 + D \tag{32.2.65}$$

where

$$C = -b^2 + ae - 2dr + 2cs - u^2 + tv, \tag{32.2.66}$$

$$\begin{aligned} D &= d^2r^2 - 2cdrs + c^2s^2 - d^2et + 2bdst - as^2t \\ &+ 2cdeu - 2bdru - 2bcsu + 2arsu + b^2u^2 \\ &- aeu^2 - c^2ev + 2bcrv - ar^2v - b^2tv + aetv. \end{aligned} \tag{32.2.67}$$

Here we have used the form (3.2.10) for J. Note that $P(\lambda)$ has only even powers of λ as expected for the characteristic polynomial of a Hamiltonian matrix. See Exercise 3.7.14.

It is well known that the coefficients of the characteristic polynomial are invariant under similarity transformations; and, according to (2.14), it is similarity transformations that are being made. Therefore C and D are invariants.

Do they have any interpretation? From (2.65) we have the relation

$$D = \det(B) = \det(JS) = [\det(J)][\det(S)] = \det(S). \tag{32.2.68}$$

Here we have used (3.1.4). It follows that D (modulo sign conventions) is the *discriminant* of the quadratic form g_2. From this perspective the invariance of D follows directly from taking the determinant of both sides of (2.11) and using (3.1.8). Also see Exercise 2.5.

The interpretation of C is a bit more complicated. Since B is traceless (see Exercise 3.7.10), use of (3.7.136), and (3.7.137), and (3.7.143) gives the result

$$C = -(1/2)\ \text{tr}(B^2) = -(1/2)(B, B)_F. \tag{32.2.69}$$

Here we have also used (21.11.15) and employed the subscript F to denote the *fundamental* representation. Since C is constructed from the invariant metric, and the quadratic Casimir operator is also constructed from this metric, by mental association C (as a function of the coefficients in S) is sometimes called the *Casimir* polynomial. From this perspective, the invariance of C is a special case of (21.11.25).

We have seen that there are two invariants for the case of 4-dimensional phase space, namely C and D. Since the set of all symmetric 4×4 matrices S is 10 dimensional, it follows that the orbit space (each equivalence class) is generically 8 dimensional. However, at certain points it has smaller dimension. See Exercise 2.6. In various regions it may also be expected to be multi-sheeted.

Eigenvalues

The eigenvalues of B [the roots of $P(\lambda) = 0$] are of the form

$$\lambda = \pm\sqrt{w} \tag{32.2.70}$$

where w is a root of the equadratic equation

$$w^2 + Cw + D = 0. \tag{32.2.71}$$

That is, w is given by the relation

$$w = [-C \pm (C^2 - 4D)^{1/2}]/2. \tag{32.2.72}$$

Figure 2.3 below shows possible eigenvalue configurations for 4 × 4 real Hamiltonian matrices depending on the values of C and D. Evidently, there are nine cases to be considered:

i. Complex quartet of eigenvalues of the form $\pm\alpha \pm i\beta$.

ii. Two pairs of pure imaginary eigenvalues $\pm i\alpha$ and $\pm i\beta$.

iii. Two pairs of real eigenvalues $\pm\alpha$ and $\pm\beta$.

iv. One pair of real eigenvalues $\pm\alpha$ and one pair of pure imaginary eigenvalues $\pm i\beta$.

v. A pair of repeated real eigenvalues $\pm\alpha$.

vi. A pair of repeated pure imaginary eigenvalues $\pm i\beta$.

vii. A pair of real eigenvalues $\pm\alpha$ and repeated zero eigenvalues $0, 0$.

viii. A pair of imaginary eigenvalues $\pm i\alpha$ and repeated zero eigenvalues $0, 0$.

ix. A quartet of zero eigenvalues $0, 0, 0, 0$.

Cases i through iv are generic, and cases v through ix are degenerate. As usual, transitions between generic configurations can only occur by passage through a degenerate configuration. Note that the results we have found are in accord with those of Exercise 3.7.14. Finally, It is instructive to compare the pair of figures 27.2.1, 3.4.3, the pair of figures 27.2.3, 3.4.4, and the pair of figures 27.2.4, 3.5.1. The first figure in each pair displays the spectrum (eigenvalues) of elements in the Lie algebra $sp(2n, \mathbb{R})$, and the second displays the spectrum of elements in the associated Lie group $Sp(2n, \mathbb{R})$. See also Figures 3.4.1 and 3.4.2.

Normal Forms

For each of the cases i through ix there is a corresponding normal form. They are listed below. Note that cases i, v, vi, and ix are special in that the two degrees of freedom cannot be uncoupled by a suitable choice of coordinates. In all other cases the normal form (when viewed as a Hamiltonian) is a sum of two terms involving different degrees of freedom, and therefore the two terms are in involution. But here is an amazing thing: In each of the

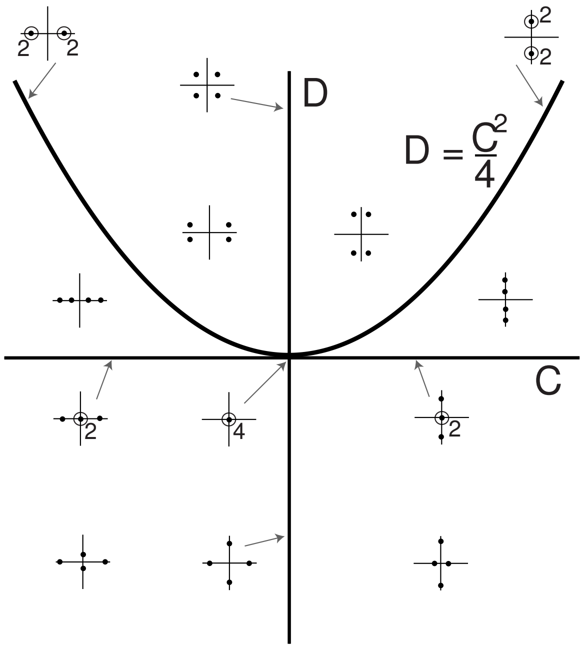


Figure 32.2.3: Eigenvalues of a 4×4 real Hamiltonian matrix as a function of the coefficients C and D in its characteristic polynomial.

coupled cases i, v, vi, and ix the normal form g_2^N is also the sum of two terms, and it can be verified in each case that the two terms are also in involution.

Case i:

$$\begin{aligned}&\lambda = \pm\alpha \pm i\beta, \text{ all signs taken independently, with } \alpha, \beta > 0,\\ &C = -2\alpha^2 + 2\beta^2,\\ &D = (\alpha^2 + \beta^2)^2,\\ &g_2^N = 2\alpha(q_1p_1 + q_2p_2) + 2\beta(q_1p_2 - q_2p_1).\end{aligned} \tag{32.2.73}$$

Case ii:

$$\begin{aligned}&\lambda = \pm i\alpha \text{ and } \pm i\beta, \text{ all signs taken independently, with } \alpha, \beta > 0,\\ &C = \alpha^2 + \beta^2,\\ &D = \alpha^2\beta^2,\\ &g_2^N = \pm\alpha(p_1^2 + q_1^2) \pm \beta(p_2^2 + q_2^2), \text{ all signs taken independently.}\end{aligned} \tag{32.2.74}$$

Case iii:

$$\begin{aligned}&\lambda = \pm\alpha \text{ and } \pm \beta, \text{ all signs taken independently, with } \alpha, \beta > 0,\\ &C = -\alpha^2 - \beta^2,\\ &D = \alpha^2\beta^2,\\ &g_2^N = 2\alpha q_1p_1 + 2\beta q_2p_2.\end{aligned} \tag{32.2.75}$$

Case *iv*:

$$\begin{aligned}&\lambda = \pm\alpha \text{ and } \pm i\beta, \text{ all signs taken independently, with } \alpha, \beta > 0,\\ &C = -\alpha^2 + \beta^2,\\ &D = -\alpha^2\beta^2,\\ &g_2^N = 2\alpha p_1 q_1 \pm \beta(p_2^2 + q_2^2).\end{aligned} \tag{32.2.76}$$

Case *v*:

$$\begin{aligned}&\lambda = \text{ repeated pair } \pm\alpha \text{ with } \alpha > 0,\\ &C = -2\alpha^2,\\ &D = \alpha^4,\\ &g_2^N = 2\alpha(q_1p_1 + q_2p_2) + 2q_1p_2 \text{ or case } iii \text{ with } \alpha = \beta.\end{aligned} \tag{32.2.77}$$

Case *vi*:

$$\begin{aligned}&\lambda = \text{ repeated pair } \pm i\beta \text{ with } \beta > 0,\\ &C = 2\beta^2,\\ &D = \beta^4,\\ &g_2^N = 2\beta(q_1p_2 - q_2p_1) \pm (q_1^2 + q_2^2) \text{ or case } ii \text{ with } \alpha = \beta.\end{aligned} \tag{32.2.78}$$

Case *vii*:

$$\begin{aligned}&\lambda = \pm\alpha \text{ and } 0, 0 \text{ with } \alpha > 0,\\ &C = -\alpha^2,\\ &D = 0,\\ &g_2^N = 2\alpha q_1p_1 \pm q_2^2 \text{ or case } iii \text{ with } \beta = 0.\end{aligned} \tag{32.2.79}$$

Case *viii*:

$$\begin{aligned}&\lambda = \pm i\alpha \text{ and } 0, 0 \text{ with } \alpha > 0,\\ &C = \alpha^2,\\ &D = 0,\\ &g_2^N = \pm\alpha(q_1^2 + p_1^2) \pm q_2^2 \text{ or case } ii \text{ with } \beta = 0.\end{aligned} \tag{32.2.80}$$

Case *ix*:

$$\begin{aligned}&\lambda = 0, 0, 0, 0,\\ &C = 0,\\ &D = 0,\end{aligned}$$

$$g_2^N = 2q_1p_2 \pm q_2^2, \text{ or} \tag{32.2.81}$$

$$g_2^N = 2q_1p_2, \text{ or} \tag{32.2.82}$$

$$g_2^N = \pm q_1^2 \pm q_2^2, \text{ all signs taken independently, or} \tag{32.2.83}$$

$$g_2^N = \pm q_1^2. \tag{32.2.84}$$

Observations

We are ready for four more observations. The first is that, up to ± signs and degeneracy complications that depend on the Jordan block structure of B, the normal form depends primarily on the eigenvalue spectrum. And, since the spectrum depends only on the invariants C and D, it follows that the normal form depends primarily on the values of C and D. Thus, as before, a necessary condition for $h_2 \sim g_2$ is that they have the same invariants.

The second observation is that, unlike the 2-dimensional case, we have not exhibited the transformation $\mathcal{R}$ that brings a g_2 to its normal form g_2^N. In general this must be done numerically, and enough is known about the problem to write computer programs for this purpose. Exercise 2.7 treats an interesting example where some of the required mathematical concepts are illustrated.

The third observation is that we have not exhibited the geometric nature of the various equivalence classes and their representative normal forms in analogy to Figure 2.2. To do so is difficult because, as already described earlier, we need to study 8-dimensional surfaces embedded in a 10-dimensional Euclidean space. See Exercise 2.6. Also, as indicted by the ± signs in cases *ii*, *iv*, and *vi* through *ix*, there is a multi-sheeted structure in parts of the space. However, in the generic situation when the eigenvalues are distinct (no degeneracy), there is one attribute of the normal forms that we can verify without too much effort. As Figure 2.2 illustrates, in the 2-dimensional generic case there is no point on a given orbit that is closer to the origin then the normal form. We may guess that the same is true in the $2n$-dimensional case. What we can easily check is that a g_2^N is at least a local minimum of $\langle g_2^{\rm tr}, g_2^{\rm tr}\rangle$.

Consider elements $g_2^{\rm tr}$ of the form

$$g_2^{\rm tr} = \mathcal{R} g_2^N \tag{32.2.85}$$

with $\mathcal{R}$ given by (1.1). These are just the elements of the equivalence class $\{g_2^N\}$. From (7.3.29) we know that transformations of the form $\exp(: f_2^c :)$ do not change the distance of an element from the origin. Therefore, we may restrict our attention to transformations of the form

$$\mathcal{R}_\epsilon = \exp(\epsilon : f_2^a :) \tag{32.2.86}$$

where, for convenience, we have explicitly included a scaling parameter ϵ. Let us define elements g_2^ϵ by the relation

$$g_2^\epsilon = \mathcal{R}_\epsilon g_2^N. \tag{32.2.87}$$

Then we find the results

$$\begin{aligned}\langle g_2^\epsilon, g_2^\epsilon\rangle &= \langle \mathcal{R}_\epsilon g_2^N, \mathcal{R}_\epsilon g_2^N\rangle = \langle g_2^N, \mathcal{R}_\epsilon^\dagger \mathcal{R}_\epsilon g_2^N\rangle = \langle g_2^N, \mathcal{R}_\epsilon^2 g_2^N\rangle = \langle g_2^N, \exp(2\epsilon : f_2^a :) g_2^N\rangle \\ &= \langle g_2^N, g_2^N\rangle + 2\epsilon\langle g_2^N, : f_2^a : g_2^N\rangle + (4\epsilon^2/2!)\langle g_2^N, : f_2^a :^2 g_2^N\rangle + O(\epsilon^3).\end{aligned} \tag{32.2.88}$$

Here we have used (7.3.30), which states that $: f_2^a :$ is Hermitian. Because $: f_2^a :$ is Hermitian, we may also write

$$\langle g_2^N, : f_2^a :^2 g_2^N\rangle = \langle : f_2^a : g_2^N, : f_2^a : g_2^N\rangle \geq 0. \tag{32.2.89}$$

It follows from (2.88) that g_2^N is at least a local minimum if we have the relation

$$\langle g_2^N, : f_2^a : g_2^N\rangle = 0. \tag{32.2.90}$$

It can be verified by explicit calculation that (2.90) holds for all the normal forms in the generic cases *i* through *iv*. Consider, for example, case *i*. Then, using (2.73) and the notation of Section 5.5.7, we may write

$$g_2^N = -2\alpha f^2 - 2\beta b^2, \tag{32.2.91}$$

$$f_2^a = \phi_1 f^1 + \phi_2 f^2 + \phi_3 f^3 + \gamma_1 g^1 + \gamma_2 g^2 + \gamma_3 g^3. \tag{32.2.92}$$

Here the ϕ_j and γ_j are arbitrary parameters. See (5.7.4) and (5.7.31). From the Poisson bracket rules (5.7.32) through (5.7.37) we find the result

$$\begin{aligned} : f_2^a : g_2^N &= [f_2^a, g_2^N] = -[g_2^N, f_2^a] \\ &= 4\beta\phi_1 f^3 - 4\beta\phi_3 f^1 + 4\beta\gamma_1 g^3 - 4\beta\gamma_3 g^1 \\ &- 4\alpha\phi_1 b^3 + 4\alpha\phi_3 b^1 + 4\alpha\gamma_2 b^0. \end{aligned} \tag{32.2.93}$$

Finally, all the terms on the right of (2.91) are orthogonal to all the terms on the right of (2.93). See Exercise 7.3.8. It follows that (2.90) is true, and therefore g_2^N is indeed at least a local minimum.

32.2.2.3 Case of General $2n$-Dimensional Phase Space

For the general case of $2n$ dimensional phase space normal-form results are also fully known, but considerably more complicated. However, based on our experience with the two and four dimensional cases, we are prepared for some statements about the general $2n$-dimensional case. The first is that nothing new, beyond what has already been seen for the 4-dimensional case, happens in the general case providing all the eigenvalues are distinct (as is generically true). When the eigenvalues are distinct (no degeneracy), the normal form g_2^N can always be taken to be a sum of terms of the form (2.73) through (2.76). Also, if zero occurs only as a doubly repeated eigenvalue, then the normal form g_2^N can be taken to be a sum of terms of the form (2.73) through (2.76) possibly augmented by a term of the form $\pm q_j^2$, as occurs for example in (2.79) and (2.80).

The second observation is that, as is already clear in the 4-dimensional case, repeated (degenerate) eigenvalues can cause complications. When the eigenvalues are degenerate there is the possibility of having the Hamiltonian analog of Jordan blocks. Fortunately, these complications are completely understood in the general case, and detailed results can be found in the literature. Moreover, it is important to note that there are two common cases where degeneracy causes no problems. In the first case, suppose we are working with a Hamiltonian that can be written in the form $h_2 = T_2(p) + V_2(q)$ where the kinetic energy term T_2 is known to be positive definite. Then standard normal mode theory shows that the Hamiltonian can always be diagonalized by a linear canonical transformation [$Sp(2n)$ element] even if some or all eigenvalue pairs are degenerate. Second, suppose that g_2 is known to be positive definite. Then g_2^N can always be taken to be a sum of harmonic oscillators, for example as in (2.74), with all signs positive even if some or all eigenvalue pairs are degenerate. This result will be proved and used in Chapter 33.

Exercises

32.2.1. Verify the equivalence relation (2.27). Find an $\mathcal{R}$ such that $\mathcal{R}q = p$ and $\mathcal{R}p = -q$.

32.2.2. Study cases *iii* and *iv* for 2-dimensional phase space and show that your results are identical to those obtained for cases *i* and *ii*. In particular, exhibit transforming matrices M analogous to (2.21) and (2.32). For example, show that for $\gamma > 0$ one may use the matrix

$$M = \begin{pmatrix} 1/\sqrt{\gamma} & -\alpha/\sqrt{\gamma} \\ 0 & \sqrt{\gamma} \end{pmatrix}. \tag{32.2.94}$$

32.2.3. Verify (2.50).

32.2.4. Solve Hamilton's equations of motion using the normal forms (2.25), (2.35), (2.26), (2.36), and (2.31) as Hamiltonians.

32.2.5. Equation (2.69) shows that the invariant C can be expressed in terms of the Lie element B and various manifestly invariant trace operations. What can be said about the discriminant invariants δ and D?

Use (3.7.115) to show that

$$\delta = \det(B) = -(1/2)\ \mathrm{tr}(B^2) = -(1/2)(B, B)_F \tag{32.2.95}$$

in the 2×2 case. (Recall that B is traceless.) Observe that *stability* is determined by the value of $(B, B)_F$. Show, by examining Figure 2.1, that stability occurs when $(B, B)_F < 0$.

Use (3.7.117) to show that

$$D = \det(B) = (1/8)[\mathrm{tr}(B^2)]^2 - (1/4)\ \mathrm{tr}(B^4) = (1/2)C^2 - (1/4)\ \mathrm{tr}(B^4) \tag{32.2.96}$$

in the 4×4 case. Show, by employing (2.69) and examining Figure 2.3, that, unlike the 2×2 case, knowledge of more that $(B, B)_F$ is required to determine stability.

In view of the ingredients of (2.95) and (2.96), δ and D could also be called Casimir polynomials.

32.2.6. This Exercise studies the dimensionality of the orbits $\{g_2^N\}$ for various normal forms g_2^N. Suppose g_2^N is some normal form. Then the orbit of g_2^N consists of all elements of the form

$$g_2^{\mathrm{tr}} = \mathcal{R} g_2^N. \tag{32.2.97}$$

Since $\mathcal{R}$ is a continuous and invertible mapping (a homeomorphism) of $\mathcal{P}_2$ into itself, the dimensionality of $\{g_2^N\}$ will be the same at every point on the orbit, and it suffices to determine the dimensionality in the vicinity of g_2^N. In this case we may take $\mathcal{R}$ to be near the identity, which means that it can be written in the form

$$\mathcal{R} = \exp(\epsilon : f_2 :) \tag{32.2.98}$$

for some small, but finite, ϵ. Correspondingly we may rewrite (2.94) in the form

$$g_2^{\mathrm{tr}} = \exp(\epsilon : f_2 :) g_2^N = g_2^N + \epsilon : f_2 : g_2^N + O(\epsilon^2). \tag{32.2.99}$$

Thus, infinitesimally, the dimensionality of $\{g_2^N\}$ is given by the number of linearly independent elements produced by terms of the form $:f_2:g_2^N$ for arbitrary choices of f_2. Note that there is the relation

$$:f_2:g_2^N=[f_2,g_2^N]=-[g_2^N,f_2]=-:g_2^N:f_2. \tag{32.2.100}$$

All terms of the form $(:g_2^N:f_2)$ for arbitrary f_2 comprise what is called the *range* (in $\mathcal{P}_2$) of the operator $:g_2^N:$. It is easy to check that the range of a linear operator is a linear vector space. (Check it!) Consequently, we may also say that the dimensionality of $\{g_2^N\}$ as a manifold is equal to the linear vector space dimensionality of the range of $:g_2^N:$.

Now carry out the calculations described below:

a) Suppose g_2^N is given by (2.73) as in case i. We are going to study the range of this $:g_2^N:$ in $\mathcal{P}_2$. Let λ^j be the eigenvalues for this case labeled by the scheme

$$\lambda^1=\alpha+i\beta, \tag{32.2.101}$$

$$\lambda^2=-\alpha+i\beta, \tag{32.2.102}$$

$$\lambda^3=-\alpha-i\beta, \tag{32.2.103}$$

$$\lambda^4=\alpha-i\beta. \tag{32.2.104}$$

Let g_1^j be the first-order polynomials defined by the relations

$$g_1^1=p_1-ip_2, \tag{32.2.105}$$

$$g_1^2=q_1-iq_2, \tag{32.2.106}$$

$$g_1^3=q_1+iq_2, \tag{32.2.107}$$

$$g_1^4=p_1+ip_2. \tag{32.2.108}$$

Evidently they are linearly independent. Verify the eigen relations

$$:g_2^N:g_1^j=2\lambda^j g_1^j. \tag{32.2.109}$$

Next let g_2^{jk} be the second-order polynomials defined by the relations

$$g_2^{jk}=g_1^j g_1^k. \tag{32.2.110}$$

Show that there are 10 such polynomials (since there is symmetry in the j,k indices), and that they are all linearly independent. They consequently form a basis for $\mathcal{P}_2$ [and, therefore, also for $sp(4)$] in the case of a four-dimensional phase space. Show that these polynomials satisfy the eigen relations

$$:g_2^N:g_2^{jk}=2(\lambda^j+\lambda^k)g_2^{jk}. \tag{32.2.111}$$

Thus, $:g_2^N:$ is diagonal in this basis. Verify that g_2^{13} and g_2^{24} are eigenvectors with eigenvalue zero, and verify that all the other eight eigenvectors have nonzero eigenvalues.

The g_2^{jk} form a complex basis for $\mathcal{P}_2$. Verify that a real basis is provided by the elements

$$g_2^1 = \operatorname{Re}(g_2^{11}) = \operatorname{Re}(g_2^{44}) = p_1^2 - p_2^2, \tag{32.2.112}$$

$$g_2^2 = \operatorname{Im}(g_2^{11}) = -\operatorname{Im}(g_2^{44}) = -2p_1p_2, \tag{32.2.113}$$

$$g_2^3 = \operatorname{Re}(g_2^{12}) = \operatorname{Re}(g_2^{34}) = q_1p_1 - q_2p_2, \tag{32.2.114}$$

$$g_2^4 = \operatorname{Im}(g_2^{12}) = -\operatorname{Im}(g_2^{34}) = -q_1p_2 - q_2p_1, \tag{32.2.115}$$

$$g_2^5 = \operatorname{Re}(g_2^{13}) = \operatorname{Re}(g_2^{24}) = q_1p_1 + q_2p_2, \tag{32.2.116}$$

$$g_2^6 = \operatorname{Im}(g_2^{13}) = -\operatorname{Im}(g_2^{24}) = q_2p_1 - q_1p_2, \tag{32.2.117}$$

$$g_2^7 = g_2^{14} = p_1^2 + p_2^2, \tag{32.2.118}$$

$$g_2^8 = \operatorname{Re}(g_2^{22}) = \operatorname{Re}(g_2^{33}) = q_1^2 - q_2^2, \tag{32.2.119}$$

$$g_2^9 = \operatorname{Im}(g_2^{22}) = -\operatorname{Im}(g_2^{33}) = -2q_1q_2, \tag{32.2.120}$$

$$g_2^{10} = g_2^{23} = q_1^2 + q_2^2. \tag{32.2.121}$$

Indeed, verify that these elements are mutually orthogonal for the inner product of Section 7.3.

Using this basis, write an arbitrary f_2 in the form

$$f_2 = \sum_{j=1}^{10} a_j g_2^j \tag{32.2.122}$$

where the a_j are arbitrary coefficients. Verify the relation

$$\begin{aligned} :g_2^N: f_2 &= 4(\alpha a_1 + \beta a_2)g_2^1 + 4(-\beta a_1 + \alpha a_2)g_2^2 + 4\beta a_4 g_2^3 - 4\beta a_3 g_2^4 \\ &+ 4\alpha a_7 g_2^7 + 4(-\alpha a_8 + \beta a_9)g_2^8 + 4(-\beta a_8 - \alpha a_9)g_2^9 \\ &- 4\alpha a_{10} g_2^{10}. \end{aligned} \tag{32.2.123}$$

Observe that all the g_2^j except g_2^5 and g_2^6 appear on the right side of (2.123). Therefore, the range of $:g_2^N:$ is potentially 8 dimensional. To be sure we must show that any linear combination of g_2^1, g_2^2, g_2^3, g_2^4, and g_2^7, g_2^8, g_2^9, g_2^{10} can be obtained on the right side of (2.123) for a suitable choice of the a_j in (2.122). Evidently there are no problems with g_2^3, g_2^4, g_2^7, and g_2^{10} since their coefficients are simply $4\beta a_4$, $-4\beta a_3$, $4\alpha a_7$, and $-4\alpha a_{10}$, respectively, and a_3, a_4, a_7, and a_{10} appear nowhere else on the right side of (2.123). For the coefficients of g_2^1 and g_2^2 we may write the matrix relation

$$\begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \alpha a_1 + \beta a_2 \\ -\beta a_1 + \alpha a_2 \end{pmatrix}. \tag{32.2.124}$$

The determinant of the matrix appearing on the left side of (2.124) has the value $(\alpha^2 + \beta^2)$. Therefore the matrix is always invertible, and we can achieve any desired combination of g_2^1 and g_2^2 on the right side of (2.123). Show that a similar argument holds for the coefficients of g_2^8 and g_2^9. We conclude that the range of $:g_2^N:$ is indeed 8 dimensional. Correspondingly, $\{g_2^N\}$ is 8 dimensional.

For any g_2^N, consider the set of all polynomials h_2^0 such that

$$: g_2^N : h_2^0 = 0. \tag{32.2.125}$$

They comprise what is called the *null space* or *kernel* (in $\mathcal{P}_2$) of $: g_2^N :$. Show that these polynomials form a Lie subalgebra with the Poisson bracket as a Lie product. (Hint: use the Jacobi identity.) Show, for the g_2^N given by (2.73), that this Lie algebra is two dimensional and is spanned by the elements $(q_1p_1 + q_2p_2)$ and $(q_1p_2 - q_2p_1)$, which are the real and imaginary parts of g_2^{13} and g_2^{24}. [Hint: Write h_2^0 in the form (2.122) and use (2.123).] Verify that g_2^N is constructed from them and that they are in involution.

Let H be the corresponding subgroup of $Sp(4)$ generated by the $: h_2^0 :$ and consider the coset space $Sp(4)/H$. Show that in some finite neighborhood of the identity any element in $\mathcal{R}$ can be written in the factored form

$$\mathcal{R} = \exp(: h_2^R :) \exp(: h_2^0 :) \tag{32.2.126}$$

where h_2^R is in the range of $: g_2^N :$ and h_2^0 is in the null space of $: g_2^N :$. The factor $\exp(: h_2^0 :)$ corresponds to an element in H, and the factor $\exp(: h_2^R :)$ corresponds to some element in the coset $Sp(4)/H$. From the relation

$$: h_2^0 : g_2^N = [h_2^0, g_2^N] = -[g_2^N, h_2^0] = - : g_2^N : h_2^0 = 0 \tag{32.2.127}$$

show that

$$\exp(: h_2^0 :) g_2^N = g_2^N \tag{32.2.128}$$

and

$$g_2^{\rm tr} = \mathcal{R} g_2^N = \exp(: h_2^R :) g_2^N. \tag{32.2.129}$$

This construction goes beyond the infinitesimal, and shows that the orbit $\{g_2^N\}$ may be identified with the coset space $Sp(4)/H$. Evidently, an analogous result holds for any phase space dimension. In the terminology of Section 5.12, $\{g_2^N\}$ is a homogeneous space and H is the stability group for g_2^N.

b) Carry out similar calculations for the remaining generic cases *ii* through *iv*. You should find that $\{g_2^N\}$ is 8 dimensional in each case.

c) As an example of a degenerate case, consider the specific subcase of case *ix* for which g_2^N is given by the relation

$$g_2^N = q_1^2. \tag{32.2.130}$$

Following the notation of (2.64), let us write an arbitrary f_2 in the form

$$\begin{aligned} f_2 &= aq_1^2 + 2bq_1p_1 + 2cq_1q_2 + 2dq_1p_2 + ep_1^2 \\ &+ 2rp_1q_2 + 2sp_1p_2 + tq_2^2 + 2uq_2p_2 + vp_2^2. \end{aligned} \tag{32.2.131}$$

Show that

$$: g_2^N : f_2 = 4bq_1^2 + 4eq_1p_1 + 4rq_1q_2 + 4sq_1p_2. \tag{32.2.132}$$

Since all the monomials on the right of (2.132) are linearly independent, inspection indicates that the range of $: g_2^N :$ is 4 dimensional. Corresponding, $\{g_2^N\}$ is 4 dimensional. Since $\mathcal{P}_2$ is 10 dimensional, in this case there must be 4 more invariants in addition to C and D.

Show that the condition $: g_2^N : h_2 = 0$, with h_2 written in the form (2.131), requires that $b = e = r = s = 0$. Therefore, the null space of $: g_2^N :$ is 6 dimensional. Verify, as expected, that the null space is a Lie subalgebra.

Show that

$$: g_2^N :^2 z_a = 0, \tag{32.2.133}$$

which implies that

$$B^2 = 0. \tag{32.2.134}$$

We say that B is *nilpotent*. Correspondingly, $\{g_2^N\}$ in this case is called a nilpotent orbit. Show that if g_2^{tr} is any element lying on a nilpotent orbit, it must satisfy the relation

$$: g_2^{\text{tr}} :^2 z_a = 0, \tag{32.2.135}$$

and conversely.

Show that the other normal forms of case *ix* also produce nilpotent orbits, and find their dimensions.

Show that each of the ladder elements $\tilde{r}(\boldsymbol{\mu})$ given by (21.5.11) through (21.5.18) lies on a nilpotent orbit. Show that any linear combination of ladder elements of the form $[a_\alpha \tilde{r}(\boldsymbol{\alpha}) + a_\beta \tilde{r}(\boldsymbol{\beta}) + a_\gamma \tilde{r}(\boldsymbol{\gamma})]$ lies on a nilpotent orbit. Are there other linear combinations of ladder operators that lie on nilpotent orbits? Represent a general element g_2 in $sp(4)$ as a linear combination of the basis elements given by (21.5.9) through (21.5.18). What are the necessary and sufficient conditions on the expansion coefficients for g_2 to lie on a nilpotent orbit?

d) Determine the dimension of $\{g_2^N\}$ for some of the other normal forms in cases *v* through *viii*. For example, show for

$$g_2^N = 2\alpha q_1 p_1 \tag{32.2.136}$$

that $\{g_2^N\}$ is 6 dimensional, and that the null space of $: g_2^N :$ has dimension 4. Is this $\{g_2^N\}$ nilpotent?

32.2.7. Consider the motion of a nonrelativistic particle of rest mass m and charge q in a uniform magnetic field $\boldsymbol{B}$ with

$$\boldsymbol{B} = \tilde{B}\boldsymbol{e}_z. \tag{32.2.137}$$

Show that this field can be generated by the vector potential

$$\boldsymbol{A} = -(1/2)(\boldsymbol{r} \times \boldsymbol{B}) = (\tilde{B}/2)(x\boldsymbol{e}_y - y\boldsymbol{e}_x). \tag{32.2.138}$$

[This choice of vector potential for $\boldsymbol{B}$ is sometimes called the *symmetric* gauge. In view of the facts that $\nabla \cdot \boldsymbol{A} = 0$ and $\boldsymbol{r} \cdot \boldsymbol{A} = 0$, this choice can also be called the Poincaré-Coulomb gauge. See Section 15.2.4 and (16.1.14).] Show that motion in this field is governed by the Hamiltonian

$$\begin{aligned} H &= (p_x + q\tilde{B}y/2)^2/(2m) + (p_y - q\tilde{B}x/2)^2/(2m) + p_z^2/(2m) \\ &= (p_x^2 + p_y^2)/(2m) + [q^2\tilde{B}^2/(8m)](x^2 + y^2) - [q\tilde{B}/(2m)](xp_y - yp_x) + p_z^2/(2m) \\ &= (p_x^2 + p_y^2)/(2m) + [q^2\tilde{B}^2/(8m)](x^2 + y^2) - [q\tilde{B}/(2m)]L_z + p_z^2/(2m). \end{aligned} \tag{32.2.139}$$

Here $L_z = xp_y - yp_x$ is the z component of the canonical angular momentum. Verify that L_z is an integral of motion.

Evidently the motion in the z direction is uncoupled from the motion in the x, y plane, and we can devote our attention to the latter. Show that this motion is governed by the Hamiltonian

$$H_{xy} = (z, Sz) \tag{32.2.140}$$

where the symbol z now stands for the phase-space variables $z = (x, p_x, y, p_y)$ and S is the symmetric matrix

$$S = \begin{pmatrix} q^2\tilde{B}^2/(8m) & 0 & 0 & -q\tilde{B}/(4m) \\ 0 & 1/(2m) & q\tilde{B}/(4m) & 0 \\ 0 & q\tilde{B}/(4m) & q^2\tilde{B}^2/(8m) & 0 \\ -q\tilde{B}/(4m) & 0 & 0 & 1/(2m) \end{pmatrix}. \tag{32.2.141}$$

Show that the corresponding Hamiltonian matrix $B = JS$ is given by the relation

$$B = \begin{pmatrix} 0 & 1/(2m) & q\tilde{B}/(4m) & 0 \\ -q^2\tilde{B}^2/(8m) & 0 & 0 & q\tilde{B}/(4m) \\ -q\tilde{B}/(4m) & 0 & 0 & 1/(2m) \\ 0 & -q\tilde{B}/(4m) & -q^2\tilde{B}^2/(8m) & 0 \end{pmatrix}. \tag{32.2.142}$$

Show that for this B the invariants C and D have the values

$$C = [q\tilde{B}/(2m)]^2 \ , \ D = 0. \tag{32.2.143}$$

Show that these invariant values correspond to case *viii* or case *ii* with $\beta = 0$, and that the eigenvalues of B are

$$\lambda = \pm iq\tilde{B}/(2m) \ , \ 0 \ , \ 0. \tag{32.2.144}$$

We now want to find the transformation $\mathcal{R}$ that brings H_{xy} to its normal form.

Introduce the notation

$$\lambda_\pm = \pm iq\tilde{B}/(2m). \tag{32.2.145}$$

Show that the vectors $w_\pm$ given by

$$w_\pm = \begin{pmatrix} 1 \\ \pm iq\tilde{B}/2 \\ \pm i \\ -q\tilde{B}/2 \end{pmatrix} \tag{32.2.146}$$

satisfy the relations

$$w_{\mp} = \overline{w}_{\pm}, \tag{32.2.147}$$

and are eigenvectors of B with eigenvalues $\lambda_{\pm}$,

$$Bw_{\pm} = \lambda_{\pm} w_{\pm}. \tag{32.2.148}$$

Show that the vectors r and s given by

$$r = \begin{pmatrix} 0 \\ -q\tilde{B}/2 \\ 1 \\ 0 \end{pmatrix}, \tag{32.2.149}$$

$$s = \begin{pmatrix} 1 \\ 0 \\ 0 \\ q\tilde{B}/2 \end{pmatrix}, \tag{32.2.150}$$

are eigenvectors of B with eigenvalue 0,

$$Br = 0\ ,\ Bs = 0. \tag{32.2.151}$$

Define scaled vectors $\hat{u}$, $\hat{v}$, $\hat{r}$, $\hat{s}$ by the relation

$$\hat{u} = (q\tilde{B})^{-1/2}\ \text{Re}(w_+) = \begin{pmatrix} (q\tilde{B})^{-1/2} \\ 0 \\ 0 \\ -(q\tilde{B})^{1/2}/2 \end{pmatrix}, \tag{32.2.152}$$

$$\hat{v} = (q\tilde{B})^{-1/2}\ \text{Im}(w_+) = \begin{pmatrix} 0 \\ (q\tilde{B})^{1/2}/2 \\ (q\tilde{B})^{-1/2} \\ 0 \end{pmatrix}, \tag{32.2.153}$$

$$\hat{r} = (q\tilde{B})^{-1/2} r = \begin{pmatrix} 0 \\ -(q\tilde{B})^{1/2}/2 \\ (q\tilde{B})^{-1/2} \\ 0 \end{pmatrix}, \tag{32.2.154}$$

$$\hat{s} = (q\tilde{B})^{-1/2} s = \begin{pmatrix} (q\tilde{B})^{-1/2} \\ 0 \\ 0 \\ (q\tilde{B})^{1/2}/2 \end{pmatrix}. \tag{32.2.155}$$

Show that these vectors obey the “symplectic” orthonormality conditions

$$(\hat{u}, J\hat{v}) = 1, \tag{32.2.156}$$

$$(\hat{r}, J\hat{s}) = 1, \tag{32.2.157}$$

$$(\hat{u}, J\hat{r}) = (\hat{v}, J\hat{r}) = (\hat{u}, J\hat{s}) = (\hat{v}, J\hat{s}) = 0. \tag{32.2.158}$$

See Sections 3.5 and 4.6 for similar constructions.

Let R be the matrix defined by the relation

$$R = (\hat{u}, \hat{v}, \hat{r}, \hat{s}) = \begin{pmatrix} (q\tilde{B})^{-1/2} & 0 & 0 & (q\tilde{B})^{-1/2} \\ 0 & (q\tilde{B})^{1/2}/2 & -(q\tilde{B})^{1/2}/2 & 0 \\ 0 & (q\tilde{B})^{-1/2} & (q\tilde{B})^{-1/2} & 0 \\ -(q\tilde{B})^{1/2}/2 & 0 & 0 & (q\tilde{B})^{1/2}/2 \end{pmatrix}. \tag{32.2.159}$$

Here each vector $\hat{u}$, $\hat{v}$, $\hat{r}$, $\hat{s}$ is to be viewed as a column vector so that the collection (2.159) forms a real 4×4 matrix. Show that R, as a consequence of (2.156) through (2.158), is a symplectic matrix. Verify the relations

$$B(\hat{u} \pm i\hat{v}) = \pm i[q\tilde{B}/(2m)](\hat{u} \pm i\hat{v}), \tag{32.2.160}$$

$$B\hat{u} = -[q\tilde{B}/(2m)]\hat{v}, \tag{32.2.161}$$

$$B\hat{v} = [q\tilde{B}/(2m)]\hat{u}. \tag{32.2.162}$$

Show that

$$\begin{aligned} BR &= (B\hat{u}, B\hat{v}, B\hat{r}, B\hat{s}) = (-[q\tilde{B}/(2m)]\hat{v}, [q\tilde{B}/(2m)]\hat{u}, 0, 0) \\ &= \begin{pmatrix} 0 & (q\tilde{B})^{1/2}/(2m) & 0 & 0 \\ -(q\tilde{B})^{3/2}/(4m) & 0 & 0 & 0 \\ -(q\tilde{B})^{1/2}/(2m) & 0 & 0 & 0 \\ 0 & -(q\tilde{B})^{3/2}/(4m) & 0 & 0 \end{pmatrix} \\ &= R \begin{pmatrix} 0 & q\tilde{B}/(2m) & 0 & 0 \\ -q\tilde{B}/(2m) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ &= RJ \begin{pmatrix} q\tilde{B}/(2m) & 0 & 0 & 0 \\ 0 & q\tilde{B}/(2m) & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \end{aligned} \tag{32.2.163}$$

Show that

$$B^{\text{tr}} = R^{-1}BR = \begin{pmatrix} 0 & q\tilde{B}/(2m) & 0 & 0 \\ -q\tilde{B}/(2m) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{32.2.164}$$

and

$$S^{\text{tr}} = J^{-1}B^{\text{tr}} = R^T SR = \begin{pmatrix} q\tilde{B}/(2m) & 0 & 0 & 0 \\ 0 & q\tilde{B}/(2m) & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{32.2.165}$$

It follows that H_{xy} belongs to case *ii* with $\beta = 0$, and has the normal form

$$H_{xy}^N = (\omega/2)(p_1^2 + q_1^2) \tag{32.2.166}$$

where ω is the *cyclotron* frequency,

$$\omega = (q\tilde{B}/m). \tag{32.2.167}$$

Note that H_{xy}^N does not depend on q_2, p_2 at all!

Let q_1^i, p_1^i, q_2^i, p_2^i be *initial* conditions at $t = 0$. For these initial conditions show that the Hamiltonian H_{yt}^N generates the trajectory

$$q_1(t) = q_1^i \cos(\omega t) + p_1^i \sin(\omega t), \tag{32.2.168}$$

$$p_1(t) = -q_1^i \sin(\omega t) + p_1^i \cos(\omega t), \tag{32.2.169}$$

$$q_2(t) = q_2^i, \tag{32.2.170}$$

$$p_2(t) = p_2^i. \tag{32.2.171}$$

Show that the old and new variables are related by the equation

$$\begin{pmatrix} x \\ p_x \\ y \\ p_y \end{pmatrix} = R \begin{pmatrix} q_1 \\ p_1 \\ q_2 \\ p_2 \end{pmatrix}, \tag{32.2.172}$$

and therefore the trajectory in the original phase-space variables is given by the equations

$$\begin{aligned} x(t) &= (q\tilde{B})^{-1/2}[q_1(t) + p_2(t)] \\ &= (q\tilde{B})^{-1/2}[q_1^i \cos(\omega t) + p_1^i \sin(\omega t) + p_2^i], \end{aligned} \tag{32.2.173}$$

$$\begin{aligned} p_x(t) &= [(q\tilde{B})^{1/2}/2][p_1(t) - q_2(t)] \\ &= [(q\tilde{B})^{1/2}/2][-q_1^i \sin(\omega t) + p_1^i \cos(\omega t) - q_2^i], \end{aligned} \tag{32.2.174}$$

$$\begin{aligned} y(t) &= (q\tilde{B})^{-1/2}[p_1(t) + q_2(t)] \\ &= (q\tilde{B})^{-1/2}[-q_1^i \sin(\omega t) + p_1^i \cos(\omega t) + q_2^i], \end{aligned} \tag{32.2.175}$$

$$\begin{aligned} p_y(t) &= [(q\tilde{B})^{1/2}/2][-q_1(t) + p_2(t)] \\ &= [(q\tilde{B})^{1/2}/2][-q_1^i \cos(\omega t) - p_1^i \sin(\omega t) + p_2^i]. \end{aligned} \tag{32.2.176}$$

Show that the orbit in the x, y plane is a circle with a radius ρ given by the relation

$$\rho^2 = [(q_1^i)^2 + (p_1^i)^2]/(q\tilde{B}), \tag{32.2.177}$$

and that the *center* of the circle has coordinates x_c, y_c given by the relations

$$x_c = (q\tilde{B})^{-1/2} p_2^i, \tag{32.2.178}$$

$$y_c = (q\tilde{B})^{-1/2} q_2^i. \tag{32.2.179}$$

Note the curious fact that x_c and y_c do *not* "commute". Evaluate the Poisson bracket $[x_c, y_c]$.

What modifications are required if the particle is relativistic?

32.2.8. Verify that (2.90) holds for all the generic normal forms.

32.2.9. This is an exercise on Krein collisions. Our aim is to study a simple example for which Krein collisions are avoided (as expected) when phase advances have the same sign, and can be seen to occur when phase advances have opposite signs. Consider the Hamiltonian

$$H = (\omega_1/2)(p_1^2 + q_1^2) + (\omega_2/2)(p_2^2 + q_2^2) + \epsilon q_1 q_2 \tag{32.2.180}$$

and the associated map $\mathcal{R}$ given by

$$\mathcal{R} = \exp(- : H :). \tag{32.2.181}$$

Since $\mathcal{R}$ is a linear map, its action on phase space is described by a matrix R. Following the discussion surrounding (10.4.24), show that this Hamiltonian can be written in the form

$$H = (1/2)(z, Sz) \tag{32.2.182}$$

where the symbol z now stands for the phase-space variables $z = (q_1, p_1, q_2, p_2)$ and S is the symmetric matrix

$$S = \begin{pmatrix} \omega_1 & 0 & \epsilon & 0 \\ 0 & \omega_1 & 0 & 0 \\ \epsilon & 0 & \omega_2 & 0 \\ 0 & 0 & 0 & \omega_2 \end{pmatrix}. \tag{32.2.183}$$

Show that the corresponding Hamiltonian matrix $B = JS$ is given by the relation

$$B = \begin{pmatrix} 0 & \omega_1 & 0 & 0 \\ -\omega_1 & 0 & -\epsilon & 0 \\ 0 & 0 & 0 & \omega_2 \\ -\epsilon & 0 & -\omega_2 & 0 \end{pmatrix}. \tag{32.2.184}$$

According to (10.4.8) there is the relation

$$R = \exp(B). \tag{32.2.185}$$

Show that, when $\epsilon = 0$, R is the matrix

$$R = \begin{pmatrix} \cos(\omega_1) & \sin(\omega_1) & 0 & 0 \\ -\sin(\omega_1) & \cos(\omega_1) & 0 & 0 \\ 0 & 0 & \cos(\omega_2) & \sin(\omega_2) \\ 0 & 0 & -\sin(\omega_2) & \cos(\omega_2) \end{pmatrix}. \tag{32.2.186}$$

Therefore, when there is no perturbation, the eigenvalues of R are $\exp(\pm i\omega_1)$ and $\exp(\pm i\omega_2)$, and the phase advances of R are ω_1 and ω_2. See Example 5.1 in Section 3.5.

What concerns us are the eigenvalues of B. Once we know them, we will also know the eigenvalues of R. In particular, if the eigenvalues of B are pure imaginary, then the eigenvalues of R will lie on the unit circle. Moreover if, under perturbation, the eigenvalues of B leave the imaginary axis to become a complex quartet, then the eigenvalues of R will leave the unit circle to form a Krein quartet as in Figure 3.5.1.

Show that for this B the invariants C and D have the values

$$C = (\omega_1^2 + \omega_2^2), \tag{32.2.187}$$

$$D = \omega_1\omega_2(\omega_1\omega_2 - \epsilon^2). \tag{32.2.188}$$

Evidently, for finite values of ω_1, ω_2 and ϵ sufficiently small, both C and D are positive. Thus, our attention should be turned to the upper right quadrant of Figure 2.3; and we are concerned with cases i and ii and the transition between them.

With (2.72) and Figure 2.3 in mind, show that

$$C^2 - 4D = (\omega_1^2 - \omega_2^2)^2 + 4\epsilon^2\omega_1\omega_2. \tag{32.2.189}$$

Show that, when $\epsilon = 0$, the eigenvalues of B are given by

$$\lambda = \pm i\omega_1,\ \pm i\omega_2. \tag{32.2.190}$$

Evidently they are pure imaginary, and they are distinct unless $\omega_1 = \omega_2$ or $\omega_1 = -\omega_2$. Suppose that ω_1 and ω_2 have the same sign. Show that in this case

$$C^2 - 4D \geq 0 \tag{32.2.191}$$

no matter what the value of ϵ. Prove, consequently, that if ω_1, ω_2 are finite and of the same sign, then the eigenvalues remain pure imaginary for sufficiently small ϵ.[1] Indeed, suppose that

$$\omega_1 = \omega_2 = \Omega. \tag{32.2.192}$$

Show that in this case that

$$\lambda = \pm i\Omega[1 \pm \epsilon/\Omega]^{1/2} \tag{32.2.193}$$

where all ± signs are to be taken independently. Thus, in this case, the eigenvalues remain pure imaginary under perturbation ($\epsilon \neq 0$ but sufficiently small) if ω_1, ω_2 are finite and of the same sign. Correspondingly, there is no Krein collision of the eigenvalues of R. Suppose, instead, that

$$\omega_1 = -\omega_2 = \Omega. \tag{32.2.194}$$

Show that in this case that

$$\lambda = \pm i\Omega[1 \pm i\epsilon/\Omega]^{1/2} \tag{32.2.195}$$

where all ± signs are to be taken independently. Now, under perturbation, the eigenvalues leave the imaginary axis to become a complex quartet. Correspondingly, the eigenvalues of R leave the unit circle to become a Krein quartet.

Suppose that ω_1 and ω_2 are opposite in sign, but not exactly equal in magnitude. What happens then under perturbation? Show that the eigenvalues of B leave the imaginary axis to become a complex quartet when

$$\epsilon \geq (1/2)|\omega_1^2 - \omega_2^2|/(|\omega_1\omega_2|)^{1/2} \tag{32.2.196}$$

[1] For sufficiently large ϵ they can be driven to the situation depicted in the lower right quadrant of Figure 2.3.

or

$$\epsilon \leq -(1/2)|\omega_1^2 - \omega_2^2|/(|\omega_1\omega_2|)^{1/2}. \tag{32.2.197}$$

We conclude from this example that, as might be expected, Krein collisions are imminent when phase advances are of opposite sign and nearly equal (even if not exactly equal) in magnitude.

Consider the case, corresponding to tunes of approximately ±1/4, for which

$$\omega_1 = 1.5,\ \omega_2 = -1.6. \tag{32.2.198}$$

Figure 2.4 displays the eigenvalues λ of B as a function of ϵ for this case. When $\epsilon = 0$, the eigenvalues have values of $\pm 1.5i$, $\pm\, 1.6i$ with all signs taken independently. As ϵ is increased, they merge in pairs and then leave the imaginary axis. Verify analytically that they merge and then leave the imaginary axis when $\epsilon \approx \pm .10$. See Figure 2.4 and the upper-right quadrant of Figure 2.3.

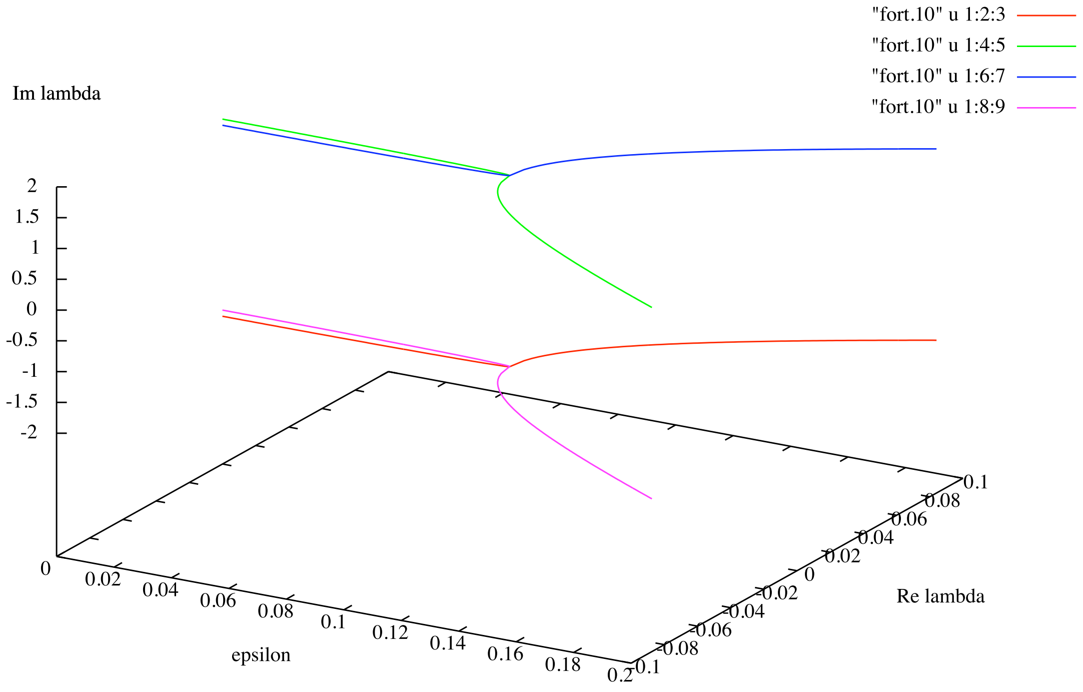


Figure 32.2.4: Eigenvalues of B as a function of ϵ when $\omega_1 = 1.5$ and $\omega_2 = -1.6$.

32.2.10. Consider quadratic polynomials in the phase-space variables for the case of a 4-dimensional phase space. For each of the normal-form cases i through ix, find the Hamiltonian matrix B associated with the specified g_2^N, verify that the invariants C and D have the indicated values, and that the eigenvalues λ have the indicated values.

32.2.11. Consider quadratic polynomials in the phase-space variables for the case of a 4-dimensional phase space. In each of the coupled cases i, v, vi, and ix the normal form g_2^N is the sum of two terms. Verify, in each case, that the two terms are in involution.

32.2.12. Consider quadratic polynomials in the phase-space variables for the case of a 4-dimensional phase space. For each of the cases i through ix, consider the motion generated by g_2^N. Do this by studying the behavior of $z'(t)$ defined by

$$z'_a(t) = \exp(-t : g_2^N :)z_a. \tag{32.2.199}$$

Show that the motion is unbounded (the origin is an unstable equilibrium point) for large $|t|$ in all cases except ii.

32.2.13. Review Exercise 2.12 above. A matrix with distinct eigenvalues can always be diagonalized by a similarity transformation. When the eigenvalues are not distinct, there are matrices for which the best that can be done by a similarity transformation is to bring them to Jordan normal form. In the 2×2 case the eigenvalues for the symplectic matrix

$$M = \begin{pmatrix} 1 & \ell \\ 0 & 1 \end{pmatrix} \tag{32.2.200}$$

are not distinct (they are both $+1$), and moreover M cannot be brought to diagonal form when $\ell \neq 0$.

When an eigenvalue collision occurs on the unit circle in the 4×4 symplectic case as in Figure 3.5.1, the eigenvalues are not distinct. Are there real 4×4 symplectic matrices for which the eigenvalues are complex, lie on the unit circle but are not distinct, and which cannot be diagonalized? What happens at the moment of collision for the two cases of $\omega_1 \simeq \omega_2$ and $\omega_1 \simeq -\omega_2$ in Exercise 2.9 above? Is R diagonalizable?

32.3 Mostly Unsolved Polynomial Orbit Problems

In this section we will describe briefly the case of orbits in $\mathcal{P}_m$ with $m > 2$. Now the situation is far more complicated because we do not have the matrix trick simplification that led to (2.14), and only limited results are available. Some results are known for 2-dimensional phase space. Much less is known for higher-dimensional phase spaces. Even the 2-dimensional case is very difficult for large m.

For the case of 4 and higher dimensional phase space no normal forms seem to be known even for g_3. However, some few invariants are known for any $\mathcal{P}_m$ and any phase-space dimension. They are sufficiently complicated that it requires several pages to write out any one of them explicitly. It is also known that in principle there are many such invariants, and an empirical estimation of their number is available. Finally, for any $\mathcal{P}_m$ and any phase-space dimension, it is known that all invariants can be computed in terms of the monomial coefficients and (when viewed as a tensor) the entries of J.

The previous section described equivalence classes and normal forms, under the action of $Sp(2n, \mathbb{R})$, for the cases of $\mathcal{P}_1$ and $\mathcal{P}_2$. The next few paragraphs provide a sample of some known results for $\mathcal{P}_3$ and $\mathcal{P}_4$.

32.3.1 Cubic Polynomials

For the case of $\mathcal{P}_3$ and 2-dimensional phase space we may write the most general g_3 in the form

$$g_3 = a_0 q^3 + 3a_1 q^2 p + 3a_2 q p^2 + a_3 p^3 \tag{32.3.1}$$

where the a_j are arbitrary coefficients. Note, as in often convenient, we have multiplied the coefficients of q^3, q^2p, qp^2, p^3 by the factors 1, 3, 3, 1. These factors are the binomial coefficients in the expansion of $(q+p)^3$. [We remark that homogeneous polynomials in two variables are called *binary forms* in the Mathematics literature. Thus, (3.1) is called a *cubic* binary form.] It can be shown in this case that there is the invariant D, called the *discriminant* of the cubic form, given by the relation

$$D = a_0^2 a_3^2 - 6a_0 a_1 a_2 a_3 + 4a_0 a_2^3 + 4a_1^3 a_3 - 3a_1^2 a_2^2. \tag{32.3.2}$$

(The *discriminant* of a binary form g_m is an invariant with the special property that its vanishing indicates that the equation $g_m = 0$ has at least one repeated root.) Associated with g_3 is a quadratic form H, called the *Hessian* of g_3, defined by the equation

$$\begin{aligned} H &= (1/36)\det(\partial^2 g_3/\partial z_i \partial z_j) \\ &= (a_0 a_2 - a_1^2)q^2 + (a_0 a_3 - a_1 a_2)qp + (a_1 a_3 - a_2^2)p^2. \end{aligned} \tag{32.3.3}$$

Here we have used our customary notation $z = (q, p)$.

With this background in mind, it can be shown for the case $D > 0$ that g_3 has the normal form

$$g_3^N = D^{1/4}(q^3 + p^3) \text{ for } D > 0. \tag{32.3.4}$$

And, if $D < 0$, g_3 has the normal form

$$g_3^N = (-D/4)^{1/4}(q^3 - 3qp^2) \text{ for } D < 0. \tag{32.3.5}$$

If the discriminant vanishes but the coefficients of the Hessian are not all zero, g_3 has the normal form

$$g_3^N = qp^2 \text{ for } D = 0 \text{ and } H \neq 0. \tag{32.3.6}$$

Finally, if both the discriminant and all coefficients in the Hessian vanish, g_3 has the normal form

$$g_3^N = q^3 \text{ for } D = 0 \text{ and } H = 0. \tag{32.3.7}$$

Evidently cases (3.4) and (3.5) are generic while cases (3.6) and (3.7) are increasingly specific. It is interesting to note that the generic normal forms (3.4) and (3.5) minimize $\langle g_3^{\text{tr}}, g_3^{\text{tr}} \rangle$. See Exercise 3.1. Note also that, under the canonical transformation $q \to p$ and $p \to -q$, (3.5) takes monkey-saddle form. See Exercise 22.5.4.

32.3.2 Quartic Polynomials

For the case of $\mathcal{P}_4$ and 2-dimensional phase space we may write the most general g_4 in the form

$$g_4 = a_0 q^4 + 4a_1 q^3 p + 6a_2 q^2 p^2 + 4a_3 q p^3 + a_4 p^4. \tag{32.3.8}$$

It can be shown that in this case there are two functionally independent invariants S and T,

$$S = a_0a_4 - 4a_1a_3 + 3a_2^2, \tag{32.3.9}$$

$$T = a_0a_2a_4 + 2a_1a_2a_3 - a_0a_3^2 - a_1^2a_4 - a_2^3. \tag{32.3.10}$$

For even degree forms there is always an invariant (with some algebraic/geometrical significance that need not concern us here) given the wonderful name *catalecticant.* In this case T is the catalecticant of g_4. The discriminant D of g_4 is functionally dependent on S and T and is given by the relation

$$D = S^3 - 27T^2. \tag{32.3.11}$$

In the case that D is positive g_4 has the normal form

$$g_4^N = \pm a(q^4 + p^4) + 6bq^2p^2 \text{ for } D > 0 \tag{32.3.12}$$

with

$$a > 0, \tag{32.3.13}$$

$$S = a^2 + 3b^2, \tag{32.3.14}$$

$$T = a^2b - b^3, \tag{32.3.15}$$

$$D = a^2(a^2 - 9b^2)^2. \tag{32.3.16}$$

In the case that D is negative g_4 has the normal form

$$g_4^N = a(q^4 - p^4) + 6bq^2p^2 \text{ for } D < 0 \tag{32.3.17}$$

with

$$a > 0, \tag{32.3.18}$$

$$S = -a^2 + 3b^2, \tag{32.3.19}$$

$$T = -a^2b - b^3, \tag{32.3.20}$$

$$D = -a^2(a^2 - 9b^2)^2. \tag{32.3.21}$$

These are the generic cases. Like g_3, there are also several specific cases for which $D = 0$, and each such case has its own normal form. We will not record them here, but results are available in the literature.

Exercises

32.3.1. Exercise on minimization.

32.4 Application to Analytic Properties

Let $\mathcal{N}$ be the two-dimensional nonlinear map defined by the relation

$$\mathcal{N} = \exp(:g_3:) \tag{32.4.1}$$

with g_3 given by (3.1). It has the action

$$Z(z) = \exp(:g_3:)z. \tag{32.4.2}$$

As usual, we write $z = (q;p)$ and $Z = (Q;P)$. What we wish to determine in this section are the analytic properties $\mathcal{N}$. That is, we wish to study the analytic properties of the quantities $Q(q,p)$ and $P(q,p)$ as functions of the variables q and p. We will see that the answer to this question depends on the normal form of g_3.

Let $\mathcal{R}$ be a linear symplectic map, and consider transformed maps of the form

$$\mathcal{N}^{\text{tr}} = \mathcal{R}\mathcal{N}\mathcal{R}^{-1} = \mathcal{R}\exp(:g_3:)\mathcal{R}^{-1} = \exp(:\mathcal{R}g_3:) = \exp(:g_3^{\text{tr}}:). \tag{32.4.3}$$

We see that a study of the analytic properties of $\mathcal{N}$ is equivalent to studying the analytic properties of $\exp(:g_3^{\text{tr}}:)$ where g_3^{tr} is any of the normal form polynomials given by (3.4) through (3.7).

Consider these cases one at a time and in order of increasing complexity. The simplest case is (3.7), for which we find that

$$\bar{q} = \exp(:q^3:)q = q, \tag{32.4.4}$$

$$\bar{p} = \exp(:q^3:)p = p + 3q^2. \tag{32.4.5}$$

Evidently $\mathcal{N}^{\text{tr}}$ in this case has no singularities save at infinity, and therefore is entire. Correspondingly, all maps in its equivalence class are also entire.

The next simplest case is (3.6), for which we find that

$$\bar{q} = \exp(:qp^2:)q = q(1-p)^2, \tag{32.4.6}$$

$$\bar{p} = \exp(:qp^2:)p = p/(1-p). \tag{32.4.7}$$

Here we have used results from Section 1.4.2. In this case $\mathcal{N}^{\text{tr}}$ has a pole on the surface $p = 1$. Correspondingly, maps in its equivalence class also have pole singularities.

The case (3.4) is next in order of increasing difficulty. Now we have

$$g_3^{\text{tr}} = \lambda(q^3 + p^3) \text{ with } \lambda = D^{1/4} \tag{32.4.8}$$

so that

$$\mathcal{N}^{\text{tr}} = \exp(\lambda :q^3 + p^3:). \tag{32.4.9}$$

Define a parameter dependent map $\mathcal{N}^{\text{tr}}(t)$ by the relation

$$\mathcal{N}^{\text{tr}}(t) = \exp(t:g_3^{\text{tr}}:) = \exp(t\lambda :q^3 + p^3:) \tag{32.4.10}$$

and write

$$\bar{q}(t) = \mathcal{N}^{\text{tr}}(t)q = \exp(t\lambda :q^3 + p^3:)q, \tag{32.4.11}$$

$$\bar{p}(t) = \mathcal{N}^{\rm tr}(t)p = \exp(t\lambda : q^3 + p^3 :)p. \tag{32.4.12}$$

We will show that there is a curve in the real q, p plane such that $\bar{q}(1) = -\infty$ and $\bar{p}(1) = +\infty$. Therefore $\mathcal{N}^{\rm tr} = \mathcal{N}^{\rm tr}(1)$ is singular on this curve.

Differentiate (4.11) and (4.12) with respect to t to obtain the equations of motion

$$\dot{\bar{q}} = \exp(t\lambda : q^3 + p^3 :)\lambda : q^3 + p^3 : q = \exp(t\lambda : q^3 + p^3 :)(-3\lambda p^2) = -3\lambda\bar{p}^2, \tag{32.4.13}$$

$$\dot{\bar{p}} = \exp(t\lambda : q^3 + p^3 :)\lambda : q^3 + p^3 : p = \exp(t\lambda : q^3 + p^3 :)(3\lambda q^2) = 3\lambda\bar{q}^2. \tag{32.4.14}$$

They have the integral

$$-\lambda(\bar{q}^3 + \bar{p}^3) = \Lambda. \tag{32.4.15}$$

Figure 4.1 shows the curves of constant Λ/λ. Also shown as arrows are the directions of the flow that follow from the equations of motion (4.13) and (4.14). Evidently $\bar{q} = \bar{p} = 0$ is the only equilibrium point, and it lies on the curve $\Lambda = 0$. Moreover, all points on the flow line $\bar{p} = -\bar{q}$ and having $\bar{q} > 0$ flow to the origin. All other points flow asymptotically to the point at infinity $\bar{q} = -\infty$, $\bar{p} = +\infty$.

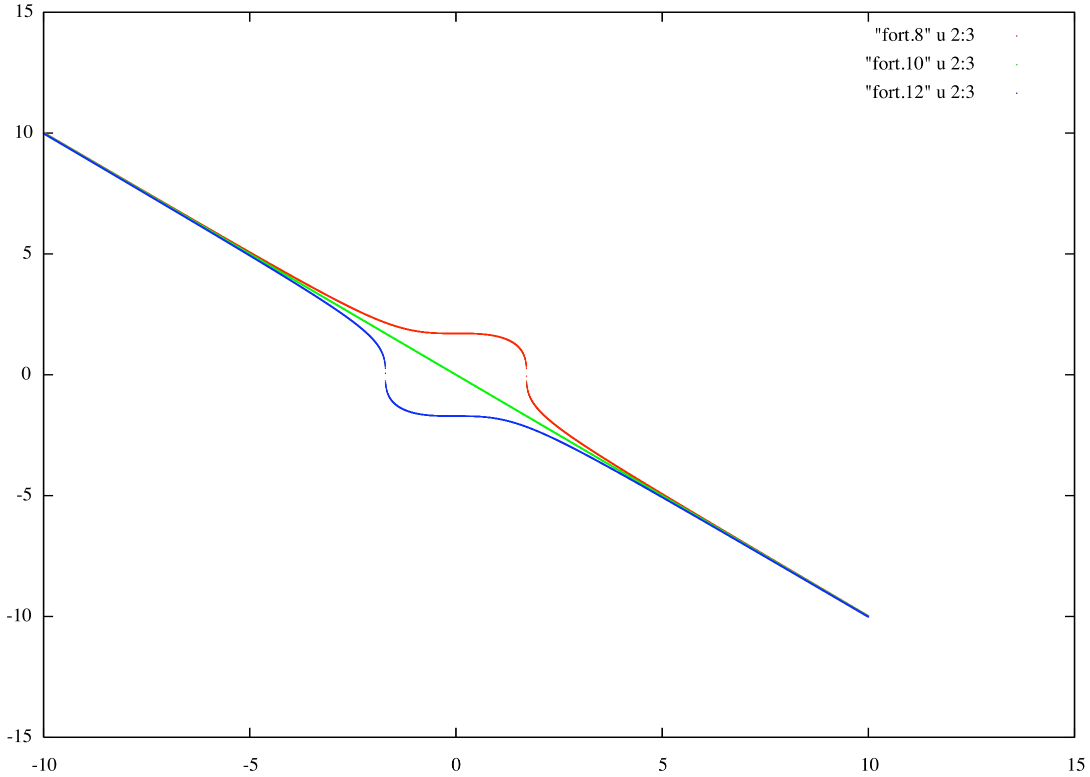


Figure 32.4.1: Curves of constant Λ/λ and flow directions for the equations of motion (4.13) and (4.14). These curves were made with $\lambda = 1$, which simply sets the scale for q and p, and $\Lambda = 0, \pm 5$.

Let us compute the *speed* at which points move along flow lines. The Euclidean distance in the $\bar{q}, \bar{p}$ plane is given by the relation

$$(ds)^2 = (d\bar{q})^2 + (d\bar{p})^2 \tag{32.4.16}$$

and therefore

$$(ds/dt)^2 = (\dot{\bar{q}})^2 + (\dot{\bar{p}})^2 = 9\lambda^2(\bar{q}^4 + \bar{p}^4). \tag{32.4.17}$$

Here we have used the equations of motion (4.13) and (4.14). We see that the speed is always positive save for the equilibrium point at the origin.

Evidently the right sides of the differential equations (4.13) and (4.14) are analytic save at infinity. Therefore, according to Section 1.3, the final conditions will be analytic functions of the initial conditions as long as the intermediate points on the flow are finite. What we wish to compute are the points for which, when regarded as initial conditions at $t = 0$, the flow reaches $\bar{q} = -\infty$, $\bar{p} = +\infty$ when $t = 1$. These points will be the frontier at which $\mathcal{N}^{\rm tr}(1)$ becomes singular.

Consider first the case $\Lambda = 0$ for which

$$\bar{p}(t) = -\bar{q}(t). \tag{32.4.18}$$

As stated earlier, points on the line $\bar{p} = -\bar{q}$ with $\bar{q} > 0$ flow along the line and into the origin; and points on the line with $\bar{q} < 0$ flow along the line to $\bar{q} = -\infty$ and $\bar{p} = +\infty$. What we will show is that on the line there is an initial condition $\bar{q}(0)$, $\bar{p}(0)$ with $\bar{q}(0) < 0$ such that

$$\bar{q}(1) = -\infty \tag{32.4.19}$$

and

$$\bar{p}(1) = +\infty \tag{32.4.20}$$

To do so, employ (4.18) in (4.13) to find the relation

$$\dot{\bar{q}} = -3\lambda\bar{q}^2 \tag{32.4.21}$$

from which it follows that

$$dt = -d\bar{q}/(3\lambda\bar{q}^2). \tag{32.4.22}$$

Integrate both sides of (4.22) to find the result

$$\int_{t^i}^{t^f} dt = -\int_{\bar{q}^i}^{\bar{q}^f} d\bar{q}/(3\lambda\bar{q}^2), \tag{32.4.23}$$

from which it follows that

$$t^f - t^i = [1/(3\lambda)](1/\bar{q})|_{\bar{q}^i}^{\bar{q}^f} = [1/(3\lambda)](1/\bar{q}^f - 1/\bar{q}^i). \tag{32.4.24}$$

Now set

$$t^i = 0,\ t^f = 1,\ \bar{q}^f = -\infty \tag{32.4.25}$$

to obtain the result

$$1 = [1/(3\lambda)](-1/\bar{q}^i) \tag{32.4.26}$$

from which it follows that

$$\bar{q}(0) = \bar{q}^i = -1/(3\lambda) \tag{32.4.27}$$

and

$$\bar{p}(0) = -\bar{q}(0) = 1/(3\lambda). \tag{32.4.28}$$

Next consider the case $\Lambda > 0$. We will again find that there is an initial condition on the curve (4.15) such that (4.19) and (4.20) hold. In this case (4.15) can be solved for $\bar{p}$ to give the result

$$\bar{p} = (-\Lambda/\lambda - \bar{q}^3)^{1/3}. \tag{32.4.29}$$

Here the negative cube root is to be extracted if the quantity $(-\Lambda/\lambda - \bar{q}^3)$ is negative, and the positive square root is to be extracted if the quantity is positive. Thus,

$$\bar{p} = -(\Lambda/\lambda + \bar{q}^3)^{1/3} < 0 \text{ when } \bar{q} > -(\Lambda/\lambda)^{1/3}, \tag{32.4.30}$$

$$\bar{p} = 0 \text{ when } \bar{q} = -(\Lambda/\lambda)^{1/3}, \tag{32.4.31}$$

$$\bar{p} = (-\Lambda/\lambda - \bar{q}^3)^{1/3} > 0 \text{ when } \bar{q} < -(\Lambda/\lambda)^{1/3}. \tag{32.4.32}$$

Observe that $\bar{p}^2$ is always ≥ 0 and is given by the relation

$$\bar{p}^2 = [(\Lambda/\lambda + \bar{q}^3)^2]^{1/3}. \tag{32.4.33}$$

Now the differential equation (4.13) takes the form

$$\dot{\bar{q}} = -3\lambda[(\Lambda/\lambda + \bar{q}^3)^2]^{1/3}, \tag{32.4.34}$$

from which we conclude

$$dt = -[1/(3\lambda)]d\bar{q}/[(\Lambda/\lambda + \bar{q}^3)^2]^{1/3}. \tag{32.4.35}$$

And integrating both sides of (4.35) yields the result

$$\int_{t^i}^{t^f} dt = -\int_{\bar{q}^i}^{\bar{q}^f} [1/(3\lambda)]d\bar{q}/[(\Lambda/\lambda + \bar{q}^3)^2]^{1/3} \tag{32.4.36}$$

from which it follows that

$$1 = -\int_{\bar{q}^i}^{-\infty} [1/(3\lambda)]d\bar{q}/[(\Lambda/\lambda + \bar{q}^3)^2]^{1/3} = \int_{-\infty}^{\bar{q}^i} [1/(3\lambda)]d\bar{q}/[(\Lambda/\lambda + \bar{q}^3)^2]^{1/3}. \tag{32.4.37}$$

Here we have again used (4.25).

What about the case $\Lambda < 0$? Here (4.29) and (4.33) continue to hold. Now we must use

$$\bar{p} = -(\Lambda/\lambda + \bar{q}^3)^{1/3} < 0 \text{ when } \bar{q} > (-\Lambda/\lambda)^{1/3}, \tag{32.4.38}$$

$$\bar{p} = 0 \text{ when } \bar{q} = (-\Lambda/\lambda)^{1/3}, \tag{32.4.39}$$

$$\bar{p} = (-\Lambda/\lambda - \bar{q}^3)^{1/3} > 0 \text{ when } \bar{q} < (-\Lambda/\lambda)^{1/3}. \tag{32.4.40}$$

With this understanding, (4.37) also continues to hold.

Taken together, (4.27) and (4.37) yield $\bar{q}^i$ as a function of Λ. And, when $\bar{q}^i$ is known and Λ is specified, (4.30) through (4.32) and (4.38) through (4.41) give $\bar{p}^i$. This curve, shown in Figure 4.2 superimposed on an enlarged portion of Figure 4.1, provides (in the real $\bar{q}, \bar{p}$ plane) the set of points at which $\mathcal{N}^{\text{tr}}(1)$ becomes singular. The map is well defined and analytic in the initial conditions for initial conditions to the right of this curve. Its status for points to left of this curve is not yet established. To do so would require, among other things some compactification procedure analogous to the Riemann sphere but in four dimensions. At this stage we do not know whether the map has only a pole or poles as was the case for (4.7), or has more complicated singularities, such as branch points that preclude single-valued analytic continuation, or singularities that preclude any analytic continuation at all. Finally we note that this one-dimensional curve is the real intersection of a two-dimensional manifold in the full four-dimensional domain of two complex variables. This manifold is specified by letting Λ be complex in (4.37) and in the relations for $\bar{p}^i$ in terms of $\bar{q}^i$ and Λ.

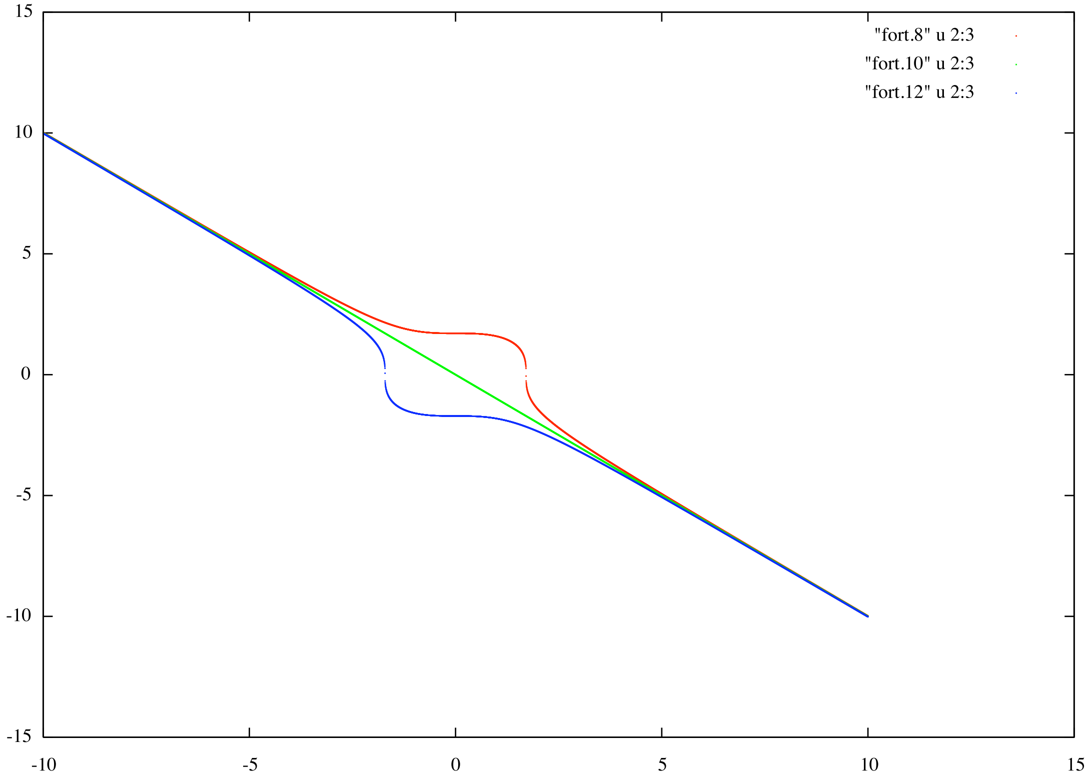


Figure 32.4.2: (Place Holder) Curve on which the map $\mathcal{N}^{\rm tr}(1)$ becomes singular. The map is well defined and analytic for phase-space points to the right of this curve. Points on the curve are sent to infinity. The possible action of the map on points to the left of the curve is unknown.

The last and most difficult case is (3.5). Now we have

$$g_3^{\rm tr} = \lambda(q^3 - 3qp^2) \text{ with } \lambda = (-D)^{1/4} \tag{32.4.41}$$

so that

$$\mathcal{N}^{\rm tr} = \exp(\lambda : q^3 - 3qp^2 :). \tag{32.4.42}$$

Again define a parameter dependent map $\mathcal{N}^{\rm tr}(t)$, now by the relation

$$\mathcal{N}^{\rm tr}(t) = \exp(t : g_3^{\rm tr} :) = \exp(t\lambda : q^3 - 3qp^2 :) \tag{32.4.43}$$

and write

$$\bar{q}(t) = \mathcal{N}^{\rm tr}(t)q = \exp(t\lambda : q^3 - 3qp^2 :)q, \tag{32.4.44}$$

$$\bar{p}(t) = \mathcal{N}^{\rm tr}(t)p = \exp(t\lambda : q^3 - 3qp^2 :)p. \tag{32.4.45}$$

We will again show that there is a curve in the real q, p plane on which $\mathcal{N}^{\rm tr} = \mathcal{N}^{\rm tr}(1)$ is singular.

Differentiate (4.44) and (4.45) with respect to t to obtain the equations of motion

$$\dot{\bar{q}} = \exp(t\lambda : q^3 - 3qp^2 :)\lambda : q^3 - 3qp^2 : q = \exp(t\lambda : q^3 - 3p^2 :)(3\lambda qp) = 3\lambda\bar{q}\bar{p}, \tag{32.4.46}$$

$$\dot{\bar{p}} = \exp(t\lambda : q^3 - 3qp^2 :)\lambda : q^3 - 3qp^2 : p = \exp(t\lambda : q^3 - 3qp^2 :)3\lambda(q^2 - p^2) = 3\lambda(\bar{q}^2 - \bar{p}^2). \tag{32.4.47}$$

They have the integral

$$-\lambda(\bar{q}^3 - 3\bar{q}\bar{p}^2) = \Lambda. \tag{32.4.48}$$

Observe that if we set $\Lambda = 0$ we find the three lines

$$\bar{q} = 0, \tag{32.4.49}$$

$$\bar{q} = (\sqrt{3})\bar{p}, \tag{32.4.50}$$

$$\bar{q} = -(\sqrt{3})\bar{p}. \tag{32.4.51}$$

Figure 4.3 shows these lines and the remaining curves of constant Λ/λ. Also shown as arrows are the directions of the flow that follow from the equations of motion (4.46) and (4.47). For the speed along these flow lines we find the result

$$(ds/dt)^2 = (\dot{\bar{q}})^2 + (\dot{\bar{p}})^2 = 9\lambda^2[(\bar{q}^2 - \bar{p}^2)^2 + (\bar{q}\bar{p})^2]. \tag{32.4.52}$$

Evidently $\bar{q} = \bar{p} = 0$ is the only equilibrium point, and it lies on the intersection of the lines (4.49) through (4.51). Moreover, we can conclude the following:

- All points on the flow line (4.49)with $p > 0$ flow into the origin. All points on that line with $p < 0$ flow to the point at infinity $(\bar{q}, \bar{p}) = (0, -\infty)$.
- All points on the flow line (4.50)with $p < 0$ flow into the origin. All points on that line with $p > 0$ flow to the point at infinity $(\bar{q}, \bar{p}) = (\infty\sqrt{3}, \infty)$.
- All points on the flow line (4.51)with $p < 0$ flow into the origin. All points on that line with $p > 0$ flow to the point at infinity $(\bar{q}, \bar{p}) = (-\infty\sqrt{3}, \infty)$.
- All other points on all other flow lines eventually flow into one of the three points at infinity listed above.

The text below requires further work

Evidently the right sides of the differential equations (3.13) and (3.14) are analytic save at infinity. Therefore, according to Section 1.3, the final conditions will be analytic functions of the initial conditions as long as the intermediate points on the flow are finite. What we wish to compute are the points for which, when regarded as initial conditions at $t = 0$, the

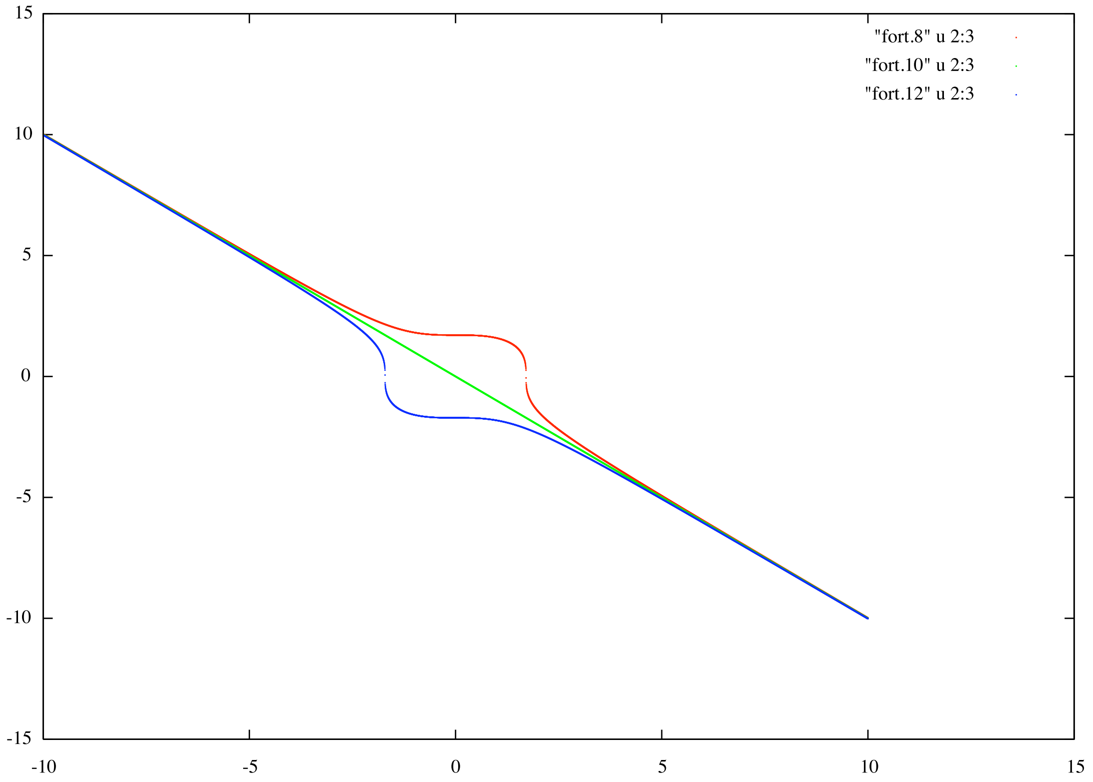


Figure 32.4.3: (Place Holder) Curves of constant Λ/λ and flow directions for the equations of motion (4.44) and (4.45). These curves were made with $\lambda = 1$, which simply sets the scale for q and p, and $\Lambda = 0, \pm 5$.

flow reaches $\bar{q} = -\infty$, $\bar{p} = +\infty$ when $t = 1$. These points will be the frontier at which $\mathcal{N}^{\text{tr}}(1)$ becomes singular.

Consider first the case $\Lambda = 0$ for which

$$\bar{p}(t) = -\bar{q}(t). \tag{32.4.53}$$

As stated earlier, points on the line $\bar{p} = -\bar{q}$ with $\bar{q} > 0$ flow along the line and into the origin; and points on the line with $\bar{q} < 0$ flow along the line to $\bar{q} = -\infty$ and $\bar{p} = +\infty$. What we will show is that on the line there is an initial condition $\bar{q}(0)$, $\bar{p}(0)$ with $\bar{q}(0) < 0$ such that

$$\bar{q}(1) = -\infty \tag{32.4.54}$$

and

$$\bar{p}(1) = +\infty \tag{32.4.55}$$

To do so, employ (3.18) in (3.13) to find the relation

$$\dot{\bar{q}} = -3\lambda\bar{q}^2 \tag{32.4.56}$$

from which it follows that

$$dt = -d\bar{q}/(3\lambda\bar{q}^2). \tag{32.4.57}$$

Integrate both sides of (3.22) to find the result

$$\int_{t^i}^{t^f} dt = -\int_{\bar{q}^i}^{\bar{q}^f} d\bar{q}/(3\lambda\bar{q}^2). \tag{32.4.58}$$

from which it follows that

$$t^f - t^i = [1/(3\lambda)](1/\bar{q})|_{\bar{q}^i}^{\bar{q}^f} = [1/(3\lambda)](1/\bar{q}^f - 1/\bar{q}^i). \tag{32.4.59}$$

Now set

$$t^i = 0,\ t^f = 1,\ \bar{q}^f = -\infty \tag{32.4.60}$$

to obtain the result

$$1 = [1/(3\lambda)](-1/\bar{q}^i) \tag{32.4.61}$$

from which it follows that

$$\bar{q}(0) = \bar{q}^i = -1/(3\lambda) \tag{32.4.62}$$

and

$$\bar{p}(0) = -\bar{q}(0) = 1/(3\lambda). \tag{32.4.63}$$

Next consider the case $\Lambda > 0$. We will again find that there is an initial condition on the curve (3.15) such that (3.19) and (3.20) hold. In this case (3.15) can be solved for $\bar{p}$ to give the result

$$\bar{p} = (-\Lambda/\lambda - \bar{q}^3)^{1/3}. \tag{32.4.64}$$

Here the negative cube root is to be extracted if the quantity $(-\Lambda/\lambda - \bar{q}^3)$ is negative, and the positive square root is to be extracted if the quantity is positive. Thus,

$$\bar{p} = -(\Lambda/\lambda + \bar{q}^3)^{1/3} < 0 \text{ when } \bar{q} > -(\Lambda/\lambda)^{1/3}, \tag{32.4.65}$$

$$\bar{p} = 0 \text{ when } \bar{q} = -(\Lambda/\lambda)^{1/3}, \tag{32.4.66}$$

$$\bar{p} = (-\Lambda/\lambda - \bar{q}^3)^{1/3} > 0 \text{ when } \bar{q} < -(\Lambda/\lambda)^{1/3}. \tag{32.4.67}$$

Observe that $\bar{p}^2$ is always ≥ 0 and is given by the relation

$$\bar{p}^2 = [(\Lambda/\lambda + \bar{q}^3)^2]^{1/3}. \tag{32.4.68}$$

Now the differential equation (3.13) takes the form

$$\dot{\bar{q}} = -3\lambda[(\Lambda/\lambda + \bar{q}^3)^2]^{1/3}, \tag{32.4.69}$$

from which we conclude

$$dt = -[1/(3\lambda)]d\bar{q}/[(\Lambda/\lambda + \bar{q}^3)^2]^{1/3}. \tag{32.4.70}$$

And integrating both sides of (3.35) yields the result

$$\int_{t^i}^{t^f} dt = -\int_{\bar{q}^i}^{\bar{q}^f} [1/(3\lambda)]d\bar{q}/[(\Lambda/\lambda + \bar{q}^3)^2]^{1/3} \tag{32.4.71}$$

from which it follows that

$$1 = -\int_{\bar{q}^i}^{-\infty} [1/(3\lambda)]d\bar{q}/[(\Lambda/\lambda + \bar{q}^3)^2]^{1/3} = \int_{-\infty}^{\bar{q}^i} [1/(3\lambda)]d\bar{q}/[(\Lambda/\lambda + \bar{q}^3)^2]^{1/3} \tag{32.4.72}$$

Here we have again used (3.25).

What about the case $\Lambda < 0$? Here (3.29) and (3.33) continue to hold. Now we must use

$$\bar{p} = -(\Lambda/\lambda + \bar{q}^3)^{1/3} < 0 \text{ when } \bar{q} > (-\Lambda/\lambda)^{1/3}, \tag{32.4.73}$$

$$\bar{p} = 0 \text{ when } \bar{q} = (-\Lambda/\lambda)^{1/3}, \tag{32.4.74}$$

$$\bar{p} = (-\Lambda/\lambda - \bar{q}^3)^{1/3} > 0 \text{ when } \bar{q} < (-\Lambda/\lambda)^{1/3}. \tag{32.4.75}$$

With this understanding, (3.37) also continues to hold.

Taken together, (3.27) and (3.37) yield $\bar{q}^i$ as a function of Λ. And, when $\bar{q}^i$ is known and Λ is specified, (3.30) through (3.32) and (3.38) through (3.41) give $\bar{p}^i$. This curve, shown in Figure 3.2 superimposed on an enlarged portion of Figure 3.1, provides (in the real $\bar{q}, \bar{q}$ plane) the set of points at which $\mathcal{N}^{\text{tr}}(1)$ becomes singular. The map is well defined and analytic in the initial conditions for initial conditions to the right of this curve. Its status for points to left of this curve is not yet established. To do so would require, among other things some compactification procedure analogous to the Riemann sphere but in four dimensions. At this stage we do not know whether the map has only a pole or poles as was the case for (3.7), or has more complicated singularities, such as branch points that preclude single-valued analytic continuation, or singularities that preclude any analytic continuation at all. Finally we note that this one-dimensional curve is the real intersection of a two-dimensional manifold in the full four-dimensional domain of two complex variables. This manifold is specified by letting Λ be complex in (3.37) and in the relations for $\bar{p}^i$ in terms of $\bar{q}^i$ and Λ.

Exercises

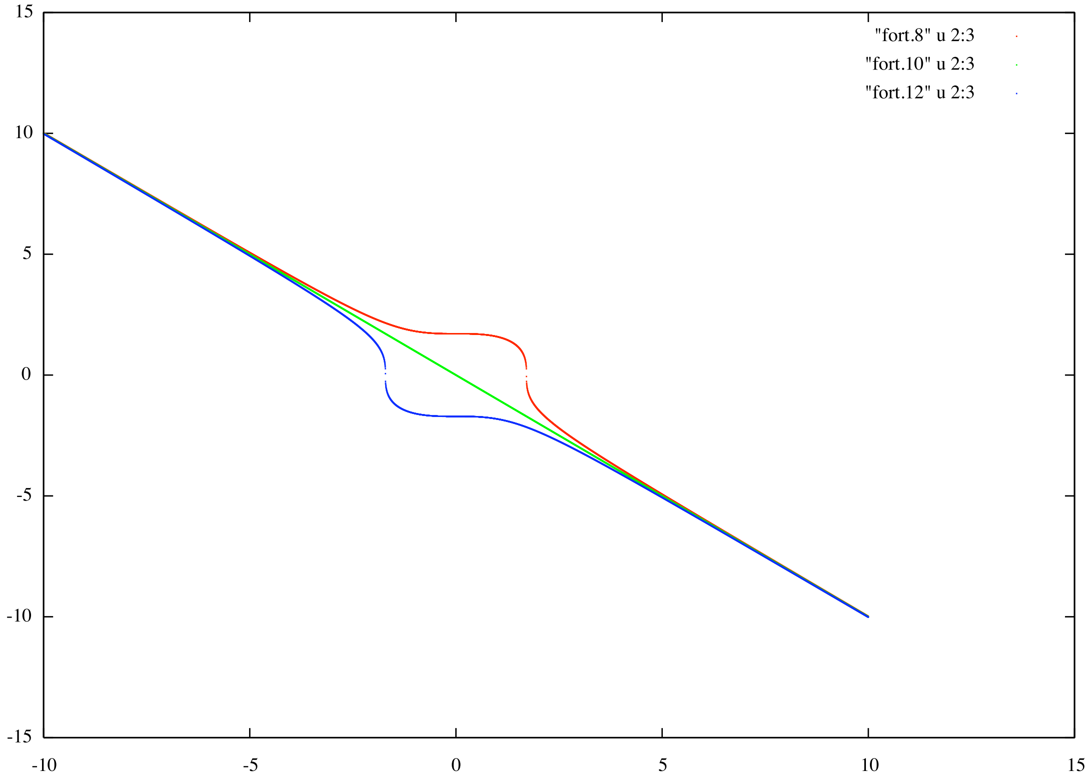


Figure 32.4.4: (Place Holder) Curve on which the map $\mathcal{N}^{\text{tr}}(1)$ becomes singular. The map is well defined and analytic for phase-space points to the right of this curve. Points on the curve are sent to infinity. The possible action of the map on points to the left of the curve is unknown.

Bibliography

Normal Forms

See also the Normal Form references given at the end of Chapter 3.

[1] J. Williamson, "On the algebraic problem concerning the normal forms of linear dynamical systems", *American Journal of Mathematics* **58**, pp. 141-163, (1936).

[2] V.I. Arnold, *Dynamical Systems III*, Springer-Verlag (1988).

[3] V.I. Arnold, *Mathematical Methods of Classical Mechanics*, Second Edition, Appendix 6, Springer-Verlag (1989).

[4] V.I. Arnold and S.P. Novikov, *Dynamical Systems IV*, Springer-Verlag (1990).

[5] N. Burgoyne and R. Cushman, "Normal Forms for Real Linear Hamiltonian Systems", *Lie Groups: History, Frontiers, and Applications*, Vol. VII [The 1976 Ames Research Center (NASA) Conference on Geometric Control Theory], eds. C. Martin and R. Hermann, Math Sci Press (1977).

[6] S-N. Chow, C. Li, D. Wang, *Normal Forms and Bifurcation of Planar Vector Fields*, Cambridge University Press (1994).

[7] R. Churchill and M. Kummer, "A Unified Approach to Linear and Nonlinear Normal Forms for Hamiltonian systems", *J. Symbolic Computation* **27**, p. 49, (1999).

Orbits

[8] D.H. Collingwood and W.M. McGovern, *Nilpotent Orbits in Semisimple Lie Algebras*, Van Nostrand Reinhold (1993).

[9] R. Peres, *Dynamical Systems and Semisimple Groups: An Introduction*, Cambridge University Press (1998).

Invariant Theory

[10] S. Borofsky, *Elementary Theory of Equations*, MacMillan Co. (1964).

[11] L.E. Dickson, *Algebraic Invariants*, John Wiley and Sons (1914).

[12] E.B. Elliott, *An Introduction to the Algebra of Quantics*, Chelsea Publishing Co. (1964).

[13] O.E. Glenn, *A Treatise on the Theory of Invariants*, Ginn and Co. (1915).

[14] J.H. Grace and A. Young, *The Algebra of Invariants*, Chelsea Publishing Co. (1903).

[15] G.B. Gurevich, *Foundations of the Theory of Algebraic Invariants*, P. Noordhoff Ltd. (1964).

[16] D. Mumford, J. Fogarty, and F. Kirwan, *Geometric Invariant Theory*, Springer-Verlag (1994).

[17] V.L. Popov, *Groups, Generators, Syzygies, and Orbits in Invariant Theory*, American Mathematical Society (1992).

[18] G. Salmon, *Lessons Introductory to the Modern Higher Algebra*, Chelsea Publishing Co. (1964).

[19] L. Smith, *Polynomial Invariants of Finite Groups*, A.K. Peters (1995).

[20] D. Stanton, Edit., *Invariant Theory and Tableaux*, Springer-Verlag (1990).

[21] B. Sturmfels, *Algorithms in Invariant Theory*, Springer-Verlag (1993).

[22] R. Goodman and N.R Wallach, *Representations and Invariants of the Classical Groups*, Cambridge University Press (1998).

[23] R. Goodman and N.R. Wallach, *Symmetry, Representations, and Invariants*, Springer (2009).

[24] P. J. Olver, *Equivalence, Invariants, and Symmetry*, Cambridge University Press (1995).

[25] P. J. Olver, *Classical Invariant Theory*, London Mathematical Society Student Texts, vol. 44, Cambridge University Press (1999).

[26] C. Procesi, *Lie Groups, An Approach through Invariants and Representations*, Springer (2007).

Chapter 33

Beam Description and Moment Transport

33.1 Preliminaries

The previous chapters dealt with *single*-particle orbit theory. In this chapter we will treat *many*-particle distributions in the approximation that the particles are *noninteracting.* Under this noninteracting assumption all results about particle distributions are derivable from properties of the single-particle transfer map $\mathcal{M}$. Here we will find it convenient to use the phase-space variable ordering

$$z = (z_1, z_2, \ldots, z_{2n-1}, z_{2n}) = (q_1, p_1, q_2, p_2, \ldots, q_n, p_n). \tag{33.1.1}$$

That is, we will employ the ordering (3.2.20) presented in Exercise 3.2.6, but will omit the prime for notational simplicity. Also, we will use the matrix J' given by (3.2.10) and (3.2.11), but will again omit the prime. See Section 3.2.

Suppose $h(z)$ is some *density* function describing a collection of particles in phase space. That is d^6N, the number of particles in a phase-space volume d^6z, is given by the relation

$$d^6N = h(z)d^6z, \tag{33.1.2}$$

and there is the result

$$N = \int d^6z\ h(z) \tag{33.1.3}$$

where N is the number of particles under consideration.

More specifically, suppose $h^i(z)$ is a function describing some *initial* distribution of particles in phase space. Next suppose the particle distribution is transported through some system described by a map $\mathcal{M}$. Then, by Liouville's theorem, the *final* distribution $h^f(z)$ at the end of the system is given by the relation

$$h^f(z) = h^i(\mathcal{M}^{-1}z). \tag{33.1.4}$$

See Subsection 6.8.1 and Exercise 6.8.2. Also recall that, as sketched in Section 6.8, the problem of determining what distribution can be sent into what under the action of some symplectic map $\mathcal{M}$, which is what (1.4) describes, is deep and only partially understood. See also Chapter 29.

33.2 Moments and Moment Transport

Suppose, as before, that $h^i(z)$ is a function describing some *initial* distribution of particles in phase space. Since $h^i(z)$ is a function, generally an infinite number of parameters are required for its specification. One way to *characterize* $h^i(z)$ is in terms of initial moments $Z^i_{abc\cdots}$ defined by the rule

$$Z^i_{abc\cdots} = \langle z_a z_b z_c \cdots \rangle^i = (1/N) \int d^6z \; h^i(z) z_a z_b z_c \cdots \, . \tag{33.2.1}$$

We use the term *characterize* advisedly rather than *specify* because in general the problem of reconstructing (uniquely determining) a function given its moments is ill posed. Nevertheless, moments may provide some useful information about $h^i(z)$.

How are initial and final moments related? To answer this question it is useful to employ a different notation for moments. Let $P_\alpha(z)$, where α is some running index, denote a complete set of homogeneous polynomials in z through terms of some fixed degree. See Chapter 36. Then one can define initial moments m^i_α by the rule

$$m^i_\alpha = (1/N) \int d^6z \; h^i(z) P_\alpha(z). \tag{33.2.2}$$

Correspondingly, the final moments are given by the relation

$$\begin{aligned} m^f_\alpha &= (1/N) \int d^6z \; h^f(z) P_\alpha(z) = (1/N) \int d^6z \; h^i(\mathcal{M}^{-1}z) P_\alpha(z) \\ &= (1/N) \int d^6\bar{z} \; h^i(\bar{z}) P_\alpha(\mathcal{M}\bar{z}). \end{aligned} \tag{33.2.3}$$

Here we have used (1.4). And, to obtain the last line, we have changed variables of integration by the rule

$$\bar{z} = \mathcal{M}^{-1} z. \tag{33.2.4}$$

Doing so required calculation of the determinant of the Jacobi matrix M associated with $\mathcal{M}$. However, it is a property of symplectic matrices that they all have determinant $+1$. Therefore the determinant of M is $+1$ and need not appear explicitly in (2.3).

Since the P_α are complete, there is an expansion of the form

$$P_\alpha(\mathcal{M}\bar{z}) = \sum_\beta \mathcal{D}_{\alpha\beta}(\mathcal{M}) P_\beta(\bar{z}) \tag{33.2.5}$$

where the $\mathcal{D}_{\alpha\beta}(\mathcal{M})$ are coefficients that can be calculated for any transfer map $\mathcal{M}$. Employing (2.5) in (2.3) gives the intermediate result

$$\begin{aligned} m^f_\alpha &= (1/N) \int d^6\bar{z} \; h^i(\bar{z}) P_\alpha(\mathcal{M}\bar{z}) = (1/N) \int d^6\bar{z} \; h^i(\bar{z}) \sum_\beta \mathcal{D}_{\alpha\beta}(\mathcal{M}) P_\beta(\bar{z}) \\ &= \sum_\beta \mathcal{D}_{\alpha\beta}(\mathcal{M}) (1/N) \int d^6\bar{z} \; h^i(\bar{z}) P_\beta(\bar{z}). \end{aligned} \tag{33.2.6}$$

It follows that moments transform *linearly* according to the rule

$$m^f_\alpha = \sum_\beta D_{\alpha\beta}(\mathcal{M}) m^i_\beta. \tag{33.2.7}$$

Note that by this method one can find the evolution of moments *without* tracking (following orbits of individual particles in) particle distributions.

33.3 Various Beam Distributions and Beam Matching

33.4 Some Properties of First-Order Moments

33.4.1 Transformation Properties

For reasons that will become clear later, see Section 5, in this subsection we will examine the transformation properties of first-order moments under the action of the inhomogeneous symplectic group $ISp(2n)$. Recall Section 9.2 for a discussion of $ISp(2n)$.

Properties under Translations

Let us first find the transformation properties of first-order moments under the action of translations. Let $\mathcal{T}$ be the *translation* map given by

$$\mathcal{T} = \exp : g_1 : \tag{33.4.1}$$

with

$$g_1(z) = -(\delta, Jz). \tag{33.4.2}$$

It has the property that

$$\mathcal{T} z_a = z_a + \delta_a. \tag{33.4.3}$$

See Section 7.7. Conversely, there is the inverse relation

$$\mathcal{T}^{-1} z_a = z_a - \delta_a. \tag{33.4.4}$$

As a special case of the Liouville relation (1.4), under the action of $\mathcal{T}$ the distribution function h becomes a transformed distribution function h' with

$$h'(z) = h(\mathcal{T}^{-1} z). \tag{33.4.5}$$

From the definition (2.1) we see that the transformed moments $\langle z_a \rangle'$ are given by the relation

$$\langle z_a \rangle' = (1/N) \int d^6\ h'(z) z_a = (1/N) \int d^6 z\ h(\mathcal{T}^{-1} z) z_a. \tag{33.4.6}$$

Introduce new variables $\bar{z}$ by the rule

$$z = \mathcal{T} \bar{z} \tag{33.4.7}$$

or, equivalently,

$$\bar{z} = \mathcal{T}^{-1} z. \tag{33.4.8}$$

The relation (4.7) implies the component relations

$$z_a = \bar{z}_a + \delta_a \tag{33.4.9}$$

Also, we see that

$$d^6 z = d^6 \bar{z}. \tag{33.4.10}$$

With these facts in mind, we see that (4.6) can be rewritten in the form

$$\begin{aligned} \langle z_a \rangle' &= (1/N) \int d^6 \bar{z}\ (\bar{z}_a + \delta_a) h(\bar{z}) \\ &= (1/N) \int d^6 \bar{z}\ \bar{z}_a h(\bar{z}) + (1/N) \int d^6 \bar{z}\ \delta_a h(\bar{z}) \\ &= \langle z_a \rangle + \delta_a, \end{aligned} \tag{33.4.11}$$

which has the more compact vector form

$$\langle z \rangle' = \langle z \rangle + \delta. \tag{33.4.12}$$

We may view the first-order moments of a distribution as specifying the *centroid* of a distribution. According to (4.12), under the action of a translation, the centroid transforms like the coordinates of a particle located at the centroid. The centroid is simply translated, as expected. We observe also that the transformation rule is the same for all distributions having the same first-order moments. Finally note that the steps leading from (4.6) to (4.12) are simply (for the translation case and for first-order moments) a more detailed recapitulation of the steps (2.3) through (2.7).

Properties under Linear Symplectic Maps

Next let us find the transformation properties of first-order moments under the action of a linear symplectic map $\mathcal{R}$ described by the symplectic matrix R. Now the Liouville relation (1.4) relating the initial distribution function h and the transformed distribution function h' takes the form

$$h'(z) = h(R^{-1} z). \tag{33.4.13}$$

From the definition (2.1) we see that the transformed moments $\langle z_a \rangle'$ are given by the relation

$$\langle z_a \rangle' = (1/N) \int d^6 z\ h'(z) z_a = (1/N) \int d^6 z\ h(R^{-1} z) z_a. \tag{33.4.14}$$

Introduce new variables $\bar{z}$ by the rule

$$z = R\bar{z} \tag{33.4.15}$$

or, equivalently,

$$\bar{z} = R^{-1} z. \tag{33.4.16}$$

The relation (4.15) implies the component relations

$$z_a = \sum_c R_{ac}\bar{z}_c. \tag{33.4.17}$$

Also, because R is symplectic and therefore must have determinant one, we find that

$$d^6z = [\det(R)]d^6\bar{z} = d^6\bar{z}. \tag{33.4.18}$$

With these facts in mind, we see that (4.14) can be rewritten in the form

$$\begin{aligned}\langle z_a\rangle' &= (1/N)\int d^6\bar{z}\ \sum_c R_{ac}\bar{z}_c h(\bar{z}) \\ &= \sum_c R_{ac}(1/N)\int d^6\bar{z}\ \bar{z}_c h(\bar{z}) \\ &= \sum_c R_{ac}\langle z_c\rangle,\end{aligned} \tag{33.4.19}$$

which has the more compact matrix form

$$\langle z\rangle' = R\langle z\rangle. \tag{33.4.20}$$

Recall that we may view the first-order moments of a distribution as specifying the centroid of a distribution. According to (4.20), under the action of a linear symplectic map, the centroid transforms like the coordinates of a particle located at the centroid. We observe, in particular, that the transformation rule is the same for all distributions having the same first-order moments. Note also that the steps leading from (4.14) to (4.20) are again (for the linear case and for first-order moments) simply a more detailed recapitulation of the steps (2.3) through (2.7).

33.4.2 Normal Form

From the work of Subsection 3.6.5 we know that $Sp(2n)$ acts transitively on phase space. See (3.6.114) through (3.6.116). Therefore, unless $\langle z\rangle = 0$, there is a sympletic matrix R such that

$$\langle z\rangle' = R\langle z\rangle = e^1. \tag{33.4.21}$$

That is, all the components of $\langle z\rangle'$ vanish save for the first, which has the value 1. Alternatively, if $\langle z\rangle$ vanishes, then $\langle z\rangle'$ also vanishes. Thus the set of first-order moments consists, under the action of $Sp(2n)$, of two *equivalence* classes: the elements that are equivalent to e^1 (which is the set all nonzero $\langle z\rangle$) and the zero element $\langle z\rangle = 0$. Correspondingly, we may view the vectors e^1 and 0 as being *normal* forms for the set of all first-order moments.

33.5 Kinematic Moment Invariants

Definition

Let m be a vector with components m_α, and let $D(\mathcal{M})$ be a matrix with entries $D_{\alpha\beta}(\mathcal{M})$. Write (2.7) in the more compact form

$$m^f = D(\mathcal{M})m^i. \tag{33.5.1}$$

A function of moments $I[m]$ is said to be a *kinematic moment invariant* if it obeys the relations

$$I[m^f] = I[m^i], \tag{33.5.2}$$

or

$$I[D(\mathcal{M})m] = I[m], \tag{33.5.3}$$

for *all* symplectic maps $\mathcal{M}$.

Rather little is known about the existence and properties of kinematic moment invariants for the set of all symplectic maps. However, kinematic moment invariants are known to exist and all kinematic moment invariants have been found when the symplectic maps $\mathcal{M}$ are restricted to be those associated with the inhomogeneous symplectic group $ISp(2n)$. Moreover, their existence is a consequence of group theory applied to $ISp(2n)$. We note that, if deviation variables are employed, the full $\mathcal{M}$ is well approximated by translation and linear maps provided excursions about the design orbit are sufficiently small.

First-Order Moments

At this point we can observe that there are no *significant* kinematic invariants in the case of first-order moments. Evidently, by definition, a kinematic invariant has the same value for all moments that belong to the same equivalence class. If only translations $\mathcal{T}$ are considered, (4.12) shows that any set of first-order moments can be transformed to any other, and therefore there is only one equivalence class. Consequently I must have a constant value. And, if only linear symplectic transformations $\mathcal{R}$ are considered, we have seen that for the case of first-order moments there are only two equivalence classes. Consequently, in this case I can have only two possible values.

Second-Order Moments

Of particular interest are kinematic moment invariants that can be constructed from the second-order moments Z_{ab} with

$$Z_{ab} = \langle z_a z_b \rangle = (1/N) \int d^6 z \; h(z) z_a z_b. \tag{33.5.4}$$

For a given particle distribution, let Z be the matrix with entries Z_{ab}. In the case of a 1-degree of freedom system phase space is 2 dimensional, and the matrix Z in this case is 2×2. It is easily verified that in this case a kinematic moment invariant [under the action of $Sp(2)$] is given by the rule

$$I[Z] = \mathrm{tr}[(ZJ_2)^2] = 2[(Z_{12})^2 - Z_{11}Z_{22}] = -2(\langle q^2 \rangle \langle p^2 \rangle - \langle qp \rangle^2). \tag{33.5.5}$$

See Exercise 6.1 where this result is verified and shown to be a consequence of group theory applied to $Sp(2)$. Note that in this case I is proportional to the *mean square emittance* ϵ^2 *defined* by the rule

$$\epsilon^2 = \langle q^2 \rangle \langle p^2 \rangle - \langle qp \rangle^2. \tag{33.5.6}$$

In the case of a 3-degree of freedom system it can be shown that there are 3 such functionally independent invariants given by the rules

$$I^{(n)}[Z] = \mathrm{tr}[(ZJ)^n], \quad n = 2, 4, 6; \tag{33.5.7}$$

and all other invariants constructed from second-order moments are functions of these invariants. See Exercises 6.2 and 6.3.

33.6 Some Properties of Second-Order Moments

In this section we will explore various properties of Z.

33.6.1 Positive Definite Property

We begin by showing that the matrix Z, which is obviously real and symmetric, is also positive definite. Since $h(z)$ is a phase-space density, it is positive or zero for all z,

$$h(z) \geq 0 \text{ for all } z; \tag{33.6.1}$$

and it follows from (6.1) and continuity that there must be some finite phase-space volume for which $h(z) > 0$. Next let u be any real six-dimensional nonzero vector. Form the function $(u, z)^2$. It has the property

$$(u, z)^2 \geq 0 \text{ for all } z. \tag{33.6.2}$$

Moreover, in the volume where $h(z) > 0$, there must be some subvolume where $(u, z)^2 > 0$. It follows that there is the result

$$\begin{aligned}(u, Zu) &= \sum_{ab} u_a Z_{ab} u_b = (1/N) \int d^6z\ h(z) \sum_{ab} u_a z_a u_b z_b \\ &= (1/N) \int d^6z\ h(z)(u, z)^2 > 0. \end{aligned} \tag{33.6.3}$$

33.6.2 Transformation Properties

Properties under Translations

From the definition (2.1) we see that under translations the transformed moments $\langle z_a z_b \rangle'$ are given by the relation

$$\langle z_a z_b \rangle' = (1/N) \int d^6\ h'(z) z_a z_b = (1/N) \int d^6z\ h(\mathcal{T}^{-1} z) z_a z_b. \tag{33.6.4}$$

As in Subsection 4.1, introduce new variables $\bar{z}$ by the rules (4.7) through (4.9) and employ (4.10). So doing reveals that (6.4) can be rewritten in the form

$$\begin{aligned}
\langle z_a z_b\rangle' &= (1/N)\int d^6\bar{z}\ (\bar{z}_a+\delta_a)(\bar{z}_b+\delta_b)h(\bar{z}) \\
&= (1/N)\int d^6\bar{z}\ \bar{z}_a\bar{z}_b h(\bar{z}) + (1/N)\int d^6\bar{z}\ \bar{z}_a\delta_b h(\bar{z}) \\
&\quad +(1/N)\int d^6\bar{z}\ \delta_a\bar{z}_b h(\bar{z}) + (1/N)\int d^6\bar{z}\ \delta_a\delta_b h(\bar{z}) \\
&= \langle z_a z_b\rangle + \langle z_a\rangle\delta_b + \delta_a\langle z_b\rangle + \delta_a\delta_b. \qquad (33.6.5)
\end{aligned}$$

Properties under Linear Symplectic Maps

Let us next find the transformation properties of second-order moments under the action of a linear symplectic map $\mathcal{R}$ described by the symplectic matrix R. From the definition (4.1) and (4.13) we see that the transformed moments $\langle z_a z_b\rangle'$ are given by the relation

$$\langle z_a z_b\rangle' = (1/N)\int d^6z\ h'(z)z_a z_b = (1/N)\int d^6z\ h(R^{-1}z)z_a z_b. \qquad (33.6.6)$$

Again introduce new variables $\bar{z}$ by the rules (4.15) through (4.17) and supplement (4.17) with the relation

$$z_b = \sum_d R_{bd}\bar{z}_d. \qquad (33.6.7)$$

Also employ the relation (4.18). With these tools we see that (6.6) can be rewritten in the form

$$\begin{aligned}
\langle z_a z_b\rangle' &= (1/N)\int d^6\bar{z}\ \sum_{cd} R_{ac}R_{bd}\bar{z}_c\bar{z}_d h(\bar{z}) \\
&= \sum_{cd} R_{ac}R_{bd}(1/N)\int d^6\bar{z}\ \bar{z}_c\bar{z}_d h(\bar{z}) \\
&= \sum_{cd} R_{ac}R_{bd}\langle z_c z_d\rangle. \qquad (33.6.8)
\end{aligned}$$

In terms of the notation employed in (5.4), the relation (6.8) can be rewritten in the component form

$$Z'_{ab} = \sum_{cd} R_{ac}R_{bd}Z_{cd} = \sum_{cd} R_{ac}Z_{cd}(R^T)_{db}, \qquad (33.6.9)$$

which has the more compact matrix form

$$Z' = RZR^T. \qquad (33.6.10)$$

This matrix relation specifies how second-order moments transform under a linear symplectic map. We observe, in particular, that the transformation rule is the same for all distributions having the same second-order moments.

33.6.3 Williamson Normal Form

Even more can be said. Since the matrix Z is real, symmetric, and positive definite, according to a theorem of Williamson there is a symplectic matrix A such that

$$AZA^T = D \tag{33.6.11}$$

where D is the *diagonal* matrix

$$D = \text{diag}\{\lambda_1, \lambda_1, \lambda_2, \lambda_2, \lambda_3, \lambda_3\} \tag{33.6.12}$$

with all $\lambda_j > 0$. The right side of (6.11) is called the *Williamson normal form* of Z.

Two things should be noted about this remarkable result. Define the matrix Z^{norm} by the rule

$$Z^{\text{norm}} = AZA^T = D. \tag{33.6.13}$$

Then, from (6.12 and (6.13), we see that there are the results

$$\langle q_j q_k \rangle^{\text{norm}} = \langle p_j p_k \rangle^{\text{norm}} = 0 \text{ if } j \neq k, \tag{33.6.14}$$

$$\langle q_j^2 \rangle^{\text{norm}} = \langle p_j^2 \rangle^{\text{norm}} = \lambda_j, \tag{33.6.15}$$

$$\langle q_j p_k \rangle^{\text{norm}} = 0. \tag{33.6.16}$$

Also, we observe that (6.11) is of the form (6.10) with $R = A$. Thus, if a beam transport system can be found whose transfer matrix is A, then this transport system will bring the second-order moments to the normal form given by (6.14) through (6.16).

33.6.4 Eigen Emittances

We will next see that two second-order moment matrices Z' and Z have the *same* Williamson normal form if they are connected by a relation of the form (6.10). Indeed, observe that we may write the relation

$$\begin{aligned} AR^{-1}Z'(AR^{-1})^T &= AR^{-1}RZR^T(R^T)^{-1}A^T \\ &= AZA^T = D. \end{aligned} \tag{33.6.17}$$

[Here we have used the result $(R^{-1})^T = (R^T)^{-1}$ which holds for any invertible matrix.] But, by the group property of symplectic matrices, the matrix AR^{-1} is symplectic if the matrices A and R are symplectic. We see from (6.17) that the symplectic matrix AR^{-1} brings Z' to Williamson normal form and, according to (6.13), this normal form is the same as that for Z. The quantities λ_j^2 are called mean-square *eigen* emittances, or simply eigen remittances. It follows that while the entries in Z evolve as a particle distribution propagates through various elements, see (6.10), the eigen emittances remain *unchanged* (in the linear approximation). Thus, given an initial particle distribution, one can compute the initial second moments $\langle z_a z_b \rangle^i$, and from them the eigen emitances. And these eigen emittances will remain unchanged (in the linear approximation) as the particle distribution evolves.

It can be shown that the eigen emittances generalize the 1-degree of freedom mean-square emittance given by (5.6) to the fully coupled case. Indeed, it can be shown that in terms of the λ_j the kinematic invariants $I^{(n)}$ given by (5.7) have the values

$$I^{(n)} = 2(-1)^{n/2}(\lambda_1^n + \lambda_2^n + \lambda_3^n). \tag{33.6.18}$$

See Exercise 6.3.

There are symplectic matrix routines that, given Z, find A and the λ_j. If only the λ_j are required, they can be found from the eigenvalues of JZ. Note that JZ is a Hamiltonian matrix. See Exercise 3.17.14.

To see that the λ_j can be found from the eigenvalues of JZ, suppose both sides of (6.11) are multiplied by J to give the result

$$JAZA^T = JD. \tag{33.6.19}$$

From the symplectic condition for A it follows that there is the relation

$$JA = (A^T)^{-1}J. \tag{33.6.20}$$

Consequently (6.19) can be rewritten in the form

$$(A^T)^{-1}JZA^T = JD, \tag{33.6.21}$$

which reveals that the matrices JZ and JD are related by a *similarity* transformation, and therefore have the same eigenvalues. See Exercise 3.7.16.

What remains is to find the eigenvalues of JD which, according to (3.2.10), (3.2.11), and (6.12) can be written in the block form

$$JD = \begin{pmatrix} \lambda_1 J_2 & & \\ & \lambda_2 J_2 & \\ & & \lambda_3 J_2 \end{pmatrix}. \tag{33.6.22}$$

Let W_2 be the unitary and (complex) symplectic 2×2 matrix

$$W_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix}. \tag{33.6.23}$$

[See (3.9.12).] It has the property

$$W_2^{-1} J_2 W_2 = iK_2 \tag{33.6.24}$$

where K_2 is the matrix

$$K_2 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{33.6.25}$$

From W_2 construct the the 6×6 matrix W given in block form by the rule

$$W = \begin{pmatrix} W_2 & & \\ & W_2 & \\ & & W_2 \end{pmatrix}. \tag{33.6.26}$$

It follows from (6.22) and (6.24) that there is the relation

$$W^{-1}JDW = \begin{pmatrix} i\lambda_1 K_2 & & \\ & i\lambda_2 K_2 & \\ & & i\lambda_3 K_2 \end{pmatrix}. \tag{33.6.27}$$

We see that the eigenvalues of JD, and hence JZ, are pure imaginary and come in the ± pairs

$$\sigma_j = \pm i\lambda_j. \tag{33.6.28}$$

Conversely, if the eigenvalues $\pm\sigma_j$ of JZ are computed, then the eigen emittances are given by the relation

$$\lambda_j = |\sigma_j]. \tag{33.6.29}$$

Suppose we combine the relations (6.21) and (6.27) to find the result

$$W^{-1}(A^T)^{-1}JZA^TW = \begin{pmatrix} i\lambda_1 K_2 & & \\ & i\lambda_2 K_2 & \\ & & i\lambda_3 K_2 \end{pmatrix} = \text{diag}\{-i\lambda_1, i\lambda_1, -i\lambda_2, i\lambda_2, -i\lambda_3, i\lambda_3\}. \tag{33.6.30}$$

Let A' be the matrix defined by the rule

$$A' = A^TW, \;\; (A')^{-1} = W^{-1}(A^T)^{-1}. \tag{33.6.31}$$

It will be symplectic since A (and hence A^T) and W are symplectic. With the aid of A' the relation (6.30) takes the form

$$(A')^{-1}JZA' = \text{diag}\{-i\lambda_1, i\lambda_1, -i\lambda_2, i\lambda_2, -i\lambda_3, i\lambda_3\}. \tag{33.6.32}$$

We observe that since Z is positive definite, JZ is a particular/special kind of Hamiltonian matrix. As a consequence of Williamson's theorem we have seen that it can be diagonalized by a similarity transformation even if its eigenvalues are not distinct; moreover the diagonalizing matrix A' can be chosen to be symplectic. And, as stated earlier, the eigenvalues of JZ are pure imaginary.

Finally, we note that multiplying both sides of (6.10) by J produces the relation

$$JZ' = JRZR^T = (R^T)^{-1}JZR^T. \tag{33.6.33}$$

Here we have used the fact that R is symplectic. We see that with the use of J the evolution rule (6.10) for Z becomes the *similarity* transformation rule (6.33) for JZ. Since eigenvalues are preserved by similarity transformations, we have found an alternative explanation of why the eigen emittances remain unchanged as a particle distribution evolves.

33.6.5 Classical Uncertainty Principle

Statement

The results of the previous subsection can be used to derive a *classical* uncertainty principle. What we will show is that there is the inequality

$$\langle q_i^2\rangle\langle p_i^2\rangle \geq \lambda_{\text{min}}^2, \; i = 1, 2, 3, \tag{33.6.34}$$

where λ_{min} is the minimum of the λ_k. No matter what is done to a beam (ignoring nonlinear and nonsymplectic effects), the products of the mean-square deviations in q_i and p_i for any plane must exceed, or at best equal, λ_{min}^2.

Proof

Begin by rewriting (6.13) in the form

$$Z = A^{-1}Z^{\text{norm}}(A^{-1})^T = N^T Z^{\text{norm}} N = N^T D N \tag{33.6.35}$$

where we have made the definition

$$N = (A^{-1})^T. \tag{33.6.36}$$

We note that N will be symplectic if A is symplectic, and conversely.

Let us compute the $\langle q_i^2\rangle$ and $\langle p_i^2\rangle$. To compute the $\langle q_i^2\rangle$ set

$$a = j \text{ with } j = 1, 3, 5 \text{ when } i = 1, 2, 3. \tag{33.6.37}$$

We then find from (6.35) that

$$\begin{aligned} \langle q_i^2\rangle &= Z_{aa} = (N^T D N)_{aa} = \sum_{cd} (N^T)_{ac} D_{cd} N_{da} \\ &= \sum_c N_{ca} D_{cc} N_{ca} = \sum_c (N_{ca})^2 D_{cc} \\ &\geq \lambda_{\text{min}} \sum_c (N_{ca})^2. \end{aligned} \tag{33.6.38}$$

Similarly, to compute the $\langle p_i^2\rangle$, upon setting

$$b = j + 1, \tag{33.6.39}$$

we find that

$$\begin{aligned} \langle p_i^2\rangle &= Z_{bb} = (N^T D N)_{bb} = \sum_{cd} (N^T)_{bc} D_{cd} N_{db} \\ &= \sum_d N_{db} D_{dd} N_{db} = \sum_d (N_{db})^2 D_{dd} \\ &\geq \lambda_{\text{min}} \sum_d (N_{db})^2. \end{aligned} \tag{33.6.40}$$

It follows that

$$\langle q_i^2\rangle\langle p_i^2\rangle \geq \lambda_{\text{min}}^2 \left[\sum_c (N_{ca})^2\right] \left[\sum_d (N_{db})^2\right]. \tag{33.6.41}$$

To proceed further, let u^a and u^b be vectors with the entries

$$u_c^a = N_{ca}, \tag{33.6.42}$$

$$u_d^b = N_{db}. \tag{33.6.43}$$

Evidently u^a and u^b are the $a^{\rm th}$ and $b^{\rm th}$ columns of N. With these definitions we may write (6.41) in the more compact form

$$\langle q_i^2\rangle\langle p_i^2\rangle \geq \lambda_{\rm min}^2 ||u^a||^2\ ||u^b||^2 \tag{33.6.44}$$

where $|| * ||$ denotes the Euclidean norm. Since N is a symplectic matrix, it follows from the symplectic condition that there is also the relation

$$(u^a, Ju^b) = 1. \tag{33.6.45}$$

See Exercise 3.6.13. It can be shown using the spectral norm for J that (6.45) in turn entails the inequality.

$$||u^a||\ ||u^b|| \geq 1. \tag{33.6.46}$$

See Exercise 3.7.1. Upon combining (6.44) and (6.46) we find the advertised result (6.34).

33.6.6 Minimum Emittance Theorem

Statement

The classical uncertainty principle shows that (in the linear approximation) no matter how a beam is transformed, the product of the spreads in position and the conjugate momentum must satisfy the relation (6.34). There is a related constraint on the mean-square emittances ϵ_i defined by

$$\epsilon_i^2 = \langle q_i^2\rangle\langle p_i^2\rangle - \langle q_i p_i\rangle^2. \tag{33.6.47}$$

What we will show is that (in the linear approximation) no matter how a beam is transformed (symplectically) there is the constraint

$$\epsilon_i^2 \geq \lambda_{\rm min}^2,\ i = 1, 2, 3. \tag{33.6.48}$$

Together the information provided by the classical uncertainty principle and the minimum emittance theorem is useful when designing a beam line to perform emittance manipulations because it sets lower limits on what one can hope to achieve.

Proof

Suppose, in the 6×6 case under consideration, that we partition Z into nine 2×2 blocks by writing

$$Z = \begin{pmatrix} Z^{11} & Z^{12} & Z^{13} \\ Z^{21} & Z^{22} & Z^{23} \\ Z^{31} & Z^{32} & Z^{33} \end{pmatrix}. \tag{33.6.49}$$

Because Z is symmetric, the blocks will satisfy the relations

$$(Z^{ij})^T = Z^{ji}. \tag{33.6.50}$$

Let R be a 6×6 matrix having the block form

$$R = \begin{pmatrix} A & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix}. \tag{33.6.51}$$

It will be symplectic if A is symplectic. Its use in (6.10) produces a Z' given by

$$Z' = \begin{pmatrix} A & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} Z^{11} & Z^{12} & Z^{13} \\ Z^{21} & Z^{22} & Z^{23} \\ Z^{31} & Z^{32} & Z^{33} \end{pmatrix} \begin{pmatrix} A^T & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix}. \tag{33.6.52}$$

Carrying out the indicated multiplication gives the result

$$Z' = \begin{pmatrix} AZ^{11}A^T & AZ^{12} & AZ^{13} \\ Z^{21}A^T & Z^{22} & Z^{23} \\ Z^{31}A^T & Z^{32} & Z^{33} \end{pmatrix}. \tag{33.6.53}$$

In particular, we see that

$$(Z')^{11} = AZ^{11}A^T. \tag{33.6.54}$$

We will now seek a symplectic A that brings Z^{11} to Williamson normal form. Define the quantity ϵ_1 by the rules

$$\epsilon_1^2 = Z_{11}^{11}Z_{22}^{11} - (Z_{12}^{11})^2 = \langle q_1^2\rangle\langle p_1^2\rangle - \langle q_1p_1\rangle^2, \tag{33.6.55}$$

$$\epsilon_1 = +\sqrt{\epsilon_1^2}. \tag{33.6.56}$$

It follows from the Schwarz inequality that there is the relation

$$\langle q_1p_1\rangle^2 \le \langle q_1^2\rangle\langle p_1^2\rangle \tag{33.6.57}$$

and therefore the right side of (6.55) can never be negative. See Exercise 6.4. Consequently ϵ_1 is well defined by (6.55) and (6.56), and is positive. Next define "beam" betatron functions α, β, γ by the rules

$$\alpha = -Z_{12}^{11}/\epsilon_1 = -\langle q_1p_1\rangle/\epsilon_1, \tag{33.6.58}$$

$$\beta = Z_{11}^{11}/\epsilon_1 = \langle q_1^2\rangle/\epsilon_1, \tag{33.6.59}$$

$$\gamma = Z_{22}^{11}/\epsilon_1 = \langle p_1^2\rangle/\epsilon_1. \tag{33.6.60}$$

In terms of these definitions, Z^{11} takes the form

$$Z^{11} = \epsilon_1 \begin{pmatrix} \beta & -\alpha \\ -\alpha & \gamma \end{pmatrix}. \tag{33.6.61}$$

From (6.59) and (6.60) there are the inequalities $\beta \ge 0$ and $\gamma \ge 0$. And, from (6.55) through (6.60), there is the relation

$$1 = \beta\gamma - \alpha^2. \tag{33.6.62}$$

Finally, define the matrix A by the rule

$$A = \begin{pmatrix} 1/\sqrt{\beta} & 0 \\ \alpha/\sqrt{\beta} & \sqrt{\beta} \end{pmatrix}. \tag{33.6.63}$$

Since A is 2×2 and evidently has unit determinant, it is symplectic. Correspondingly, the R given by (6.51) is symplectic. And, from the definitions made and executing the matrix multiplications $AZ^{11}A^T$ indicated in (6.54), we find that

$$(Z')^{11} = AZ^{11}A^T = \mathrm{diag}(\epsilon_1, \epsilon_1). \tag{33.6.64}$$

See Exercise 6.5.

We will now exploit these results. From (6.64) we find that

$$\langle q_1^2 \rangle' = (Z')^{11}_{11} = \epsilon_1, \tag{33.6.65}$$

$$\langle p_1^2 \rangle' = (Z')^{11}_{22} = \epsilon_1. \tag{33.6.66}$$

It follows that

$$\langle q_1^2 \rangle' \, \langle p_1^2 \rangle' = \epsilon_1^2. \tag{33.6.67}$$

But we also have the relation (6.34). We conclude that there is the inequality

$$\epsilon_1^2 \geq \lambda_{\min}^2, \tag{33.6.68}$$

in accord with (6.48). Analogous results hold for the other planes.

Sharpening

We close this subsection by noting that the minimum emittance theorem (6.48) sharpens the classical uncertainty principle (6.34). Indeed, combining (6.47) and (6.48) produces the result

$$\langle q_i^2 \rangle \langle p_i^2 \rangle \geq \lambda_{\min}^2 + \langle q_i p_i \rangle^2, \; i = 1, 2, 3. \tag{33.6.69}$$

We see that to minimize $\langle q_i^2 \rangle \langle p_i^2 \rangle$ we must insure that $\langle q_i p_i \rangle$ vanishes.

33.6.7 Nonexistence of Maximum Emittances

The classical uncertainty principle (6.34) and the minimum emittance theorem (6.48) show that the mean square emittances are bounded from below under the action of linear symplectic maps. We will now see that they are *not* bounded from above if the phase space has 4 or more dimensions.

Consider the 4-dimensional case, and suppose initially a particle distribution has all quadratic moments zero save for the moments $\langle q_1^2 \rangle$, $\langle p_1^2 \rangle$, $\langle q_2^2 \rangle$, and $\langle p_2^2 \rangle$. (From the work of Subsection 6.3 we know that there is always a linear symplectic transformation that will bring the quadratic moments to this form.) In this case the mean square emittances ϵ_i^2 are given by the relation

$$\epsilon_i^2 = \langle q_i^2 \rangle \langle p_i^2 \rangle. \tag{33.6.70}$$

Let $\mathcal{R}$ be the linear symplectic map

$$\mathcal{R} = \exp(\nu : q_1 p_2 :). \tag{33.6.71}$$

Here ν is some some real parameter. It is easily verified that this map has the properties

$$\bar{q}_1 = \mathcal{R} q_1 = q_1, \tag{33.6.72}$$

$$\bar{p}_1 = \mathcal{R} p_1 = p_1 + \nu p_2, \tag{33.6.73}$$

$$\bar{q}_2 = \mathcal{R} q_2 = q_2 - \nu q_1, \tag{33.6.74}$$

$$\bar{p}_2 = \mathcal{R} p_2 = p_2. \tag{33.6.75}$$

It is also easily verified that there are the transformed moment relations

$$\langle \bar{q}_1^2 \rangle = \langle q_1^2 \rangle, \tag{33.6.76}$$

$$\langle \bar{p}_1^2 \rangle = \langle (p_1 + \nu p_2)^2 \rangle = \langle p_1^2 \rangle + \nu^2 \langle p_2^2 \rangle, \tag{33.6.77}$$

$$\langle \bar{q}_1 \bar{p}_1 \rangle = 0; \tag{33.6.78}$$

$$\langle \bar{q}_2^2 \rangle = \langle (q_2 - \nu q_1)^2 \rangle = \langle q_2^2 \rangle + \nu^2 \langle q_1^2 \rangle, \tag{33.6.79}$$

$$\langle \bar{p}_2^2 \rangle = \langle p_2^2 \rangle, \tag{33.6.80}$$

$$\langle \bar{q}_2 \bar{p}_2 \rangle = 0. \tag{33.6.81}$$

Correspondingly, the transformed mean square emittance $\bar{\epsilon}_1^2$ satisfies the relation

$$\begin{aligned} \bar{\epsilon}_1^2 &= \langle \bar{q}_1^2 \rangle \langle \bar{p}_1^2 \rangle - \langle \bar{q}_1 \bar{p}_1 \rangle^2 \\ &= \langle q_1^2 \rangle (\langle p_1^2 \rangle + \nu^2 \langle p_2^2 \rangle) \\ &= \epsilon_1^2 + \nu^2 \langle q_1^2 \rangle \langle p_2^2 \rangle. \end{aligned} \tag{33.6.82}$$

Similarly the transformed mean square emittance $\bar{\epsilon}_2^2$ satisfies the relation

$$\bar{\epsilon}_2^2 = \epsilon_2^2 + \nu^2 \langle q_1^2 \rangle \langle p_2^2 \rangle. \tag{33.6.83}$$

We conclude that both $\bar{\epsilon}_1^2$ and $\bar{\epsilon}_2^2$ can be made arbitrarily large by making $|\nu|$ arbitrarily large.

33.6.8 Second-Order Moments about the Beam Centroid

Definition

Define second-order moments *About* the *Beam Centroid*, denoted as Z_{ab}^{ABC}, by the rule

$$Z_{ab}^{\mathrm{ABC}} = \langle (z_a - \langle z_a \rangle)(z_b - \langle z_b \rangle) \rangle = (1/N) \int d^6 z \; h(z)(z_a - \langle z_a \rangle)(z_b - \langle z_b \rangle). \tag{33.6.84}$$

Executing the indicated operations gives the result

$$\begin{aligned} Z_{ab}^{\rm ABC} &= \langle (z_a - \langle z_a \rangle)(z_b - \langle z_b \rangle)\rangle \\ &= \langle z_a z_b \rangle - \langle z_a \langle z_b \rangle\rangle - \langle\langle z_a \rangle z_b\rangle + \langle\langle z_a\rangle\langle z_b\rangle\rangle \\ &= \langle z_a z_b\rangle - \langle z_a\rangle\langle z_b\rangle \\ &= Z_{ab} - Z_{ab}^{\rm OBC}. \end{aligned} \tag{33.6.85}$$

Here $Z_{ab}^{\rm OBC}$ denotes the set of second-order moments *Of* the *Beam Centroid* defined by the rule

$$Z_{ab}^{\rm OBC} = \langle z_a\rangle\langle z_b\rangle. \tag{33.6.86}$$

Intuitively, $Z^{\rm OBC}$ may be viewed as the collection of second-order moments of a beam distribution consisting of a single *macro* particle located at the beam centroid. We also observe, in passing, that (6.85) can be rewritten in the form

$$Z = Z^{\rm ABC} + Z^{\rm OBC}, \tag{33.6.87}$$

which is analogous to the fact that the inertia tensor of a rigid body about some specified origin is the sum of its inertia tensor about its center of mass plus the inertia tensor of its center of mass about the specified origin.

Properties under Translations

What are the transformation properties of $Z^{\rm ABC}$ under the action of a translation $\mathcal{T}$? Starting from the definition (6.83) we find the chain of equalities

$$\begin{aligned} (Z_{ab}^{\rm ABC})' &= (\langle (z_a - \langle z_a\rangle')(z_b - \langle z_b\rangle')\rangle)' \\ &= (1/N)\int d^6z\ h'(z)(z_a - \langle z_a\rangle')(z_b - \langle z_b\rangle') \\ &= (1/N)\int d^6z\ h(\mathcal{T}^{-1}z)(z_a - \langle z_a\rangle')(z_b - \langle z_b\rangle') \\ &= (1/N)\int d^6\bar z\ h(\bar z)(\bar z_a + \delta_a - \langle z_a\rangle')(\bar z_b + \delta_b - \langle z_b\rangle') \\ &= (1/N)\int d^6\bar z\ h(\bar z)(\bar z_a - \langle z_a\rangle)(\bar z_b - \langle z_b\rangle) \\ &= Z_{ab}^{\rm ABC}. \end{aligned} \tag{33.6.88}$$

Here we have used (4.9) and (4.10) to change variables and have used (4.12) to obtain the relation

$$(\bar z_a + \delta_a - \langle z_a\rangle')(\bar z_b + \delta_b - \langle z_b\rangle') = (\bar z_a - \langle z_a\rangle)(\bar z_b - \langle z_b\rangle). \tag{33.6.89}$$

We see that $Z^{\rm ABC}$ is *invariant* under translations. (For an alternate proof of this result, see Exercise 6.6.) The quantity $Z^{\rm ABC}$ describes an *intrinsic* property of the beam distribution in that it does not depend on the location of the beam relative to the design orbit. Note that, according to (6.85) and(6.86), each component $Z_{ab}^{\rm ABC}$ of $Z^{\rm ABC}$ depends on first and second moments. Therefore, each component is a moment invariant under the action of translations.

Properties under Linear Symplectic Maps

What are the transformation properties of $Z^{\rm ABC}$ under the action of a linear symplectic map $\mathcal{R}$? Evidently, according to (6.85), we may write the relation

$$(Z^{\rm ABC})' = Z' - (Z^{\rm OBC})'. \tag{33.6.90}$$

We see from (4.19) that there is the relation

$$(Z^{\rm OBC}_{ab})' = \langle z_a\rangle'\langle z_b\rangle' = \sum_{cd} R_{ac}R_{bd}\langle z_c\rangle\langle z_d\rangle, \tag{33.6.91}$$

which can be written in the more compact form

$$(Z^{\rm OBC})' = RZ^{\rm OBC}R^T. \tag{33.6.92}$$

Recall also the relation (6.10). It follows that there is the result

$$(Z^{\rm ABC})' = RZR^T - RZ^{\rm OBC}R^T = RZ^{\rm ABC}R^T. \tag{33.6.93}$$

We conclude that Z, $Z^{\rm ABC}$, and $Z^{\rm OBC}$ all transform in the same manner.

Positive Definiteness

What about positive definiteness? First consider $Z^{\rm OBC}$. As in Subsection 6.1, u be any real nonzero vector. It follows from (6.86) that there is the result

$$(u, Z^{\rm OBC}u) = \sum_{ab} u_a Z^{\rm OBC}_{ab} u_b = \sum_{ab} u_a\langle z_a\rangle\langle z_b\rangle u_b = (u, \langle z\rangle)^2 \geq 0. \tag{33.6.94}$$

We see that $(u, Z^{\rm OBC}u)$ can never be negative. However if $(u, \langle z\rangle) = 0$, which is certainly possible, then $(u, Z^{\rm OBC}u) = 0$. Therefore $Z^{\rm OBC}$ is *not* positive *definite*.

Even more can be said. Let R be a symplectic matrix that has the property (4.21). Then we see from (6.91) that in this case

$$(Z^{\rm OBC})' = D \tag{33.6.95}$$

where D is a diagonal matrix with all entries zero save for D_{11} which has the value $D_{11} = 1$. Correspondingly, DJ has all entries zero save that $(DJ)_{12} = 1$. Consequently, the eigenvalues of DJ and JD, and hence of $Z^{\rm OBC}J$ and $JZ^{\rm OBC}$, all vanish. Recall Exercises 3.7.18 and 3.7.16. Moreover, there is the relation

$$(DJ)^2 = 0, \tag{33.6.96}$$

from which it follows that

$$(Z^{\rm OBC}J)^2 = 0. \tag{33.6.97}$$

We conclude from (5.7) and (6.97) that for $Z^{\rm OBC}$ there is the result

$$I^{(n)}[Z^{\rm OBC}] = 0. \tag{33.6.98}$$

The second-order moment invariants of a beam distribution consisting of a single macro particle all vanish.

What about second-order moments about the beam centroid, those described by Z^{ABC}? Calculation/insight shows that Z^{ABC} *is* positive definite. See Exercise 6.7. Therefore it has a Williamson normal form and is characterized by eigen emittances. Note that because Z^{ABC} is invariant under translations, these eigen emittances are *invariant* under the action of the *full* group $ISp(2n)$.

We also observe that in general these eigen emittances may differ from those of Z, but they may be of interest if the beam centroid is quite far from the design orbit. See Exercise 6.8. However in practice it is probably desirable, as one of the criteria for beam matching, to arrange to have the beam centroid coincide with the design orbit. This is also natural from an instrumentation perspective since beam position monitors essentially record the spatial coordinates of the beam centroid. Recall Section 3.

33.6.9 Summary of What We Have Learned

The information provided by the classical uncertainty principle and the minimum emittance theorem is useful when designing a beam line to perform emittance manipulations on a beam because it sets lower limits on what one can hope to achieve. It should also be useful in analyzing the results of beam cooling experiments. In this case one can measure all quadratic moments before and after a cooling channel. Next compute the eigen emittances of Z before and Z after. Ideally, one would like to find that all the λ_j^2 have decreased, or at least the minimum of the λ_j^2 has decreased.

We have seen that, in considering what can be achieved under beam transport (in the linear approximation), what counts are the eigen emittances, and these can be viewed as properties of the *initial* particle distribution. Moreover, according to (6.14) through (6.16), the best that can be achieved are the spread relations

$$\langle q_i^2\rangle\langle p_i^2\rangle = \lambda_i^2,\ i = 1, 2, 3 \tag{33.6.99}$$

where the λ_i are the eigen emittances in some order. Thus, in the combined context of both source and beam-line design, the challenge is to produce an initial particle distribution having optimal eigen emittances and to then transform the initial particle distribution in such a way that the optimal spread relations are realized in the desired planes. The next sections will describe various methods for producing initial particle distributions having optimal eigen emittances and how to then transform these distributions in such a way that the optimal spread relations are realized in the desired planes.

Exercises

33.6.1. Verify (5.5). Next suppose that Z' and Z are related by (6.10) and that R is symplectic. Verify that

$$I[Z'] = \mathrm{tr}[(Z'J_2)^2] = \mathrm{tr}[(RZR^TJ_2)^2]. \tag{33.6.100}$$

Next verify that

$$\begin{aligned} \mathrm{tr}[(RZR^TJ_2)^2] &= \mathrm{tr}[RZR^TJ_2RZR^TJ_2] = \mathrm{tr}[RZJ_2ZR^TJ_2] \\ &= \mathrm{tr}[ZJ_2ZR^TJ_2R] = \mathrm{tr}[ZJ_2ZJ_2] \\ &= \mathrm{tr}[(ZJ_2)^2] = I[Z]. \end{aligned} \tag{33.6.101}$$

Here we have used that assumption that R is symplectic and the trace property (3.6.130). Combining (6.100) and (6.101) shows that I is invariant,

$$I[Z'] = I[Z]. \tag{33.6.102}$$

How might we have known that there should be an invariant? Consider the space of all quadratic polynomial functions of the two-dimensional phase-space variables q, p. For present purposes, a convenient basis for these polynomials is given by the monomials

$$c^1 = q^2, \tag{33.6.103}$$

$$c^2 = qp, \tag{33.6.104}$$

$$c^3 = p^2. \tag{33.6.105}$$

According to Section 24.2, these polynomials carry the $sp(2)$ representation $\Gamma(2)$. Also, comparison of (2.5) and (2.7) shows that polynomials P_α and moments m_α have the *same* transformation properties. Therefore, and in particular, second-order moments of the two-dimensional phase-space variables q, p also carry the representation $\Gamma(2)$.

Next we observe that, for $sp(2)$, there is the Clebsch-Gordan series result

$$\Gamma(2) \otimes \Gamma(2) = \Gamma(0) \oplus \Gamma(2) \oplus \Gamma(4). \tag{33.6.106}$$

This is just the $sp(2)$ analog of the familiar statement that spin 1 and spin 1 combine to make spin 0, spin 1, and spin 2. Even more familiar, it is the analog of the statement that two vectors can be combined to make a scalar by use of the dot product, or can be combined to make another vector by use of the cross product, or can be combined to make a tensor by use of the tensor product. By definition, an entity (if is nonzero) that carries the representation $\Gamma(0)$ will be an invariant. Therefore, according to (6.106), there is at least the hope/possibility of constructing an invariant out of quadratic products of second-order moments. Note that the contents of (5.5) are indeed quadratic products of second-order moments.

How can we construct an entity that carries the representation $\Gamma(0)$? We have already seen some such constructions in Section 24.11, which you should review. You will now have the pleasure of making a similar construction for the problem at hand.

Begin by finding the symmetric matrices S^j associated with the c^j by the rule

$$c^j = (z, S^j z). \tag{33.6.107}$$

Show that these matrices are given by the relations

$$S^1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \tag{33.6.108}$$

$$S^2 = (1/2)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{33.6.109}$$

$$S^3 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}. \tag{33.6.110}$$

Next find the associated $sp(2)$ matrices C^j defined by the rule

$$C^j = JS^j. \tag{33.6.111}$$

Show that these matrices are given by the relations

$$C^1 = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}, \tag{33.6.112}$$

$$C^2 = (1/2)\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{33.6.113}$$

$$C^3 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}. \tag{33.6.114}$$

From these $sp(2)$ matrices construct the associated down-index metric tensor g by the rule

$$g_{jk} = \mathrm{tr}(C^j C^k). \tag{33.6.115}$$

Show that g has the entries

$$g = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 1/2 & 0 \\ -1 & 0 & 0 \end{pmatrix}. \tag{33.6.116}$$

With g in hand, construct the up-index metric tensor $\hat{g}$ by the rule

$$\hat{g}^{jk} = (g^{-1})_{jk}. \tag{33.6.117}$$

Show that $\hat{g}$ has the entries

$$\hat{g} = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 2 & 0 \\ -1 & 0 & 0 \end{pmatrix}. \tag{33.6.118}$$

Finally, based on arguments provided in Section 24.11, the quantity I defined by

$$I = \sum_{jk} \langle c^j \rangle \hat{g}^{jk} \langle c^k \rangle \tag{33.6.119}$$

should be invariant. Verify that

$$I = \sum_{jk} \langle c^j \rangle \hat{g}^{jk} \langle c^k \rangle = -2\langle c^1 \rangle \langle c^3 \rangle + 2\langle c^2 \rangle^2 = -2(\langle q^2 \rangle \langle p^2 \rangle - \langle qp \rangle^2), \tag{33.6.120}$$

which agrees with (5.6). If we look at (6.120) from a group-theory perspective, we see that the quantities $\hat{g}^{jk}$ are the $Sp(2)$ Clebsch-Gordan coefficients that couple $\Gamma(2)$ and $\Gamma(2)$ down to $\Gamma(0)$.

There is one further proof of the invariance of I as given by (5.5), actually of ϵ^2 as given by (5.6), that historically probably came first. Begin by verifying, in the case of a two-dimensional phase space, that the second-order moment matrix Z has the form

$$Z = \begin{pmatrix} \langle q^2 \rangle & \langle qp \rangle \\ \langle qp \rangle & \langle p^2 \rangle \end{pmatrix}. \tag{33.6.121}$$

Consequently, the determinant of Z has the value

$$\det Z = \langle q^2 \rangle \langle p^2 \rangle - \langle qp \rangle^2 = \epsilon^2. \tag{33.6.122}$$

Next look at the transformation rule (6.10) in the two-dimensional phase space case. Verify that taking the determinant of both sides of (6.10) and recalling Section 3.3.3 give the result

$$\det Z' = \det RZR^T = (\det R)(\det Z)(\det R^T) = \det Z, \tag{33.6.123}$$

thereby demonstrating the invariance of the mean-square emittance in the two-dimensional phase space case.

33.6.2. The aim of this exercise is to show that $I^{(n)}[Z]$ as given by (5.7) is invariant. First, as warmup steps, verify the relations

$$I^{[n]} = 0 \text{ for odd } n, \tag{33.6.124}$$

$$\mathrm{tr}[(ZJ)^n] = \mathrm{tr}[(JZ)^n]. \tag{33.6.125}$$

Deduce from (6.33) that

$$(JZ')^n = [(R^T)^{-1} JZR^T]^n = (R^T)^{-1}(JZ)^n R^T. \tag{33.6.126}$$

Next verify from (6.126) that

$$\mathrm{tr}[(JZ')^n] = \mathrm{tr}[(JZ)^n]. \tag{33.6.127}$$

You have shown that

$$I^{(n)}[Z'] = \mathrm{tr}[(JZ')^n] = \mathrm{tr}[(Jz)^n] = I^{(n)}[Z]. \tag{33.6.128}$$

33.6.3. The aim of this exercise is to prove (6.18) and to remark on one of its consequences. To do so, begin by verifying the following chain of equalities:

$$I^{(n)}[Z] = I^{(n)}[AZA^T] = I^{(n)}[D] = \mathrm{tr}[(DJ)^n] = \mathrm{tr}[(JD)^n]. \tag{33.6.129}$$

Here we have used the invariance of $I^{(n)}$, the relation (6.11), and the relation (6.125). Next use (6.22) to show that

$$(JD)^2 = -\mathrm{diag}\{\lambda_1^2, \lambda_1^2, \lambda_2^2, \lambda_2^2, \lambda_3^2, \lambda_3^2\}, \tag{33.6.130}$$

from which it follows that

$$(JD)^n = (-1)^{n/2}\mathrm{diag}\{\lambda_1^n, \lambda_1^n, \lambda_2^n, \lambda_2^n, \lambda_3^n, \lambda_3^n\}. \tag{33.6.131}$$

Here we assume n is even since the case of odd n has already been covered in (6.124). Finally, verify that

$$\mathrm{tr}[(JD)^n] = 2(-1)^{n/2}(\lambda_1^n + \lambda_2^n + \lambda_2^n) \tag{33.6.132}$$

thereby proving (6.18).

As a parting comment we remark that, if desired, relations of the form (6.18) can be solved for the λ_j in terms of the $I^{[n]}$, and that the solution is given in terms of radicals for the cases where the phase-space dimension is less than or equal to 8. Find explicit results when the phase-space dimension is 2 or 4.

33.6.4. Let $f(z)$ and $g(z)$ be any two real *polynomial* functions. Such functions form a vector space. By using the phase-space density $h(z)$, which is assumed to fall off sufficiently fast at infinity, define a scalar product (f, g) by the rule

$$(f, g) = (1/N) \int d^6z\; h(z)f(z)g(z). \tag{33.6.133}$$

Verify that (6.133) satisfies all the requirements to be a scalar product including the positive-definite conditions

$$(f, f) \geq 0, \tag{33.6.134}$$

$$(f, f) = 0 \Leftrightarrow f = 0. \tag{33.6.135}$$

Note also that, in terms of moment notation, there is the relation

$$(f, g) = \langle fg \rangle. \tag{33.6.136}$$

Prove the Schwarz inequality in this context, and use it to verity the result (6.57). See Exercise 3.7.1.

33.6.5. Verify (6.62). Let I be the 2×2 identity matrix, and let Z^{11} and A be the matrices (6.61) and (6.63), respectively. Verify the matrix multiplication result

$$AZ^{11}A^T = \epsilon_1 I. \tag{33.6.137}$$

33.6.6. The aim of this exercise is to provide an alternate proof of (6.88).

33.6.7. Suppose Z and Z' are two matrices related by (6.10). Under the assumption that R is nonsingular, but not necessarily symplectic, show that Z' is symmetric and positive definite if the same is true for Z.

A further task is to show that Z^{ABC} is positive definite. First provide a proof along the lines of that in Subsection 6.1. Next $\cdots$.

33.6.8. The purpose of this exercise is to illustrate by a simple example that translation can change an emittance. Consider, for simplicity, the case of a two-dimensional phase space, and suppose a beam initially has the moments

$$\langle q^2 \rangle = \langle p^2 \rangle = \lambda, \tag{33.6.138}$$

$$\langle qp \rangle = \langle q \rangle = \langle p \rangle = 0. \tag{33.6.139}$$

Then the initial mean-square emittance ϵ^2 has the value

$$\epsilon^2 = \lambda^2. \tag{33.6.140}$$

Next consider the effect of a translation that simply augments q by an amount δ. Then, to compute the final emittance, we need the quantities

$$\langle q^2\rangle' = \langle (q+\delta)^2\rangle = \langle (q^2 + 2q\delta + \delta^2)\rangle = \langle q^2\rangle + 2\delta\langle q\rangle + \delta^2 = \lambda + \delta^2, \tag{33.6.141}$$

$$\langle p^2\rangle' = \langle p^2\rangle = \lambda \tag{33.6.142}$$

$$\langle qp\rangle' = \langle (q+\delta)p\rangle = \langle qp\rangle + \delta\langle p\rangle = 0. \tag{33.6.143}$$

From these quantities show that the transformed mean-square emittance is given by the relation

$$(\epsilon^2)' = \langle q^2\rangle'\langle p^2\rangle' - (\langle qp\rangle')^2 = (\lambda + \delta^2)\lambda = \lambda^2 + \delta^2\lambda. \tag{33.6.144}$$

Upon comparing (6.140) and (6.144), we see that the mean-square emittance has been *increased* by an amount $\delta^2\lambda$.

33.7 Construction of Initial Distributions with Small/Optimized Eigen Emittances

33.8 Realization of Eigen Emittances as Mean-Square Emittances

Bibliography

Normal Forms

[1] J. Williamson, "On the algebraic problem concerning the normal forms of linear dynamical systems", *American Journal of Mathematics* **58**, pp. 141-163, (1936).

[2] R. Churchill and M. Kummer, "A Unified Approach to Linear and Nonlinear Normal Forms for Hamiltonian systems", *J. Symbolic Computation* **27**, p. 49, (1999).

Moments and Emittances

[3] A. Dragt et al., *Ann. Rev. Nucl. Part. Sci.* **38**, 455 (1988).

[4] A. Dragt et al., *MaryLie 3.0 Users' Manual* (2003). See www.physics.umd.edu/dsat/.

[5] A. Dragt et al., in *Frontiers of Particle Beams; Observation, Diagnosis and Correction*, edited by M. Month and S. Turner, Lecture Notes in Physics Vol. 343, p. 94, Springer Verlag (1989).

[6] A. Dragt et al., *Phys. Rev. A* **45**, 2572 (1992).

[7] M.A. de Gosson, *The principles of Newtonian and quantum mechanics: the need for Planck's constant, h*, Imperial College Press, London (2001).

[8] M.A. de Gosson, "The symplectic camel and phase space quantization", *J. Phys. A: Math. Gen.* **34** p. 10085 (2001).

[9] M.A. de Gosson, "The 'symplectic camel principle' and semiclassical mechanics", *J. Phys A: Math. Gen.* **35**, p. 6825 (2002).

[10] M.A. de Gosson, "Uncertainty Principle, Phase-Space Ellipsoids, and Weyl Calculus", *Operator Theory: Advances and Applications*, Vol. 164, p. 121, Birkhäuser Verlag (2006).

[11] M.A. de Gosson, *Symplectic geometry and quantum mechanics*, Birkhäuser Verlag (2006).

[12] M.A. de Gosson and F. Luef, "Symplectic capacities and the geometry of uncertainty: The irruption of symplectic topology in classical and quantum mechanics", *Physics Reports* 484, 131-179, Elsevier (2009).

[13] M.A. de Gosson, “The symplectic egg in classical and quantum mechanics”, *American Journal of Physics* **81**, 328 (2013).

Chapter 34

Optimal Evaluation of Symplectic Maps

34.1 Overview of Symplectic Map Approximation

Several previous chapters have been devoted to the subject of describing, computing, manipulating, and analyzing symplectic maps. This chapter is devoted to the subject of *applying* symplectic maps to phase-space data. That is, we are given a symplectic map $\mathcal{M}$ in some some form and a general phase-space point z, and we wish to find the phase-space point $\bar{z}$ given by

$$\bar{z} = \mathcal{M}z. \tag{34.1.1}$$

This task is more difficult and more complicated than one might suppose.

In practice, generally the only symplectic maps we can deal with in an explicit way are *truncated* Taylor series of the form (7.5.5) or (7.6.1) or (7.7.13). These Taylor series will obey the symplectic condition to the order through they have been calculated, i.e. they are symplectic jets.[1] But usually they will fail to obey the symplectic condition exactly because of the missing higher-order terms. Of course, we can always make the Lie factorization (7.7.23) and then truncate the infinite product at some order. So doing will still yield a symplectic map. However, its evaluation will generally involve summing infinite series of the form (5.4.1) and (5.42). When working numerically, at best these series can be evaluated to machine precision. But usually this is impractical because of the great effort involved if this is to be done very often. Usually we must truncate the Lie series, in which case we are again left with a symplectic jet which generally does not satisfy the symplectic condition exactly.

Put another way, suppose the map $\mathcal{M}$ is factored in the form

$$\mathcal{M} = \exp(:f_1:)\mathcal{R}\mathcal{N} \tag{34.1.2}$$

where

$$\mathcal{N} = \exp(:f_3:)\exp(:f_4:)\exp(:f_5:)\cdots . \tag{34.1.3}$$

Then we can evaluate $\exp(:f_1:)$ exactly because it simply produces a translation, and we can evaluate the linear part $\mathcal{R}$ to produce a matrix R that is symplectic to machine precision

[1] See Section 7.5.

using the methods of Chapter 4. However, there is generally no easy way to evaluate the action of the nonlinear part $\mathcal{N}$. In particular, approximating its action as a jet generally violates the symplectic condition.

Failure to satisfy the symplectic condition exactly may not produce serious errors when tracking particles through *single-pass* systems such as beam lines or electron microscopes or spot forming systems or linear accelerators or linear colliders. However, failure to satisfy the symplectic condition is very serious when one tries to model the long-term behavior of particles in *circulating* devices such as synchrotrons or damping rings or storage rings.

As a simple example, consider the two-dimensional symplectic map $\mathcal{M}$ given by the relation

$$\mathcal{M} = \mathcal{R}\mathcal{N} \tag{34.1.4}$$

with linear part

$$\mathcal{R} = \exp(-(\theta/2) : p^2 + q^2 :) \tag{34.1.5}$$

and nonlinear part

$$\mathcal{N} = \exp(: qp^2 :). \tag{34.1.6}$$

The map $\mathcal{R}$ can be evaluated exactly, see (1.2.48) and (1.2.49). And, thanks to its simplicity, so can the map $\mathcal{N}$. There is the result

$$\bar{q} = \mathcal{N}q = q(1-p)^2, \tag{34.1.7}$$

$$\bar{p} = \mathcal{N}p = p/(1-p). \tag{34.1.8}$$

See Section 1.4.2. Therefore $\mathcal{M}$ can also be evaluated exactly.

Figure 1.1 shows the result of applying $\mathcal{M}$ repeatedly to seven initial conditions for the case $\theta/2\pi = 0.22$. That is, seven initial conditions have been selected and their orbits have been found under the repeated action of $\mathcal{M}$. One initial condition is near the origin, and its orbit appears to lie on a closed curve that is nearly elliptical. (It would be nearly circular had the horizontal and vertical scales been equal.) This is to be expected because the effect of the nonlinear part $\mathcal{N}$ is small on such orbits so that such orbits are essentially those of $\mathcal{R}$. By contrast, the other initial conditions are successively farther from the origin where the effect of $\mathcal{N}$ becomes ever more significant. Their orbits appear to lie on closed curves that, the farther they are from the origin, are more and more noticeably distorted from circular by nonlinearities.

Now suppose the nonlinear map $\mathcal{N}$ is *truncated*, to become the map $\mathcal{N}^{\rm tr}$, by retaining only the first two terms in its Taylor expansion,

$$\mathcal{N}^{\rm tr} = \mathcal{I} + : qp^2 : . \tag{34.1.9}$$

The truncated map $\mathcal{N}^{\rm tr}$ has the effect

$$\bar{q} = \mathcal{N}^{\rm tr}q = (\mathcal{I} + : qp^2 :)q = q + [qp^2, q] = q - 2qp, \tag{34.1.10}$$

$$\bar{p} = \mathcal{N}^{\rm tr}p = (\mathcal{I} + : qp^2 :)p = p + [qp^2, p] = p + p^2. \tag{34.1.11}$$

Evidently $\mathcal{N}^{\rm tr}$ is a degree-two symplectic jet map.

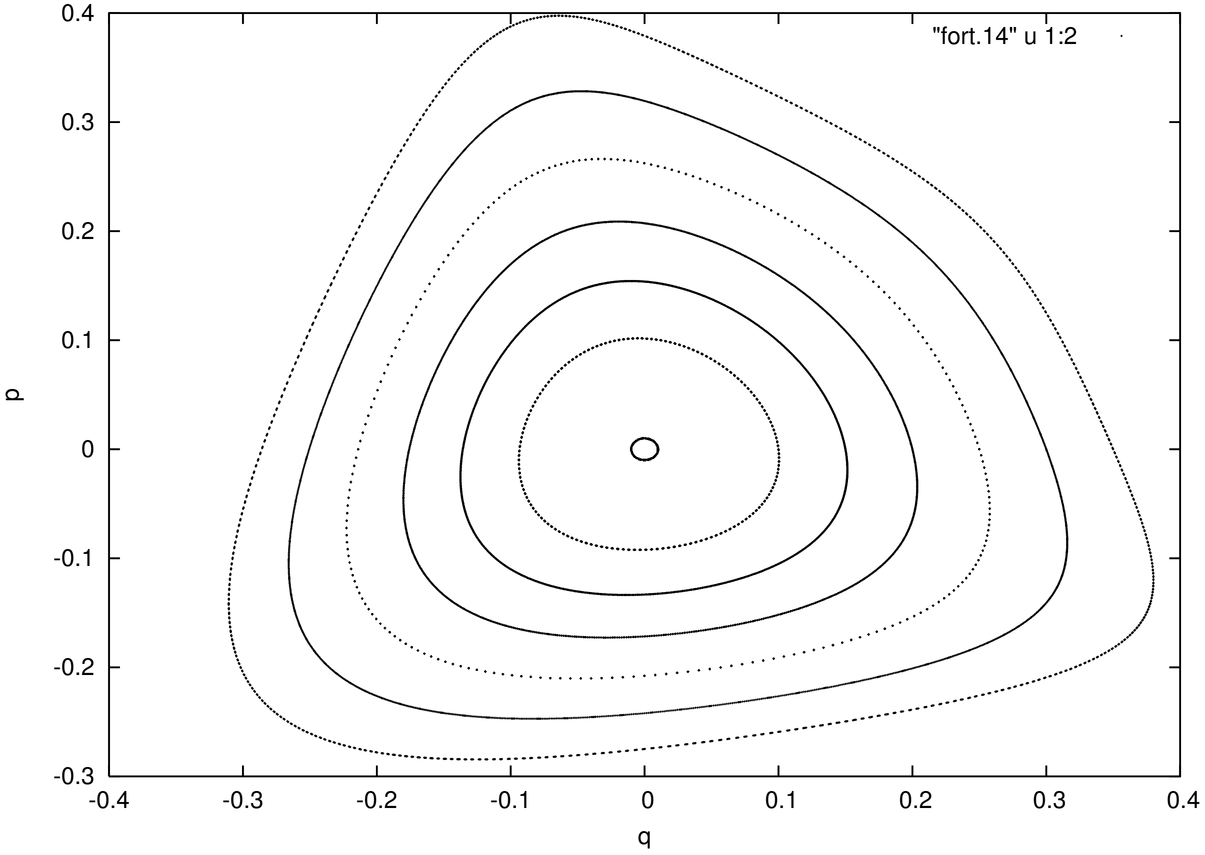


Figure 34.1.1: Phase-space portrait, in the case $\theta/2\pi = 0.22$, resulting from applying the map $\mathcal{M}$ repeatedly (2000 times) to the seven initial conditions $(q,p) = (.01,0)$, $(.1,0)$, $(.15,0)$, $(.2,0)$, $(.25,0)$, $(.3,0)$, and $(.35,0)$ to find their orbits.

Next define a corresponding map $\mathcal{M}^{\rm tr}$ by writing

$$\mathcal{M}^{\rm tr} = \mathcal{R}\mathcal{N}^{\rm tr}. \tag{34.1.12}$$

Figure 1.2 shows the orbits of $\mathcal{M}^{\rm tr}$ for two initial conditions, one near the origin and one quite far away. Inspection of the figure shows that orbits are no longer distorted circles, but instead appear to spiral into the origin. This motion into the origin occurs because $\mathcal{N}^{\rm tr}$, and consequently $\mathcal{M}^{\rm tr}$, is not symplectic. See Exercise 1.1. Indeed, following the discussion of Section 22.1, the map $\mathcal{M}^{\rm tr}$ must have a factorization of the form

$$\begin{aligned}\mathcal{M}^{\rm tr} = \; & \exp(\mathcal{G}_4)\exp(\mathcal{G}_5)\exp(\mathcal{G}_6)\cdots\times \\ & \exp(-(\theta/2) : p^2+q^2 :)\exp(: qp^2 :)\exp(: f_4 :)\exp(: f_5 :)\cdots\end{aligned} \tag{34.1.13}$$

with the non-Hamiltonian vector field $\mathcal{G}_4$ being nonzero.

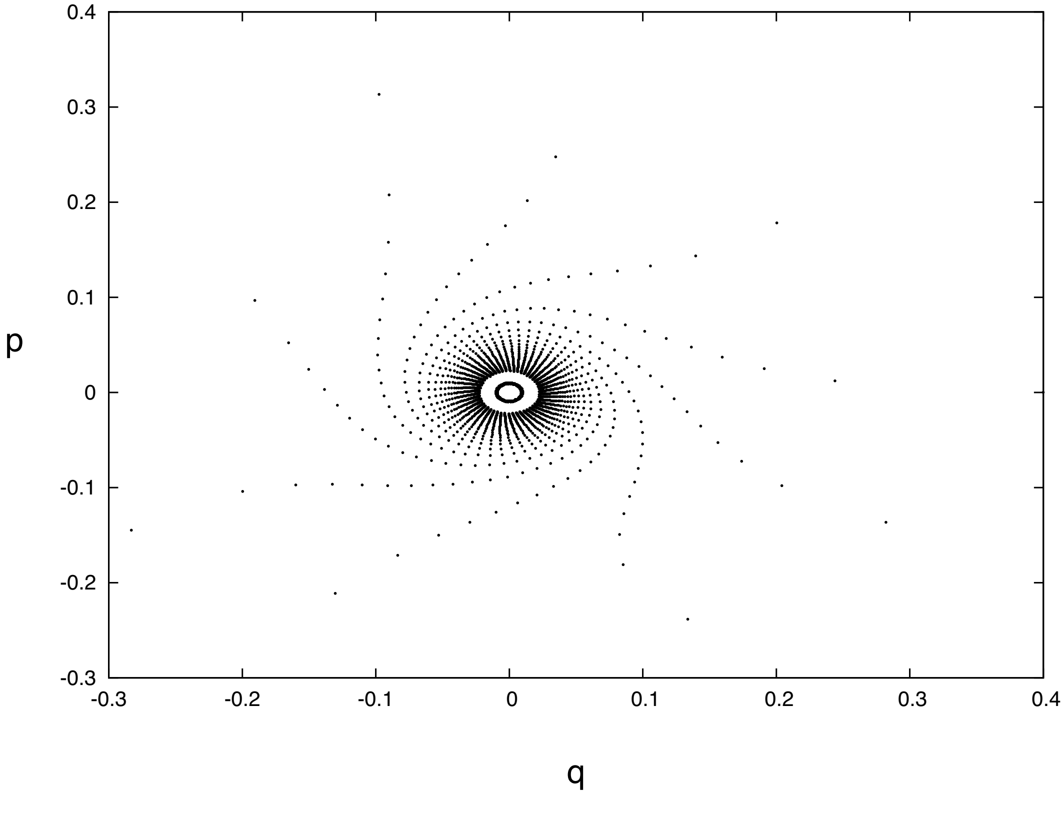


Figure 34.1.2: Phase-space portrait, in the case $\theta/2\pi = 0.22$, resulting from applying the map $\mathcal{M}^{\rm tr}$ repeatedly (2000 times) to the two initial conditions $(q,p) = (.01,0)$ and $(.4,0)$ to find their orbits. The orbits appear to spiral into the origin.

Suppose we also retain the next term in the Lie series for $\exp(: qp^2 :)$ to form the degree-three symplectic jet map

$$\mathcal{N}^{\rm tr3} = \mathcal{I}+ : qp^2 : + : qp^2 :^2 /2. \tag{34.1.14}$$

The truncated map $\mathcal{N}^{\rm tr3}$ has the effect

$$\begin{aligned}\bar{q} &= \mathcal{N}^{\rm tr3}q = (\mathcal{I} + :qp^2: + :qp^2:^2/2)q \\ &= q + [qp^2, q] + [qp^2, [qp^2, q]]/2 = q - 2qp + qp^2, \end{aligned} \tag{34.1.15}$$

$$\begin{aligned}\bar{p} &= \mathcal{N}^{\rm tr3}p = (\mathcal{I} + :qp^2: + :qp^2:^2/2)p \\ &= p + [qp^2, p] + [qp^2, [qp^2, p]]/2 = p + p^2 + p^3. \end{aligned} \tag{34.1.16}$$

Again define a corresponding map $\mathcal{M}^{\rm tr3}$ by writing

$$\mathcal{M}^{\rm tr3} = \mathcal{R}\mathcal{N}^{\rm tr3}. \tag{34.1.17}$$

Figure 1.3 shows the orbits of $\mathcal{M}^{\rm tr3}$ for four initial conditions relatively near the origin. Now orbits move away from the origin. And the farther they are from the origin, the faster they move further away from the origin. Indeed, the orbits of initial conditions somewhat farther from the origin move very far from the origin under 2000 applications of $\mathcal{M}^{\rm tr3}$. This motion away from the origin occurs because $\mathcal{N}^{\rm tr3}$, and consequently $\mathcal{M}^{\rm tr3}$, is again not symplectic, although more nearly symplectic than $\mathcal{N}^{\rm tr}$ because $\mathcal{N}^{\rm tr3}$ is a degree-three symplectic jet whereas $\mathcal{N}^{\rm tr}$ is a degree-two symplectic jet. Now, following the discussion of Section 22.1, the map $\mathcal{M}^{\rm tr3}$ must have a factorization of the form

$$\begin{aligned}\mathcal{M}^{\rm tr3} = \; & \exp(\mathcal{G}_5)\exp(\mathcal{G}_6)\exp(\mathcal{G}_7)\cdots \times \\ & \exp(-(\theta/2):p^2+q^2:)\exp(:qp^2:)\exp(:f_5:)\exp(:f_6:)\cdots \end{aligned} \tag{34.1.18}$$

with the non-Hamiltonian vector field $\mathcal{G}_5$ being nonzero.

We have learned that violation of the symplectic condition can lead both to spurious damping (motion toward the origin) and spurious growth (motion away from the origin).

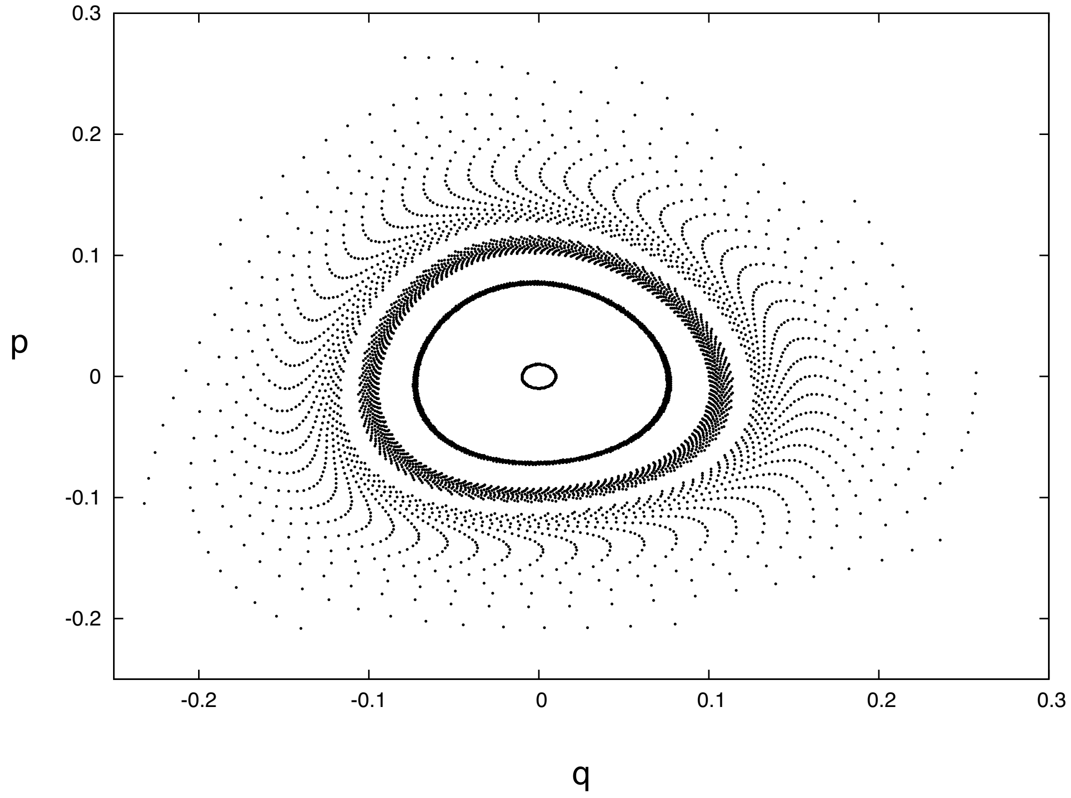


Figure 34.1.3: Phase-space portrait, in the case $\theta/2\pi = 0.22$, resulting from applying the map $\mathcal{M}^{\mathrm{tr3}}$ repeatedly (2000 times) to the four initial conditions $(q,p) = (.01,0)$, $(.075,0)$, $(.1,0)$, and $(.125,0)$ to find their orbits. The orbits appear to move away from origin.

Exercises

34.1.1. Verify that (1.10) and (1.11) are truncated Taylor expansions of (1.7) and (1.8). Show that the map $\mathcal{N}^{\rm tr}$ given by (1.10) and (1.11) satisfies the relation

$$[\bar{q}, \bar{p}] = 1 - 4p^2, \tag{34.1.19}$$

and is therefore a symplectic jet but not a symplectic map.

34.1.2. Find $\mathcal{G}_4$ and f_4 in the factorization (1.13) for the map $\mathcal{M}^{\rm tr}$

34.2 Symplectic Completion of Symplectic Jets

34.2.1 Criteria

34.2.2 Monomial Approximation

34.2.3 Generating Function Approximation

34.2.4 Cremona Maps

Kick Approximation

Jolt Approximation

34.3 Connection Between Mixed-Variable Generating Functions and Lie Generators

Sections 6.5 through 6.7 described the parameterization of symplectic maps in terms of mixed-variable generating functions. Chapters 7 through 9 described, among other things, the parameterization of symplectic maps in terms of Lie generators. The purpose of this section is to study the relation between these two parameterizations.

In particular, suppose $\mathcal{N}$ is a *nonlinear* symplectic map of the form

$$\mathcal{N} = \exp(:f_3:)\exp(:f_4:)\exp(:f_5:)\cdots . \tag{34.3.1}$$

Select some Darboux matrix α. Then, generally, there will be some source function $g(u)$ that will produce the same map using (6.7.21). We will see that the source function $g(u)$ has a homogeneous polynomial expansion of the form

$$g = g_2 + g_3 + g_4 + \cdots . \tag{34.3.2}$$

What we wish to do is to find the relation between the f_m and the g_m.

34.3.1 Method of Calculation

The task we have posed is algebraically complicated. As a first step, with the aid of the CBH series, we combine all the exponents appearing in (3.1) into one grand exponent $: e :$. Thus, we may also write

$$\mathcal{N} = \exp(: e :) \tag{34.3.3}$$

with

$$e = e_3 + e_4 + e_5 + \cdots \tag{34.3.4}$$

where

$$\begin{aligned} e_3 &= f_3, \\ e_4 &= f_4, \\ e_5 &= f_5 + [f_3, f_4], \\ e_6 &= f_6 +, \text{ etc.} \end{aligned} \tag{34.3.5}$$

Next define a function h by the rule

$$h = -e = h_3 + h_4 + h_5 + \cdots \tag{34.3.6}$$

so that $\mathcal{N}$ can be written in the form

$$\mathcal{N} = \exp(- : h :). \tag{34.3.7}$$

That is, $\mathcal{N}$ can be viewed as the map generated by integrating from $t = 0$ to $t = 1$ the equations of motion arising from the *time-independent* Hamiltonian h. Our intermediate goal now is to find a relation between the h_m and the g_m.

This goal can be achieved with the aid of the results of Section 6.7.3.2. There we learned that the source function and the Hamiltonian are related by (6.7.131) and (6.7.145). In this instance, we should set $t^i = 0$, $t = 1$, and $H(\zeta, \tau) = h(\zeta)$ to give the result

$$g(u) = g(u, t = 1) = (1/2)(\hat{Z}, \alpha^T S \alpha \hat{Z}) + (1/2)A'(u, t = 1) \tag{34.3.8}$$

with

$$A'(u, t = 1) = \int_0^1 d\tau[(\zeta, J\dot{\zeta}) + 2h(\zeta)]. \tag{34.3.9}$$

It is this result that we will manipulate and evaluate to bring it into usable form.

Let us begin with the first term appearing on the right side of (3.8). For this term we undo (6.7.144) to write

$$(1/2)(\hat{Z}, \alpha^T S \alpha \hat{Z}) = (1/2)(U, u) = (1/2)(u, U) \tag{34.3.10}$$

with

$$U = A^\alpha Z + B^\alpha z. \tag{34.3.11}$$

Now work on $(1/2)A'(u, t = 1)$, the second term on the right side of (3.8). Consider the first term appearing in the integrand of (3.9). We know that $\zeta(\tau)$ is given by the relation

$$\zeta(\tau) = \exp(-\tau : h :)z \tag{34.3.12}$$

and therefore

$$\dot{\zeta} = -\exp(-\tau : h :) : h : z = -\exp(-\tau : h :)[h, z]. \tag{34.3.13}$$

Consequently we find that

$$(\zeta, J\dot{\zeta}) = -(\exp(-\tau : h :)z, J\exp(-\tau : h :)[h, z]) = -\exp(-\tau : h :)(z, J[h, z]). \tag{34.3.14}$$

Let us evaluate $(z, J[h, z])$. In terms of components, and employing the convention of summing over repeated indices, we find that

$$\begin{aligned}(z, J[h, z]) = z_a J_{ab}[h, z_b] &= z_a J_{ab}(\partial h/\partial z_c)J_{cd}(\partial z_b/\partial z_d)\\ &= z_a J_{ab}(\partial h/\partial z_c)J_{cd}\delta_{bd}\\ &= z_a J_{ab}(\partial h/\partial z_c)J_{cb}\\ &= z_a J_{ab}(J^T)_{bc}(\partial h/\partial z_c)\\ &= z_a(JJ^T)_{ac}(\partial h/\partial z_c)\\ &= z_a(\partial h/\partial z_a).\end{aligned} \tag{34.3.15}$$

The right side of (3.15) cries out for Euler's homogeneous function theorem. By this theorem, for each homogeneous component appearing in (3.6), we have the result

$$z_a(\partial h_m/\partial z_a) = m h_m. \tag{34.3.16}$$

Therefore we may also write

$$(z, J[h, z]) = \sum_{m=3}^{\infty} m h_m. \tag{34.3.17}$$

It follows that

$$(\zeta, J\dot{\zeta}) = -(\exp(-\tau : h :)\sum_{m=3}^{\infty} m h_m. \tag{34.3.18}$$

We also note that the second term appearing in the integrand of (3.9) can be rewritten in the form

$$2h(\zeta) = 2h(\exp(-\tau : h :)z) = 2\exp(-\tau : h :)h(z). \tag{34.3.19}$$

Consequently, the full integrand can be rewritten as

$$(\zeta, J\dot{\zeta}) + 2h(\zeta) = \exp(-\tau : h :)\sum_{m=3}^{\infty}(2-m)h_m. \tag{34.3.20}$$

Correspondingly, the integral takes the form

$$A'(u, t = 1) = \int_0^1 d\tau[\exp(-\tau : h :)\sum_{m=3}^{\infty}(2-m)h_m. \tag{34.3.21}$$

Since the τ behavior has been isolated, this integral can be evaluated to give the result

$$A'(u, t = 1) = \mathrm{iex}(- : h :)\sum_{m=3}^{\infty}(2-m)h_m. \tag{34.3.22}$$

We can now put all our results together to obtain the relation

$$g(u) = (1/2)(u, U) + (1/2)\text{iex}(- : h :) \sum_{m=3}^{\infty} (2 - m)h_m. \tag{34.3.23}$$

Recall the relation (6.7.18), which can be rewritten in the form

$$u = C^\alpha Z + D^\alpha z = C^\alpha \mathcal{N} z + D^\alpha z = C^\alpha \exp(- : h :)z + D^\alpha z. \tag{34.3.24}$$

This relation is to solved to give z in terms of u. This $z(u)$ must then be substituted into the right side of (3.23) to yield $g(u)$. Finally, g must be expanded in homogeneous polynomials as in (3.2).

34.3.2 Computing g_2

We begin with the computation g_2. The relation (3.24) has the expansion

$$\begin{aligned} u &= C^\alpha \exp(- : h :)z + D^\alpha z = C^\alpha(\mathcal{I} - : h : + \cdots)z + D^\alpha z \\ &= (C^\alpha + D^\alpha)z + C^\alpha(- : h : + \cdots)z. \end{aligned} \tag{34.3.25}$$

Thus, because the quantity $[(- : h : + \cdots)z]$ consists of terms that are of order 2 and higher, the expansion (3.24) has the inverse expansion

$$z = z^{(1)}(u) + O(u^2) \tag{34.3.26}$$

with

$$z^{(1)}(u) = (C^\alpha + D^\alpha)^{-1}u. \tag{34.3.27}$$

Observe that the second set of terms on the right side of (3.23) is of order 3 and higher in z. It follows that second set of terms contributes only terms of order u^3 and higher. Therefore second degree terms, the ones required for g_2, can only come from the quantity $(1/2)(u, U)$, the first term on the right side of (3.23). We have from (3.11) the expansion

$$\begin{aligned} U &= A^\alpha Z + B^\alpha z = A^\alpha \mathcal{N} z + B^\alpha z \\ &= A^\alpha \exp(- : h :)z + B^\alpha z = A^\alpha(\mathcal{I} - : h : + \cdots)z + B^\alpha z \\ &= (A^\alpha + B^\alpha)z + A^\alpha(- : h : + \cdots)z \\ &= (A^\alpha + B^\alpha)z + O(z^2). \end{aligned} \tag{34.3.28}$$

Next substitute (3.26) and (3.27) into (3.28) to yield the expansion

$$U = (A^\alpha + B^\alpha)(C^\alpha + D^\alpha)^{-1}u + O(u^2). \tag{34.3.29}$$

Also observe that the matrix product appearing in (3.29) can be written in the Möbius transformation form

$$(A^\alpha + B^\alpha)(C^\alpha + D^\alpha)^{-1} = (A^\alpha I + B^\alpha)(C^\alpha I + D^\alpha)^{-1} = T_\alpha(I). \tag{34.3.30}$$

Therefore (3.29) can be rewritten as

$$U = Wu + O(u^2) \tag{34.3.31}$$

with

$$W = T_\alpha(I), \tag{34.3.32}$$

and it follows that

$$(1/2)(u, U) = (1/2)(u, Wu) + O(u^3). \tag{34.3.33}$$

Thus we have the result

$$g_2(u) = (1/2)(u, Wu), \tag{34.3.34}$$

which is what we should have expected. Consult Exercise 6.7.1. We see that W is well defined provided

$$\det(C^\alpha + D^\alpha) \neq 0, \tag{34.3.35}$$

which was also required to write (3.27).

34.3.3 Low Order Results: Computing g_3 and g_4

Let us push on to compute g_3 and g_4. To do so, we will need to retain various higher-order terms in the expressions we have already encountered. We might think that we need to retain higher order terms in (3.25) or (3.26), in (3.28), and in the second term on the right side of (3.23). In fact, this would be one way to proceed. However, at this stage, it is also possible to avoid dealing with (3.28) entirely, thereby achieving a considerable simplification.

Again Euler comes to the rescue. We thought we had to deal with (3.28) because it appeared to be needed to compute the first term on the right side of (3.23), namely $(1/2)(u, U)$. However, using (6.7.14), we may write

$$(u, U) = \sum_a u_a(\partial g/\partial u_a). \tag{34.3.36}$$

Therefore, if we decompose g into homogeneous polynomials as in (3.2) by writing

$$g = \sum_{n=2}^{\infty} g_n, \tag{34.3.37}$$

we find by Euler's theorem the result

$$(1/2)(u, U) = (1/2)\sum_a u_a(\partial g/\partial u_a) = (1/2)\sum_{n=2}^{\infty}\sum_a u_a(\partial g_n/\partial u_a) = (1/2)\sum_{n=2}^{\infty} n g_n. \tag{34.3.38}$$

Moreover, we find that

$$g - (1/2)(u, U) = \sum_{n=2}^{\infty}(1 - n/2)g_n. \tag{34.3.39}$$

Consequently, when (6.7.14) is taken into account, the defining relation (3.23) can also be written in the form

$$\sum_{n=3}^{\infty}(2-n)g_n = \text{iex}(-:h:)\sum_{m=3}^{\infty}(2-m)h_m. \tag{34.3.40}$$

It is this form that we will employ to compute g_3 and g_4.

We must still retain higher-order terms in (3.25). Doing so gives the result

$$\begin{aligned} u &= (C^\alpha + D^\alpha)z + C^\alpha(-:h:+:h:^2/2!+\cdots)z \\ &= (C^\alpha + D^\alpha)z + C^\alpha(-:h_3:+:h_3:^2/2!-:h_4:+\cdots)z \\ &= (C^\alpha + D^\alpha)z + C^\alpha(-[h_3,z]+(1/2)[h_3,[h_3,z]]-[h_4,z]+\cdots). \end{aligned} \tag{34.3.41}$$

Let us rewrite this relation in the implicit form

$$z = (C^\alpha + D^\alpha)^{-1}u - (C^\alpha + D^\alpha)^{-1}C^\alpha\{-[h_3,z]+(1/2)[h_3,[h_3,z]]-[h_4,z]+\cdots\}. \tag{34.3.42}$$

We can now invert the relation by iteration. In lowest order we have the results (3.26) and (3.27). In next order, we find

$$z = z^{(2)}(u) + O(u^3). \tag{34.3.43}$$

with

$$z^{(2)}(u) = (C^\alpha + D^\alpha)^{-1}u - (C^\alpha + D^\alpha)^{-1}C^\alpha\{-[h_3,z]\}|_{z=z^{(1)}}. \tag{34.3.44}$$

Also, we need to expand the right side of (3.40). Through terms of degree 4 we have the result

$$\begin{aligned} \text{iex}(-:h:)\sum_{m=3}^{\infty}(2-m)h_m &= (\mathcal{I}-:h:/2+\cdots)(-h_3-2h_4-\cdots) \\ &= -h_3-2h_4+O(z^5). \end{aligned} \tag{34.3.45}$$

Next we need to express both sides of (3.45) as functions of u using (3.26) and (3.43). Doing so we find the result

$$\{\text{iex}(-:h:)\sum_{m=3}^{\infty}(2-m)h_m\}|_{z=z^{(2)}} = -(h_3)|_{z=z^{(2)}} - 2(h_4)|_{z=z^{(1)}} + O(u^5). \tag{34.3.46}$$

Note that because $z^{(2)}(u)$ contains quadratic terms in u, see (3.44), the first quantity on the right of (3.46) will contribute terms that are both of degree 3 and 4 in u. Let us see what they are. Rewrite (3.44) in the form

$$z^{(2)}(u) = z^{(1)}(u) + \Delta \tag{34.3.47}$$

where

$$\Delta = -(C^\alpha + D^\alpha)^{-1}C^\alpha\{-[h_3,z]\}|_{z=z^{(1)}}. \tag{34.3.48}$$

With this notation, and inspired by Taylor, we find the result

$$(h_3)|_{z=z^{(2)}} = h_3(z^{(1)}+\Delta) = h_3(z^{(1)}) + \sum_a \Delta_a(\partial h_3/\partial z_a)|_{z=z^{(1)}} + O(\Delta^2). \tag{34.3.49}$$

Also, the quantity $[h_3, z]$ appearing in Δ can be evaluated,

$$\begin{aligned} [h_3, z_a] &= \sum_{bc} (\partial h_3/\partial z_b) J_{bc} (\partial z_a/\partial z_c) = \sum_{bc} (\partial h_3/\partial z_b) J_{bc} \delta_{ac} \\ &= \sum_b (\partial h_3/\partial z_b) J_{ba} = -\sum_b J_{ab} (\partial h_3/\partial z_b). \end{aligned} \qquad (34.3.50)$$

This result can be written more compactly in vector notation as

$$[h_3, z] = -J(\partial h_3/\partial z). \qquad (34.3.51)$$

Correspondingly, Δ takes the more compact form

$$\Delta = K(\partial h_3/\partial z) \qquad (34.3.52)$$

where K is the matrix

$$K = -(C^\alpha + D^\alpha)^{-1} C^\alpha J. \qquad (34.3.53)$$

Finally, (3.49) takes the compact form

$$(h_3)|_{z=z^{(2)}} = (h_3)|_{z=z^{(1)}} + ((\partial h_3/\partial z), K(\partial h_3/\partial z))|_{z=z^{(1)}} + O(u^5), \qquad (34.3.54)$$

and (3.46) becomes

$$\begin{aligned} \{\mathrm{iex}(- : h :) \sum_{m=3}^{\infty} (2-m) h_m\}|_{z=z^{(2)}} &= -(h_3)|_{z=z^{(1)}} - ((\partial h_3/\partial z), K(\partial h_3/\partial z))|_{z=z^{(1)}} \\ &\quad -2(h_4)|_{z=z^{(1)}} + O(u^5). \end{aligned} \qquad (34.3.55)$$

Now we are ready to equate terms of like degree in (3.40). Equating terms of degree 3 gives the result

$$-g_3(u) = -h_3(z)|_{z=z^{(1)}}, \qquad (34.3.56)$$

or

$$g_3(u) = h_3(z)|_{z=z^{(1)}}. \qquad (34.3.57)$$

And equating terms of degree 4 gives the result

$$-2g_4(u) = -((\partial h_3/\partial z), K(\partial h_3/\partial z))|_{z=z^{(1)}} - 2h_4(z)|_{z=z^{(1)}} \qquad (34.3.58)$$

or

$$g_4(u) = (1/2)((\partial h_3/\partial z), K(\partial h_3/\partial z))|_{z=z^{(1)}} + h_4(z)|_{z=z^{(1)}} \qquad (34.3.59)$$

In terms of the f_m, these relations can be written in the form

$$g_3(u) = -f_3(z)|_{z=z^{(1)}}, \qquad (34.3.60)$$

$$g_4(u) = (1/2)((\partial f_3/\partial z), K(\partial f_3/\partial z))|_{z=z^{(1)}} - f_4(z)|_{z=z^{(1)}}. \qquad (34.3.61)$$

34.3.4 Two Examples

Eventually we will want to find higher-order results to determine the $g_n(u)$ for, say, $n \leq 8$ and some select Darboux matrices α. Before doing so, let us see what can be said so far for familiar choices of α. If we look at the Darboux α matrices for the generating functions F_1 and F_4, see Table 6.7.1, we observe that the matrices $(C^\alpha + D^\alpha)$ are singular. Hence these Darboux matrices cannot be used for our purposes. By contrast, the matrices $(C^\alpha + D^\alpha)$ for the three Darboux matrices associated with F_2, F_3, and F_+ are invertible. We will study the cases of F_2 and F_+. The case of F_3 is similar to that of F_2.

The Case of F_2

First consider the case of F_2. We find from (6.7.56) that

$$C^\alpha + D^\alpha = I^{2n} \tag{34.3.62}$$

and

$$C^\alpha = \begin{pmatrix} 0 & 0 \\ 0 & I^n \end{pmatrix}. \tag{34.3.63}$$

It follows from (3.27) that

$$z^{(1)}(u) = u. \tag{34.3.64}$$

Also, from (6.7.55) and (3.32), we find that

$$W = \begin{pmatrix} 0 & I^n \\ I^n & 0 \end{pmatrix}. \tag{34.3.65}$$

Finally, from (3.41), (3.47), and (3.63), we find that

$$K = -C^\alpha J = \begin{pmatrix} 0 & 0 \\ I^n & 0 \end{pmatrix}. \tag{34.3.66}$$

We are now ready to compute $g_2(u)$ through $g_4(u)$. As in Exercise 6.7.5, partition u into position-like and momentum-like components by writing

$$u = (v; w). \tag{34.3.67}$$

Then, from (3.42), (3.48), and (3.49) we have for $g_2(u)$ the result

$$g_2(u) = (v, w), \tag{34.3.68}$$

which can also be written in the form

$$g_2(u) = (q, p)|_{z=u}. \tag{34.3.69}$$

Next, from (3.45) and (3.47), we find that

$$g_3(u) = -f_3(u). \tag{34.3.70}$$

Finally, from (3.46), (3.47), and taking into account the form of K given by (3.49), we find the results

$$g_4(u) = -f_4(u) + (1/2)(\partial f_3/\partial q, \partial f_3/\partial p)|_{z=u}. \tag{34.3.71}$$

The Case of F_+

Next consider the case of F_+. We find from (6.7.67) that

$$C^\alpha + D^\alpha = \sqrt{2} I^{2n} \tag{34.3.72}$$

and

$$C^\alpha = (1/\sqrt{2}) I^{2n}. \tag{34.3.73}$$

It follows that in this case

$$z^{(1)}(u) = (1/\sqrt{2})u. \tag{34.3.74}$$

Also, from (6.7.67) and (3.15), we find that

$$W = 0. \tag{34.3.75}$$

Finally, from (3.41), (3.55), and (6.7.67), we find that

$$K = -(1/2)J. \tag{34.3.76}$$

We are again ready to compute $g_2(u)$ through $g_4(u)$. From (3.42) and (3.56) we find that

$$g_2(u) = 0. \tag{34.3.77}$$

Next, from (3.45) and (3.55), we find that

$$g_3(u) = -f_3(u/\sqrt{2}). \tag{34.3.78}$$

Finally we observe from (3.57) that

$$((\partial f_3/\partial z), K(\partial f_3/\partial z)) = -(1/2)((\partial f_3/\partial z), J(\partial f_3/\partial z)) = 0 \tag{34.3.79}$$

since J is an antisymmetric matrix. It follows from (3.46) and (3.60) that in this case g_4 takes the simple form

$$g_4(u) = -f_4(u/\sqrt{2}). \tag{34.3.80}$$

34.3.5 Exploration

Let us explore what maps are produced when the source function consists only of quadratic and cubic terms,

$$g(u) = g_2(u) + g_3(u). \tag{34.3.81}$$

For simplicity, we will explore only the use of F_2 and F_+ generating functions, and work with only a two-dimensional phase space so that $z = (q; p)$ and $Z = (Q; P)$.

Use of F_2

General Discussion

Let us begin with the use of F_2. In that case, we will consider generating functions of the form

$$F_2(q,P) = qP + aq^3 + bq^2P + cqP^2 + dP^3 \tag{34.3.82}$$

with arbitrary coefficients a through d. Then use of (6.7.54) produces the implicit relations

$$p = P + 3aq^2 + 2bqP + cP^2, \tag{34.3.83}$$

$$Q = q + bq^2 + 2cqP + 3dP^2. \tag{34.3.84}$$

Since these equations are quadratic, they can be solved exactly, and we will do so shortly. First, however, let us find the first few terms in the Taylor expansions of $Q(q,p)$ and $P(q,p)$ in powers of q and p. Rewrite (3.64) in the form

$$P = p - 3aq^2 - 2bqP - cP^2. \tag{34.3.85}$$

Now we can expand Q and P in powers of q and p by iteration of (3.65) and (3.66). In lowest approximation, they have the solution

$$Q = q + O(z^2), \tag{34.3.86}$$

$$P = p + O(z^2). \tag{34.3.87}$$

Now substitute (3.67) and (3.68) into (3.65) and (3.66) to get the improved solution

$$Q = q + bq^2 + 2cqp + 3dp^2 + O(z^3), \tag{34.3.88}$$

$$P = p - 3aq^2 - 2bqp - cp^2 + O(z^3) \tag{34.3.89}$$

For our present purposes we will be content with expansions that retain terms through degree three. This can be achieved by substituting (3.69) and (3.70) into (3.65) and (3.66). Doing so gives the results

$$Q = q + bq^2 + 2cqp + 3dp^2 + *q^3 + *q^2p + *qp^2 + *p^3 + O(z^4), \tag{34.3.90}$$

$$P = p - 3aq^2 - 2bqp - cp^2 + *q^3 + *q^2p + *qp^2 + *p^3 + O(z^4). \tag{34.3.91}$$

For comparison, let us evaluate the Taylor series for the transformation

$$Z = \mathcal{N}z \tag{34.3.92}$$

where

$$\mathcal{N} = \exp(:f_3:) \tag{34.3.93}$$

with

$$f_3(z) = -aq^3 - bq^2p - cqp^2 - dp^3. \tag{34.3.94}$$

Then it is easily verified that the Lie transformation

$$Z = \exp(:f_3:)z = \sum_{m=0}^{\infty}(1/m!) :f_3:^m z = z + :f_3: z + (1/2!) :f_3:^2 z + O(z^4) \quad (34.3.95)$$

gives the result

$$Q = q + bq^2 + 2cqp + 3dp^2 + *q^3 + *q^2p + *qp^2 + *p^3 + O(z^4), \quad (34.3.96)$$

$$P = p - 3aq^2 - 2bqp - cp^2 + *q^3 + *q^2p + *qp^2 + *p^3 + O(z^4). \quad (34.3.97)$$

We see that the linear and quadratic terms in (3.71) and (3.72) agree with those in (3.77) and (3.78), respectively. However, the cubic terms do not.

These results are to be expected based on the findings of Subsection 27.3.3. There we saw that g_3 and f_3 should be related by (3.53), and that is what has been done in writing (3.63) and (3.76). Therefore the quadratic terms in in (3.71) and (3.72) should agree with those in (3.77) and (3.78). With regard to cubic terms, in writing (3.62) we have implicitly made the requirement

$$g_n(u) = 0 \text{ for } n \geq 4. \quad (34.3.98)$$

And, in writing (3.74), we have implicitly made the requirement

$$f_n(z) = 0 \text{ for } n \geq 4. \quad (34.3.99)$$

But we see from (3.54) that in general these requirements are incompatible. Therefore we expect differences in the cubic (and higher-order) terms.

As promised, let us now solve the relations (3.65) and (3.66) exactly. We must distinguish two cases:

The Case When $c = 0$

If $c = 0$, (3.66) has the immediate solution

$$P = (p - 3aq^2)/(2bq + 1). \quad (34.3.100)$$

And substituting this result into (3.65) gives the complementary result

$$Q = q + bq^2 + 3d(p - 3aq^2)^2/(2bq + 1)^2. \quad (34.3.101)$$

The Case When $c \neq 0$

When $c \neq 0$, the relation (3.66) is quadratic in P and has the solution

$$P = [1/(2c)]\{-(2bq + 1) + [(2bq + 1)^2 + 4c(p - 3aq^2)]^{1/2}\}. \quad (34.3.102)$$

The implicit relation (3.64) can be solved to give the explicit relation

$$P = [1/(2c)]\{-(1 + 2bq) + [1 + 4bq + cp + 3(b^2 - 4ac)q^2]^{1/2}\}. \quad (34.3.103)$$

And (3.81) can then be substituted into (3.65) to give the complementary explicit relation

$$Q = . \tag{34.3.104}$$

We see that, as functions of q and p, Q and P *generically* have branch points.[2] They occur on the surface

$$1 + 4bq + cp + 3(b^2 - 4ac)q^2 = 0. \tag{34.3.105}$$

By contrast, we know from the work of Section 25.3 that the map given by (3.73) and (3.74) generally has poles.

Case When Only $a \neq 0$

Let us consider some special cases. First suppose that only $a \neq 0$. Then the Lie transformation series (3.76) terminates and gives the exact result

$$Q = q, \tag{34.3.106}$$

$$P = p - 3aq^2. \tag{34.3.107}$$

Also, in this case, solution of the implicit relations (3.64) and (3.65) gives identical results. Thus in this case, which is easily verified to be that of a kick map, the use of F_2 and the exact Lie transformation give the same result. It is an easy calculation to show that the same holds true when only $d \neq 0$.

Jolt Case

Next assume that f_3 is of the form

$$f_3 = (\alpha q - \beta p)^3 \tag{34.3.108}$$

which amounts to setting

$$a = -\alpha^3, \tag{34.3.109}$$

$$b = 3\alpha^2\beta, \tag{34.3.110}$$

$$c = -3\alpha\beta^2, \tag{34.3.111}$$

$$d = \beta^3. \tag{34.3.112}$$

Then it is easily verified that the Lie transformation series (3.76) also terminates and gives the exact result

$$Q = q + 3\beta(\alpha q - \beta p)^2, \tag{34.3.113}$$

$$P = p + 3\alpha(\alpha q - \beta p)^2. \tag{34.3.114}$$

See Section 22.3. In fact, $\mathcal{N}$ in this case is a jolt map. See Exercise *. By contrast, solution of the implicit relations (3.64) and (3.65) in this case gives the results

$$Q =, \tag{34.3.115}$$

[2] However.

$$P = . \qquad (34.3.116)$$

We see that the Q and P given by (3.91) and (3.92) are entire functions of q and p. By contrast, the relations (3.93) and (3.94) show that the map produced by F_2 in this case has branch points. They are located on the surface

$$= . \qquad (34.3.117)$$

Case When Only $c \neq 0$

As another example, suppose f_3 is of the form

$$f_3 = -cqp^2 \qquad (34.3.118)$$

as in (1.6). So doing amounts to assuming that only $c \neq 0$ in (3.75). As already seen, setting $c = -1$ in (3.73) and (3.74) leads to the relation

$$Q = \mathcal{N}q = q(1-p)^2, \qquad (34.3.119)$$

$$P = \mathcal{N}p = p/(1-p). \qquad (34.3.120)$$

By contrast, the implicit equations (3.64) and (3.65) in this case have the explicit solution

$$Q = q(1+4cp)^{1/2}, \qquad (34.3.121)$$

$$P = [1/(2c)][(1+4cp)^{1/2} - 1]. \qquad (34.3.122)$$

We see that in this case the map produced by F_2 has a branch point (when $c = -1$) on the surface $p = 1/4$ while, according to (3.98), the exact Lie transformation map has a pole on the surface $p = 1$.

The last case to be considered in this vein is that of $b \neq 0$ and all other coefficients in (3.63) or (3.75) set to zero. You, dear reader, will have the pleasure of doing so in Exercise *.

F_2 Symplectic Completion of $\mathcal{N}^{\rm tr}$

Let $\mathcal{N}^{\rm sc}$ be the map given by (3.99) and (3.100) when $c = -1$. We may view $\mathcal{N}^{\rm sc}$ as a *symplectic completion* of the degree-two symplectic jet map $\mathcal{N}^{\rm tr}$ given by (1.9). Corespondingly, we define the associated map $\mathcal{M}^{\rm sc}$ by the relation

$$\mathcal{M}^{\rm sc} = \mathcal{R}\mathcal{N}^{\rm sc}. \qquad (34.3.123)$$

Figure 3.1 shows the result of applying $\mathcal{M}^{\rm sc}$ repeatedly to four initial conditions for the case $\theta/2\pi = 0.22$. Note that, unlike the cases of Figures 1.2 and 1.3, points on the orbit no longer spiral into or out of the origin. Moreover, the behavior of the orbits is similar to that shown in Figure 1.1.

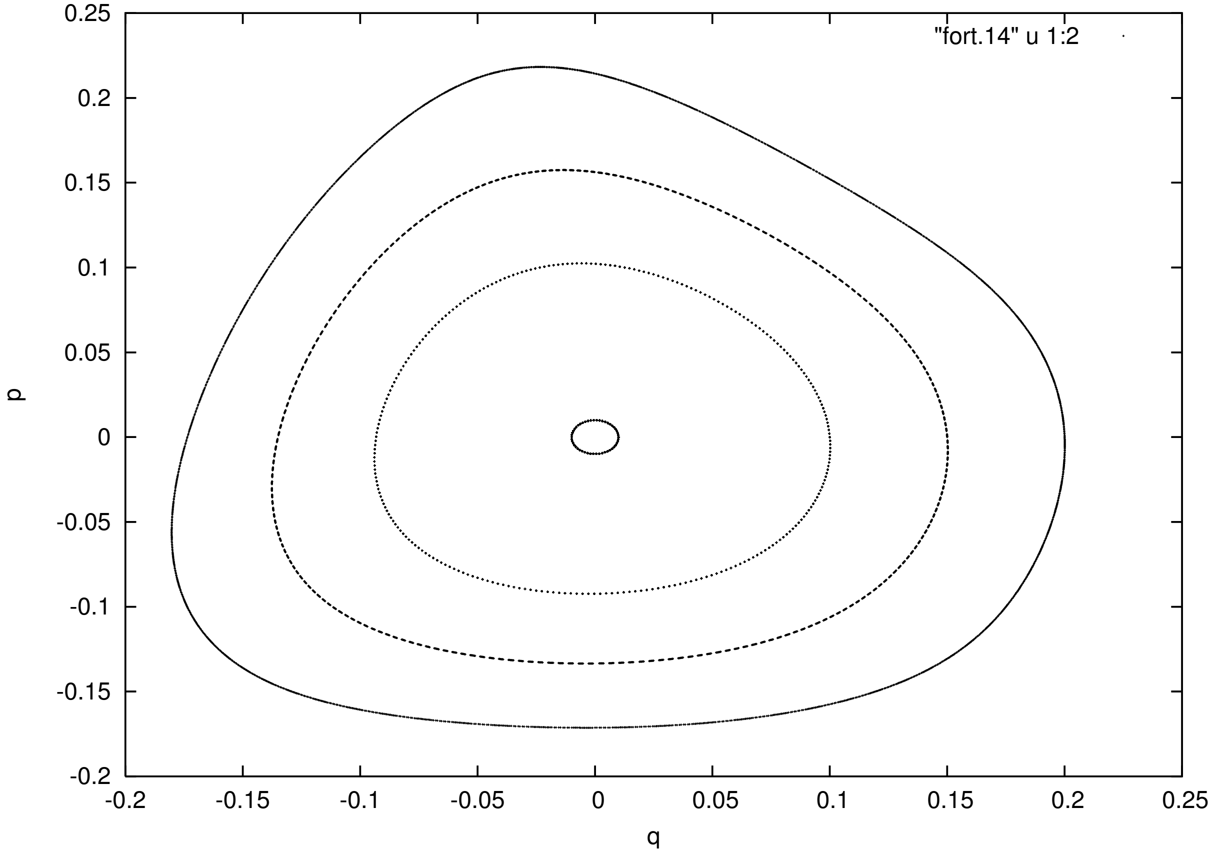


Figure 34.3.1: Phase-space portrait, in the case $\theta/2\pi = 0.22$, resulting from applying the map $\mathcal{M}^{\rm sc}$ repeatedly (2000 times) to the four initial conditions $(q,p) = (.01,0)$, $(.1,0)$, $(.15,0)$, and $(.2,0)$ to find their orbits.

Root Trick

Of course, we know that in this case use of $\mathcal{N}^{\rm sc}$ can at best make sense for $|p| < 1/4$. What can be done for points farther from the origin? For a map of the form (3.3) we can easily extract a square root. In accord with the relation (3.6), we will write

$$\mathcal{N}^{1/2} = \mathcal{N}(1/2). \tag{34.3.124}$$

Also, let $\mathcal{N}^{\rm sc}(1/2)$ denote the map given by (3.99) and (3.100) with $c = -1/2$. We may view $\mathcal{N}^{\rm sc}(1/2)$ as the F_2 symplectification of the degree-two jet map $\mathcal{N}^{\rm tr}(1/2)$. We see, from (3.99) and (3.100) with $c = -1/2$, that the map $\mathcal{N}^{\rm sc}(1/2)$ is well defined for $|p| < 1/2$.

Now watch closely. We know that

$$\mathcal{M} = \mathcal{R}\mathcal{N}^{1/2}\mathcal{N}^{1/2}. \tag{34.3.125}$$

Therefore, it makes sense to consider what we will call the *improved symplectically completed* map $\mathcal{M}^{\rm isc}$ defined by the relation

$$\mathcal{M}^{\rm isc} = \mathcal{R}\mathcal{N}^{\rm sc}(1/2)\mathcal{N}^{\rm sc}(1/2). \tag{34.3.126}$$

This map will be defined over a larger region of phase space and will be a better approximation to $\mathcal{M}$. Figure 3.2 shows the result of applying $\mathcal{M}^{\rm isc}$ repeatedly to seven initial conditions for the case $\theta/2\pi = 0.22$. Again points on the orbits neither spiral into or out of the origin, and the orbits more nearly approximate those of Figure 1.1.

Use of the Poincaré Generating Function F_+

General Discussion

We now repeat much of the work above, but this time for the case where F_+ is used. Now (3.58) holds and, in accord with (3.59), the source function g_3 will have the form

$$g_3(u) = (1/\sqrt{2})^3(av^3 + bv^2w + cvw^2 + dw^3). \tag{34.3.127}$$

In this case use of (6.7.21) with α given by (6.7.67) gives the implicit relations

$$Q = q + (1/4)[b(Q + q)^2 + 2c(Q + q)(P + p) + 3d(P + p)^2], \tag{34.3.128}$$

$$P = p - (1/4)[3a(Q + q)^2 + 2b(Q + q)(P + p) + c(P + p)^2]. \tag{34.3.129}$$

Since these equations are quadratic, they can again be solved exactly. However, the solutions are far too long to record. In section 3.5 we will see they again involve square roots, and therefore the functions $Q(q, p)$ and $P(q, p)$, like the F_2 case, have square root branch-point singularities.

But, again, we can obtain Taylor expansions of Q and P in terms if q and p by iteration. Doing so gives, as first and second passes, the expansions (3.67) through (3.70). And the third pass gives the expansions (3.77) and (3.78). That is, unlike the F_2 case, use of F_+ gives expansions that agree with the exact result through terms of third order.

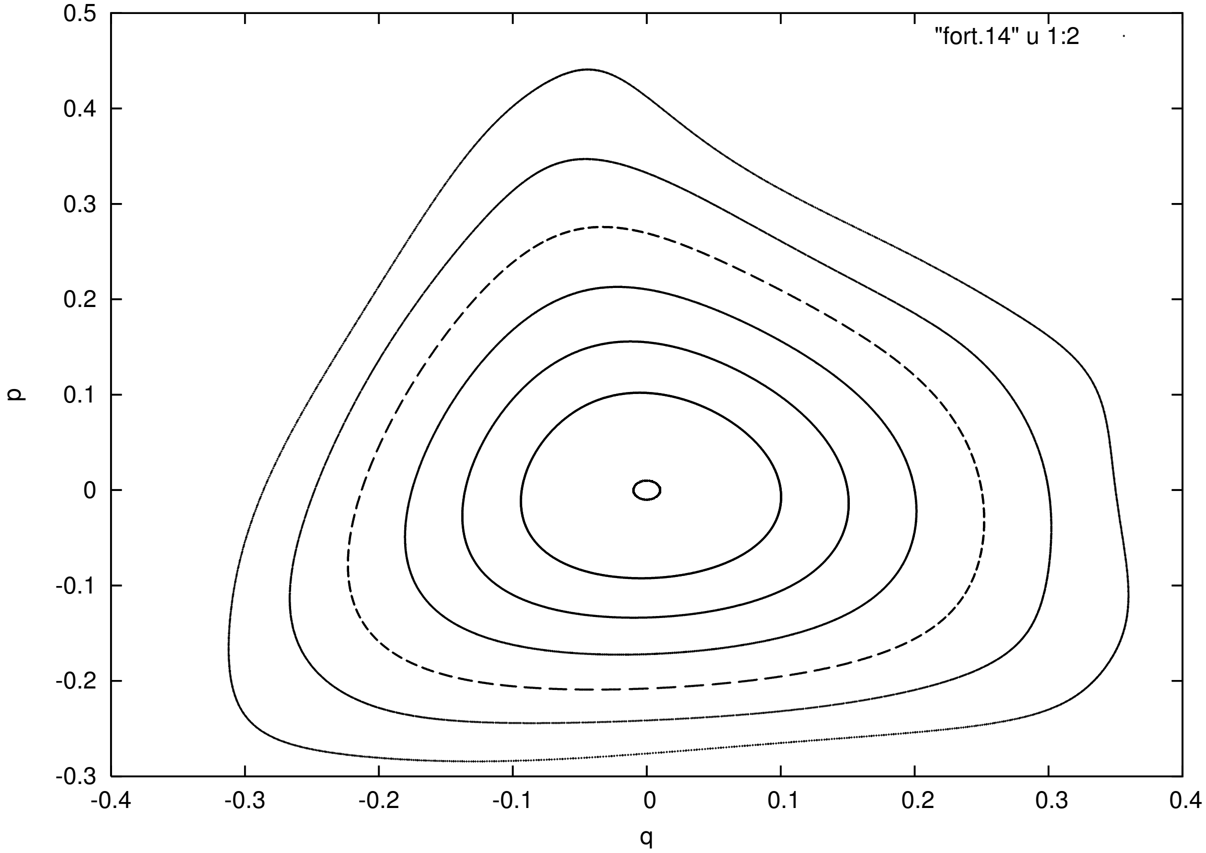


Figure 34.3.2: Phase-space portrait, in the case $\theta/2\pi = 0.22$, resulting from applying the map $\mathcal{M}^{\rm isc}$ repeatedly (2000 times) to the seven initial conditions $(q,p) = (.01,0)$, $(.1,0)$, $(.15,0)$, $(.2,0)$, $(.25,0)$, $(.3,0)$, and $(.35,0)$ to find their orbits.

This accuracy is again to be expected based on the findings of Subsection 27.3.3. There we saw that g_3 and f_3 should be related by (3.59), and that is what has been done in writing (3.105). Therefore the maps produced by the use of F_+ should agree with those produced by the $\mathcal{N}$ given by (3.73) through quadratic terms. With regard to cubic terms, we see from (3.61) that the implicit assumptions (3.79), which are still in effect, imply that f_4 vanishes. Therefore, the cubic terms must also agree. Later we will see that quartic and higher-order terms need not agree.

Cases When Only a or $d \neq 0$

Let us again consider some special cases. First suppose that only $a \neq 0$. Then (3.106) and (3.107) can be solved immediately to give the results (3.84) and (3.85). And if only $d \neq 0$, (3.106) and (3.107) have the solution

$$Q = q + 3dp^2, \tag{34.3.130}$$

$$P = p. \tag{34.3.131}$$

Thus, like F_2, use of F_+ also gives exact results for kick maps.

Jolt Case

Suppose the values of a through d given by (3.114) through (3.117) are employed in (3.106) and (3.107); and that also the quantities Q and P appearing in (3.106) and (3.107) are replaced by the right sides of equations (3.91) and (3.92). Upon doing so one finds that the resulting two equations (which now involve only the quantities α, β, q, and p) are satisfied *identically* for all values of α, β, q, and p. It follows that, unlike the case of F_2, the use of F_+ gives exact results for jolt maps as well. Why this should be so is explained in a subsequent section.

Case When Only $c \neq 0$

If only $c \neq 0$, (3.106) and (3.107) have the solution

$$Q = q(1 + 2cp)^{1/2}/[2 - (1 + 2cp)^{1/2}], \tag{34.3.132}$$

$$P = -[p + (2/c)] + (2/c)(1 + 2cp)^{1/2}. \tag{34.3.133}$$

We see that in this case use of F_+ produces a map that has a branch point on the surface

$$p = -1/(2c). \tag{34.3.134}$$

By contrast, according to (3.100), use of F_2 in this case produces a map that has a branch point on the surface

$$p = -1/(4c). \tag{34.3.135}$$

Therefore the branch-point surface for F_+ is farther from the origin than that of F_2. In particular, for $c = -1$, it is located at $p = 1/2$. But note that it is still closer to the origin than the pole of the exact map which, we have seen, is on the surface $p = 1$.

Again the last case to be considered in this vein is that of only $b \neq 0$. This case is treated in Exercise *.

F_+ Symplectic Completion of $\mathcal{N}^{\mathrm{tr}}$

Let $\mathcal{N}^{\mathrm{psc}}$ be the map given by (3.113) and (3.114) when $c = -1$. We may view $\mathcal{N}^{\mathrm{psc}}$ as the *Poincaré symplectic completion* of the degree-two symplectic jet map $\mathcal{N}^{\mathrm{tr}}$ given by (1.9). Corespondingly, we define the associated map $\mathcal{M}^{\mathrm{psc}}$ by the relation

$$\mathcal{M}^{\mathrm{psc}} = \mathcal{R}\mathcal{N}^{\mathrm{psc}}. \tag{34.3.136}$$

Figure 3.3 shows the result of applying $\mathcal{M}^{\mathrm{psc}}$ repeatedly to seven initial conditions for the case $\theta/2\pi = 0.22$. Again there is no spurious spiraling into or out of the origin. Note also that we have been able to apply this map to a larger region of phase space than we could for $\mathcal{N}^{\mathrm{sc}}$. Compare Figures 3.1 and 3.3.

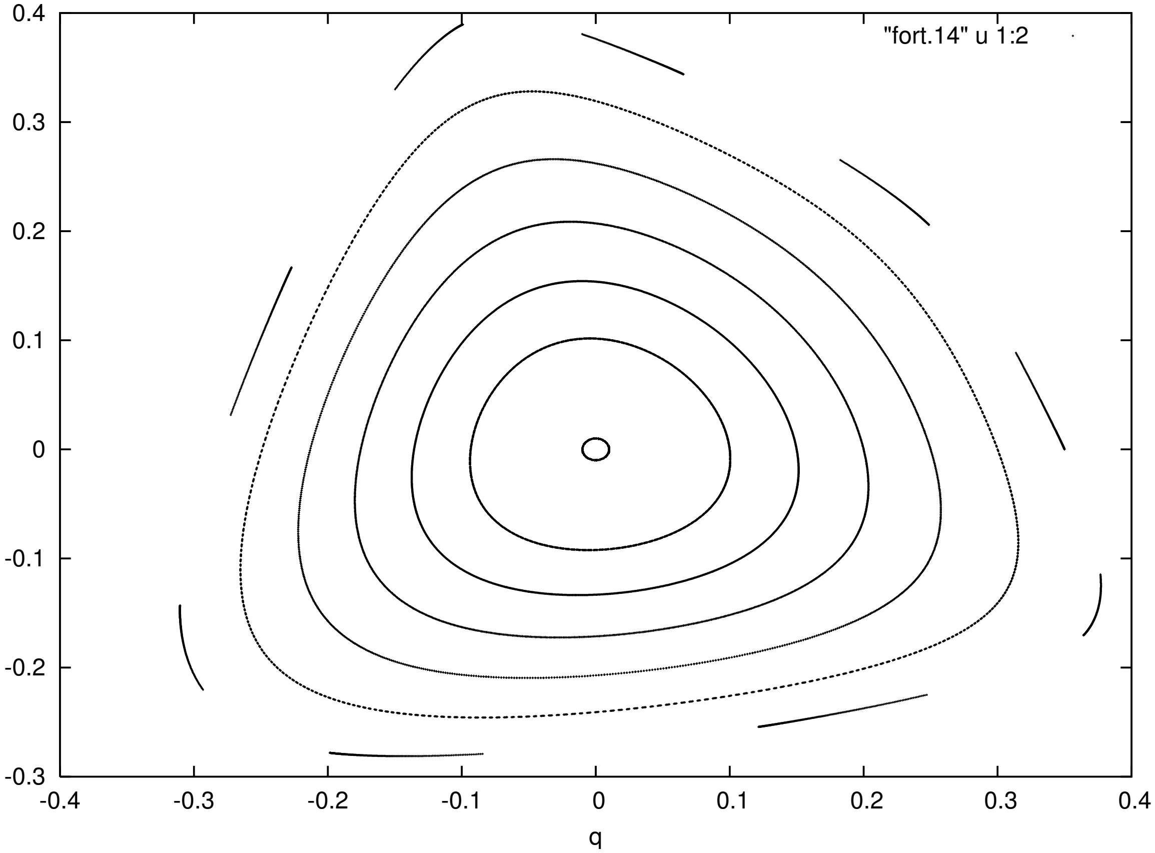


Figure 34.3.3: Phase-space portrait, in the case $\theta/2\pi = 0.22$, resulting from applying the map $\mathcal{M}^{\mathrm{psc}}$ repeatedly (2000 times) to the to the seven initial conditions $(q, p) = (.01, 0)$, $(.1, 0)$, $(.15, 0)$, $(.2, 0)$, $(.25, 0)$, $(.3, 0)$, and $(.35, 0)$ to find their orbits.

Root Trick for F_+

Evidently the root trick can be applied to any symplectification procedure. Here we will explore its use for our example of Poincaré symplectification. Let $\mathcal{N}^{\mathrm{psc}}(1/2)$ denote the map given by (3.113) and (3.114) with $c = -1/2$. We may view $\mathcal{N}^{\mathrm{psc}}(1/2)$ as the F_+

symplectification of the degree-two jet map $\mathcal{N}^{\rm tr}(1/2)$. We see, from (3.115) with $c = -1/2$, that the map $\mathcal{N}^{\rm psc}(1/2)$ has a branch point on the surface $p = 1$, the same surface on which the origin map has a pole. We will now study the behavior of the *improved Poincaré symplectically completed* map $\mathcal{M}^{\rm ipsc}$ defined by the relation

$$\mathcal{M}^{\rm ipsc} = \mathcal{R}\mathcal{N}^{\rm psc}(1/2)\mathcal{N}^{\rm psc}(1/2). \tag{34.3.137}$$

Figure 3.4 shows the result of applying $\mathcal{M}^{\rm ipsc}$ repeatedly to seven initial conditions for the case $\theta/2\pi = 0.22$. We see that now the orbits approximate those of Figure 1.1 remarkably well. Presumably one reason for this improvement is the larger domain of analyticity for $\mathcal{N}^{\rm psc}(1/2)$.

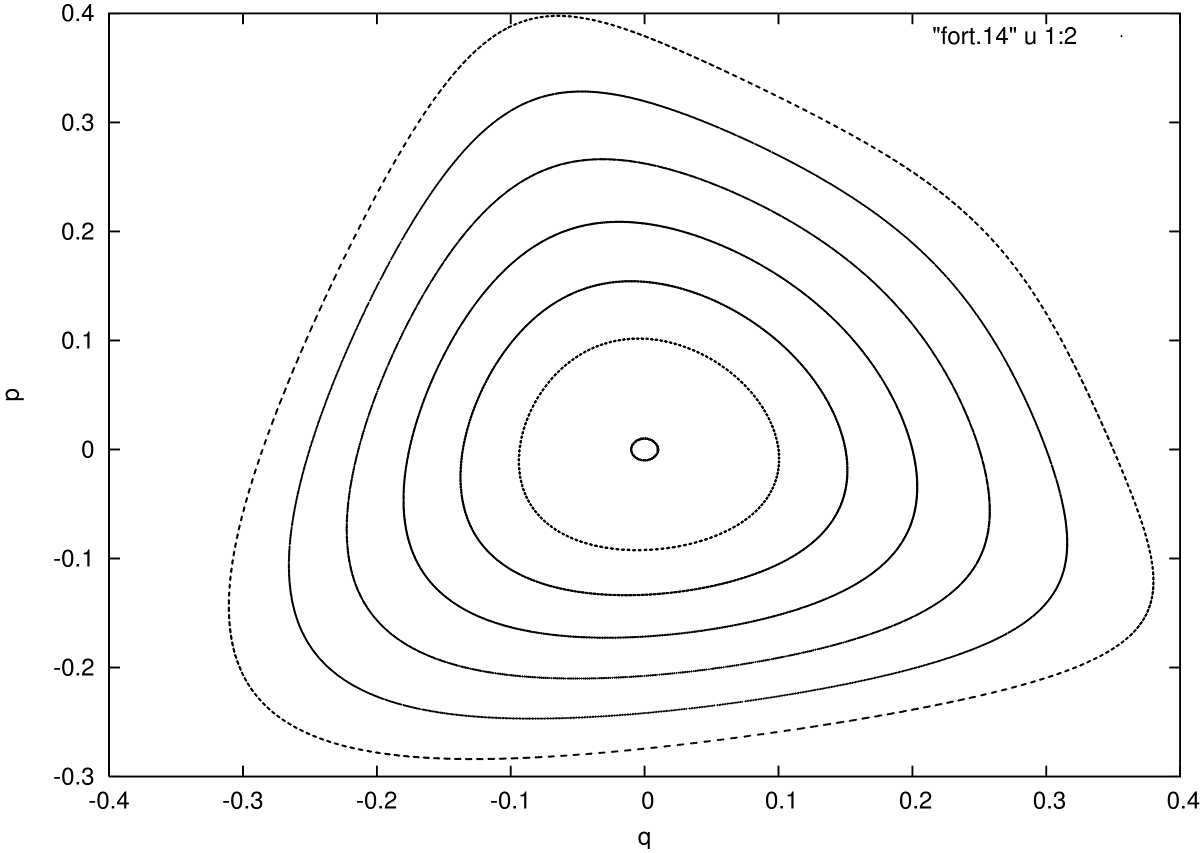


Figure 34.3.4: Phase-space portrait, in the case $\theta/2\pi = 0.22$, resulting from applying the map $\mathcal{M}^{\rm ipsc}$ repeatedly (2000 times) to the seven initial conditions $(q, p) = (.01, 0)$, $(.1, 0)$, $(.15, 0)$, $(.2, 0)$, $(.25, 0)$, $(.3, 0)$, and $(.35, 0)$ to find their orbits.

34.3.6 Comments and Comparisons

At this point some comments and comparisons are in order. Based on our experience so far, we may draw the following (sometimes tentative) conclusions:

1. The symplectification of symplectic jets overcomes the problem of spurious spiraling into or out of the origin.

2. The use of either F_2 or F_+ (Poincaré) generating functions gives exact results for jets that are kick maps. Kick maps are, of course, symplectic, and their symplectification by use of either F_2 or F_+ (Poincaré) generating functions leaves them unchanged.

3. The use of F_+ (Poincaré) generating functions gives exact results for jets that are jolt maps. Such jet maps are also also exactly symplectic, and their Poincaré symplectification also leaves them unchanged. Such is not the case for F_2 symplectification. Given a jolt map, it generally converts this map into some other map. While the result of this conversion is a symplectic map, it is not generally the original jolt map. Put colloquially, Poincaré symplectification has the good sense to leave a good thing alone, but F_2 symplectification generally does not. In fact, in this case F_2 symplectification replaces a map with no singularities by a map with singularities.

4. The root trick enlarges the applicable domain and improves accuracy. It might also appear to require more work because now a symplectified nonlinear map has to be evaluated twice. However, the iterative method employed to solve numerically the implicit equations associated with the use of $\mathcal{N}^{\rm psc}(1/2)$ is expected to converge faster than that for $\mathcal{N}^{\rm psc}(1)$ because $\mathcal{N}^{\rm psc}(1/2)$ is less nonlinear; and this gain in convergence speed is likely to be greater than the loss associated with evaluating $\mathcal{N}^{\rm psc}(1/2)$ twice.

5. Compared to a F_2 symplectified map, a Poincaré symplectified map has a larger domain of applicability.

6. Compared to a F_2 symplectified map, a Poincaré symplectified map has higher-order accuracy.

Let us explore item 5 above in some more detail. In Section 26.2 we listed the normal forms for cubic polynomials in two variables, namely those given by (26.2.4) through (26.2.7). It is instructive to compare F_2 and Poincaré symplectification for each. We have already considered the cases (26.2.6) and (26.2.7). We now consider the two remaining cases.

Case When $a = 1$, $b = -3$, $c = d = 0$

We begin with the case (26.2.5), which is the easier of the two. In this case, by suitable rescaling, there is no loss of generality in taking a and b as the the only nonzero coefficients and giving them the values $a = 1$ and $b = -3$.

For these values use of (3.81) and (3.82) gives the map

$$Q = q - 3q^2, \tag{34.3.138}$$

$$P = (p - 3q^2)/(1 - 6q). \tag{34.3.139}$$

We see that for the F_2 case the map is singular on the surface

$$? = . \tag{34.3.140}$$

For these same values of a and b, the implicit relations (3.106) and (3.107), produced by the use of F_+, take the form

$$Q = q - (3/4)(Q + q)^2, \tag{34.3.141}$$

$$P = p - (1/4)[3(Q+q)^2 - 6(Q+q)(P+p)]. \tag{34.3.142}$$

These implicit relations have the solution

$$Q = (2/3)\{-[1+(3/2)q] + [1+6q]^{1/2}\}, \tag{34.3.143}$$

$$P = . \tag{34.3.144}$$

We see that for the F_+ case the map is singular on the surface

$$? = . \tag{34.3.145}$$

Case When $a = d = 1$, $b = c = 0$

For the remaining case (26.2.4) there is no loss of generality in taking a and d as the only nonzero coefficients and giving them the values $a = d = 1$.

For these values use of (3.81) and (3.82) gives the map

$$Q =, \tag{34.3.146}$$

$$P = . \tag{34.3.147}$$

We see that for the F_2 case the map is singular on the surface

$$? = . \tag{34.3.148}$$

For these same values of a and d, the implicit relations (3.106) and (3.107), produced by the use of F_+, take the form

$$Q = q + (3/4)(P+p)^2, \tag{34.3.149}$$

$$P = p - (3/4)(Q+q)^2. \tag{34.3.150}$$

These implicit relations have the solution

$$Q =, \tag{34.3.151}$$

$$P = . \tag{34.3.152}$$

We see that for the F_+ case the map is singular on the surface

$$? = . \tag{34.3.153}$$

Let us also explore item 6 above in some more detail. To do so it is convenient to make some definitions. First, suppose $\mathcal{N}$ is some nonlinear map. We will define $\mathrm{n}(\mathcal{N}, z)$, the local *nonlinearity* of $\mathcal{N}$, by the rule

$$\mathrm{n}(\mathcal{N}, z) = ||(\mathcal{N} - \mathcal{I})z||/||z||. \tag{34.3.154}$$

Here $||\ ||$ denotes the vector norm of a phase-space vector in the usual Euclidean metric. The quantity $n(\mathcal{N}, z)$ measures how much a phase-space point z moves under the action of $\mathcal{N}$ normalized by its distance from the origin.

It may be the case that $\mathcal{N}$ is a symplectic jet. In that case, it is useful to have some measure of the violation of the symplectic condition associated with the action of $\mathcal{N}$. One possibility is to define $\mathrm{sv}(\mathcal{N}, z)$, the local *symplectic violation*, by the rule

$$\mathrm{sv}(\mathcal{N}, z) = ||([\mathcal{N}z_a, \mathcal{N}z_b] - J_{ab})||. \tag{34.3.155}$$

Here $|| \; ||$ denotes some matrix norm, say the maximum column sum norm.

Given two nonlinear maps $\mathcal{N}_1$ and $\mathcal{N}_2$, we will also want to have some measure of the *difference* between them. One way to do so is to introduce the quantity $\mathrm{d}(\mathcal{N}_1, \mathcal{N}_2, z)$ by the rule

$$\mathrm{d}(\mathcal{N}_1, \mathcal{N}_2, z) = ||\mathcal{N}_1 z - \mathcal{N}_2 z||. \tag{34.3.156}$$

34.4 Use of Poincaré Generating Function

34.4.1 Determination of Poincaré Generating Function in Terms of H

Suppose we are given a time-independent Hamiltonian H. Use it to generate the symplectic map

$$\mathcal{M}(\tau) = \exp(-\tau : H :) \tag{34.4.1}$$

Let $F_+(\Sigma, \tau)$ be the Poincaré generating function associated with $\mathcal{M}(\tau)$. We want to find a formula for F_+ in terms of H. To do so we will seek a Taylor expansion of $F_+(\Sigma, \tau)$ in powers of τ.

We first note that F_+ is odd in τ. From (2.1) we have the relation

$$\mathcal{M}(-\tau) = \exp(+\tau : H :) = \mathcal{M}^{-1}(\tau). \tag{34.4.2}$$

Next we observe that (6.6.45), which can be written in the form

$$Z = z + J\partial_\Sigma F_+|_{\Sigma=(Z+z)/2}, \tag{34.4.3}$$

can be rewritten in the form

$$z = Z - J\partial_\Sigma F_+|_{\Sigma=(Z+z)/2}, \tag{34.4.4}$$

which reveals that if the Poincaré generating function associated with $\mathcal{M}(\tau)$ is $F_+(\Sigma, \tau)$, then the Poincaré generating function associated with $\mathcal{M}^{-1}(\tau)$ is $-F_+(\Sigma, \tau)$. Consequently, we conclude that

$$F_+(\Sigma, -\tau) = -F_+(\Sigma, \tau). \tag{34.4.5}$$

Since F_+ is odd in τ, only odd powers of τ can occur in its Taylor expansion so that we may write

$$F_+(\Sigma, \tau) = F_+^{(1)}(\Sigma)\tau + F_+^{(3)}(\Sigma)\tau^3 + F_+^{(5)}(\Sigma)\tau^5 + \cdots . \tag{34.4.6}$$

The first term in the expansion is H itself,

$$F_+^{(1)}(\Sigma) = H(\Sigma). \tag{34.4.7}$$

Let $\mathcal{S}(H)$ denote the Hessian of H,

$$\mathcal{S}(H)_{ab} = \partial_a \partial_b H. \tag{34.4.8}$$

Then the next term in the expansion is given by the equation

$$F_+^{(3)}(\Sigma) = (1/24)(\partial H, J\mathcal{S}(H)J\partial H). \tag{34.4.9}$$

The successive terms are ever more complicated to state in explicit form. For $F_+^{(5)}$ there is the intermediate result

$$F_+^{(5)}(\Sigma) = \tag{34.4.10}$$

and the final result

$$F_+^{(5)}(\Sigma) = . \tag{34.4.11}$$

34.4.2 Application to Quadratic Hamiltonian

As a preliminary application of these results, let us first consider the simple case where

$$H(z) = h_2(z) = (1/2)(z, Sz). \tag{34.4.12}$$

Then we have the relations

$$\partial_a H = S_{ab} z_b, \tag{34.4.13}$$

$$\mathcal{S}(H) = S. \tag{34.4.14}$$

Correspondingly, we find the results

$$F_+^{(1)}(\Sigma) = H(\Sigma) = (1/2)(\Sigma, S\Sigma), \tag{34.4.15}$$

$$F_+^{(3)}(\Sigma) = (1/24)(\partial_a H J_{ab} \mathcal{S}(H)_{bc} J_{cd} \partial_d H) = (1/24)(\Sigma, SJSJS\Sigma), \tag{34.4.16}$$

$$F_+^{(5)}(\Sigma) = . \tag{34.4.17}$$

The net result is that $F_+(\Sigma, \tau)$ has the expansion

$$F_+(\Sigma,\tau) = (\Sigma, [(1/2)\tau S + (1/24)\tau^3 S(JS)^2 + ()\tau^5 S(JS)^4 + \cdots]\Sigma) \tag{34.4.18}$$

However, thanks to (), we already know that in this case

$$\begin{aligned} F_+(\Sigma,\tau) &= (1/2)(\Sigma, W'\Sigma) = (\Sigma, -J\tanh[\tau JS/2]\Sigma) \\ &= (\Sigma, [-J\tau JS/2 + J(1/3)(\tau JS/2)^3 + J(2/15)(\tau JS/2)^5/3 + \cdots]\Sigma) \\ &= (\Sigma, [(1/2)\tau S + (1/24)\tau^3 S(JS)^2 + (1/240)\tau^5 S(JS)^4 + \cdots]\Sigma). \end{aligned} \tag{34.4.19}$$

Here we have used the series

$$\tanh x = x - (1/3)x^3 + (2/15)x^5 + \cdots . \tag{34.4.20}$$

Evidently the expansions () and () agree.

34.4.3 Application to Symplectic Approximation

As a second application, suppose H has a homogeneous polynomial expansion of the form

$$H = h_3 + h_4 + h_5 + \cdots . \tag{34.4.21}$$

In this case we wish to obtain a homogeneous polynomial expansion for $F_+(\Sigma, \tau)$ of the form

$$F_+(\Sigma, \tau) = F_+^3(\Sigma, \tau) + F_+^4(\Sigma, \tau) + F_+^5(\Sigma, \tau) + \cdots . \tag{34.4.22}$$

We will now find that each $F_+^m(\Sigma, \tau)$ is also polynomial in the variable τ. Therefore, we may set $\tau = 1$, and drop τ from our variable list. Then we will have the relation

$$\mathcal{M} = \exp(- : H :). \tag{34.4.23}$$

For this symplectic map there will be the Poincaré generating function

$$F_+(\Sigma) = F_+^3(\Sigma) + F_+^4(\Sigma) + F_+^5(\Sigma) + \cdots \tag{34.4.24}$$

with

$$F_+^m(\Sigma) = F_+^m(\Sigma, \tau = 1). \tag{34.4.25}$$

Upon equating like powers of Σ on both sides of () and (), we find, through terms of degree 8, the results

$$F_+^3 = h_3, \tag{34.4.26}$$

$$F_+^4 = h_4, \tag{34.4.27}$$

$$F_+^5 = h_5 + (1/24)(\partial h_3, J\mathcal{S}(h_3)J\partial h_3), \tag{34.4.28}$$

$$F_+^6 = h_6 + (1/24)(\partial h_3, J\mathcal{S}(h_4)J\partial h_3) + (1/12)(\partial h_3, J\mathcal{S}(h_3)J\partial h_4), \tag{34.4.29}$$

$$F_+^7 = h_7 + (1/24)(\partial h_4, J\mathcal{S}(h_3)J\partial h_4) + (1/12)(\partial h_3, J\mathcal{S}(h_4)J\partial h_4), \tag{34.4.30}$$

$$F_+^8 = h_8 + (1/24)(\partial h_4, J\mathcal{S}(h_4)J\partial h_4). \tag{34.4.31}$$

Suppose we know h_3 through h_n and wish to 'evaluate'

$$\mathcal{M}^{[n]} = \exp(- : H^{[n]} :) \tag{34.4.32}$$

where

$$H^{[n]} = h_3 + h_4 + \cdots + h_n. \tag{34.4.33}$$

For this purpose, let us use a corresponding Poincaré generating function also truncated beyond terms of degree n. That is, we use the function $F_+^{[n]}$ defined by the rule

$$F_+^{[n]} = F_+^3 + F_+^4 + \cdots + F_+^n \tag{34.4.34}$$

Let $\mathcal{M}_+^{[n]}$ be the symplectic map produced by the use of $F_+^{[n]}$. By construction it will have the single-exponent Lie representation

$$\mathcal{M}_+^{[n]} = \exp(- : h_3 : - : h_4 : - \cdots - : h_n : + : g_{n+1} : + : g_{n+2} : + \cdots). \tag{34.4.35}$$

That is, the exponent of $\mathcal{M}_+^{[n]}$ will agree with that of $\mathcal{M}^{[n]}$ through terms of degree n, but there will generally be additional terms $g_{n+1},\ g_{n+2},\ \cdots$ which reflect the fact that the maps $\mathcal{M}_+^{[n]}$ and $\mathcal{M}$ are generally not identical. We will see that these additional terms depend on the given/known h_m, and that this dependence has three desirable properties.

Suppose, for example, that $n = 4$ so that

$$\mathcal{M}^{[4]} = \exp(- : h_3 : - : h_4 :), \tag{34.4.36}$$

$$F_+^{[4]} = F_+^3 + F_+^4, \tag{34.4.37}$$

and

$$\mathcal{M}_+^{[4]} = \exp(- : h_3 : - : h_4 : + : g_5 : + : g_6 : + \cdots). \tag{34.4.38}$$

Evidently truncating the series (2.34) beyond terms of degree 4 is equivalent to including all terms in the series and requiring that

$$F_+^m = 0 \text{ for all } m > 4. \tag{34.4.39}$$

Inspection of (2.28) through (2.31) shows that the requirement (2.39) produces, through terms of degree 8, the relations

$$-h_5 = (1/24)(\partial h_3, J\mathcal{S}(h_3)J\partial h_3), \tag{34.4.40}$$

$$-h_6 = (1/24)(\partial h_3, J\mathcal{S}(h_4)J\partial h_3) + (1/12)(\partial h_3, J\mathcal{S}(h_3)J\partial h_4), \tag{34.4.41}$$

$$-h_7 = (1/24)(\partial h_4, J\mathcal{S}(h_3)J\partial h_4) + (1/12)(\partial h_3, J\mathcal{S}(h_4)J\partial h_4), \tag{34.4.42}$$

$$-h_8 = (1/24)(\partial h_4, J\mathcal{S}(h_4)J\partial h_4), \tag{34.4.43}$$

and we see that

$$g_m = -h_m \text{ for all } m > 4 \tag{34.4.44}$$

with the h_m for $m > 4$ defined by in terms of h_3 and h_4 by the relations (2.40) through (2.43).

Now we are ready to examine in some detail the properties of the dependence of the $g_{n+1},\ g_{n+2},\ \cdots$ on the $h_3,\ h_4,\ \cdots\ h_n$. To do so, it is useful to introduce a somewhat more elaborate notation. Let us employ, in place of $\mathcal{M}_+^{[n]}$, the symbols $\mathcal{M}_+^{[n]}\{H^{[n]}\}$ to indicate that the map $\mathcal{M}_+^{[n]}$ depends on $h_3,\ h_4,\ \cdots\ h_n$. The first property is this: Suppose all the h_m are replaced by $-h_m$. Then we see, from () through () and () through (), that all the F_+^m and all the g_m are replaced by $-F_+^m$ and $-g_m$, respectively. Consequently, there is the relation

$$\mathcal{M}_+^{[n]}\{-H^{[n]}\} = (\mathcal{M}_+^{[n]}\{H^{[n]}\})^{-1}. \tag{34.4.45}$$

In words, if $\mathcal{M}_+^{[n]}$ is the symplectic approximation to $\mathcal{M}^{[n]}$, then $(\mathcal{M}_+^{[n]})^{-1}$ is the symplectic approximation to $(\mathcal{M}^{[n]})^{-1}$. We may invert and then symplectically approximate, or symplectically approximate and then invert. The result of both procedures is the same. We may say that symplectic approximation by the use of a Poincaré generating function is invariant under the operation of map inversion.

The second property is more subtle. Suppose $\mathcal{R}$ is a linear symplectic map with associated symplectic matrix R. Suppose the h_m are transformed under the action of $\mathcal{R}$ to become the homogeneous polynomials $h_m^{\rm tr}$ by the rule

$$h_m^{\rm tr}(z) = \mathcal{R}h_m(z) = h_m(Rz). \tag{34.4.46}$$

Also, let $g_{n+1}^{\rm tr}(z)$, $g_{n+2}^{\rm tr}(z)$, $\cdots$ be the functions obtained by applying the rules defining the $g_{n+1}(z)$, $g_{n+2}(z)$, $\cdots$ to the $h_m^{\rm tr}$. Then there is also the result

$$\begin{aligned} g_{n+1}^{\rm tr}(z) &= \mathcal{R}g_{n+1}, \\ g_{n+2}^{\rm tr}(z) &= \mathcal{R}g_{n+2}, \text{ etc.} \end{aligned} \tag{34.4.47}$$

Suppose we assume, for the moment, that (2.47) is correct. It follows that there is then the relation

$$\mathcal{M}_+^{[n]}\{\mathcal{R}H^{[n]}\} = \mathcal{R}\mathcal{M}_+^{[n]}\{H^{[n]}\}(\mathcal{R})^{-1}. \tag{34.4.48}$$

Of course, we also have the relation

$$\mathcal{R}\mathcal{M}^{[n]}(\mathcal{R})^{-1} = \exp(- : \mathcal{R}H^{[n]} :). \tag{34.4.49}$$

In words, if we conjugate the map $\mathcal{M}^{[n]}$ with $\mathcal{R}$ and then symplectically approximate the result, the outcome is the same as first symplectically approximating $\mathcal{M}^{[n]}$ and then conjugating with $\mathcal{R}$. We may say that symplectic approximation by the use of a Poincaré generating function is invariant under the operation of conjugation with the linear symplectic map $\mathcal{R}$.

34.5 Use of Other Generating Functions

34.6 Cremona Approximation

Decomposition of the $sp(6, \mathbb{R})$ representation $\Gamma(\ell, 0, 0)$ into representations of $su(3)$.

Bibliography

Symplectic Completion of Symplectic Jets and Cremona Maps

[1] A. Dragt and D. Abell, "Symplectic Maps and Computation of Orbits in Particle Accelerators", in *Integration Algorithms and Classical Mechanics*, J. Marsden, G. Patrick, and W. Shadwick, Edit., American Mathematical Society (1996).

[2] D. Abell, *Analytic Properties and Cremona Approximation of Transfer Maps for Hamiltonian Systems*, University of Maryland Physics Department Ph.D. Thesis (1995).

[3] J.E. Fornaess and N. Sibony, "Complex Dynamics in Higher Dimension", in *Several Complex Variables*, M. Schneider and Y.-T. Siu, Edit., Cambridge University Press (1999).

[4] D. Abell, E. McIntosh, and F. Schmidt "Fast Symplectic Map Tracking for the CERN Large Hadron Collider", *Physical Review Special Topics, Accelerators and Beams* **6**, 064001 (2003).

Solution for Monomial Hamiltonian

[5] F.J. Testa, *J. Math Phys.* **14**, p. 1097 (1973).

[6] P.J. Channell, *Explicit Integration of Kick Hamiltonians in Three Degrees of Freedom*, Accelerator Theory Note AT-6:ATN-86-6, Los Alamos National Laboratory (1986).

[7] I. Gjaja, *Particle Accelerators*, vol. 43 (3), pp. 133-144 (1994).

[8] L. Michelotti, *Comment on the exact evaluation of symplectic maps*, Fermilab preprint (1992).

Generating Functions

[9] C.R. Menyuk, "Some Properties of the Discrete Hamiltonian Method", *Physica D* **11**, p. 109 (1984).

[10] A.J. Dragt, F. Neri, G. Rangarajan, D.R. Douglas, L.M. Healy, and R.D. Ryne, "Lie Algebraic Treatment of Linear and Nonlinear Beam Dynamics", *Ann. Rev. Nucl. Part. Sci.* **38**, pp. 455-496 (1988).

Chapter 35

Orbit Stability, Long-Term Behavior, and Dynamic Aperture

Chapter 36

Reversal Symmetry

The concept of reversibility, and that reversibility has various implications for charged-particle and light optics (and the study of general dynamical systems), are part of the common lore of those working in these fields, and its role and value are generally understood at least on an intuitive level. The purpose of this chapter is to explore reversal symmetry systematically. Reversal symmetry is defined; it is shown that the transfer maps for most common beam-line elements are reversal symmetric including nonlinear effects; and various linear and nonlinear consequences of reversal symmetry are worked out in some detail.

Section 1 defines the operation of reversal and works out some of its properties. Section 2 describes some of the applications of these properties. In particular it defines what is meant for a transfer map to be reversal symmetric, and shows that the transfer maps for many common beam-line elements are reversal symmetric. Section 3 works out some of the general consequences of reversal symmetry for straight and circular machines, and Section 4 treats some special cases. Section 5 studies the consequences of reversal symmetry for closed orbits in a circular machine, and Section 6 studies the consequences for the *Courant-Snyder* functions in a circular machine. A final section treats various nonlinear consequences of reversal symmetry. It seems remarkable that such a simple concept should be so rich in consequences.

36.1 Reversal Operator

We will work with a coordinate system that is particularly useful for charged-particle optics. We write

$$z = (x, p_x; y, p_y; \tau, p_\tau). \tag{36.1.1}$$

The quantities x and y are transverse deviations from a design trajectory, and p_x and p_y are their conjugate momenta. The quantity τ is the difference (time deviation) between the arrival/departure time of a given particle and a particle on the design trajectory. Finally, p_τ is the *negative* of the energy difference between that of the given particle and that of a particle on the design trajectory. Note that this choice of variables presumes that some *coordinate* (it could be Cartesian or angular or path length along some design trajectory) is taken to play the role of the *independent* (time-like) variable. When this is done, its conjugate momentum does not appear in the associated Hamiltonian, and τ and p_τ are both *dependent* variables. Finally, we note that in accelerator physics it is common to scale the transverse coordinates x and y by some convenient scale length ℓ, to scale the transverse momenta p_x and p_y by some "design" momentum p^0, to scale p_t (which is the negative of the energy) by p^0c (where c is the speed of light), and to scale the time t by ℓ/c. In this chapter we do not do so. In particular, for the purpose of this chapter, p_τ is defined in terms of p_t simply by subtracting off the design value of p_t without any scaling, and τ is defined in terms of t simply by subtracting off the time of flight for the design orbit, again without any scaling factor.

Let z be any point in phase space as specified by (1.1). *Define* a "reversal" operator $\mathcal{R}$ acting on phase space by the rule

$$\mathcal{R}z = z^r \tag{36.1.2}$$

with

$$z^r = (x, -p_x; y, -p_y; -\tau, p_\tau). \tag{36.1.3}$$

The reversal operation is analogous to time reversal, but differs from it in two essential ways. First, recall that the definitions (1.1) through (1.3) presume that some coordinate is playing the role of the independent (time-like) variable, its conjugate momentum is absent, and τ and p_τ are dependent variables. Second, the magnetic field does not change sign. It is for these reasons that, as we will see, a transfer map can violate what we will define as reversal symmetry even though the fundamental laws that govern charged-particle motion, the electromagnetic field, and the electromagnetic interaction are all invariant under time reversal as usually defined.

It is easily verified that, when acting on phase space, the effect of $\mathcal{R}$ can be described by a matrix R with

$$\mathcal{R}z_a = (Rz)_a \tag{36.1.4}$$

where R is the matrix

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{36.1.5}$$

We note for future reference that J and R have the properties

$$JR = \begin{pmatrix} 0 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \tag{36.1.6}$$

$$J^2 = -I, \tag{36.1.7}$$

$$R^2 = I, \tag{36.1.8}$$

$$(JR)^2 = I, \tag{36.1.9}$$

$$RJR = -J, \tag{36.1.10}$$

$$RJ = -JR. \tag{36.1.11}$$

Here we use the form (3.2.10) for J. We also remark that since R is a symmetric matrix, $R^T = R$, (1.10) can also be written in the form

$$R^T JR = -J. \tag{36.1.12}$$

Therefore, R is an *antisymplectic* matrix. See Exercise 3.12.8. Correspondingly, we will say that $\mathcal{R}$ is an *antisymplectic* map. We have learned that, in our setting of Classical Mechanics, evolution (with some coordinate playing the role of the independent variable) is described by a symplectic map acting on phase space, and reversal is described by an antisymplectic map. This terminology is analogous to that employed in Quantum Mechanics where time evolution is described by a unitary transformation acting on Hilbert space, and time reversal is described by what is called an *antiunitary* transformation.

We have seen that $\mathcal{R}$ is an antisymplectic map. The same is true of the maps $\mathcal{M}\mathcal{R}$ and $\mathcal{R}\mathcal{M}$ if $\mathcal{M}$ is symplectic. To prove this, let $\mathcal{N}$ be the map given by the product

$$\mathcal{N} = \mathcal{M}\mathcal{R}. \tag{36.1.13}$$

By the chain rule its Jacobian matrix N is given by the relation

$$N = MR, \tag{36.1.14}$$

and we find the result

$$N^T JN = RM^T JMR = RJR = -J. \tag{36.1.15}$$

Similarly, if $\mathcal{N}$ is the map given by the product

$$\mathcal{N} = \mathcal{R}\mathcal{M}, \tag{36.1.16}$$

we find the results

$$N = RM \tag{36.1.17}$$

and

$$N^T JN = M^T RJRM = M^T(-J)M = -J. \tag{36.1.18}$$

We also note the converse conclusion: If $\mathcal{N}$ is an antisymplectic map, then the maps $\mathcal{R}\mathcal{N}$ and $\mathcal{N}\mathcal{R}$ are symplectic. Finally, there is an immediate generalization: The product of a symplectic map and an antisymplectic map is antisymplectic, and the product of two antisymplectic maps is symplectic.

Extend $\mathcal{R}$ to phase-space functions $f(z)$ by the rule

$$\mathcal{R}f = f^r \tag{36.1.19}$$

with

$$f^r(z) = f(z^r). \tag{36.1.20}$$

Evidently $\mathcal{R}$ is a linear operator with the property

$$\mathcal{R}^2 = \mathcal{I} \text{ or } \mathcal{R}^{-1} = \mathcal{R}. \tag{36.1.21}$$

For any two functions g and h there is also the property

$$\mathcal{R}(gh) = (\mathcal{R}g)(\mathcal{R}h). \tag{36.1.22}$$

Finally we note that since the application of $\mathcal{R}$ commutes with the operation of scaling variables, everything we will conclude about reversal properties of maps given in terms of unscaled variables will also hold for maps given in terms of scaled variables.

Let us determine the effect of reversal on Lie operators. We claim that $\mathcal{R}$ has the property

$$\mathcal{R} : f : \mathcal{R} = - : \mathcal{R}f := - : f^r : \tag{36.1.23}$$

for any Lie operator $: f :$. To prove this claim, let $\mathcal{R} : f : \mathcal{R}$ act on any function g. The calculation is a bit delicate, and is best done in stages and pieces. To begin, we have the result

$$\begin{aligned}
\mathcal{R} : f : \mathcal{R}g &= \mathcal{R}[f, \mathcal{R}g] \\
&= \mathcal{R}\{\sum_j (\partial f/\partial q_j)(\partial(\mathcal{R}g)/\partial p_j) - (\partial f/\partial p_j)(\partial(\mathcal{R}g)/\partial q_j)\} \\
&= \sum_j \{\mathcal{R}(\partial f/\partial q_j)\}\{\mathcal{R}(\partial(\mathcal{R}g)/\partial p_j)\} \\
&- \{\mathcal{R}(\partial f/\partial p_j)\}\{\mathcal{R}(\partial(\mathcal{R}g)/\partial q_j)\}.
\end{aligned} \tag{36.1.24}$$

Here we have used (1.22). Next, it follows from (1.2) and (1.3) that there are the operator relations

$$\mathcal{R}(\partial/\partial z_a) = (\partial/\partial z_a)\mathcal{R} \text{ for } a = 1, 3, 6; \tag{36.1.25}$$

$$\mathcal{R}(\partial/\partial z_a) = -(\partial/\partial z_a)\mathcal{R} \text{ for } a = 2, 4, 5. \tag{36.1.26}$$

Therefore we find for the $j = 1$ terms in (1.24) the results

$$\begin{aligned}
\{\mathcal{R}(\partial f/\partial x)\}\{\mathcal{R}(\partial(\mathcal{R}g)/\partial p_x)\} &= \{\partial(\mathcal{R}f)/\partial x)\}(-1)\{\mathcal{R}^2(\partial g/\partial p_x)\} \\
&= -(\partial f^r/\partial x)(\partial g/\partial p_x),
\end{aligned} \tag{36.1.27}$$

$$\begin{aligned}
-\{\mathcal{R}(\partial f/\partial p_x)\}\{\mathcal{R}(\partial(\mathcal{R}g)/\partial x)\} &= +\{\partial(\mathcal{R}f)/\partial p_x\}\{\mathcal{R}^2(\partial g/\partial x)\} \\
&= (\partial f^r/\partial p_x)(\partial g/\partial x).
\end{aligned} \tag{36.1.28}$$

Analogous results hold for the $j = 2$ terms, which involve the y, p_y pair. And for the $j = 3$ terms, which involve the τ, p_τ pair, we find the results

$$\begin{aligned}
\{\mathcal{R}(\partial f/\partial \tau)\}\{\mathcal{R}(\partial(\mathcal{R}g)/\partial p_\tau)\} &= -\{\partial(\mathcal{R}f)/\partial \tau\}\{\mathcal{R}^2(\partial g/\partial p_\tau)\} \\
&= -(\partial f^r/\partial \tau)(\partial g/\partial p_\tau),
\end{aligned} \tag{36.1.29}$$

$$\begin{aligned}
-\{\mathcal{R}(\partial f/\partial p_\tau)\}\{\mathcal{R}(\partial(\mathcal{R}g)/\partial \tau)\} &= -\{\partial(\mathcal{R}f/\partial p_\tau)\}(-1)\{\mathcal{R}^2(\partial g/\partial \tau)\} \\
&= (\partial f^r/\partial p_\tau)(\partial g/\partial \tau).
\end{aligned} \tag{36.1.30}$$

Now put all these results into (1.24) to obtain the relation

$$\begin{aligned}\mathcal{R} : f : \mathcal{R}g &= \sum_j \{\mathcal{R}(\partial f/\partial q_j)\}\{\mathcal{R}(\partial(\mathcal{R}g)/\partial p_j)\} \\ &- \{\mathcal{R}(\partial f/\partial p_j)\}\{\mathcal{R}(\partial(\mathcal{R}g)/\partial q_j)\} \\ &= -\sum_j (\partial f^r/\partial q_j)(\partial g/\partial p_j) - (\partial f^r/\partial p_j)(\partial g/\partial q_j) \\ &= -[f^r, g] = - : f^r : g. \end{aligned} \tag{36.1.31}$$

Evidently (1.23) is the operator version of (1.31).

Let us next determine the effect of reversal on Lie transformations. From (1.21) and (1.23) we find the additional property

$$\begin{aligned}\mathcal{R} : f :^n \mathcal{R} &= \mathcal{R} : f :: f :: f : \cdots : f : \mathcal{R} \\ &= \mathcal{R} : f : \mathcal{R}\mathcal{R} : f : \mathcal{R}\mathcal{R} : f : \mathcal{R} \cdots \mathcal{R} : f : \mathcal{R} \\ &= (\mathcal{R} : f : \mathcal{R})^n = (-1)^n : \mathcal{R}f :^n = (-1)^n : f^r :^n . \end{aligned} \tag{36.1.32}$$

Suppose $\mathcal{M}$ is a map that, for some f, can be written in the single exponent form

$$\mathcal{M} = \exp(: f :) = \sum_{n=0}^{\infty} : f :^n /n!. \tag{36.1.33}$$

Then, from (1.32) and (1.33), we find the result

$$\begin{aligned}\mathcal{R}\mathcal{M}\mathcal{R} &= \sum_{n=0}^{\infty} \mathcal{R} : f :^n \mathcal{R}/n! = \sum_{n=0}^{\infty} (-1)^n : \mathcal{R}f :^n /n! \\ &= \exp(- : \mathcal{R}f :) = \exp(- : f^r :). \end{aligned} \tag{36.1.34}$$

The stage is set to define the effect of reversal on maps. Suppose $\mathcal{M}$ is any map that sends initial points z^i to final points z^f,

$$z^f = \mathcal{M}z^i. \tag{36.1.35}$$

Reverse both z^i and z^f to yield $\mathcal{R}z^i$ and $\mathcal{R}z^f$. We *define* the reversed map $\mathcal{M}^r$ to be that map which sends $\mathcal{R}z^f$ to $\mathcal{R}z^i$,

$$\mathcal{M}^r\mathcal{R}z^f = \mathcal{R}z^i. \tag{36.1.36}$$

See Figure 1.1. Combining (1.35) and (1.36) gives the result

$$\mathcal{M}^r\mathcal{R}\mathcal{M}z^i = \mathcal{R}z^i. \tag{36.1.37}$$

Equivalently, we have the operator relation

$$\mathcal{M}^r\mathcal{R}\mathcal{M} = \mathcal{R}. \tag{36.1.38}$$

This relation can be solved for $\mathcal{M}^r$ to give the intermediate result

$$\mathcal{M}^r = \mathcal{R}(\mathcal{R}\mathcal{M})^{-1} = \mathcal{R}\mathcal{M}^{-1}\mathcal{R}^{-1}, \tag{36.1.39}$$

and use of (1.21) in (1.39) gives the final equivalent definition for $\mathcal{M}^r$:

$$\mathcal{M}^r = \mathcal{R}\mathcal{M}^{-1}\mathcal{R}. \tag{36.1.40}$$

Note that, in analogy to (1.21), reversing a map twice leaves it unchanged:

$$\begin{aligned}(\mathcal{M}^r)^r &= (\mathcal{R}\mathcal{M}^{-1}\mathcal{R})^r = \mathcal{R}(\mathcal{R}\mathcal{M}^{-1}\mathcal{R})^{-1}\mathcal{R} \\ &= \mathcal{R}\mathcal{R}^{-1}\mathcal{M}\mathcal{R}^{-1}\mathcal{R} = \mathcal{M}.\end{aligned} \tag{36.1.41}$$

From (1.21), (1.38), and (1.40) we deduce the chain of relations

$$\mathcal{M}^r\mathcal{R}\mathcal{M}\mathcal{R} = \mathcal{I}, \tag{36.1.42}$$

$$\mathcal{M}^r\mathcal{R}(\mathcal{M}^{-1})^{-1}\mathcal{R} = \mathcal{I}, \tag{36.1.43}$$

$$\mathcal{M}^r(\mathcal{M}^{-1})^r = \mathcal{I}, \tag{36.1.44}$$

$$(\mathcal{M}^{-1})^r = (\mathcal{M}^r)^{-1}. \tag{36.1.45}$$

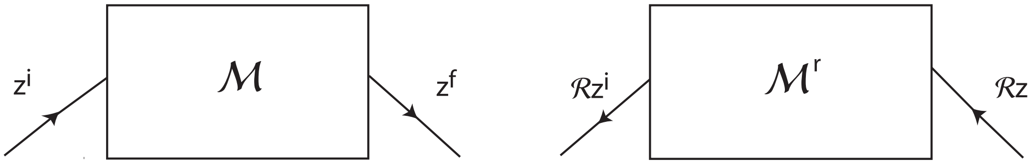


Figure 36.1.1: Actions of a map $\mathcal{M}$ and its reversed counterpart $\mathcal{M}^r$.

We also observe that if $\mathcal{M}$ can be written in the single exponent form (1.33), then use of (1.34) and (1.40) shows that there is the relation

$$\mathcal{M}^r = \mathcal{R}\exp(- : f :)\mathcal{R} = \exp(: f^r :). \tag{36.1.46}$$

If $\mathcal{M}$ is a symplectic map, so is $\mathcal{M}^r$. To prove this, let M^r be the Jacobian matrix of $\mathcal{M}^r$. From (1.40) and the chain rule we find the result

$$M^r = RM^{-1}R. \tag{36.1.47}$$

Let us check whether M^r is a symplectic matrix. We find from (3.1.2), (1.47), and (1.11) the result

$$\begin{aligned}(M^r)^TJM^r &= R(M^{-1})^TRJRM^{-1}R \\ &= -R(M^{-1})^TJM^{-1}R = -RJR = J.\end{aligned} \tag{36.1.48}$$

We see that M^r is a symplectic matrix, and hence $\mathcal{M}^r$ is a symplectic map. The observant reader will have noticed that the same conclusion could have been reached immediately from the discussion surrounding equations (1.13) through (1.18).

Suppose we combine (1.47) with the symplectic condition. From the symplectic condition (3.1.2) we deduce that

$$M^{-1} = -JM^TJ, \tag{36.1.49}$$

and hence (1.47) can also be written in the form

$$M^r = JRM^TJR = JRM^T(JR)^{-1}. \tag{36.1.50}$$

Here we have also used (1.11) and (1.9).

From (1.50) it follows that M and M^r have the *same spectrum*. Indeed, let P and P^r be the *characteristic polynomials* of M and M^r,

$$P(\lambda) = \det(M - \lambda I), \tag{36.1.51}$$

$$P^r(\lambda) = \det(M^r - \lambda I). \tag{36.1.52}$$

Then, by use of (1.50), we have the result

$$\begin{aligned} P^r(\lambda) &= \det(M^r - \lambda I) = \det[JRM^T(JR)^{-1} - \lambda I] \\ &= \det[(JR)(M^T - \lambda I)(JR)^{-1}] \\ &= \det(JR)\det[(JR)^{-1}]\det(M^T - \lambda I) \\ &= \det(M - \lambda I) = P(\lambda). \end{aligned} \tag{36.1.53}$$

The last task for this section is to determine the effect of reversal on a relation involving the action of a map on a function. Suppose the function h is the result of the symplectic map $\mathcal{M}$ acting on the function g,

$$h = \mathcal{M}g. \tag{36.1.54}$$

Letting $\mathcal{R}$ act on both sides of (1.54) and using (1.21) and (1.42) give the results

$$\mathcal{R}h = \mathcal{R}\mathcal{M}g = \mathcal{R}\mathcal{M}\mathcal{R}\mathcal{R}g = (\mathcal{M}^r)^{-1}\mathcal{R}g, \tag{36.1.55}$$

or

$$h^r = (\mathcal{M}^r)^{-1}g^r. \tag{36.1.56}$$

36.2 Applications

Suppose $\mathcal{M}$ is a product of several maps $\mathcal{M}_1$ to $\mathcal{M}_n$,

$$\mathcal{M} = \mathcal{M}_1\mathcal{M}_2\mathcal{M}_3\cdots\mathcal{M}_n. \tag{36.2.1}$$

Then, from (1.21) and (1.22), there is the result

$$\begin{aligned} \mathcal{M}^r &= \mathcal{R}(\mathcal{M}_1\mathcal{M}_2\mathcal{M}_3\cdots\mathcal{M}_n)^{-1}\mathcal{R} \\ &= \mathcal{R}(\mathcal{M}_n^{-1}\cdots\mathcal{M}_3^{-1}\mathcal{M}_2^{-1}\mathcal{M}_1^{-1})\mathcal{R} \\ &= \mathcal{R}\mathcal{M}_n^{-1}\mathcal{R}\cdots\mathcal{R}\mathcal{M}_3^{-1}\mathcal{R}\mathcal{R}\mathcal{M}_2^{-1}\mathcal{R}\mathcal{R}\mathcal{M}_1^{-1}\mathcal{R} \\ &= \mathcal{M}_n^r\cdots\mathcal{M}_3^r\mathcal{M}_2^r\mathcal{M}_1^r. \end{aligned} \tag{36.2.2}$$

Thus, the reverse of a product of maps is the product of the reverses of the individual maps taken in *opposite* order.

Define a map $\mathcal{M}$ to be *reversal symmetric* if it equals its reverse,

$$\mathcal{M}^r = \mathcal{M}. \tag{36.2.3}$$

For a reversal symmetric map (1.38) can be rewritten in the forms

$$\mathcal{R}\mathcal{M}\mathcal{R} = \mathcal{M}^{-1} \text{ or } \mathcal{M}\mathcal{R}\mathcal{M} = \mathcal{R}; \tag{36.2.4}$$

$$\mathcal{R}\mathcal{M}\mathcal{R}\mathcal{M} = \mathcal{I}, \tag{36.2.5}$$

$$\mathcal{M}\mathcal{R}\mathcal{M}\mathcal{R} = \mathcal{I}, \text{ or} \tag{36.2.6}$$

$$(\mathcal{R}\mathcal{M})^2 = (\mathcal{M}\mathcal{R})^2 = \mathcal{I}. \tag{36.2.7}$$

Here we have used (1.21). A map whose square is the identity is called an *involution.* We have seen that $\mathcal{R}\mathcal{M}$ and $\mathcal{M}\mathcal{R}$ are involutions if $\mathcal{M}$ is reversal symmetric. According to (1.21), $\mathcal{R}$ is also an involution. Finally, since $\mathcal{M}^k$ will be reversal symmetric if $\mathcal{M}$ is reversal symmetric, the maps $\mathcal{R}\mathcal{M}^k$ and $\mathcal{M}^k\mathcal{R}$ for any k are also involutions,

$$(\mathcal{R}\mathcal{M}^k)(\mathcal{R}\mathcal{M}^k) = (\mathcal{R}\mathcal{M}^k\mathcal{R})(\mathcal{M}^k) = \mathcal{M}^{-k}\mathcal{M}^k = \mathcal{I}, \tag{36.2.8}$$

$$(\mathcal{M}^k\mathcal{R})(\mathcal{M}^k\mathcal{R}) = \mathcal{M}^k(\mathcal{R}\mathcal{M}^k\mathcal{R}) = \mathcal{M}^k\mathcal{M}^{-k} = \mathcal{I}. \tag{36.2.9}$$

Moreover, there are the obvious identities

$$\mathcal{M} = \mathcal{R}\mathcal{R}\mathcal{M} = (\mathcal{R})(\mathcal{R}\mathcal{M}), \tag{36.2.10}$$

$$\mathcal{M} = \mathcal{M}\mathcal{R}\mathcal{R} = (\mathcal{M}\mathcal{R})(\mathcal{R}), \tag{36.2.11}$$

$$\mathcal{M}^k = \mathcal{R}\mathcal{R}\mathcal{M}^k = (\mathcal{R})(\mathcal{R}\mathcal{M}^k), \tag{36.2.12}$$

$$\mathcal{M}^k = \mathcal{M}^k\mathcal{R}\mathcal{R} = (\mathcal{M}^k\mathcal{R})(\mathcal{R}). \tag{36.2.13}$$

They show that if $\mathcal{M}$ is reversal symmetric, then $\mathcal{M}$ and $\mathcal{M}^k$ for any k can be written as the product of two involutions. The discovery and classification of the fixed points (closed orbits) of a map are greatly simplified if the map can be written as the product of two involutions. See Section 7.

Suppose $\mathcal{M}$ can be written in the single exponent form (1.33), and is reversal symmetric. Then we see from (1.46) that the generator f must satisfy the relation

$$f^r = f. \tag{36.2.14}$$

Suppose $\mathcal{M}$ can be written as a product of several maps $\mathcal{M}_1$ to $\mathcal{M}_n$ and their reverses,

$$\mathcal{M} = \mathcal{M}_1\mathcal{M}_2 \cdots \mathcal{M}_n\mathcal{M}_n^r \cdots \mathcal{M}_2^r\mathcal{M}_1^r. \tag{36.2.15}$$

Then, simple calculation shows that $\mathcal{M}$ is reversal symmetric. Indeed, from (2.2) and (1.40) we find the result

$$\begin{aligned} \mathcal{M}^r &= (\mathcal{M}_1^r)^r(\mathcal{M}_2^r)^r \cdots (\mathcal{M}_n^r)^r\mathcal{M}_n^r \cdots \mathcal{M}_2^r\mathcal{M}_1^r \\ &= \mathcal{M}_1\mathcal{M}_2 \cdots \mathcal{M}_n\mathcal{M}_n^r \cdots \mathcal{M}_2^r\mathcal{M}_1^r = \mathcal{M}. \end{aligned} \tag{36.2.16}$$

We next claim, based on end-to-end symmetry, that the transfer maps $\mathcal{M}$ for many common beamline elements are reversal symmetric. These elements include drifts, bends (including combined-function bends) with equal entry and exit angles, quadrupoles, sextupoles, octupoles, etc. This statement holds even if fringe-field and multipole effects are included provided the element in question has end-to-end symmetry. We will also show that the transfer map for a short on-phase RF cavity (a cavity that maintains bunching, but provides no net acceleration) is reversal symmetric. Finally, we note that the transfer map for a solenoid with end-to-end symmetry is *not* reversal symmetric. Instead, the reversed map for such a solenoid is the map for that solenoid with opposite magnetic field. That is, if $\mathcal{M}[\boldsymbol{B}(\boldsymbol{r})]$ is the map for such a solenoid with magnetic field $\boldsymbol{B}(\boldsymbol{r})$, there is the relation

$$\mathcal{M}^r[\boldsymbol{B}(\boldsymbol{r})] = \mathcal{M}[-\boldsymbol{B}(\boldsymbol{r})]. \tag{36.2.17}$$

Here we have used a square-bracket notation to indicate that the map $\mathcal{M}$ is a *functional* of the magnetic field $\boldsymbol{B}(\boldsymbol{r})$.

Imagine integrating (10.1.8) to find the map $\mathcal{M}$ for some beamline element. Divide the integration interval into $2N$ equal segments each of "duration" h. Label the intervals $1, 2, \cdots N$ followed by $\tilde{N}, \cdots \tilde{2}, \tilde{1}$. Then $\mathcal{M}$ can be written in the product form

$$\mathcal{M} = \mathcal{M}_1\mathcal{M}_2 \cdots \mathcal{M}_N\mathcal{M}_{\tilde{N}} \cdots \mathcal{M}_{\tilde{2}}\mathcal{M}_{\tilde{1}} \tag{36.2.18}$$

where $\mathcal{M}_j$ is the map for the j^{th} segment. The segments N and $\tilde{N}$ are on either side of the center of the element, and the segments 1 and $\tilde{1}$ are at the leading and trailing ends, etc. For each map $\mathcal{M}_j$ we have an approximation of the form

$$\mathcal{M}_j = \exp(-h : H_j :) + O(h^2) \tag{36.2.19}$$

where H_j is the Hamiltonian evaluated at the center of the j^{th} segment. Let us compute $(\mathcal{M}_j)^r$. From (1.34), (1.46), and (2.19) we find the result

$$(\mathcal{M}_j)^r = \exp[-h : (H_j)^r :] + O(h^2). \tag{36.2.20}$$

We now make the symmetry assumption

$$(H_{\tilde{j}})^r = H_j \text{ for } j = 1, 2, \cdots N. \tag{36.2.21}$$

It then follows that

$$(\mathcal{M}_{\tilde{j}})^r = \mathcal{M}_j + O(h^2) \tag{36.2.22}$$

and, by (1.41),

$$\mathcal{M}_{\tilde{j}} = (\mathcal{M}_j)^r + O(h^2). \tag{36.2.23}$$

Correspondingly, we may rewrite (2.18) in the form

$$\mathcal{M} = \mathcal{M}_1\mathcal{M}_2 \cdots \mathcal{M}_N(\mathcal{M}_N)^r \cdots (\mathcal{M}_2)^r(\mathcal{M}_1)^r + O(Nh^2). \tag{36.2.24}$$

Here, as a worst case estimate, we assume that all the $O(h^2)$ terms in (2.18) add constructively to produce a possible term of order Nh^2 in (2.24). Comparison of (2.15), (2.16), and (2.24) gives the result

$$\mathcal{M}^r = \mathcal{M} + O(Nh^2). \tag{36.2.25}$$

Now let the number N of segments approach infinity and the duration h of each approach zero. Then in this limit, $Nh^2 \to 0$, and we see that $\mathcal{M}^r$ must equal $\mathcal{M}$ exactly, and hence $\mathcal{M}$ is reversal symmetric.

There is a related result that is also of use. Let us write (2.18) in the form

$$\mathcal{M} = \mathcal{M}_\ell \mathcal{M}_t \tag{36.2.26}$$

where $\mathcal{M}_\ell$, the *leading* half of $\mathcal{M}$, is given by the product

$$\mathcal{M}_\ell = \mathcal{M}_1 \mathcal{M}_2 \cdots \mathcal{M}_N, \tag{36.2.27}$$

and $\mathcal{M}_t$, the *trailing* half of $\mathcal{M}$, is given by the product

$$\mathcal{M}_t = \mathcal{M}_{\tilde{N}} \cdots \mathcal{M}_{\tilde{2}} \mathcal{M}_{\tilde{1}}. \tag{36.2.28}$$

From (2.27) there is the relation

$$(\mathcal{M}_\ell)^r = (\mathcal{M}_N)^r \cdots (\mathcal{M}_2)^r (\mathcal{M}_1)^r, \tag{36.2.29}$$

and by combining this relation with (2.23) we obtain the estimate

$$(\mathcal{M}_\ell)^r = \mathcal{M}_{\tilde{N}} \cdots \mathcal{M}_{\tilde{2}} \mathcal{M}_{\tilde{1}} + O(Nh^2) = \mathcal{M}_t + O(Nh^2). \tag{36.2.30}$$

Again let the number N of segments approach infinity and the duration h of each approach zero so that $Nh^2 \to 0$. By so doing we conclude that the estimate (2.30) must in fact be the equality

$$(\mathcal{M}_\ell)^r = \mathcal{M}_t. \tag{36.2.31}$$

A few words need to be said about the symmetry assumption (2.21). Consider first *static* elements for which H does not depend on τ. In this case it is only necessary to examine how H depends on p_x and p_y. For a drift H_j is an even function (depends only on p_x^2 and p_y^2) and, of course, independent of the segment j. Therefore (2.21) holds. The same is true for the *body* of any multipole (including dipoles and combined-function dipoles), and therefore (2.21) again holds.

At the ends of a multipole H can have odd terms in p_x and p_y. For example, for a quadrupole, the Hamiltonian is of the form

$$H = -[(p_t/c)^2 - m^2c^2 - (p_x - qA_x)^2 - (p_y - qA_y)^2]^{1/2} - qA_z. \tag{36.2.32}$$

Here we have abandoned the notation (1.1). Instead, z is now a Cartesian coordinate in the longitudinal direction, and we take it to be the *independent* variable. Also, p_t is the negative of the total energy. The vector potential $\boldsymbol{A}$ for a quadruple has an expansion (shown through fourth order) of the form

$$A_x = \frac{g'(z)}{4}(x^3 - xy^2) + \cdots \,, \tag{36.2.33}$$

$$A_y = -\frac{g'(z)}{4}(y^3 - x^2y) + \cdots \,, \tag{36.2.34}$$

$$A_z = -\frac{g(z)}{2}(x^2 - y^2) + \frac{g''(z)}{12}(x^4 - y^4) + \cdots \,. \tag{36.2.35}$$

Here $g(z)$ is the on-axis field gradient, and the quantities $g'(z)$ and $g''(z)$ are derivatives of g with respect to z. We note that once $g(z)$ is specified (and quadrupole symmetry is imposed), then all other terms are determined by the Maxwell equations. Inspection of (2.32) shows that H is unchanged by the substitution $(p_x, p_y) \to (-p_x, -p_y)$ provided there is also the substitution $(A_x, A_y) \to (-A_x, -A_y)$. Suppose, for convenience, we choose the z coordinate so that $z = 0$ is at the center of the quadrupole. Then, for what we would intuitively call a symmetric quadrupole in the sense of having end-to-end symmetry, $g(z)$ should be an *even* function of z,

$$g(-z) = g(z). \tag{36.2.36}$$

From (2.36) we deduce that $g''(z)$, $g^{iv}(z)$, etc. are then also *even* functions of z; and $g'(z)$, $g'''(z)$, etc. are *odd* functions of z. It follows from (2.33) through (2.35) that, for a quadrupole with end-to-end symmetry, A_x and A_y are odd functions of z,

$$A_x(x, y, -z) = -A_x(x, y, z), \tag{36.2.37}$$

$$A_y(x, y, -z) = -A_y(x, y, z), \tag{36.2.38}$$

and A_z is an even function,

$$A_z(x, y, -z) = A_z(x, y, z). \tag{36.2.39}$$

[Note that the conditions (2.37) through (2.39) imply for the magnetic field the symmetry relations $B_{x,y}(x, y, -z) = B_{x,y}(x, y, z)$ and $B_z(x, y, -z) = -B_z(x, y, z)$]. From (2.32) and (2.37) through (2.39) we conclude that

$$H^r(-z) = H(z), \tag{36.2.40}$$

and therefore (2.21) is again satisfied. Thus, our intuitive sense of symmetry for a quadrupole coincides with the precise definition (2.21) for the Hamiltonian, which in turn implies the reversal symmetry condition (2.3) for the associated transfer map.

The same can be shown to be true for any multipole, including skew multipoles, with end-to-end symmetry. Finally, the same can be shown to be true for any dipole, with or without additional multipoles intended or otherwise, provided the magnet (including all multipole and fringe fields) has end-to-end symmetry. [Note that the Hamiltonian (2.32) can also be used for curved elements providing the bending angle is less than π. And for larger bend angles an analogous treatment can be formulated using cylindrical coordinates.] That is, in all these cases the relations (2.37) through (2.39) hold, and they imply the relation (2.40). By contrast, the transfer map for a dipole with unequal entry and exit angles, or for a combined function dipole with excessive quadrupole field at one end, will not be reversal symmetric.

Consider next the case of a short RF cavity phased to act as a buncher. Such an idealized cavity can be described by a map $\mathcal{M}$ of the form (1.32) with f given by the relation

$$f = (V/\omega) \cos \omega\tau. \tag{36.2.41}$$

Here V and ω are the voltage and frequency of the cavity. We see that f is *even* in τ, and therefore $f^r = f$. It follows that $\mathcal{M}$ is reversal symmetric. The case of a finite length RF cavity with realistic electromagnetic fields awaits investigation. Finally, it is evident that the transfer map for a short RF cavity phased to operate as an accelerating element rather than a bunching element [$\cos\omega\tau$ replaced by $\cos(\omega\tau+\phi)$] is not reversal symmetric.

We have seen that the transfer maps for many common beam-line elements are reversal symmetric. Suppose that $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ are reversal symmetric maps. Then, by (2.1) and (2.2), the maps $\mathcal{M}$ given by products of the form

$$\mathcal{M} = \mathcal{M}_1\mathcal{M}_2\mathcal{M}_1, \tag{36.2.42}$$

$$\mathcal{M} = \mathcal{M}_1\mathcal{M}_2\mathcal{M}_3\mathcal{M}_2\mathcal{M}_1 \tag{36.2.43}$$

will be reversal symmetric. For example, $\mathcal{M}_1$ could be the map for a drift and $\mathcal{M}_2$ could be the map for a quadrupole. It follows that the map for a quadrupole sandwiched between two equal length drifts, given by (2.42), is reversal symmetric. Or $\mathcal{M}_1$ and $\mathcal{M}_3$ could be maps for quadrupoles and $\mathcal{M}_2$ could be the map for a drift. Then (2.43) would be the map for a quadrupole triplet, and we conclude that such maps are reversal symmetric. In the case of solenoids with end-to-end symmetry we could consider maps of the form

$$\mathcal{M} = \mathcal{M}_1\mathcal{M}_2\mathcal{M}_3 \tag{36.2.44}$$

where $\mathcal{M}_1$ is a solenoid map, $\mathcal{M}_2$ is a drift or quadrupole, and $\mathcal{M}_3$ is a map for an identical solenoid except for a reversed field. These maps (2.44) would also be reversal symmetric.

It frequently happens that a given map $\mathcal{M}$ is not reversal symmetric, but is *conjugate* to a map $\mathcal{N}$ that is reversal symmetric. That is, given $\mathcal{M}$, there exists a conjugating map $\mathcal{A}$ such that $\mathcal{M}$ can be written in the form

$$\mathcal{M} = \mathcal{A}^{-1}\mathcal{N}\mathcal{A} \tag{36.2.45}$$

where $\mathcal{N}$ is reversal symmetric. Consider, for example, the case of a FODO cell. Its map $\mathcal{M}$ is given by the product

$$\mathcal{M} = \mathcal{F}\mathcal{O}\mathcal{D}\mathcal{O} \tag{36.2.46}$$

where $\mathcal{F}$ and $\mathcal{D}$ are the maps for (horizontally) *focusing* and *defocussing* quadrupoles and $\mathcal{O}$ is the map for a *drift* (or a reversal symmetric dipole). Evidently $\mathcal{M}$ is not reversal symmetric although, according to our previous discussion, its factors are. However, we know that $\mathcal{F}$ (including all multipole and fringe-field effects) can be written as the product

$$\mathcal{F} = \mathcal{F}_\ell\mathcal{F}_t \tag{36.2.47}$$

where $\mathcal{F}_\ell$ and $\mathcal{F}_t$ are the maps for the leading and trailing halves of the focusing quadrupole. Moreover, there is the relation

$$\mathcal{F}_\ell^r = \mathcal{F}_t. \tag{36.2.48}$$

As a result of (2.47) $\mathcal{M}$ can be written in the form

$$\mathcal{M} = \mathcal{F}_\ell\mathcal{F}_t\mathcal{O}\mathcal{D}\mathcal{O} = \mathcal{F}_\ell\mathcal{F}_t\mathcal{O}\mathcal{D}\mathcal{O}\mathcal{F}_\ell\mathcal{F}_\ell^{-1} = \mathcal{A}^{-1}\mathcal{N}\mathcal{A} \tag{36.2.49}$$

with

$$\mathcal{A}^{-1} = \mathcal{F}_\ell \tag{36.2.50}$$

and

$$\mathcal{N} = \mathcal{F}_t \mathcal{O}\mathcal{D}\mathcal{O}\mathcal{F}_\ell. \tag{36.2.51}$$

Let us compute the reverse of $\mathcal{N}$. We find, using (2.1) and (2.48), the result

$$\mathcal{N}^r = \mathcal{F}_\ell^r \mathcal{O}^r \mathcal{D}^r \mathcal{O}^r \mathcal{F}_t^r = \mathcal{F}_t \mathcal{O}\mathcal{D}\mathcal{O}\mathcal{F}_\ell = \mathcal{N}. \tag{36.2.52}$$

Therefore, $\mathcal{N}$ is reversal symmetric. This example illustrates that the one-turn map for a ring is often reversal symmetric providing the surface of section (location at which the one-turn map is computed) is properly chosen.

As a second illustration, suppose that (through some order) $\mathcal{M}$ can be brought to normal form. (For example, assume that the eigenvalues of the linear part of $\mathcal{M}$ lie on the unit circle and that the corresponding tunes are not resonant through some order.) Then we may take $\mathcal{A}$ to be the normalizing map, and $\mathcal{N}$ to be the normal form of $\mathcal{M}$. The map $\mathcal{N}$ can be written in terms of a single exponent,

$$\mathcal{N} = \exp(:h:). \tag{36.2.53}$$

In the case of a static ring (no RF) h takes the form

$$\begin{aligned} h &= -(\phi_x/2)(p_x^2 + x^2) - (\phi_y/2)(p_y^2 + y^2) + b^1 p_\tau^2 + b^2 p_\tau^3 + b^3 p_\tau^4 \\ &\quad + a_{xx}(p_x^2 + x^2)^2 + a_{xy}(p_x^2 + x^2)(p_y^2 + y^2) + a_{yy}(p_y^2 + y^2)^2 \\ &\quad + c_x^1 (p_x^2 + x^2) p_\tau + c_x^2 (p_x^2 + x^2) p_\tau^2 \\ &\quad + c_y^1 (p_y^2 + y^2) p_\tau + c_y^2 (p_y^2 + y^2) p_\tau^2 + \cdots . \end{aligned} \tag{36.2.54}$$

We see that h is a power series in the quantities K_x, K_y, and p_τ where

$$K_x = (p_x^2 + x^2), \tag{36.2.55}$$

$$K_y = (p_y^2 + y^2). \tag{36.2.56}$$

The quantities $[\phi/(2\pi)]$ are (fractional) tunes, the quantities b are related to phase slip (momentum compaction), the quantities a are related to anharmonicities, and the quantities c are related to chromaticities. In the dynamic case h takes the form

$$\begin{aligned} h &= -(\phi_x/2)K_x - (\phi_y/2)K_y - (\phi_\tau/2)K_\tau \\ &\quad + a_{xx}K_x^2 + a_{yy}K_y^2 + a_{\tau\tau}K_\tau^2 + a_{xy}K_x K_y \\ &\quad + a_{x\tau}K_x K_\tau + a_{y\tau}K_y K_\tau + \cdots . \end{aligned} \tag{36.2.57}$$

Now h is a power series in K_x, K_y, and K_τ with K_x, K_y defined as above and

$$K_\tau = (p_\tau^2 + \tau^2). \tag{36.2.58}$$

We see that in both cases $h^r = h$, and therefore $\mathcal{N}$ is reversal symmetric. Of course, the normal form procedure usually leads to divergent series and therefore these normal form results are only approximate in the formal sense of holding to any order, but usually not exactly.

We have defined a map $\mathcal{M}$ to be reversal symmetric if it satisfies the condition (2.3). We now define a map to be reversal *antisymmetric* if it satisfies the condition

$$\mathcal{M}^r = \mathcal{M}^{-1}. \tag{36.2.59}$$

Suppose $\mathcal{M}$ can be written in the single exponent form (1.33), and is reversal antisymmetric. Then we see from (1.46) that the generator f must satisfy the relation

$$f^r = -f. \tag{36.2.60}$$

Note that reversal antisymmetric maps form a subgroup. In particular, if two maps $\mathcal{M}_1$ and $\mathcal{M}_2$ are reversal antisymmetric, so is their product $\mathcal{M}_1\mathcal{M}_2$:

$$(\mathcal{M}_1\mathcal{M}_2)^r = \mathcal{M}_2^r\mathcal{M}_1^r = (\mathcal{M}_2^{-1})(\mathcal{M}_1^{-1}) = (\mathcal{M}_1\mathcal{M}_2)^{-1}. \tag{36.2.61}$$

Correspondingly, functions that satisfy (2.60) form a Lie algebra under Poisson bracketing. Indeed, if f and g are *any* two functions, there is the general relation

$$\mathcal{R}[f,g] = \mathcal{R} : f : g = \mathcal{R} : f : \mathcal{R}\mathcal{R}g = - : f^r : g^r = -[f^r, g^r]. \tag{36.2.62}$$

And, if f and g satisfy (2.60), there is the result

$$\mathcal{R}[f,g] = -[f,g]. \tag{36.2.63}$$

Thus, if f and g change sign under reversal, so does their Poisson bracket.

We close this section by showing that under certain circumstances a symplectic map $\mathcal{M}$ can be written uniquely as the product of a reversal symmetric symplectic map and a reversal antisymmetric symplectic map. Given any symplectic map $\mathcal{M}$, define an associated symplectic map $\mathcal{B}$ by the relation

$$\mathcal{B} = \mathcal{M}\mathcal{M}^r. \tag{36.2.64}$$

Then, by (2.1) and (2.2), $\mathcal{B}$ is reversal symmetric. Assume that it is possible to find a *square root* $\mathcal{S}$ for $\mathcal{B}$, and that this square root is also reversal symmetric. That is, assume there is an $\mathcal{S}$ satisfying

$$\mathcal{S}^r = \mathcal{S} \tag{36.2.65}$$

such that

$$\mathcal{B} = \mathcal{S}^2. \tag{36.2.66}$$

For example, suppose that $\mathcal{B}$ can be written in the single exponent form

$$\mathcal{B} = \exp(: f :) \tag{36.2.67}$$

and that f has the property (2.14). Then we may define $\mathcal{S}$ by the relation

$$\mathcal{S} = \exp(: f : /2). \tag{36.2.68}$$

Next define a symplectic map $\mathcal{A}$ by the relation

$$\mathcal{A} = \mathcal{S}^{-1}\mathcal{M}. \tag{36.2.69}$$

Then $\mathcal{A}$ will be reversal antisymmetric. Indeed, we have the results

$$\mathcal{A}^r = \mathcal{M}^r \mathcal{S}^{-1}, \tag{36.2.70}$$

$$\mathcal{A}\mathcal{A}^r = \mathcal{S}^{-1}\mathcal{M}\mathcal{M}^r\mathcal{S}^{-1} = \mathcal{S}^{-1}\mathcal{B}\mathcal{S}^{-1} = \mathcal{S}^{-1}\mathcal{S}^2\mathcal{S}^{-1} = \mathcal{I}, \tag{36.2.71}$$

and therefore $\mathcal{A}$ is reversal antisymmetric. Finally, (2.69) can be rewritten in the form

$$\mathcal{M} = \mathcal{S}\mathcal{A}. \tag{36.2.72}$$

We see that, under the assumptions made, $\mathcal{M}$ can be written as the product of a reversal symmetric map $\mathcal{S}$ and a reversal antisymmetric map $\mathcal{A}$, and we have found expressions for each.

Exercises

36.2.1. Review the first part of Exercise 6.2.9. Let $\mathcal{S}_3$ be the set of all reversal antisymmetric symplectic maps. Show that $\mathcal{S}_3$ forms a group, and that there is the inclusion relation

$$\mathcal{S}_0 \supset \mathcal{S}_1 \supset \mathcal{S}_2 \supset \mathcal{S}_3. \tag{36.2.73}$$

36.3 General Consequences for Straight and Circular Machines

Suppose $\mathcal{M}$ has a factorization of the form

$$\mathcal{M} = \mathcal{L}\ \exp: f_3 : \exp : f_4 : \cdots . \tag{36.3.1}$$

Here $\mathcal{L}$ is the linear part of $\mathcal{M}$. Then, we find for $\mathcal{M}^r$ the factorization

$$\begin{aligned} \mathcal{M}^r &= \cdots [\exp(: f_4 :)]^r [\exp(: f_3 :)]^r \mathcal{L}^r \\ &= \cdots \exp(: f_4^r :) \exp(: f_3^r :)\mathcal{L}^r. \end{aligned} \tag{36.3.2}$$

Here we have used (2.2) and (1.46). Also, if the action of $\mathcal{L}^r$ is described by the matrix L^r, then the operator relation

$$\mathcal{L}^r = \mathcal{R}\mathcal{L}^{-1}\mathcal{R} \tag{36.3.3}$$

yields the matrix relation

$$L^r = RL^{-1}R \tag{36.3.4}$$

consistent with (1.47). Also, as done before for (1.49), we deduce from the symplectic condition (3.1.2) that (3.3) can also be written in the form

$$L^r = JRL^TJR. \tag{36.3.5}$$

Let us (in the 6×6 case) write L using the standard matrix notation,

$$L = \begin{pmatrix} L_{11} & L_{12} & L_{13} & L_{14} & L_{15} & L_{16} \\ L_{21} & L_{22} & L_{23} & L_{24} & L_{25} & L_{26} \\ L_{31} & L_{32} & L_{33} & L_{34} & L_{35} & L_{36} \\ L_{41} & L_{42} & L_{43} & L_{44} & L_{45} & L_{46} \\ L_{51} & L_{52} & L_{53} & L_{54} & L_{55} & L_{56} \\ L_{61} & L_{62} & L_{63} & L_{64} & L_{65} & L_{66} \end{pmatrix}. \tag{36.3.6}$$

Then, upon evaluating (3.5), we find that L^r is the matrix

$$L^r = \begin{pmatrix} L_{22} & L_{12} & L_{42} & L_{32} & -L_{62} & -L_{52} \\ L_{21} & L_{11} & L_{41} & L_{31} & -L_{61} & -L_{51} \\ L_{24} & L_{14} & L_{44} & L_{34} & -L_{64} & -L_{54} \\ L_{23} & L_{13} & L_{43} & L_{33} & -L_{63} & -L_{53} \\ -L_{26} & -L_{16} & -L_{46} & -L_{36} & L_{66} & L_{56} \\ -L_{25} & -L_{15} & -L_{45} & -L_{35} & L_{65} & L_{55} \end{pmatrix}. \tag{36.3.7}$$

Now suppose that $\mathcal{M}$ is reversal symmetric so that (2.3) holds. Taken together (2.3) and (3.2) give the result

$$\cdots \exp(:f_4^r:)\exp(:f_3^r:)\mathcal{L}^r = \mathcal{L}\exp(:f_3:)\exp(:f_4:)\cdots . \tag{36.3.8}$$

The right side of (3.8) can be rewritten in the form

$$\begin{aligned} \mathcal{L}\exp(:f_3:)\exp(:f_4:)\cdots &= \mathcal{L}\exp(:f_3:)\exp(:f_4:)\cdots\mathcal{L}^{-1}\mathcal{L} \\ &= \cdots\exp(:g_4:)\exp(:g_3:)\mathcal{L}, \end{aligned} \tag{36.3.9}$$

where the quantities $g_3, g_4, \cdots$ are yet to be determined. Upon comparing (3.8) and (3.9) we see that reversal symmetry requires the relations

$$\mathcal{L}^r = \mathcal{L}, \tag{36.3.10}$$

$$f_m^r = g_m. \tag{36.3.11}$$

The linear part $\mathcal{L}$ of $\mathcal{M}$ must be reversal symmetric, and the Lie generators of the nonlinear part must satisfy (3.11).

To work out the implications of (3.11) for the nonlinear part in more detail, we must find the g_m in terms of the f_m. From (3.9) we conclude that there is the relation

$$\exp(:\mathcal{L}f_3:)\exp(:\mathcal{L}f_4:)\cdots = \cdots\exp(:g_4:)\exp(:g_3:). \tag{36.3.12}$$

Now use the Baker-Campbell-Hausdorff series (8.2.28) and (8.2.29) to combine the exponents on each side of (3.12) and equate terms of like degree. Doing so yields the relations

$$g_3 = \mathcal{L}f_3, \tag{36.3.13}$$

$$g_4 = \mathcal{L}f_4, \tag{36.3.14}$$

$$g_5 = \mathcal{L}f_5 + [\mathcal{L}f_3, \mathcal{L}f_4], \tag{36.3.15}$$

$$g_6 = \mathcal{L}f_6 + [\mathcal{L}f_3, \mathcal{L}f_5] + (1/2)[\mathcal{L}f_3, [\mathcal{L}f_3, \mathcal{L}f_4]], \text{ etc.} \tag{36.3.16}$$

Finally, employ (3.11) in the relations (3.13) through (3.16) to find the results

$$f_3^r = \mathcal{L}f_3, \tag{36.3.17}$$

$$f_4^r = \mathcal{L}f_4, \tag{36.3.18}$$

$$f_5^r = \mathcal{L}f_5 + [\mathcal{L}f_3, \mathcal{L}f_4], \tag{36.3.19}$$

$$f_6^r = \mathcal{L}f_6 + [\mathcal{L}f_3, \mathcal{L}f_5] + (1/2)[\mathcal{L}f_3, [\mathcal{L}f_3, \mathcal{L}f_4]], \text{ etc.} \tag{36.3.20}$$

Let us explore the consequences of reversal symmetry for $\mathcal{L}$. The relation (3.15) implies the associated matrix relation

$$L^r = L. \tag{36.3.21}$$

In view of (3.6) and (3.7), and in the 6×6 case, simple enumeration shows that reversal symmetry places on the entries in L the 15 restrictions listed below:

$$L_{11} = L_{22}, \tag{36.3.22}$$

$$L_{13} = L_{42}, \tag{36.3.23}$$

$$L_{14} = L_{32}, \tag{36.3.24}$$

$$L_{15} = -L_{62}, \tag{36.3.25}$$

$$L_{16} = -L_{52}, \tag{36.3.26}$$

$$L_{23} = L_{41}, \tag{36.3.27}$$

$$L_{24} = L_{31}, \tag{36.3.28}$$

$$L_{25} = -L_{61}, \tag{36.3.29}$$

$$L_{26} = -L_{51}, \tag{36.3.30}$$

$$L_{33} = L_{44}, \tag{36.3.31}$$

$$L_{35} = -L_{64}, \tag{36.3.32}$$

$$L_{36} = -L_{54}, \tag{36.3.33}$$

$$L_{45} = -L_{63}, \tag{36.3.34}$$

$$L_{46} = -L_{53}, \tag{36.3.35}$$

$$L_{55} = L_{66}. \tag{36.3.36}$$

Of course, there are also restrictions on L that follow from the symplectic condition. Suppose that $\mathcal{M}$ is a static (time independent) map. From the work of Section 19.4 we know that in the static case L must have the form

$$L = \begin{pmatrix} L_{11} & L_{12} & L_{13} & L_{14} & 0 & (\check{L}\delta)_1 \\ L_{21} & L_{22} & L_{23} & L_{24} & 0 & (\check{L}\delta)_2 \\ L_{31} & L_{32} & L_{33} & L_{34} & 0 & (\check{L}\delta)_3 \\ L_{41} & L_{42} & L_{43} & L_{44} & 0 & (\check{L}\delta)_4 \\ -\delta_2 & \delta_1 & -\delta_4 & \delta_3 & 1 & L_{56} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{36.3.37}$$

Recall that we have previously defined a matrix $\hat{L}$ by the rule

$$\hat{L} = \begin{pmatrix} L_{11} & L_{12} & L_{13} & L_{14} & 0 & 0 \\ L_{21} & L_{22} & L_{23} & L_{24} & 0 & 0 \\ L_{31} & L_{32} & L_{33} & L_{34} & 0 & 0 \\ L_{41} & L_{42} & L_{43} & L_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{36.3.38}$$

and that we have used the notation

$$\hat{L} = \begin{pmatrix} \check{L} & 0 \\ 0 & I \end{pmatrix} \tag{36.3.39}$$

where $\check{L}$ is the 4×4 matrix

$$\check{L} = \begin{pmatrix} L_{11} & L_{12} & L_{13} & L_{14} \\ L_{21} & L_{22} & L_{23} & L_{24} \\ L_{31} & L_{32} & L_{33} & L_{34} \\ L_{41} & L_{42} & L_{43} & L_{44} \end{pmatrix}. \tag{36.3.40}$$

See (19.*.*). Moreover, we know that the linear map $\hat{\mathcal{L}}$ associated with the matrix $\hat{L}$ is a symplectic map, and hence $\hat{L}$ and $\check{L}$ are symplectic matrices.

Suppose, now, that the static linear map described by (3.37) is also reversal symmetric so that the conditions (3.22) through (3.36) also hold. Then, from the form (3.37) for L, it is evident that the five conditions (3.25), (3.29), (3.32), (3.34), and (3.36) on the matrix elements L_{15} through L_{55} are automatically satisfied. By contrast, the four conditions (3.26), (3.30), (3.33), and (3.35) on the matrix elements L_{16} through L_{46} yield the relations

$$(\check{L}\delta)_1 = -\delta_1, \tag{36.3.41}$$

$$(\check{L}\delta)_2 = \delta_2, \tag{36.3.42}$$

$$(\check{L}\delta)_3 = -\delta_3, \tag{36.3.43}$$

$$(\check{L}\delta)_4 = \delta_4. \tag{36.3.44}$$

In the spirit of (3.40), let us use J and R as given by (3.2.10) and (1.5) to define associated 4×4 matrices $\check{J}$ and $\check{R}$,

$$\check{J} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}, \tag{36.3.45}$$

$$\check{R} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \tag{36.3.46}$$

With the aid of $\check{R}$ the relations (3.41) through (3.44) can be written in the more compact forms

$$\check{L}\delta = -\check{R}\delta \tag{36.3.47}$$

or

$$\check{R}\check{L}\delta = -\delta. \tag{36.3.48}$$

Moreover, the remaining six reversal symmetry relations (3.22), (3.23), (3.24), (3.27), (3.28), and (3.31) for the entries of $\check{L}$ can be written in the compact form

$$\check{L} = \check{J}\check{R}\check{L}^T\check{J}\check{R}. \tag{36.3.49}$$

Finally, the entries of $\check{L}$ must also satisfy the symplectic condition

$$\check{L}^T\check{J}\check{L} = \check{J}. \tag{36.3.50}$$

The imposition of reversal symmetry on $\mathcal{L}$ also implies an associated relation that must be satisfied by the f_1 in (19.*.*). This relation can be obtained by matrix and vector manipulation using (19.*.*) and (3.48). It is also instructive to obtain the condition on f_1 by Lie manipulation starting with (19.*.*). Rewrite (19.*.*) in the explicit form

$$\mathcal{L} = \exp(:\xi p_\tau^2:)\exp(:p_\tau f_1:)\hat{\mathcal{L}}, \tag{36.3.51}$$

and apply reversal to find the equivalent relation

$$\begin{aligned}\mathcal{L}^r &= \hat{\mathcal{L}}^r\exp(:p_\tau f_1^r:)\exp(:\xi p_\tau^2:)\\ &= \exp(:\xi p_\tau^2:)\hat{\mathcal{L}}^r\exp(:p_\tau f_1^r:)(\hat{\mathcal{L}}^r)^{-1}\hat{\mathcal{L}}^r\\ &= \exp(:\xi p_\tau^2:)\exp(:p_\tau\hat{\mathcal{L}}^r f_1^r:)\hat{\mathcal{L}}^r.\end{aligned} \tag{36.3.52}$$

Here we have also used (8.2.27). Now require that $\mathcal{L}$ satisfy the reversal symmetry condition (3.10). Comparison of (3.51) and (3.52) then gives the relation

$$\hat{\mathcal{L}}^r = \hat{\mathcal{L}}, \tag{36.3.53}$$

which is equivalent to (3.49), and the relation

$$\hat{\mathcal{L}}^r f_1^r = f_1. \tag{36.3.54}$$

In view of (3.53), the relation (3.54) can also be written in the form

$$f_1^r = \hat{\mathcal{L}}^{-1}f_1, \tag{36.3.55}$$

which is equivalent to (3.47) or (3.48).

36.4 Consequences for some Special Cases

Let us explore the implication of reversal symmetry for some special cases. Again assume that $\mathcal{M}$ is static, and suppose further that $\check{L}$ does not couple the x, p_x, and y, p_y degrees of freedom. In this case L as given by (3.37) takes the form

$$L = \begin{pmatrix} a & b & 0 & 0 & 0 & \check{\delta}_1 \\ c & d & 0 & 0 & 0 & \check{\delta}_2 \\ 0 & 0 & e & f & 0 & \check{\delta}_3 \\ 0 & 0 & g & h & 0 & \check{\delta}_4 \\ -\delta_2 & \delta_1 & -\delta_4 & \delta_3 & 1 & L_{56} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{36.4.1}$$

where

$$\check{\delta}_a = (\check{L}\delta)_a. \tag{36.4.2}$$

And, by combining (3.6), (3.7), and (3.48), L^r takes the form

$$L^r = \begin{pmatrix} d & b & 0 & 0 & 0 & -\delta_1 \\ c & a & 0 & 0 & 0 & \delta_2 \\ 0 & 0 & h & f & 0 & -\delta_3 \\ 0 & 0 & g & e & 0 & \delta_4 \\ -\check{\delta}_2 & -\check{\delta}_1 & -\check{\delta}_4 & -\check{\delta}_3 & 1 & L_{56} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{36.4.3}$$

Imposing the reversal symmetry condition (3.21) yields the relations

$$\check{\delta}_1 = -\delta_1, \tag{36.4.4}$$

$$\check{\delta}_2 = \delta_2, \tag{36.4.5}$$

$$\check{\delta}_3 = -\delta_3, \tag{36.4.6}$$

$$\check{\delta}_4 = \delta_4, \tag{36.4.7}$$

which echo (3.41) through (3.44), and the additional relations

$$a = d, \tag{36.4.8}$$

$$e = h. \tag{36.4.9}$$

Moreover, the symplectic condition (3.50) provides the relations

$$ad - bc = 1, \tag{36.4.10}$$

$$eh - fg = 1. \tag{36.4.11}$$

Suppose we further assume that, say in the x, p_x plane, the system is *imaging* (and therefore $b = 0$) or *telescopic* (and therefore $c = 0$). Then it follows from (4.8) and (4.10) that

$$a = d = \pm 1. \tag{36.4.12}$$

Similar conclusions hold for the y, p_y plane. We have learned that if a system is reversal symmetric and imaging or telescopic, then (up to a sign) the system must have unit magnification or unit "parallelization".

Suppose instead that the eigenvalues of $\check{L}$ lie on the unit circle (but differ from ± 1) and that the x, p_x and y, p_y degrees of freedom are uncoupled. Then $\hat{\mathcal{L}}$ can be written in the form

$$\hat{\mathcal{L}} = \exp(: f_2 :) \tag{36.4.13}$$

where

$$\begin{aligned} f_2 = \; & - \;\; (\phi_x/2)(\beta_x p_x^2 + 2\alpha_x x p_x + \gamma_x x^2) \\ & - \;\; (\phi_y/2)(\beta_y p_y^2 + 2\alpha_y y p_y + \gamma_y y^2). \end{aligned} \tag{36.4.14}$$

Here $[\phi_x/(2\pi)]$ and $[\phi_y/2\pi]$ are the horizontal and vertical (fractional) tunes, and $\alpha_x, \beta_x, \gamma_x$ and $\alpha_y, \beta_y, \gamma_y$ are the horizontal and vertical Courant-Snyder (Twiss) functions. Then reversal symmetry requires the relation (2.14) from which it follows that

$$\alpha_x = \alpha_y = 0. \tag{36.4.15}$$

Thus, the Courant-Snyder ellipses for both the x and y degrees of freedom are *upright* if $\mathcal{M}$ is reversal symmetric.

36.5 Consequences for Closed Orbit in a Circular Machine

There is a further factorization of a static linear sympletic map $\mathcal{L}$ as given by (19.*.*) that is of use. Let $\mathcal{A}_d$ be the map given by the relation

$$\mathcal{A}_d = \exp(: p_\tau g_1 :) \tag{36.5.1}$$

where g_1 is yet to be determined (but does not depend on τ and p_τ). Here we use the subscript d to indicate that $\mathcal{A}_d$ is analogous to $\mathcal{D}$ as given by (19.*.*). Consider the map $\mathcal{N}_d$ given by

$$\mathcal{N}_d = \mathcal{A}_d \mathcal{L} \mathcal{A}_d^{-1}. \tag{36.5.2}$$

From (19.*.*) we have the result

$$\mathcal{N}_d = \mathcal{A}_d \mathcal{C} \mathcal{D} \hat{\mathcal{L}} \mathcal{A}_d^{-1} = \mathcal{C} \mathcal{A}_d \mathcal{D} \hat{\mathcal{L}} \mathcal{A}_d^{-1} = \mathcal{C} \mathcal{A}_d \mathcal{D} \hat{\mathcal{L}} \mathcal{A}_d^{-1} \hat{\mathcal{L}}^{-1} \hat{\mathcal{L}}. \tag{36.5.3}$$

At this point we ask if there is a choice of $\mathcal{A}_d$ such that

$$\mathcal{A}_d \mathcal{D} \hat{\mathcal{L}} \mathcal{A}_d^{-1} \hat{\mathcal{L}}^{-1} = \mathcal{C}(\xi') \tag{36.5.4}$$

where $\mathcal{C}(\xi')$ is a map of the form (19.*.*). With the aid of (5.13.14), (8.2.27) and the Baker-Campbell-Hausdorff series (3.7.34) we find the result

$$\begin{aligned} & \mathcal{A}_d \mathcal{D} \hat{\mathcal{L}} \mathcal{A}_d^{-1} \hat{\mathcal{L}}^{-1} = \\ & \exp(: p_\tau g_1 :) \exp(: p_\tau f_1 :) \hat{\mathcal{L}} \exp(- : p_\tau g_1 :) \hat{\mathcal{L}}^{-1} = \\ & \exp(: p_\tau g_1 :) \exp(: p_\tau f_1 :) \exp(- : p_\tau \hat{\mathcal{L}} g_1) \\ & = \exp\{: p_\tau (g_1 + f_1 - \hat{\mathcal{L}} g_1) :\} \times \\ & \exp\{: (p_\tau^2/2)(-[f_1, \hat{\mathcal{L}} g_1] + [g_1, f_1] - [g_1, \hat{\mathcal{L}} g_1]) :\}. \end{aligned} \tag{36.5.5}$$

We see that (5.4) can be achieved providing a g_1 can be found such that

$$(\hat{\mathcal{L}} - \mathcal{I})g_1 = f_1. \tag{36.5.6}$$

Inspection of (3.39) and (3.40) shows that (5.6) can be solved (uniquely) to obtain g_1 in terms of f_1 providing $\check{L}$ as given by (3.40) does not have $+1$ as an eigenvalue. That is, neither transverse tune is integer. For the record, we also note that when g_1 is specified by (5.6), the map $\mathcal{C}(\xi')$ is given by the relation

$$\xi' = [g_1, f_1]/2. \tag{36.5.7}$$

By combining (5.3) and (5.4), we find the relation

$$\mathcal{N}_d = \mathcal{C}(\eta)\hat{\mathcal{L}} \tag{36.5.8}$$

with $\mathcal{C}(\eta)$ given by the product

$$\mathcal{C}(\eta) = \mathcal{C}(\xi)\mathcal{C}(\xi') = \mathcal{C}(\xi + \xi'). \tag{36.5.9}$$

Finally, solving (5.2) for $\mathcal{L}$ and use of (5.8) yields the factorization

$$\mathcal{L} = \mathcal{C}(\eta)\mathcal{A}_d^{-1}\hat{\mathcal{L}}\mathcal{A}_d. \tag{36.5.10}$$

The quantities η and g_1 appearing in $\mathcal{C}(\eta)$ and $\mathcal{A}_d$ have a physical interpretation. In analogy with (19.*.*), write g_1 in the form

$$g_1 = \Delta_2 x - \Delta_1 p_x + \Delta_4 y - \Delta_3 p_y. \tag{36.5.11}$$

The operator relation (5.10) is equivalent to the matrix relation

$$L = A_d\hat{L}A_d^{-1}C(\eta) = A_d C(\eta)\hat{L}A_d^{-1} \tag{36.5.12}$$

where $C(\eta)$ and A_d are the matrices associated with $\mathcal{C}(\eta)$ and $\mathcal{A}_d$. In analogy with (19.*.*) and (19.*.*) they are given by the equations

$$C(\eta) = C(\xi + \xi'), \tag{36.5.13}$$

$$A_d = D(\Delta), \tag{36.5.14}$$

$$A_d^{-1} = D(-\Delta). \tag{36.5.15}$$

[Note that in writing (5.12) we have employed the fact that $C(\eta)$ commutes with A_d^{-1}, $\hat{L}$, and A_d.] Now let $z^0(p_\tau)$ be the vector with all zero entries save for p_τ in the last entry,

$$z^0(p_\tau) = (0, 0, 0, 0, 0, p_\tau). \tag{36.5.16}$$

It evidently has the properties

$$\hat{L}z^0(p_\tau) = z^0(p_\tau), \tag{36.5.17}$$

$$C(\eta)z^0(p_\tau) = \bar{z}^0(p_\tau) \tag{36.5.18}$$

where $\bar{z}^0(p_\tau)$ has the entries

$$\bar{z}^0(p_\tau) = (0, 0, 0, 0, \eta p_\tau, p_\tau). \tag{36.5.19}$$

Next let $z^c(p_\tau)$ be the vector defined by the relation

$$z^c(p_\tau) = A_d z^c(p_\tau). \tag{36.5.20}$$

Carrying out the indicated multiplication shows that it has the entries

$$z^c(p_\tau) = (\Delta_1 p_\tau, \Delta_2 p_\tau, \Delta_3 p_\tau, \Delta_4 p_\tau, 0, p_\tau). \tag{36.5.21}$$

Also, by construction, it has the property

$$z^0(p_\tau) = A_d^{-1} z^c(p_\tau). \tag{36.5.22}$$

Finally, let us apply L to $z^c(p_\tau)$. From (5.12), (5.17) through (5.20), and (5.22) we find the result

$$\begin{aligned} L z^c(p_\tau) &= A_d C'' \hat{L} A_d^{-1} z^c(p_\tau) = A_d C'' \hat{L} z^0(p_\tau) \\ &= A_d C'' z^0(p_\tau) = A_d z^0(p_\tau) = \bar{z}^c(p_\tau) \end{aligned} \tag{36.5.23}$$

where $\bar{z}^c(p_\tau)$ is the vector

$$\bar{z}^c(p_\tau) = A_d \bar{z}^0(p_\tau). \tag{36.5.24}$$

From (19.*.*), (5.14), (5.19), and (5.24) we find that $\bar{z}^c(p_\tau)$ has the entries

$$\bar{z}^c(p_\tau) = (\Delta_1 p_\tau, \Delta_2 p_\tau, \Delta_3 p_\tau, \Delta_4 p_\tau, \eta p_\tau, p_\tau). \tag{36.5.25}$$

Comparison of (5.21) and (5.25) shows that $z^c(p_\tau)$ and $\bar{z}^c(p_\tau)$ *agree* in all entries except for the fifth. We conclude that the quantities $\Delta_a p_\tau$, for $a = 1$ to 4, are the transverse phase-space coordinates for the *closed* orbit (in the linear approximation to $\mathcal{M}$ described by $\mathcal{L}$). Thus, the quantities Δ_1 and Δ_3 are dispersions, and Δ_2 and Δ_4 are their momentum counterparts. And, as is evident from the fifth components of $z^c(p_\tau)$ and $\bar{z}^c(p_\tau)$, the quantity ηp_τ is the differential time-of-flight on the closed orbit. Correspondingly, η is the *phase slip* factor.

What can be said about g_1 and $\mathcal{A}_d$ if $\mathcal{L}$ is reversal symmetric? Rewrite (5.6) in the form

$$f_1 = \hat{\mathcal{L}} g_1 - g_1. \tag{36.5.26}$$

Applying reversal produces the equivalent relation [see (1.56)]

$$f_1^r = (\hat{\mathcal{L}}^r)^{-1} g_1^r - g_1^r, \tag{36.5.27}$$

and multiplying both sides of this relation by $\hat{\mathcal{L}}^r$ yields the result

$$\hat{\mathcal{L}}^r f_1^r = g_1^r - \hat{\mathcal{L}}^r g_1^r. \tag{36.5.28}$$

Now suppose that $\mathcal{L}$ is reversal symmetric so that (3.53) and (3.54) hold. Then (5.28) becomes

$$f_1 = g_1^r - \hat{\mathcal{L}} g_1^r, \tag{36.5.29}$$

and this result when combined with (5.26) yields the relation

$$(\hat{\mathcal{L}} - \mathcal{I})g_1 = -(\hat{\mathcal{L}} - \mathcal{I})g_1^r. \tag{36.5.30}$$

We have already assumed that $(\hat{\mathcal{L}} - \mathcal{I})$ is invertible, and therefore (5.30) implies the relation

$$g_1^r = -g_1. \tag{36.5.31}$$

This is the condition for $\mathcal{A}_d$ to be reversal antisymmetric. We have learned that if $\mathcal{L}$ is reversal symmetric, it can be written in the form (5.10) with $\mathcal{A}_d$ being reversal antisymmetric,

$$\mathcal{A}_d^r = \mathcal{A}_d^{-1}. \tag{36.5.32}$$

Moreover, comparison of (5.11) and (5.31) shows that there is the relation

$$\Delta_2 = \Delta_4 = 0 \tag{36.5.33}$$

if $\mathcal{L}$ is reversal symmetric. In view of (5.21), this relation can also be written in the form

$$Rz^c(p_\tau) = z^c(p_\tau). \tag{36.5.34}$$

If $\mathcal{L}$ is reversal symmetric the off-energy closed orbit has no transverse momentum components (at the ring location for which $\mathcal{L}$ is computed).

How do the quantities Δ_a and η vary from place to place around a ring? We will see shortly that η is constant. First let us consider the Δ_a. Suppose that in the linear approximation the one-turn map $\mathcal{L}$ for a ring can be written as the product of $2n$ maps in the form

$$\mathcal{L} = \mathcal{L}_1\mathcal{L}_2 \cdots \mathcal{L}_n\mathcal{L}_{\tilde{n}} \cdots \mathcal{L}_{\tilde{2}}\mathcal{L}_{\tilde{1}} \tag{36.5.35}$$

with

$$\mathcal{L}_{\tilde{j}} = \mathcal{L}_j^r. \tag{36.5.36}$$

Figure 5.1 illustrates such a ring. In view of (5.36), we may say that the ring has location 0 as a symmetry point. From (2.15) and (2.16) we see that $\mathcal{L}$ is reversal symmetric, i.e. (3.10) holds. Our task is to compute the closed-orbit quantities Δ_a at various other locations (Poincaré surfaces of section) j and $\tilde{j}$ around the ring.

Introduce the maps $\mathcal{S}^j$ and $\mathcal{S}^{\tilde{j}}$ defined by the rules

$$\mathcal{S}^j = \mathcal{L}_1\mathcal{L}_2 \cdots \mathcal{L}_j, \tag{36.5.37}$$

$$\mathcal{S}^{\tilde{j}} = \mathcal{L}_{\tilde{j}} \cdots \mathcal{L}_{\tilde{2}}\mathcal{L}_{\tilde{1}}, \tag{36.5.38}$$

and let S^j and $S^{\tilde{j}}$ be their corresponding matrices. Let z^{cj} be the closed-orbit vector at location j. In analogy to (5.21) we write it in the form

$$z^{cj}(p_\tau) = (\Delta_1^j p_\tau, \Delta_2^j p_\tau, \Delta_3^j p_\tau, \Delta_4^j p_\tau, *, p_\tau) \tag{36.5.39}$$

where the entry labeled $*$ need not concern us. Also, from (5.21) and (5.34), the closed-orbit vector at location 0, denote it by $z^{c0}(p_\tau)$, is given by the relation

$$z^{c0}(p_\tau) = z^c(p_\tau) = (\Delta_1^0 p_\tau, 0, \Delta_3^0 p_\tau, 0, 0, p_\tau) \tag{36.5.40}$$

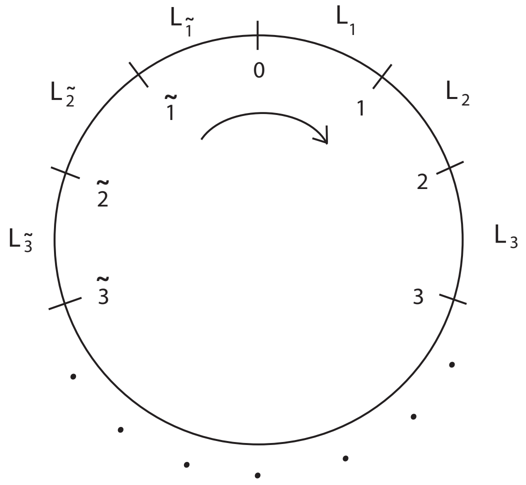


Figure 36.5.1: Schematic drawing of a ring showing the locations (Poincaré surfaces of section) $0; 1, \tilde{1}; 2, \tilde{2}; 3, \tilde{3}$; etc.

with $\Delta_1^0 = \Delta_1$ and $\Delta_3^0 = \Delta_3$. Then, inspection of Figure 5.1 shows that the closed-orbit vector at location j must be the result of propagating the closed-orbit initial conditions from location 0 to location j. That is, there is the vector and matrix relation

$$z^{cj}(p_\tau) = S^j z^{c0}. \tag{36.5.41}$$

Similarly, let $z^{c\tilde{j}}$ be the closed-orbit vector at location $\tilde{j}$, and write it in the form

$$z^{c\tilde{j}}(p_\tau) = (\Delta_1^{\tilde{j}} p_\tau, \Delta_2^{\tilde{j}} p_\tau, \Delta_3^{\tilde{j}} p_\tau, \Delta_4^{\tilde{j}} p_\tau, *, p_\tau). \tag{36.5.42}$$

Then, by following the closed orbit *backwards* from location 0 to location $\tilde{j}$, we see that there is the relation

$$z^{c\tilde{j}} = (S^{\tilde{j}})^{-1} z^{c0}. \tag{36.5.43}$$

We are now prepared to compare z^{cj} and $z^{c\tilde{j}}$. From (5.36) through (5.38) it is easily checked that there is the operator relation

$$(\mathcal{S}^j)^r = \mathcal{S}^{\tilde{j}} \tag{36.5.44}$$

and hence also the corresponding matrix relation

$$(S^j)^r = S^{\tilde{j}}. \tag{36.5.45}$$

It follows that (5.43) can be written in the form

$$\begin{aligned} z^{c\tilde{j}} &= ((S^j)^r)^{-1} z^{c0} = (R(S^j)^{-1}R)^{-1} z^{c0} \\ &= RS^jRz^{c0} = RS^jz^{c0} = Rz^{cj}. \end{aligned} \tag{36.5.46}$$

Here we have also used (1.8), (3.4), and (5.34). From (5.46) we conclude that there are the relations

$$\Delta_1^{\tilde{j}} = \Delta_1^j, \tag{36.5.47}$$

$$\Delta_2^{\tilde{j}} = -\Delta_2^j, \tag{36.5.48}$$

$$\Delta_3^{\tilde{j}} = \Delta_3^j, \tag{36.5.49}$$

$$\Delta_4^{\tilde{j}} = -\Delta_4^j. \tag{36.5.50}$$

We have learned that Δ_1 and Δ_3 are "even" functions of position about the symmetry location 0, and Δ_2 and Δ_4 are "odd" functions.

36.6 Consequences for Courant-Snyder Functions in a Circular Machine

We next turn to the task of determining the behavior of the Courant-Snyder lattice functions (as well as the phase slip η) for a ring with a reversal symmetric one-turn map, and for a ring that also has a symmetry point. Some preparatory work is required. The relation (5.10) can be rewritten in the form

$$\mathcal{L} = \mathcal{A}_d^{-1}[\mathcal{C}(\eta)\hat{\mathcal{L}}]\mathcal{A}_d. \tag{36.6.1}$$

In this form we see that a general $\mathcal{L}$ has been expressed as the similarity transform of the simpler map $\mathcal{C}(\eta)\hat{\mathcal{L}}$. Let us now further simplify $\hat{\mathcal{L}}$ itself. With reference to (3.39) and (3.40), suppose that the eigenvalues of $\check{L}$ lie on the unit circle and are distinct. Then, according to normal form theory, there is a symplectic matrix $\check{A}_b$ such that

$$\check{A}_b^{-1}\check{L}\check{A}_b = \check{N}_b \tag{36.6.2}$$

where $\check{N}_b$ takes the simple (*normal*) form

$$\check{N}_b(\phi_1,\phi_2) = \begin{pmatrix} \cos\phi_1 & \sin\phi_1 & 0 & 0 \\ -\sin\phi_1 & \cos\phi_1 & 0 & 0 \\ 0 & 0 & \cos\phi_2 & \sin\phi_2 \\ 0 & 0 & -\sin\phi_2 & \cos\phi_2 \end{pmatrix}. \tag{36.6.3}$$

Here the quantities ϕ are the (eigen) phase advances of $\check{L}$ and the $[\phi/(2\pi)]$ are the (eigen) tunes. The subscript b indicates the connection that $\check{A}_b$ and $\check{N}_b$ have with *betatron* oscillations.

In the spirit of (3.39), let $\hat{A}_b$ and $\hat{N}_b$ denote the matrices

$$\hat{A}_b = \begin{pmatrix} \check{A}_b & 0 \\ 0 & I \end{pmatrix}, \tag{36.6.4}$$

$$\hat{N}_b = \begin{pmatrix} \check{N}_b & 0 \\ 0 & I \end{pmatrix}, \tag{36.6.5}$$

and let $\mathcal{A}_b$ and $\mathcal{N}_b$ denote the corresponding symplectic maps. Then the matrix relation (6.2) is equivalent to the map relation

$$\mathcal{A}_b\hat{\mathcal{L}}\mathcal{A}_b^{-1} = \mathcal{N}_b. \tag{36.6.6}$$

We note that $\mathcal{N}_b$ can be written in the Lie form

$$\mathcal{N}_b = \exp(:h_2:) \tag{36.6.7}$$

with

$$h_2 = -(\phi_1/2)(p_x^2 + x^2) - (\phi_2/2)(p_y^2 + y^2). \tag{36.6.8}$$

The relation (6.6) can also be written in the form

$$\hat{\mathcal{L}} = \mathcal{A}_b^{-1}\mathcal{N}_b\mathcal{A}_b, \tag{36.6.9}$$

and then inserted into (6.1) to give the relation

$$\mathcal{L} = \mathcal{A}_d^{-1}\mathcal{C}(\eta)\mathcal{A}_b^{-1}\mathcal{N}_b\mathcal{A}_b\mathcal{A}_d. \tag{36.6.10}$$

From (19.*.*) and (6.4) it is evident that $\mathcal{C}(\eta)$ and $\mathcal{A}_b^{-1}$ commute. Therefore (6.10) can be cast in the still simpler form

$$\mathcal{L} = \mathcal{A}^{-1}\mathcal{N}\mathcal{A} \tag{36.6.11}$$

where

$$\mathcal{A} = \mathcal{A}_b\mathcal{A}_d, \tag{36.6.12}$$

$$\mathcal{N} = \mathcal{C}(\eta)\mathcal{N}_b. \tag{36.6.13}$$

Inspection of (19.*.*), (5.13), (6.7), and (6.8) shows that $\mathcal{N}$ can be written in the form

$$\mathcal{N} = \exp(:h_2:) \tag{36.6.14}$$

with

$$h_2 = -(\phi_1/2)(p_x^2 + x^2) - (\phi_2/2)(p_y^2 + y^2) - (\eta/2)p_\tau^2. \tag{36.6.15}$$

Let us again explore the effect of reversal. Applying reversal to (6.6) produces the equivalent result

$$(\mathcal{A}_b^{-1})^r\hat{\mathcal{L}}^r\mathcal{A}_b^r = \mathcal{N}_b^r, \tag{36.6.16}$$

which can be rewritten in the form

$$(\mathcal{A}_b^r)^{-1}\hat{\mathcal{L}}^r\mathcal{A}_b^r = \mathcal{N}_b. \tag{36.6.17}$$

Here we have used (2.2) and the fact that $\hat{\mathcal{N}_b}$ is manifestly reversal symmetric. Let us solve (6.16) and (6.17) for $\hat{\mathcal{L}}$ and $\hat{\mathcal{L}}^r$ to yield the relations

$$\hat{\mathcal{L}} = \mathcal{A}_b^{-1}\mathcal{N}_b\mathcal{A}_b, \tag{36.6.18}$$

$$\hat{\mathcal{L}}^r = \mathcal{A}_b^r\mathcal{N}_b(\mathcal{A}_b^r)^{-1}. \tag{36.6.19}$$

Now suppose that $\mathcal{L}$ is reversal symmetric, in which case $\hat{\mathcal{L}}$ is also reversal symmetric. See (3.53). Then comparison of (6.18) and (6.19) gives the relation

$$\mathcal{A}_b^{-1}\mathcal{N}_b\mathcal{A}_b = \mathcal{A}_b^r\mathcal{N}_b(\mathcal{A}_b^r)^{-1}, \tag{36.6.20}$$

which can be rewritten in the forms

$$(\mathcal{A}_b^r)^{-1}\mathcal{A}_b^{-1}\mathcal{N}_b\mathcal{A}_b\mathcal{A}_b^r = \mathcal{N}_b, \tag{36.6.21}$$

$$(\mathcal{A}_b\mathcal{A}_b^r)^{-1}\mathcal{N}_b(\mathcal{A}_b\mathcal{A}_b^r) = \mathcal{N}_b, \tag{36.6.22}$$

$$\mathcal{N}_b(\phi_1,\phi_2)(\mathcal{A}_b\mathcal{A}_b^r) = (\mathcal{A}_b\mathcal{A}_b^r)\mathcal{N}_b(\phi_1,\phi_2). \tag{36.6.23}$$

In the last form we have made explicit the dependence of $\mathcal{N}_b$ on ϕ_1 and ϕ_2.

The relation (6.23) states that the product $(\mathcal{A}_b\mathcal{A}_b^r)$ *commutes* with $\mathcal{N}_b$. Inspection of (6.7) and (6.8) shows that (6.23) can hold only if the product $(\mathcal{A}_b\mathcal{A}_b^r)$ is of the form

$$\mathcal{A}_b\mathcal{A}_b^r = \mathcal{N}_b(\phi_1',\phi_2') \tag{36.6.24}$$

for some values of ϕ_1' and ϕ_2'. We now show that, without loss of the desired relations (6.6) through (6.8), we can require the condition

$$\mathcal{A}_b\mathcal{A}_b^r = \mathcal{I} \text{ or } \mathcal{A}_b^r = \mathcal{A}_b^{-1}. \tag{36.6.25}$$

That is, if $\mathcal{L}$ is reversal symmetric, there is a reversal antisymmetric $\mathcal{A}_b$ that accomplishes the desired goals (6.6) through (6.8). Indeed, given an $\mathcal{A}_b$ that satisfies (6.24), define an associated map $\bar{\mathcal{A}}_b$ by the rule

$$\bar{\mathcal{A}}_b = \mathcal{N}_b(-\phi_1'/2,-\phi_2'/2)\mathcal{A}_b. \tag{36.6.26}$$

From (6.6) we see that it satisfies the desired normalizing relation,

$$\begin{aligned}\bar{\mathcal{A}}_b\hat{\mathcal{L}}\bar{\mathcal{A}}_b^{-1} &= \mathcal{N}_b(-\phi_1'/2,-\phi_2'/2)\mathcal{A}_b\hat{\mathcal{L}}\mathcal{A}_b^{-1}[\mathcal{N}_b(-\phi_1'/2,-\phi_2'/2)]^{-1}\\ &= \mathcal{N}_b(-\phi_1'/2,-\phi_2'/2)\mathcal{N}_b(\phi_1,\phi_2)[\mathcal{N}_b(-\phi_1'/2,-\phi_2'/2)]^{-1}\\ &= \mathcal{N}_b(\phi_1,\phi_2).\end{aligned} \tag{36.6.27}$$

Also, it has the desired product relation,

$$\begin{aligned}\bar{\mathcal{A}}_b\bar{\mathcal{A}}_b^r &= \mathcal{N}_b(-\phi_1'/2,-\phi_2'/2)\mathcal{A}_b\mathcal{A}_b^r\mathcal{N}_b(-\phi_1'/2,-\phi_2'/2)\\ &= \mathcal{N}_b(-\phi_1'/2,-\phi_2'/2)\mathcal{N}_b(\phi_1',\phi_2')\mathcal{N}_b(-\phi_1'/2,-\phi_2'/2)\\ &= \mathcal{I}.\end{aligned} \tag{36.6.28}$$

The generalized Courant-Snyder quadratic invariants I_x and I_y (which include the possibility of coupling between the x, p_x and y, p_y degrees of freedom) are defined by the rules

$$I_x = \mathcal{A}_b^{-1}(p_x^2 + x^2), \tag{36.6.29}$$

$$I_y = \mathcal{A}_b^{-1}(p_y^2 + y^2). \tag{36.6.30}$$

In the no-coupling case they take the form

$$I_x = \beta_x p_x^2 + 2\alpha_x x p_x + \gamma_x x^2, \text{ etc.} \tag{36.6.31}$$

The relations (6.29) and (6.30) have the reversed counterparts

$$I_x^r = \mathcal{A}_b^r(p_x^2 + x^2), \text{ etc.} \tag{36.6.32}$$

Here we have used (1.56). Now assume that $\mathcal{L}$ is reversal symmetric so that $\mathcal{A}_b$ can be taken to satisfy (6.25). In this case we find the relations

$$I_x^r = \mathcal{A}_b^r(p_x^2 + x^2) = \mathcal{A}_b^{-1}(p_x^2 + x^2) = I_x, \text{ etc.} \tag{36.6.33}$$

Thus, if $\mathcal{L}$ is reversal symmetric, the generalized Courant-Snyder invariants (at the location for which $\mathcal{L}$ is computed) are free of terms that are *odd* in the momenta.

We now explore how I_x, I_y, and η vary about a ring that has a symmetry point. Refer again to Figure 5.1. Let $\mathcal{L}^0$ be the one-turn map starting at location 0,

$$\mathcal{L}^0 = \mathcal{L}_1\mathcal{L}_2\cdots\mathcal{L}_n\mathcal{L}_{\tilde{n}}\cdots\mathcal{L}_{\tilde{2}}\mathcal{L}_{\tilde{1}} = \mathcal{L}. \tag{36.6.34}$$

Similarly, let $\mathcal{L}^1$ and $\mathcal{L}^{\tilde{1}}$ be the maps starting at the locations 1 and $\tilde{1}$. For these maps we find the results

$$\begin{aligned} \mathcal{L}^1 &= \mathcal{L}_2\mathcal{L}_3\cdots\mathcal{L}_n\mathcal{L}_{\tilde{n}}\cdots\mathcal{L}_{\tilde{2}}\mathcal{L}_{\tilde{1}}\mathcal{L}_1 \\ &= \mathcal{L}_1^{-1}\mathcal{L}_1\mathcal{L}_2\mathcal{L}_3\cdots\mathcal{L}_n\mathcal{L}_{\tilde{n}}\cdots\mathcal{L}_{\tilde{2}}\mathcal{L}_{\tilde{1}}\mathcal{L}_1 \\ &= \mathcal{L}_1^{-1}\mathcal{L}\mathcal{L}_1, \end{aligned} \tag{36.6.35}$$

$$\begin{aligned} \mathcal{L}^{\tilde{1}} &= \mathcal{L}_{\tilde{1}}\mathcal{L}_1\mathcal{L}_2\cdots\mathcal{L}_n\mathcal{L}_{\tilde{n}}\cdots\mathcal{L}_{\tilde{3}}\mathcal{L}_{\tilde{2}} \\ &= \mathcal{L}_{\tilde{1}}\mathcal{L}_1\mathcal{L}_2\cdots\mathcal{L}_n\mathcal{L}_{\tilde{n}}\cdots\mathcal{L}_{\tilde{1}}\mathcal{L}_{\tilde{2}}\mathcal{L}_{\tilde{1}}\mathcal{L}_{\tilde{1}}^{-1} \\ &= \mathcal{L}_{\tilde{1}}\mathcal{L}\mathcal{L}_{\tilde{1}}^{-1}. \end{aligned} \tag{36.6.36}$$

In the same way we find for $\mathcal{L}^2$ and $\mathcal{L}^{\tilde{2}}$ the results

$$\mathcal{L}^2 = \mathcal{L}_2^{-1}\mathcal{L}_1^{-1}\mathcal{L}\mathcal{L}_1\mathcal{L}_2 = (\mathcal{L}_1\mathcal{L}_2)^{-1}\mathcal{L}(\mathcal{L}_1\mathcal{L}_2), \tag{36.6.37}$$

$$\mathcal{L}^{\tilde{2}} = \mathcal{L}_{\tilde{2}}\mathcal{L}_{\tilde{1}}\mathcal{L}\mathcal{L}_{\tilde{1}}^{-1}\mathcal{L}_{\tilde{2}}^{-1} = (\mathcal{L}_{\tilde{2}}\mathcal{L}_{\tilde{1}})\mathcal{L}(\mathcal{L}_{\tilde{2}}\mathcal{L}_{\tilde{1}})^{-1}. \tag{36.6.38}$$

Finally, in terms of the maps $\mathcal{S}^j$ and $\mathcal{S}^{\tilde{j}}$ defined by (5.37) and (5.38), the relations (6.34) through (6.38) etc. take the general form

$$\mathcal{L}^j = (\mathcal{S}^j)^{-1}\mathcal{L}\mathcal{S}^j, \tag{36.6.39}$$

$$\mathcal{L}^{\tilde{j}} = (\mathcal{S}^{\tilde{j}})\mathcal{L}(\mathcal{S}^{\tilde{j}})^{-1}. \tag{36.6.40}$$

Insert the representation (6.11) into (6.39) and (6.40) to obtain the relations

$$\mathcal{L}^j = (\mathcal{S}^j)^{-1}\mathcal{A}^{-1}\mathcal{N}\mathcal{A}(\mathcal{S}^j), \tag{36.6.41}$$

$$\mathcal{L}^{\tilde{j}} = (\mathcal{S}^{\tilde{j}})\mathcal{A}^{-1}\mathcal{N}\mathcal{A}(\mathcal{S}^{\tilde{j}})^{-1}. \tag{36.6.42}$$

These relations can be rewritten in the form

$$\mathcal{L}^j = (\mathcal{A}^j)^{-1}\mathcal{N}^j\mathcal{A}^j, \tag{36.6.43}$$

$$\mathcal{L}^{\tilde{j}} = (\mathcal{A}^{\tilde{j}})^{-1}\mathcal{N}^{\tilde{j}}\mathcal{A}^{\tilde{j}}, \tag{36.6.44}$$

where

$$\mathcal{A}^{j} = \mathcal{A}\mathcal{S}^{j}, \tag{36.6.45}$$

$$\mathcal{N}^{j} = \mathcal{N}, \tag{36.6.46}$$

$$\mathcal{A}^{\tilde{j}} = \mathcal{A}(\mathcal{S}^{\tilde{j}})^{-1}, \tag{36.6.47}$$

$$\mathcal{N}^{\tilde{j}} = \mathcal{N}. \tag{36.6.48}$$

Comparison of (6.46) and (6.48) immediately gives the result

$$\mathcal{N}^{\tilde{j}} = \mathcal{N}^{j}, \tag{36.6.49}$$

from which we conclude that ϕ_1, ϕ_2 and η are *global* properties of a ring. [See (6.15).] Their values are independent of the choice of the surface of section. Note that in arriving at this conclusion no assumptions were required about reversal symmetry.

Next, in analogy to (6.12), make the factorizations

$$\mathcal{A}^{j} = \mathcal{A}_b^{j}\mathcal{A}_d^{j}, \tag{36.6.50}$$

$$\mathcal{A}^{\tilde{j}} = \mathcal{A}_b^{\tilde{j}}\mathcal{A}_d^{\tilde{j}}. \tag{36.6.51}$$

Here the maps $\mathcal{A}_d^{j}$ and $\mathcal{A}_d^{\tilde{j}}$ are defined by the analogs of (5.1) and (5.11):

$$\mathcal{A}_d^{j} = \exp(:p_\tau g_1^{j}:) \tag{36.6.52}$$

with

$$g_1^{j} = \Delta_2^{j}x - \Delta_1^{j}p_x + \Delta_4^{j}y - \Delta_3^{j}p_y; \tag{36.6.53}$$

$$\mathcal{A}_d^{\tilde{j}} = \exp(:p_\tau g_1^{\tilde{j}}:) \tag{36.6.54}$$

with

$$g^{\tilde{j}} = \Delta_2^{\tilde{j}}x - \Delta_1^{\tilde{j}}p_x + \Delta_4^{\tilde{j}}y - \Delta_3^{\tilde{j}}p_y. \tag{36.6.55}$$

So far we have not necessarily assumed that $\mathcal{L}$ is reversal symmetric and that the ring has 0 as a symmetry point. Do so now. Then (5.47) through (5.50) can be used to rewrite (6.53) in the form

$$g_1^{j} = -\Delta_2^{\tilde{j}}x - \Delta_1^{\tilde{j}}p_x - \Delta_4^{\tilde{j}}y - \Delta_3^{\tilde{j}}p_y. \tag{36.6.56}$$

Now we see that there is the relation

$$(g_1^{j})^r = -g_1^{\tilde{j}}, \tag{36.6.57}$$

from which it follows that

$$(\mathcal{A}_d^{j})^r = (\mathcal{A}_d^{\tilde{j}})^{-1}. \tag{36.6.58}$$

We also know that, since $\mathcal{L}$ is reversal symmetric, the maps $\mathcal{A}_d$ and $\mathcal{A}_b$ are reversal antisymmetric. See (5.32) and (6.25). It follows from the group property for reversal antisymmetric maps that $\mathcal{A}$ as given by (6.12) is also reversal antisymmetric,

$$\mathcal{A}^r = \mathcal{A}^{-1}. \tag{36.6.59}$$

We are now prepared to study the relation between $\mathcal{A}_b^j$ and $\mathcal{A}_b^{\tilde{j}}$. Solving (6.50) and (6.51) for $\mathcal{A}_b^j$ and $\mathcal{A}_b^{\tilde{j}}$ and using (6.45) and (6.47) give the results

$$\mathcal{A}_b^j = \mathcal{A}^j(\mathcal{A}_d^j)^{-1} = \mathcal{A}\mathcal{S}^j(\mathcal{A}_d^j)^{-1}, \tag{36.6.60}$$

$$\mathcal{A}_b^{\tilde{j}} = \mathcal{A}^{\tilde{j}}(\mathcal{A}_d^{\tilde{j}})^{-1} = \mathcal{A}(\mathcal{S}^{\tilde{j}})^{-1}(\mathcal{A}_d^{\tilde{j}})^{-1}. \tag{36.6.61}$$

Let us compute the map $((\mathcal{A}_b^j)^r)^{-1}$. From (6.60) we find the results

$$(\mathcal{A}_b^j)^r = ((\mathcal{A}_d^j)^{-1})^r(\mathcal{S}^j)^r\mathcal{A}^r, \tag{36.6.62}$$

$$\begin{aligned} ((\mathcal{A}_b^j)^r)^{-1} &= (\mathcal{A}^r)^{-1}((\mathcal{S}^j)^r)^{-1}(\mathcal{A}_d^j)^r \\ &= \mathcal{A}(\mathcal{S}^{\tilde{j}})^{-1}(\mathcal{A}_d^{\tilde{j}})^{-1} \\ &= \mathcal{A}_b^{\tilde{j}}. \end{aligned} \tag{36.6.63}$$

Here we have used (5.44), (6.58), and (6.59). Note that (6.63) can also be written in the form

$$(\mathcal{A}_b^j)^r = (\mathcal{A}_b^{\tilde{j}})^{-1}, \tag{36.6.64}$$

which is analogous to the relation (6.58) for $(\mathcal{A}_d^j)^r$.

We are ready for the coup de maître. The horizontal Courant-Snyder (eigen) invariants at the locations j and $\tilde{j}$ are given by the expressions

$$I_x^j = (\mathcal{A}_b^j)^{-1}(p_x^2 + x^2), \tag{36.6.65}$$

$$I_x^{\tilde{j}} = (\mathcal{A}_b^{\tilde{j}})^{-1}(p_x^2 + x^2); \tag{36.6.66}$$

and there are analogous expressions for their vertical counterparts. Now apply the reversal operation to (6.65). Doing so gives the result

$$(I_x^j)^r = (\mathcal{A}_b^j)^r(p_x^2 + x^2) = (\mathcal{A}_b^{\tilde{j}})^{-1}(p_x^2 + x^2) = I_x^{\tilde{j}}. \tag{36.6.67}$$

Similarly, there is the relation

$$(I_y^j)^r = I_y^{\tilde{j}}. \tag{36.6.68}$$

Consider the coefficients of the monomials in I_x^j or I_y^j that are *even* under $\mathcal{R}$. The relations (6.67) and (6.68) show that these coefficients are also *even* functions of position about the symmetry location 0. For example, in the no-coupling case (6.31), there are the relations

$$\beta_x^{\tilde{j}} = \beta_x^j, \tag{36.6.69}$$

$$\gamma_x^{\tilde{j}} = \gamma_x^j. \tag{36.6.70}$$

Next consider the coefficients of the monomials in I_x^j or I_y^j that are *odd* under $\mathcal{R}$. The relations (6.67) and (6.68) show that these coefficients are also *odd* functions of position about the symmetry location 0. For example, in the non-coupling case (6.31) there is the relation

$$\alpha_x^{\tilde{j}} = -\alpha_x^j. \tag{36.6.71}$$

36.7 Some Nonlinear Consequences

So far we have mostly explored the consequences of reversal symmetry for $\mathcal{L}$, the linear part of $\mathcal{M}$. We now explore some of the consequences of reversal symmetry for the full map $\mathcal{M}$. One line of inquiry would be to generalize the results of the previous Sections 5 and 6 to include nonlinear terms. Equations (5.25) and (5.39) give the linear terms of a power series (in p_τ) expansion of $z^{cj}(p_\tau)$. The higher order terms in this expansion could be found and their dependence on the location j could be explored. Similarly a full normalization of $\mathcal{M}$ [analogous to that given in (6.11) for its linear part $\mathcal{L}$] could be found. The *betatron* part of the associated normalizing map could then be used in (6.65), etc., to find the generalized Courant-Snyder invariants that take into account all nonlinear effects through any desired order. These invariants contain, in addition to quadratic monomials, monomials of degree 3 and higher. When considered as functions of location j, the coefficients of these monomials yield *nonlinear* lattice functions. Just how these coefficients depend on j could also be explored.

We will leave these generalizations to the reader. Instead, we will devote this section to the exploration of how reversal symmetry affects the dynamic aperture of a ring including the location of fixed points, and how it limits the kind of nonlinearities that can occur in $\mathcal{M}$.

Consider the map $\mathcal{M}$ on two-dimensional phase space given by the product

$$\mathcal{M}(\theta) = \exp[-(\phi/4) :p^2 + q^2:] \exp(:q^3:) \exp[-(\phi/4) :p^2 + q^2 :]. \tag{36.7.1}$$

As described in Section 1.2.3, this map consists of a $\phi/2$ phase advance, followed by a sextupole kick, followed again by a $\phi/2$ phase advance. It may be viewed as describing horizontal betatron motion in an idealized storage ring with a single thin *sextupole* insertion S, and an *observation* point O (Poincaré surface of section) located diametrically across the ring from the sextupole insertion. Recall Figure 1.2.8. We verified in Section 18.8.4 that this map is a variant of the usual Hénon map, and differs from it only by a linear change of variables. Now we note that $\mathcal{M}$ as given by (7.1) is reversal symmetric.

Figure 7.1, which is a replication of Figure 1.2.9, shows the dynamic aperture for our variant of the Hénon map for the case $\phi/(2\pi) = .22$. Points in the black area of the q, p (mapping) plane remain there under repeated application of the map. [Actually, the points shown remain there for at least 10,000 iterations ($\mathcal{M}^n$ with $n \leq 10,000$).] By contrast, any point launched in the white area eventually iterates away to infinity. Inspection of the figure suggests symmetry about the q axis. That is, all features of the figure are invariant under reversal.

By construction, the map sends the origin into itself (the origin is a fixed point of $\mathcal{M}$), and the origin is unchanged under reversal. Not shown, because it is unstable and also outside the viewing window, there is also a fixed point of $\mathcal{M}$ at $p = 0$ and $q = .7158$, and this point is also unchanged under reversal. Next observe that there are 5 islands. They surround 5 fixed points of the map $\mathcal{M}^5$. [Note that the tune $\phi/(2\pi) = .22$ is close to 1/5.] These fixed points appear to be located symmetrically about the q axis. Also, each island is surrounded by 6 smaller islands. These smaller islands surround fixed points of $\mathcal{M}^{30}$, and these fixed points appear to be located symmetrically about the q axis. Finally, the whole

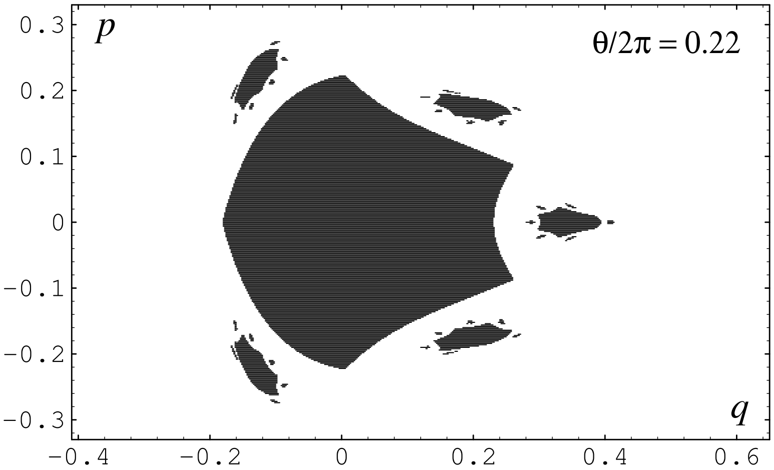


Figure 36.7.1: The dynamic aperture of the Hénon map for the case $\phi/(2\pi) = 0.22$.

dynamic aperture for our variant of the Hénon map appears to be symmetrical about the q axis. That is, the dynamic aperture appears to be invariant under reversal.

To explore these conjectures, let us first think about fixed points of $\mathcal{M}$ and its powers. Suppose $\mathcal{M}$ is a general map in any number of phase-space dimensions, and suppose z^f is a *fixed* point,

$$\mathcal{M}z^f = z^f. \tag{36.7.2}$$

Suppose also that $\mathcal{M}$ is reversal symmetric so that (2.5) holds. Applying $\mathcal{RMR}$ to both sides of (7.2) yields the result

$$\mathcal{RMRM}z^f = \mathcal{RMR}z^f. \tag{36.7.3}$$

Use of (2.5) in (7.3) gives the relation

$$z^f = \mathcal{RMR}z^f, \tag{36.7.4}$$

which, in view of (1.21), can be rewritten as

$$\mathcal{M}(\mathcal{R}z^f) = (\mathcal{R}z^f). \tag{36.7.5}$$

We see that if z^f is a fixed point of $\mathcal{M}$, so is the point $\mathcal{R}z^f$. An analogous result holds for powers of $\mathcal{M}$ because $\mathcal{M}^n$ will be reversal symmetric when $\mathcal{M}$ is reversal symmetric: If z^f is a fixed point of $\mathcal{M}^n$, so is the point $\mathcal{R}z^f$.

We define a fixed point z^f to be *symmetric* if it satisfies the relation

$$\mathcal{R}z^f = z^f. \tag{36.7.6}$$

For example, for the Hénon map of Figure 7.1, the origin, and the point $(q,p) = (.7158, 0)$ not shown, are symmetric fixed points of $\mathcal{M}$; and $(q,p) = (.3458, 0)$ is a symmetric fixed point of $\mathcal{M}^5$. From appearances there are two symmetric fixed points of $\mathcal{M}^{30}$: those on the q axis and surrounding the fixed point of $\mathcal{M}^5$ also on the q axis. The discovery (or ruling out the existence) of symmetric fixed points is easier than the discovery of general fixed points. Let Fix $(\mathcal{R})$ be the set of points z satisfying

$$\mathcal{R}z = z. \tag{36.7.7}$$

From the definition (1.2) and (1.3) of $\mathcal{R}$ it is evident that (for a 6-dimensional phase space) the set Fix $(\mathcal{R})$ is 3 dimensional. [In general, if the dimension of phase space is $2m$, the dimension of Fix $(\mathcal{R})$ is m.] Consequently, the dimension of the space to be searched to find a symmetric fixed point is only half as large as the dimension of the full phase space that must be searched to find a general fixed point.

We next remark that we have found that $\mathcal{R}$ and $\mathcal{R}\mathcal{M}^k$ and $\mathcal{M}^k\mathcal{R}$ (for any k) are all involutions. See (2.8) and (2.9). Moreover, they are all antisymplectic. We have also seen that Fix $(\mathcal{R})$ is m dimensional. Based on a theorem of *Bochner* and *Montgomery*, it can be shown that the sets Fix $(\mathcal{R}\mathcal{M}^k)$ and Fix $(\mathcal{M}^k\mathcal{R})$ are also m dimensional if phase space is $2m$ dimensional. That is, the set of points obeying

$$\mathcal{R}\mathcal{M}^k z = z \tag{36.7.8}$$

is m dimensional for any value of k, and so is the set of points obeying

$$\mathcal{M}^k\mathcal{R}z = z. \tag{36.7.9}$$

See References 19, 22, and 23 in the Bibliography at the end this chapter.

Suppose that z^f is the fixed point of some power $\mathcal{M}$, say $\mathcal{M}^n$, but not of some lower power. Suppose that z^f is also symmetric. Then more can be said. For example, suppose $n = 2$. Then we have the relations

$$\mathcal{M}^2 z^f = z^f, \tag{36.7.10}$$

$$\mathcal{M}^2\mathcal{M}z^f = \mathcal{M}\mathcal{M}^2 z^f = \mathcal{M}z^f. \tag{36.7.11}$$

Thus z^f and $\mathcal{M}z^f$ are *two distinct* fixed points of $\mathcal{M}^2$. By assumption z^f is symmetric. What about $\mathcal{M}z^f$? We have the relations

$$\mathcal{R}\mathcal{M}z^f = \mathcal{R}\mathcal{M}\mathcal{R}\mathcal{R}z^f = \mathcal{M}^{-1}z^f = \mathcal{M}^{-1}\mathcal{M}^2 z^f = \mathcal{M}z^f. \tag{36.7.12}$$

Here we have used (1.21), (1.40), (7.6), and (7.10). We see that $\mathcal{M}z^f$ is also a symmetric fixed point. It is easy to show that this result can be generalized to the case of any even n: If z^f is a symmetric fixed point of $\mathcal{M}^n$, then $\mathcal{M}^{n/2}z^f$ is also a symmetric fixed point of $\mathcal{M}^n$. Thus, symmetric fixed points of $\mathcal{M}^n$ occur in pairs when n is even. As an example, we have now proved that the two fixed points of $\mathcal{M}^{30}$ that appear to lie on the q axis in Figure 5.1 are indeed symmetric.

The case of odd n is a bit more complicated. Suppose for example that $n = 3$. Then, if z^f is a fixed point of $\mathcal{M}^3$, so are the two other points $\mathcal{M}z^f$ and $\mathcal{M}^2 z^f$. By assumption z^f is symmetric. For $\mathcal{M}z^f$ and $\mathcal{M}^2 z^f$ we find the following results:

$$\mathcal{R}\mathcal{M}z^f = \mathcal{R}\mathcal{M}\mathcal{R}\mathcal{R}z^f = \mathcal{M}^{-1}z^f \tag{36.7.13}$$

from which it follows that

$$(\mathcal{M}^2\mathcal{R})(\mathcal{M}z^f) = (\mathcal{M}z^f); \tag{36.7.14}$$

$$\mathcal{R}\mathcal{M}^2 z^f = \mathcal{R}\mathcal{M}^2\mathcal{R}\mathcal{R}z^f = \mathcal{M}^{-2}z^f \tag{36.7.15}$$

from which it follows that

$$(\mathcal{M}\mathcal{R})(\mathcal{M}^2 z^f) = \mathcal{M}^{-1}z^f = \mathcal{M}^{-1}\mathcal{M}^3 z^f = (\mathcal{M}^2 z^f). \tag{36.7.16}$$

We know that $\mathcal{R}$ and $\mathcal{M}\mathcal{R}$ and $\mathcal{M}^2\mathcal{R}$ are involutions. Thus, the relations (7.6), (7.14), and (7.16) are all analogous: Some particular fixed point is invariant under some particular involution.

Finally, suppose the set of points that satisfies (7.9) for some $k = k_1$ *intersects* the set of points that satisfies (7.9) for some other $k = k_2$. That is, suppose there is a common point w such that

$$\mathcal{M}^{k_1}\mathcal{R}w = w, \tag{36.7.17}$$

$$\mathcal{M}^{k_2}\mathcal{R}w = w. \tag{36.7.18}$$

Then we have the relation

$$\mathcal{M}^{k_2}w = \mathcal{M}^{k_2}\mathcal{M}^{k_1}\mathcal{R}w = \mathcal{M}^{k_1}\mathcal{M}^{k_2}\mathcal{R}w = \mathcal{M}^{k_1}w, \tag{36.7.19}$$

and therefore

$$\mathcal{M}^{k_1-k_2}w = w. \tag{36.7.20}$$

We have found a fixed point, namely w, of the map $\mathcal{M}^{k_1-k_2}$! Note that this fixed point need not be symmetric. For some problems it is possible to construct the m-dimensional manifolds Fix $(\mathcal{M}^k\mathcal{R})$ that satisfy (7.9) for various values of k, and then determine their intersections to find fixed points of $\mathcal{M}^n$. See References 12 and 19 in the Bibliography at the end of this chapter.

So far we have studied fixed points. Next consider general points. What can be said about symmetry for them? We can think about this question for a general map $\mathcal{M}$ in any number of phase-space dimensions as follows: Suppose $\mathcal{M}$ sends the origin into itself. Let N be some large integer, and let Σ be some set in phase space such that all the points $\mathcal{M}^n z$ with $z \in \Sigma$ and $n \in [1, 2N]$ have some desirable property such as being near the origin. Let Γ be the set $\mathcal{M}^N\Sigma$. It is the set of all points obtained by letting $\mathcal{M}^N$ act on all points in Σ. Evidently Γ is a set such that all the points $\mathcal{M}^n z$ with $z \in \Gamma$ and $n \in [-N, N]$ have the same desirable property. Now suppose that $\mathcal{M}$ is reversal symmetric. Then Γ is also reversal symmetric,

$$\Gamma^r = \Gamma. \tag{36.7.21}$$

That is, if the phase-space point z is in Γ, so is the point z^r.

To see the truth of this assertion, suppose $z \in \Gamma$. Consider the set of points $\mathcal{M}^n z^r$ for $n \in [1, N]$. From (1.21) and (1.39) we have the result

$$\begin{aligned}
\mathcal{M}^n z^r &= \mathcal{M}^n\mathcal{R}z = \mathcal{R}\mathcal{R}\mathcal{M}^n\mathcal{R}z \\
&= \mathcal{R}(\mathcal{R}\mathcal{M}\mathcal{R})^n z = \mathcal{R}(\mathcal{M}^r)^{-n}z \\
&= \mathcal{R}\mathcal{M}^{-n}z.
\end{aligned} \tag{36.7.22}$$

Here, in the last step, we have used the assumption that $\mathcal{M}$ is reversal symmetric. Now we know that the sequence of points $\mathcal{M}^{-n}z$ is well behaved since $z \in \Gamma$. Therefore the sequence of points $\mathcal{R}\mathcal{M}^{-n}z$ is well behaved. (Note that, for any point $\bar{z}$, the points $\bar{z}$ and $\bar{z}^r$ are equidistant from the origin.) It follows from (7.22) that the sequence of points $\mathcal{M}^n z^r$ is well behaved.

Next consider the set of points $\mathcal{M}^{-n}z^r$. In this case we have the result

$$\begin{aligned} \mathcal{M}^{-n}z^r &= \mathcal{M}^{-n}\mathcal{R}z = \mathcal{R}\mathcal{R}\mathcal{M}^{-n}\mathcal{R}z \\ &= \mathcal{R}(\mathcal{R}\mathcal{M}^{-1}\mathcal{R})^n z = \mathcal{R}(\mathcal{M}^r)^n z \\ &= \mathcal{R}\mathcal{M}^n z. \end{aligned} \tag{36.7.23}$$

By hypothesis the sequence of points $\mathcal{M}^n z$ is well behaved, and therefore by (7.23) the sequence $\mathcal{M}^{-n}z^r$ is well behaved.

We have learned that the points $\mathcal{M}^n z^r$ are well behaved for $n \in [-N, N]$. It follows that $z^r \in \Gamma$, and hence (7.21) is correct.

We end this section with an exploration of what restrictions reversal invariance places on the nonlinear part of $\mathcal{M}$. Technically, we have already found these restrictions. They are given by the relations (3.17) through (3.20). What we want to do here is to explore these restrictions in more detail using some of the results of previous sections. For brevity we will treat only the case of static maps, which is actually somewhat more complicated than the dynamic case.

Rather than working with $\mathcal{M}$, it is convenient to work with the related map $\mathcal{M}'$ defined by the relation

$$\mathcal{M}' = \mathcal{A}\mathcal{M}\mathcal{A}^{-1} \tag{36.7.24}$$

with $\mathcal{A}$ given by (6.12). From the representation (3.1) and (6.11) it follows that $\mathcal{M}'$ has the representation

$$\begin{aligned} \mathcal{M}' &= \mathcal{A}\mathcal{L}\exp(:f_3:)\exp(:f_4:)\cdots\mathcal{A}^{-1} \\ &= [\mathcal{A}\mathcal{L}\mathcal{A}^{-1}][\mathcal{A}\exp(:f_3:)\exp(:f_4:)\cdots\mathcal{A}^{-1}] \\ &= \mathcal{N}\exp(:g_3:)\exp(:g_4:)\cdots \end{aligned} \tag{36.7.25}$$

where the g_m are given by the relation

$$g_m = \mathcal{A}f_m. \tag{36.7.26}$$

Next use the Baker-Campbell-Hausdorff series to combine the exponents of the nonlinear terms on the right side of (7.25),

$$\exp(:g_3:)\exp(:g_4:)\exp(:g_5:)\cdots = \exp(:h:) \tag{36.7.27}$$

where

$$h = h_3 + h_4 + h_5 + h_6 + \cdots . \tag{36.7.28}$$

with

$$h_3 = g_3, \tag{36.7.29}$$

$$h_4 = g_4, \tag{36.7.30}$$

$$h_5 = g_5 + (1/2)[g_3, g_4], \tag{36.7.31}$$

$$h_6 = g_6 + (1/2)[g_3, g_5] + (1/12)[g_3, [g_3, g_4]], \text{ etc.} \tag{36.7.32}$$

The net result is that $\mathcal{M}'$ can be written in the form form

$$\mathcal{M}' = \mathcal{N} \exp(: h :). \tag{36.7.33}$$

If $\mathcal{M}$ is reversal symmetric, $\mathcal{M}'$ will also be reversal symmetric,

$$(\mathcal{M}')^r = (\mathcal{A}^{-1})^r \mathcal{M}^r \mathcal{A}^r = \mathcal{A}\mathcal{M}\mathcal{A}^{-1} = \mathcal{M}'. \tag{36.7.34}$$

Here we have used (6.59), which holds if $\mathcal{M}$ is reversal symmetric. Now employ the representation (7.33) in (7.34) to obtain the condition

$$\exp(: h^r :)\mathcal{N} = \mathcal{N} \exp(: h :) \tag{36.7.35}$$

or

$$\exp(: h^r :) = \mathcal{N} \exp(: h :)\mathcal{N}^{-1} \tag{36.7.36}$$

from which it follows that

$$h^r = \mathcal{N} h. \tag{36.7.37}$$

Here we have used (8.2.27) and the fact that $\mathcal{N}$ is reversal symmetric as is evident from (6.14) and (6.15).

To explore the implications of (7.37) it is convenient to expand h in a *static resonance* basis. Introduce polynomials $R_{abcde}(z)$ and $I_{abcde}(z)$ defined by the relations

$$R_{abcde}(z) = \text{ Re } [(x + ip_x)^a (x - ip_x)^b (y + ip_y)^c (y - ip_y)^d p_\tau^e], \tag{36.7.38}$$

$$I_{abcde}(z) = \text{ Im } [(x + ip_x)^a (x - ip_x)^b (y + ip_y)^c (y - ip_y)^d p_\tau^e], \tag{36.7.39}$$

Here the quantities (exponents) a through e are integers. The first few such polynomials of interest are given by the relations

$$R_{11001} = (x^2 + p_x^2)p_\tau, \tag{36.7.40}$$

$$R_{00111} = (y^2 + p_y^2)p_\tau, \tag{36.7.41}$$

$$I_{11001} = I_{00111} = 0, \tag{36.7.42}$$

$$R_{30000} = x^3 - 3xp_x^2, \tag{36.7.43}$$

$$I_{30000} = 3x^2 p_x - p_x^3, \tag{36.7.44}$$

$$R_{20100} = x^2 y - p_x^2 y - 2xp_x p_y, \tag{36.7.45}$$

$$I_{20100} = x^2 p_y - p_x^2 p_y + 2xp_y y. \tag{36.7.46}$$

It is evident from (7.38) and (7.39) that the R_{abcde} and I_{abcde} form a basis for all static polynomials; and it is also evident that they have the simple reversal properties

$$R_{abcde}^r = R_{abcde}, \tag{36.7.47}$$

$$I_{abcde}^r = -I_{abcde}. \tag{36.7.48}$$

They also have simple properties under the action of $\mathcal{N}$, which will allow us to exploit (7.37). From (6.15) we have the results

$$\mathcal{N}(x \pm ip_x) = [\exp(\pm i\phi_1)](x \pm ip_x), \tag{36.7.49}$$

$$\mathcal{N}(y \pm ip_y) = [\exp(\pm i\phi_2)](y \pm ip_y), \tag{36.7.50}$$

$$\mathcal{N}p_\tau = p_\tau. \tag{36.7.51}$$

Introduce the *resonance* phase advances ψ_{abcd} defined by the rules

$$\psi_{abcd} = (a-b)\phi_1 + (c-d)\phi_2. \tag{36.7.52}$$

It follows from (8.3.52) and (7.49) through (7.52) that there are the relations

$$\mathcal{N}R_{abcde} = (\cos\psi_{abcd})R_{abcde} - (\sin\psi_{abcd})I_{abcde}, \tag{36.7.53}$$

$$\mathcal{N}I_{abcde} = (\sin\psi_{abcd})R_{abcde} + (\cos\psi_{abcd})I_{abcde}. \tag{36.7.54}$$

We are now ready to work out the implications of (7.37). Expand h in terms of the static resonance basis by writing

$$h = \sum_{abcde} A_{abcde}R_{abcde} + B_{abcde}I_{abcde}. \tag{36.7.55}$$

Then, by (7.47) and (7.48), we have the relation

$$h^r = \sum_{abcde} A_{abcde}R_{abcde} - B_{abcde}I_{abcde}; \tag{36.7.56}$$

and by (7.53) and (7.54) we have the relation

$$\begin{aligned}\mathcal{N}h &= \sum_{abcde}[A_{abcde}\cos\psi_{abcd} + B_{abcde}\sin\psi_{abcd}]R_{abcde} \\ &+ [-A_{abcde}\sin\psi_{abcd} + B_{abcde}\cos\psi_{abcd}]I_{abcde}.\end{aligned} \tag{36.7.57}$$

Upon comparing (7.56) and (7.57) we see that (7.37) implies the restrictions

$$A_{abcde}(-1+\cos\psi_{abcd}) + B_{abcde}\sin\psi_{abcd} = 0, \tag{36.7.58}$$

$$-A_{abcde}\sin\psi_{abcd} + B_{abcde}(1+\cos\psi_{abcd}) = 0. \tag{36.7.59}$$

These restrictions are equivalent, and yield the final result

$$B_{abcde} = [\tan(\psi_{abcd}/2)]A_{abcde}. \tag{36.7.60}$$

What we have learned is that if $\mathcal{M}$ is reversal symmetric, then the net Lie generator of its nonlinear part has only about half the number of linearly independent nonlinear basis generators that it otherwise might have. For example, suppose we were to employ, in a reversal symmetric system, various nonlinear correctors (sextupoles, octupoles, etc.) to drive various A_{abcde} to zero in the representation (7.55). Then, according to (7.60), the corresponding B_{abcde} would also automatically be driven to zero. Applications of this principal include the construction of high-order achromats.

Bibliography

Charged-Particle and Light Optics

[1] Gluckstern, R.L. and Holsinger, R.F., "Variable Strength Focussing with Permanent Magnet Quadrupoles", Proceedings of the 1980 Conference on Charged Particle Optics, Giessen, Germany, September 1980, *Nuclear Instruments and Methods in Physics Research*, **187**, 119 (1981).

[2] Wiedemann, H., *Particle Accelerator Physics - Basic Principles and Linear Beam Dynamics*, Springer-Verlag (1993). See pages 126, 127, 184, 186, and 189.

[3] Berz, M., *Modern Map Methods in Particle Beam Physics*, Volume 108 of *Advances in Imaging and Electron Physics*, Academic Press (1999). See Section 4.1.2 and Chapter 6.

[4] Carey, D.C., *The Optics of Charged Particle Beams*, Harwood Academic (1987). See references in this book's Index to *Antisymmetric reflection* and *Symmetry*.

[5] Wollnik, H., *Optics of Charged Particles*, Academic Press (1987). See references in this book's Index to *Mirror symmetric cell*.

[6] Dragt, A.J. et al., *MaryLie 3.0 Users' Manual, A Program for Charged Particle Beam Transport Based on Lie Algebraic Methods*, University of Maryland Physics Department Technical Report (2003). The reversal operation described by (1.3), (1.20), (3.2), and (3.5) has been implemented since 1985 as the MaryLie command with type code *rev*, but has not been completely documented until now.

[7] Wan, W., "Theory and Applications of Arbitrary-Order Achromats", Michigan State University Ph.D. Thesis (1995).

[8] Wan, W. and Berz, M., "Analytical theory of arbitrary-order achromats", *Physical Review E*, **54**, pp. 2870-2883 (1996).

[9] Wan, W. and Berz, M., "Design of a fifth-order achromat", *Nuclear Instruments and Methods in Physics Research*, Section A **423**, pp. 1-6 (1999).

[10] Gerrard, A. and Burch, J.M., *Introduction to Matrix Methods in Optics*, pp. 167-171, Wiley (1975) and Dover (1994).

Dynamical Systems

[11] The entire issue of *Physica D, Nonlinear Phenomena*, Volume 112, Nos. 1-2, pages 1-328 (1998) is devoted to the subject of *Reversal Symmetry in Dynamical Systems*. See, in particular, the opening review article *Time-reversal symmetry in dynamical systems: a survey* by J.S.W. Lamb and J.A.G. Roberts, and its extensive bibliography.

[12] Lamb, J. S. W., "Reversing Symmetries in Dynamical Systems", Doctoral Thesis (1994).

[13] Arnol'd, V.I., "Reversible Systems", *Second International Workshop on Nonlinear and Turbulent Processes in Physics*, Volume 3, ISBN 3-7186-0218-0, p. 1161, Harwood (1984)

[14] McLachlan, R.I., Quispel, G.R.W., and Turner, G.S., "Numerical Integrators that Preserve Symmetries and Reversing Symmetries", *SIAM J. Numer. Anal.* **35**, pp. 586-599 (1998)

[15] McLachlan, R.I., and Perlmutter, M., "Energy Drift in Reversible Time Integration", *J. Phys. A: Math. Gen.* **37**, Number 45, pp. L593-L598 (2004).

[16] Devaney, R.L., "Reversible Diffeomorphisms and Flows", *Transactions of the American Mathematical Society*, **218**, pp. 89-113 (1976).

[17] Mancini, S., Manoel, M., and Teixeira, M.A., "Divergent Diagrams and Simultaneous Conjugacy of Involutions", *Discrete and Continuous Dynamical Systems*, **12**, pp. 657-674 (2005).

[18] Meyer, K.R., "Hamiltonian Systems with a Discrete Symmetry", *Journal of Differential Equations*, **41**, pp. 228-238 (1981).

[19] DeVogelaere, R., "On the structure of periodic solutions of conservative systems, with applications", *Contribution to the theory of nonlinear oscillations*, Volume 4, pp. 53-84, Lefschetz, S. (ed), Princeton University Press (1958).

[20] MacKay, R.S., *Renormalisation in area-preserving maps* (annotated version of Ph.D. thesis, Princeton University, 1982), Advanced series in nonlinear dynamics, volume 6, World Scientific, Singapore (1993).

Quantum Mechanics

[21] Sakurai, J.J., *Modern Quantum Mechanics Revised Edition*, Tuan, S.F., ed., p. 266, Addison-Wesley (1994).

Group Theory

[22] Montgomery, D. and Zippin, L., *Topological Transformation Groups*, Chapter 5, Theorem 1, p. 206, Interscience (1955).

[23] Duistermaat, J.J. and Kolk, J.A.C., *Lie Groups*, Section 2.2, p. 96, Springer (2000).

[24] A. G. O'Farrell and I. Short, *Reversibility in Dnamics and Group Theory*, Cambridge University Press (2015).

Chapter 37

Standard First- and Higher-Order Optical Modules

Chapter 38

Analyticity and Convergence

We have learned from Poincaré's Theorem 1.3.3 that trajectories will be analytic functions of the initial conditions in some domain if the right sides of the equations of motion (1.3.4) are analytic. Appendix F shows that for particle motion in electric and magnetic fields this desired analyticity is realized under very general circumstances. Correspondingly, the Taylor map (7.5.5) will converge in some domain about the origin.

The purpose of this chapter is to describe in some detail what is meant by analyticity and to apply the results of Analytic Function Theory to various problems of interest. Section 38.1 reviews briefly what is needed for the case of one complex variable, and much of its material should already be familiar to the reader. Sections 38.2 and 38.3 treat the case of several complex variables, and may be less familiar. The remaining sections discuss various applications.

38.1 Analyticity in One Complex Variable

In the theory of analytic functions of one complex variable, there are two common ways to *define* analyticity due to Riemann and Weierstrass, respectively. Riemann's approach assumes that the function in question, call it $f(z)$, is defined in some domain (open set) $\mathcal{D}$ of the complex z plane. It then says that f is analytic at the point z if it is *complex* (or *totally*) differentiable there. Complex or totally differentiable means that the limit

$$f'(z) = \lim_{\zeta \to 0} [f(z+\zeta) - f(z)]/\zeta \tag{38.1.1}$$

exists and is the same for all possible ways (e.g. directions) in which the complex variable ζ can approach zero. This definition leads directly to the Cauchy-Riemann equations. Further, f is said to be analytic in $\mathcal{D}$ if f is analytic at each point in $\mathcal{D}$.

In comparison to Weierstrass' definition of analyticity, which will be presented shortly, Riemann's definition is remarkably succinct. It also has some other advantages. For example, it immediately follows from the chain rule for differentiation that the composition of two analytic functions produces a function that is again analytic. That is, if f and g are analytic functions, then $h(z)$ defined by the rule

$$h(z) = g(f(z)) \tag{38.1.2}$$

is also an analytic function. (Of course, for z values of interest, the range of f must also be in the domain of g.) We have the mantra "an analytic function of an analytic function is analytic".

By contrast, Weierstrass defines analyticity in terms of representations by convergent Taylor series. His definition is more involved and, in order to discuss Taylor series, we first must remind ourselves of some properties of infinite sequences and infinite series. We will do this by recalling definitions and stating theorems without proof. Further information may be found in the references listed at the end of this chapter.

Definition 1.1: Consider an infinite sequence c_m of real or complex numbers. We say that this sequence *converges* to a number c,

$$\lim_{m\to\infty} c_m = c,$$

if, given any $\epsilon > 0$, there exists a positive integer N such that

$$|c_m - c| < \epsilon \text{ if } m > N.$$

Definition 1.2: Suppose the sequence c_m has the property that, given any $\epsilon > 0$, there exists an N such that

$$|c_\ell - c_m| < \epsilon \text{ if } \ell, m > N.$$

Such a sequence is called *Cauchy* or *fundamental.*

Theorem 1.1: A sequence is convergent if and only if it is Cauchy.

Definition 1.3: Consider an infinite series $\sum\limits_{n=0}^{\infty} b_n$. We say that this series converges to the (finite) value s if the sequence of partial sums s_m, with

$$s_m = \sum_{n=0}^{m} b_n, \tag{38.1.3}$$

converges to the value s.

Definition 1.4: If an infinite series does not converge, we say that it is *divergent.* By this definition, divergence includes the possibility that the sequence of partial sums increases (in absolute value) without bound, or has more than one limit point. For example, the sequence of partial sums for the series 1, 1/2, 1/3, 1/4 $\cdots$ grows without bound, and the sequence of partial sums for the series 1, -1, 1, -1, $\cdots$ has the limit points 1 and 0.

Theorem 1.2: If the infinite series $\sum b_n$ converges, then its terms b_n must tend to zero,

$$\lim_{n\to\infty} b_n = 0.$$

Theorem 1.3: Suppose the terms b_n are all real and alternate in sign. Suppose also that

$$\lim_{n\to\infty} b_n = 0.$$

Then the series $\sum b_n$ converges.

Definition 1.5: We say that the infinite series $\sum b_n$ converges *absolutely* if the series $\sum |b_n|$ converges.
Theorem 1.4: If the infinite series $\sum b_n$ converges absolutely, then it also converges in the sense of Definition 1.3.
Definition 1.6: Suppose $g(m)$ is a function that provides an invertible mapping of the (non-negative) integers onto themselves (a bijection). Then we say that the series

$$\sum_{m=0}^{\infty} b_{g(m)} \tag{38.1.4}$$

is a *rearrangement* of the series $\sum b_n$.
Theorem 1.5: If the infinite series $\sum b_n$ converges absolutely, then all *rearrangements* of the series $\sum b_n$ converge (in the sense of Definition 1.3 and also absolutely), and they all converge to the same value s.
Definition 1.7: Supposes that the infinite series $\sum b_n$ converges, but not absolutely. That is, the series $\sum |b_n|$ diverges. Then we say that the series $\sum b_n$ converges *conditionally.*
Theorem 1.6 (Riemann): Suppose that the b_n are all real and the series $\sum b_n$ converges conditionally. Then, there are rearrangments of the series that diverge. Moreover, given any (real) number s', there are rearrangements of the series that converge to s'.
Theorem 1.7 (Levy): Suppose the (possibly complex) series $\sum b_n$ is conditionally convergent. Then there is at least one line in the complex plane such that, upon selecting any point w' on that line, there is a rearrangement of the series $\sum b_n$ that converges to w'.
Theorem 1.8 (Steinitz): All the possible sums of a conditionally convergent complex series obtained by rearrangement lie on a straight line or else cover the whole complex plane.

Definitions 1.5 through 1.7 and Theorems 1.4 through 1.8 illustrate the importance of absolute convergence and the need to handle conditionally convergent series with care.

We are now prepared to discuss Taylor series. Without loss of generality, we will for the most part restrict our attention to Taylor series about the origin. Suppose the terms b_n in an infinite series are of the form $a_n z^n$ where z is a complex variable and (a_n) is a sequence of complex numbers. In this case we have the result
Theorem 1.9 (Abel): Suppose that for some (possibly complex) nonzero value $z = z'$ the series $\sum a_n z^n$ converges. Then the series converges absolutely for $|z| < |z'|$.
Moreover, suppose a *radius of convergence* R is defined in terms of the coefficients a_n by the rule

$$1/R = \limsup_{n\to\infty} |a_n|^{1/n}. \tag{38.1.5}$$

Then we have the more specific result
Theorem 1.10 (Cauchy-Hadamard): The series $\sum a_n z^n$ converges absolutely for every z with $|z| < R$, and the convergence is uniform in every closed disk $|z| \leq \rho < R$. (*Uniform* means that given any closed disk and any ϵ as in Definition 1.1, there is an associated N that suffices for all points z in the disk.) If $|z| > R$, the terms in the series are unbounded, and (by Theorem 1.2) the series is divergent. Suppose the series is differentiated term by term any number of times. Then the resulting series is also absolutely convergent for $|z| < R$, and uniformly convergent in every closed disk $|z| \leq \rho < R$.

At this point we observe that if the radius of convergence R of a Taylor series is nonzero, then (by Theorem 1.10) we obtain an infinitely differentiable function f in some neighborhood of the origin by writing

$$f(z) = \sum_{n=0}^{\infty} a_n z^n. \tag{38.1.6}$$

Moreover, we have the relations

$$a_n = (1/n!) f^{(n)}(0). \tag{38.1.7}$$

We are ready for Weierstrass' definition of analyticity. According to his definition, a function f is analytic in the domain $\mathcal{D}$ if, for any point $z^0 \in \mathcal{D}$, the function $f(z)$ has a convergent Taylor expansion about z^0,

$$f(z) = \sum_{n=0}^{\infty} a_n(z^0)(z - z^0)^n. \tag{38.1.8}$$

[Here we have used the notation $a_n(z^0)$ to indicate that the expansion coefficients in general depend on the expansion point z^0.] The definition of Weierstrass has the advantage that initially f need not be defined for complex values of z. For example, when working around the origin, we may at first require only that f have an expansion of the form (1.6) with z real and near zero. Then the series (1.6) automatically *extends* the definition of f to complex values of z. Moreover, it can be verified that the expansion point z^0, which was initially the origin in (1.6), can subsequently be moved to any point in the open disk $|z| < R$ and (26.1.8) will hold. Thus, with Weierstrass' definition, a function that is represented by a Taylor series about the origin is automatically analytic within its disk of convergence. Conversely, if a function is analytic in the disc $|z| < R$, then it will have a Taylor expansion about the origin that converges within that disc.

Contour integration is a fundamental tool in the Theory of Functions of a Single Complex Variable. Applications include the integral formula specified by
Theorem 1.11 (Cauchy): An analytic function f has the representation

$$f(z) = (1/2\pi i) \oint dz' f(z')/(z' - z). \tag{38.1.9}$$

Also, its Taylor coefficients can be found by contour integration. Suppose f is analytic about the origin in a disc of radius R. Let $R' = R - \epsilon$ where ϵ is any small positive number. Then the Taylor coefficients of f about the origin are given by the integrals

$$a_n = (1/2\pi i) \oint_{|z|=R'} dz f(z)/z^{n+1}. \tag{38.1.10}$$

Use of this representation provides the bound

$$|a_n| \leq K(R')^{-n} \tag{38.1.11}$$

where the constant K is given by the relation

$$K = \text{ max } |f(z)| \text{ over the points } |z| = R'. \tag{38.1.12}$$

Suppose f is some given analytic function, and we wish to find the radius of convergence for its Taylor expansion about some point, say the origin. One procedure would be to find its Taylor coefficients and then form the limit (1.5). However, there is an easier way. Simply find the location of the singularity of f that is closest to the origin. Consistent with Theorems 1.10 and 1.11, its distance from the origin will be R.

Finally, we remark that Cauchy's integral formula can be obtained from Riemann's definition of analyticity; and this formula, in turn, can be used to prove the existence of convergent Taylor expansions. Thus we have the key result that Riemann's and Weierstrass' definitions of analyticity are mathematically equivalent.

Historically, the adjective *holomorphic* was used to describe the case where a function was analytic as defined by Riemann, and the adjective *analytic* was used for analyticity as defined by Weierstrass. Now, because these two definitions are mathematically equivalent, these two adjective are commonly used interchangeably.

38.2 Analyticity in Several Complex Variables

The Theory of Functions of Several Complex Variables proceeds in a way that is somewhat analogous to the case of one complex variable, but it is much richer and much less explored. Again there are two ways to define analyticity. According to Riemann, a function $f(z) = f(z_1, z_2, \cdots z_m)$ of m complex variables is analytic if it is analytic (complex differentiable) in each variable *separately*.

As before, Riemann's definition immediately shows that the composition of two analytic functions produces an analytic function. For example, let $f(z_1, z_2, \cdots z_m)$ be an analytic function of m complex variables, and let $g(w)$ be an analytic function of the single complex variable w. Then, by the chain rule, $h(z_1, z_2, \cdots z_m)$ defined by

$$h(z_1, z_2, \cdots z_m) = g(f(z_1, z_2, \cdots z_m)) \tag{38.2.1}$$

is also an analytic function of the m complex variables $z_1, z_2, \cdots z_m$ providing the range of f is in the domain of g for the z values of interest.

According to Weierstrass, a function of m complex variables is analytic in the neighborhood of a point $z^0 = (z_1^0, z_2^0, \cdots z_m^0)$ if it has a convergent multiple Taylor expansion of the form

$$f(z) = f(z_1, z_2, \cdots z_m) = \sum [a_{n_1, n_2, \cdots n_m}(z^0)][(z_1 - z_1^0)^{n_1}(z_2 - z_2^0)^{n_2} \cdots (z_m - z_m^0)^{n_m}]. \tag{38.2.2}$$

The definition of Weierstrass again has the advantage that initially f need not be defined for complex values of the z_ℓ. For example, when working around the origin, we may at first require only that f have an expansion of the form (2.2) with $z^0 = 0$ and the z_ℓ real and near zero. Then, as will become evident in our subsequent discussion, the series expansion automatically extends the definition of f to complex values of the z_ℓ. Moreover, we also get analyticity in an open set.

The handling of *multiple* Taylor series requires some care. For notational simplicity, we mostly limit our discussion to the case of two complex variables and to expansions about the origin. The extension to more variables and general expansion points will generally be obvious.

In the case of double Taylor series about the origin, we have expressions of the form

$$f(z) = f(z_1, z_2) = \sum_{jk} a_{jk} z_1^j z_2^k. \tag{38.2.3}$$

Since this is a double Taylor series, there are many ways of listing its terms sequentially, and the meaning of the infinite sum of these terms is not well defined as it stands. By an *ordering* or *arrangement* of these terms, we mean some procedure for putting the pairs of integers j, k into one-to-one correspondence (a bijection) with the integers $n = 0, 1, 2, \cdots$. Since the set of pairs of integers and the set of integers both have the same cardinality, such a correspondence is always possible. We will denote this correspondence by writing

$$j = j(n), \tag{38.2.4}$$

$$k = k(n).$$

In particular, for any correspondence and given any pair j', k', there is always a unique finite number m such that

$$j' = j(m),$$

$$k' = k(m). \tag{38.2.5}$$

With these ideas in mind we see that, to be precise and without serious loss of generality, we may consider in place of the double series (2.3) a *simple* series of the form

$$f(z) = f(z_1, z_2) = \sum_{n=0}^{\infty} a_{j(n),k(n)} z_1^{j(n)} z_2^{k(n)} = \sum_{n=0}^{\infty} b_n(z) \tag{38.2.6}$$

with

$$b_n(z) = a_{j(n),k(n)} z_1^{j(n)} z_2^{k(n)}. \tag{38.2.7}$$

The treatment of such series is amenable to the methods of Definition 1.3. Figure 2.1 illustrates two of many possible orderings for the case of a double series. In ordering a, successive terms are taken sequentially from the two sides of ever larger squares. In ordering b, successive terms are taken sequentially from the long sides of ever larger triangles.

We can now discuss the convergence properties of multiple Taylor series. First, we define the *convergence set* $\mathcal{T}$ of a Taylor series to be the set of all points for which the series (2.6) converges for some ordering. We have already seen that, according to Theorem 1.10, the convergence set for a Taylor series in one complex variable is a disc. What can be said about the "shape" of the convergence set $\mathcal{T}$ for a Taylor series in several complex variables? We begin with a several complex variable analog of Theorem 1.9.

Theorem 2.1: Suppose that for some pair of (possibly complex) nonzero values z_1', z_2' and some ordering the series (2.6) converges. Define real positive numbers R_1', R_2' by the relations

$$R_1' = |z_1'|, \ R_2' = |z_2'|. \tag{38.2.8}$$

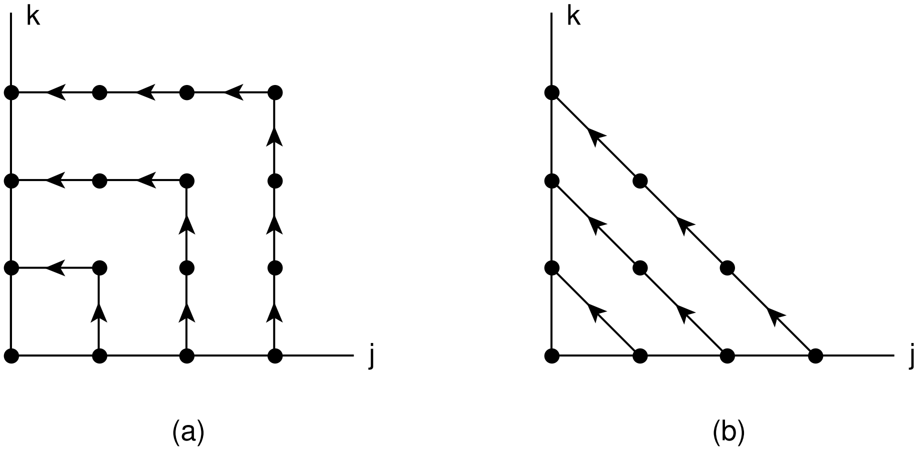


Figure 38.2.1: Two of many possible orderings for the terms in a double series.

Let $\mathcal{P}$ denote the domain (called a *polydisc* or *polycylinder* or *polycircle* about the origin with radii R'_ℓ)

$$|z_1| < R'_1,\ |z_2| < R'_2. \tag{38.2.9}$$

Then, the series (2.6) converges absolutely for all $z_1, z_2 \in \mathcal{P}$, and (by Theorem 1.5) for any ordering. Moreover, by a result of *Fubini*, the iterated series

$$f(z) = f(z_1, z_2) = \sum_{j=0}^{\infty} z_1^j \sum_{k=0}^{\infty} a_{jk} z_2^k = \sum_{k=0}^{\infty} z_2^k \sum_{j=0}^{\infty} a_{jk} z_1^j \tag{38.2.10}$$

also converge (including absolutely) for $z_1, z_2 \in \mathcal{P}$. Finally, all the various series and orderings converge to the same value.

The theorem just quoted, like Theorem 1.9, can be extended. To do so requires the introduction of some notation and further definitions. With regard to notation, let $\boldsymbol{\tau} = (\tau_1, \tau_2 \cdots \tau_m)$ be a collection of m complex numbers. Then by the symbol $\boldsymbol{\tau}z$ we mean the collection of variables $(\tau_1 z_1, \tau_2 z_2, \cdots \tau_m z_m)$. We are ready to give

Definition 2.1: Suppose $\mathcal{R}$ is a set having the property that if z is in $\mathcal{R}$, then so is $\boldsymbol{\tau}z$ providing the τ_ℓ satisfy $|\tau_\ell| \leq 1$. Such a set is called a *complete Reinhardt* set with center at the origin.

A complete Reinhardt set can be descripted pictorially by a *Reinhardt diagram*. Figure 2.2 shows a Reinhardt diagram for the case of two complex variables. Note that the diagram displays the real numbers $|z_1|$ and $|z_2|$. The complete Reinhardt set corresponds to the shaded area. It may also include portions of the axes, called *thorns*, that correspond to the dark lines that extend from the shaded area. A thorn is a polydisc all of whose radii are zero save one.

We next need to define a *logarithmically convex* complete Reinhardt set. Suppose we replot Figure 2.2 in terms of the variables $w_\ell = \log |z_\ell|$. Figure 2.3 shows the resulting *logarithmic image* of the shaded area in Figure 2.2. This image is *convex* if any two points in it can be joined by a straight line that is also in the image. With these concepts in mind we may state

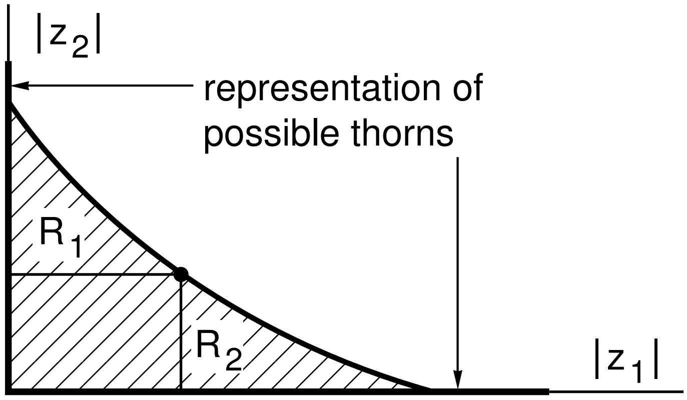


Figure 38.2.2: Possible Reinhardt diagram for the case of two complex variables. The complete Reinhardt set consists of those points $z = (z_1, z_2)$ for which the pairs $|z_1|$, $|z_2|$ lie in the shaded area, or the darkened portions of the axes representing possible thorns. The quantities R_1, R_2 on the boundary of the shaded area are conjugate radii.

Definition 2.2: A complete Reinhardt set is logarithmically convex if the logarithmic image of the "shaded" portion of its corresponding Reinhardt diagram (ignoring possible thorns) is convex.

We are now prepared to describe the characteristic features of the convergence set $\mathcal{T}$ of a Taylor series in several complex variables. The result will be given for a Taylor series about the origin,

$$f(z) = f(z_1, z_2, \cdots z_m) = \sum a_{n_1,n_2,\cdots n_m} z_1^{n_1} z_2^{n_2} \cdots z_m^{n_m}, \tag{38.2.11}$$

but can be easily extended to any expansion point z^0 by simple translation.

Theorem 2.2: The (absolute) convergence set of a multiple Taylor series of the form (2.11) is the *interior* of some logarithmically convex complete Reinhardt set about the origin. (This interior is called a logarithmically convex complete Reinhardt *domain*.) The series also converges absolutely in possible thorns of the set save perhaps at their "tips". Suppose the series is differentiated term by term any number of times with respect to any number of its variables. Then the resulting series is also absolutely convergent. Moreover, let R_ℓ be the value of $|z_\ell|$ at a point on the boundary of the set excluding points for which some $|z_\ell| = 0$. (Note that this requirement excludes thorns.) See Figure 2.2. Then the R_ℓ, called *conjugate radii*, satisfy the condition [a generalization of (1.5)]

$$\limsup_{|n|\to\infty} [|a_{n_1,n_2,\cdots n_m}| R_1^{n_1} R_2^{n_2} \cdots R_m^{n_m}]^{1/|n|} = 1. \tag{38.2.12}$$

Here we have introduced the notation

$$|n| = n_1 + n_2 + \cdots + n_m. \tag{38.2.13}$$

For points on the boundary of the logarithmically convex complete Reinhardt set the series may or may not converge, either absolutely or conditionally. Finally, for any point z outside the set, the terms in the series are unbounded, and hence the series is divergent for any ordering.

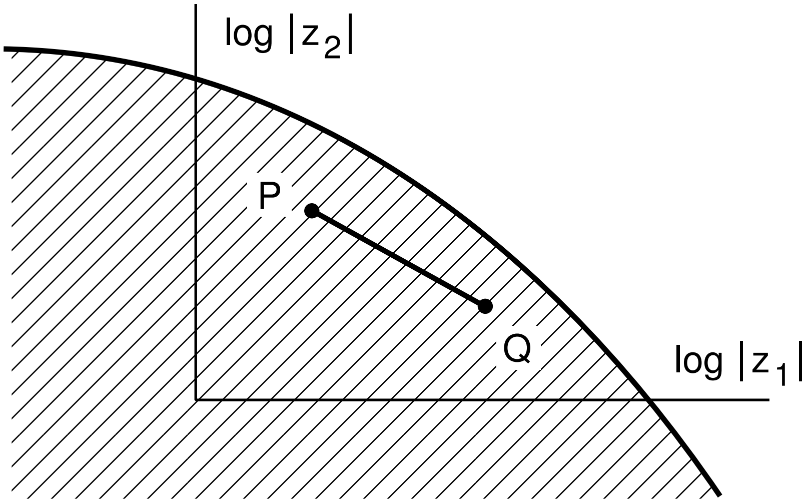


Figure 38.2.3: The logarithmic image of the shaded area in Figure 2.2. This image is convex if any two points P, Q in the image can be joined by a straight line that is also in the image.

According to this theorem, a convergent multiple Taylor series about the origin defines an infinitely differentiable function (with respect to all variables) in a logarthmically convex complete Reinhardt domain about the origin. Moreover, we have the relations

$$a_{n_1,n_2,\cdots n_m} = (1/n_1!)(1/n_2!)\cdots(1/n_m!)\partial_1^{n_1}\partial_2^{n_2}\cdots\partial_m^{n_m} f(z)|_{z=0}\,. \tag{38.2.14}$$

Finally, it can be shown that the expansion point z^0, which was initially the origin in (2.11), can subsequently be moved to any interior point in the logarithmically convex complete Reinhardt domain and (2.2) will hold for z sufficiently near z^0. Thus, a function of several complex variables that is represented by a multiple Taylor series about the origin is automatically analytic within its logarithmically convex complete Reinhardt domain of convergence. Conversely, if a function is analytic in a logarithmically convex complete Reinhardt domain about the origin, then it will have a Taylor expansion about the origin that converges within this logarithmically convex complete Reinhardt domain.

Contour integration also plays an important role in the Theory of Several Complex Variables. For example, there is a generalized Cauchy's Theorem which shows that functions of several complex variables have integral representations of the form

$$\begin{aligned} &f(z) = f(z_1, z_2, \cdots z_m) = \\ &(1/2\pi i)^m \oint\oint\cdots\oint dz_1' dz_2' \cdots dz_m' \times \\ &f(z')/(z_1' - z_1)(z_2' - z_2)\cdots(z_m' - z_m). \end{aligned} \tag{38.2.15}$$

Also, the Taylor coefficients (2.14) can be found by contour integration. Let ϵ be any small positive number. Define quantities R_ℓ' by the relation

$$R_\ell' = R_\ell - \epsilon \tag{38.2.16}$$

where the R_ℓ are some set of conjugate radii. Then the Taylor coefficients of f about the origin are given by the integrals

$$a_{n_1,n_2,\cdots n_m} = (1/2\pi i)^m \oint\limits_{|z_1|=R_1'} \oint\limits_{|z_2|=R_2'} \cdots \oint\limits_{|z_m|=R_m'} dz_1 dz_2 \cdots dz_m f(z)/(z_1^{n_1+1} z_2^{n_1+1} \cdots z_m^{n_m+1}). \tag{38.2.17}$$

From this representation we deduce the Cauchy bound

$$|a_{n_1,n_2,\cdots n_m}| \leq K(R_1')^{-n_1}(R_2')^{-n_2} \cdots (R_m')^{-n_m} \tag{38.2.18}$$

where the constant K is given by the relation

$$K = \max|f(z)| \text{ over the points } |z_1| = R_1', |z_2| = R_2', \cdots |z_m| = R_m'. \tag{38.2.19}$$

Finally, starting from Riemann's definition of analyticity, Hartogs showed that analyticity in each variable separately implies continuity as well with respect to the set of all variables. The generalized Cauchy's Theorem (2.15) then follows; and from it follows the existence of convergent multiple Taylor expansions. Thus, Hartogs obtained the celebrated result that Riemann's and Weierstrass' definitions of analyticity are also equivalent in the much more difficult case of several complex variables.

We have seen at the end of Section 31.1 that, in the case of a function of one complex variable, it is possible to determine the radius of convergence of its Taylor series simply from a knowledge of the locations of the singularities of the function. The same is true for a function of several complex variables: the logarithmically convex complete Reinhardt domain of convergence of its multiple Taylor series expansion can be found from a knowledge of the locations of its singularities. However, the calculation is considerably more involved.

Consider, for simplicity, the case of a function $f(z_1, z_2)$ of two complex variables and its Taylor expansion about the origin. First we examine the two functions $f_1(z_1) = f(z_1, 0)$ and $f_2(z_2) = f(0, z_2)$. They are analytic functions of a single complex variable, and the radii of convergence of their Taylor expansions can be found using the method described at the end of Section 31.1. These two radii determine the tips of the two possible thorns represented in Figure 2.2.

Next we determine the conjugate radii R_1 and R_2. To do so we carry out the following steps:

1. Write z_1 in the form
$$z_1 = \hat{z}_1(\phi_1) = R_1 \exp(i\phi_1). \tag{38.2.20}$$

2. Hold R_1 fixed and positive, and initially small.

3. Examine the function
$$g_2(z_2, \phi_1) = f(\hat{z}_1, z_2). \tag{38.2.21}$$

4. Find the locations of the singularities of g_2 (viewed as a function of z_2) in the complex z_2 plane for each value of $\phi_1 \in [0, 2\pi)$.

5. Follow these singularities as ϕ_1 varies from 0 to 2π.

6. Define R_2 by the relation
$$R_2 = \min_{\phi_1} |z_2^c(\phi_1)| \tag{38.2.22}$$
where $|z_2^c(\phi_1)|$ is the distance from the origin of the singularity of g_2 that is *closest* to the origin for each value of ϕ_1.

7. Repeat steps 1 through 6 above, while slowly increasing the value of R_1, until R_2 becomes zero.

The result of this process is the set of R_1, R_2 values that describe the boundary of the logarithmically convex complete Reinhardt domain. See Figure 2.4. Alternatively, we may carry out a similar process, but interchange the roles of z_1 and z_2. See Figure 2.5. It can be shown that both processes yield identical sets of R_1, R_2 values.

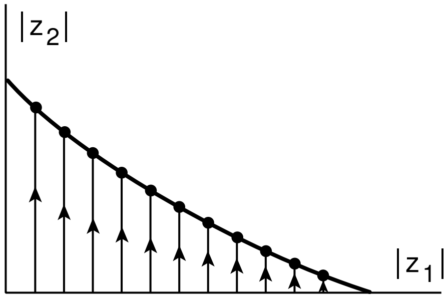


Figure 38.2.4: Determining the conjugate radii by fixing R_1 and searching for the closest singularity in z_2 as ϕ_1 varies to yield R_2.

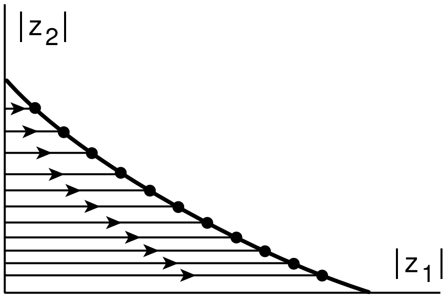


Figure 38.2.5: Determining the conjugate radii by fixing R_2 and searching for the closest singularity in z_1 as ϕ_2 varies to yield R_1.

Example 2.1: Consider the function of three variables defined by the equation
$$\psi(\boldsymbol{r}) = [x_1^2 + (x_2 - 1)^2 + x_3^2]^{-1/2}. \tag{38.2.23}$$

Up to a normalization, ψ is the electrostatic potential due to a point charge located at unit distance from the origin along the x_2 axis. See Figure 2.6.

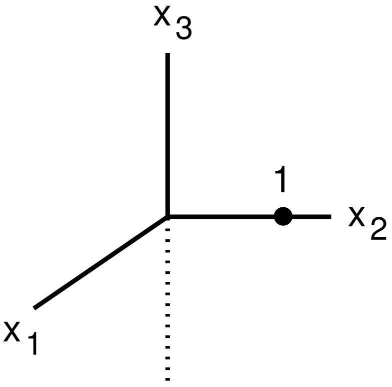


Figure 38.2.6: A point charge located at unit distance from the origin along the x_2 axis.

Suppose we expand $\psi(\boldsymbol{r})$ in a triple Taylor series about the origin,

$$\psi(\boldsymbol{r}) = \sum_{jk\ell} a_{jk\ell} x_1^j x_2^k x_3^\ell. \tag{38.2.24}$$

To examine the convergence of this series, we consider the function

$$f(z) = [z_1^2 + (z_2 - 1)^2 + z_3^2]^{-1/2}. \tag{38.2.25}$$

It is singular when the argument of the square root vanishes,

$$z_1^2 + (z_2 - 1)^2 + z_3^2 = 0. \tag{38.2.26}$$

The Reinhardt diagram in this case is 3-dimensional. An extension of the method just described shows that its boundary is given by the relation

$$R_1 = \min_{\phi_2,\phi_3} |\{-[R_2 \exp(i\phi_2) - 1]^2 - [R_3 \exp(i\phi_3)]^2\}^{1/2}| \text{ with } R_2, R_3 \in [0,1], \tag{38.2.27}$$

or equivalently,

$$R_1^2 = \min_{\phi_2,\phi_3} |[R_2 \exp(i\phi_2) - 1]^2 + [R_3 \exp(i\phi_3)]^2| \text{ with } R_2, R_3 \in [0,1]. \tag{38.2.28}$$

This result follows, esssentially, from solving (2.26) for z_1. Now it is easily verified from geometric considerations in the complex plane that the minimum sought in (2.28) occurs when $\phi_2 = 0$ and $\phi_3 = \pi/2$. See Exercise 2.4. Consequently, we have the result

$$R_1^2 = (R_2 - 1)^2 - R_3^2 \text{ or } R_1^2 + R_3^2 = (R_2 - 1)^2. \tag{38.2.29}$$

This is just the equation for a cone. See Figure 2.7.
Example 2.2: Suppose instead we hold x_3 fixed (and real) and simply expand $\psi(\boldsymbol{r})$ in a double Taylor series in x_1 and x_2 about the origin,

$$\psi(x_1, x_2; x_3) = \sum_{jk} a_{jk}(x_3) x_1^j x_2^k. \tag{38.2.30}$$

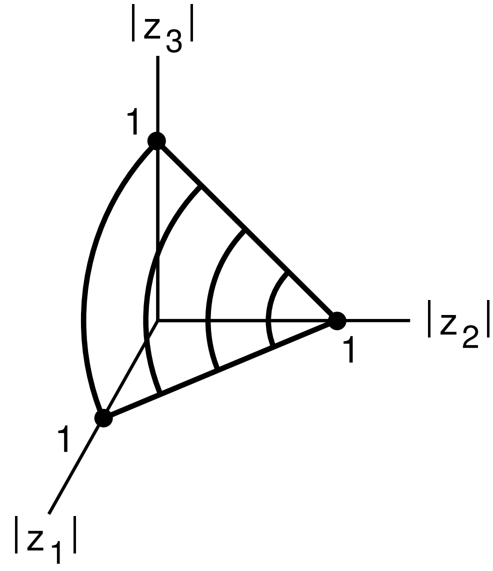


Figure 38.2.7: The Reinhardt diagram for the series (2.24), which represents the function $f(z)$ given by (2.25), is a cone.

To study the convergence of this series we must consider the function

$$f(z;x_3) = [z_1^2 + (z_2-1)^2 + x_3^2]^{-1/2}, \tag{38.2.31}$$

which is singular when

$$z_1^2 + (z_2-1)^2 + x_3^2 = 0. \tag{38.2.32}$$

For this simpler 2-dimensional example direct application of (2.20) through (2.22) gives the result

$$R_2 = \min_{\phi_1} |1 \pm [-R_1^2 \exp(2i\phi_1) - x_3^2]^{1/2}|. \tag{38.2.33}$$

In general the evaluation of (2.33) requires numerical computation. However for the special case $x_3 = 0$, there is the simple result

$$R_1 + R_2 = 1 \text{ when } x_3 = 0. \tag{38.2.34}$$

Figure 2.8 shows, for three values of x_3, the Reinhardt diagrams in the $|z_1|$, $|z_2|$ plane of the series (2.30) that represents $f(z;x_3)$. It is easily verified that the boundaries of these diagrams in the $|z_1| = 0$ and $|z_2| = 0$ planes are hyperbolas,

$$R_2^2 = 1 + x_3^2 \text{ when } |z_1| = 0, \tag{38.2.35}$$

$$R_1^2 = 1 + x_3^2 \text{ when } |z_2| = 0. \tag{38.2.36}$$

Some observations are in order. Note that the double series (2.30) is an iterated version of the triple series (2.24),

$$\psi = \sum_{jk\ell} a_{jk\ell} x_1^j x_2^k x_3^\ell = \sum_{jk} x_1^j x_2^k \sum_{\ell} a_{jk\ell} x_3^\ell = \sum_{jk} a_{jk}(x_3) x_1^j x_2^k, \tag{38.2.37}$$

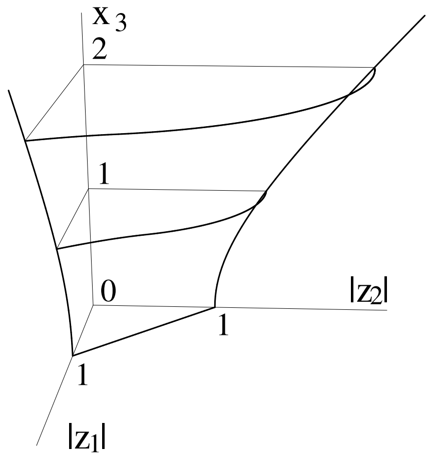


Figure 38.2.8: Reinhardt diagrams, for three values of x_3, of the series (2.30) that represents $f(z; x_3)$. Diagrams for negative values of x_3 are not shown since the diagrams for $\pm x_3$ are identical.

with

$$a_{jk}(x_3) = \sum_{\ell} a_{jk\ell} x_3^{\ell}. \tag{38.2.38}$$

Next we observe from Figure 2.8 that, unlike the previous example of Figure 2.7, the size of the convergence set *increases* with increasing $|x_3|$. We see that summation over some of the variables in a multiple series can yield a series with an extended convergence domain.

We close this section by describing briefly a remarkable feature of analytic functions of *one* complex variable, and a further remarkable feature of functions of *many* complex variables that distinguishes them from functions of a single complex variable. These features have to do with *analytic continuation*.

We begin with functions of a single complex variable, which we call w. Suppose $g(w)$ is such a function that is defined and analytic in some arbitrarily shaped simply-connected domain $\mathcal{D}$. Then we may attempt to extend g beyond $\mathcal{D}$ by analytic continuation. In its simplest description, analytic continuation consists of two steps:

1. Choose some path that leads out of $\mathcal{D}$.

2. Make successive related Taylor expansions of g in overlapping discs along the path. In this process, the Taylor coefficients for the expansion in an adjacent subsequent disc are found from those for the previous disc by multiply differentiating the Taylor expansion for the previous disc and evaluating the result at the new expansion point. That is, two successive Taylor series must give identical values for g in the region where their expansion discs overlap.

See Figure 2.9. However, this process may fail. (The convergence radii of successive Taylor expansions may shrink to zero.) It can be shown that there are functions that are anaytic in $\mathcal{D}$, but cannot be extended in an analytic way beyond the boundary of $\mathcal{D}$. Such functions are said to have a *natural boundary*, which, in this case, is the boundary of $\mathcal{D}$. [Basically, a function (of a single complex variable) with a natural boundary has a dense set of singularities on the boundary that prevent its analytic extension beyond the boundary.] Thus, given any arbitrarily shaped boundary in the complex plane, it is possible to find a function that has this boundary as a natural boundary.

A function that is specified and analytic on a small domain is called a *germ*. What we have learned is the remarkable fact that a germ specifies the full function. If a function of a single variable is defined and analytic is some domain ever so small (it could even be a small line segment), then the function is uniquely extendable over the whole domain to which it can be continued until it is defined everywhere (perhaps in a multiple sheeted way) or a natural boundary is encountered. Moreover, any boundary in the complex plane is a potential natural boundary for some analytic function.

The case of analytic functions of many complex variables is very different. Again, one can begin with a function defined in a small domain, a germ, and then attempt to extend the domain by analytic continuation. But, remarkably, now there are restrictions to what could possibly be natural boundaries.

Specifically, suppose again that $\mathcal{D}$ is some arbitrary domain, but now in C^m, the space of m complex variables. Suppose also that $f(z_1, z_2, \cdots z_m)$ is some function that is analytic in $\mathcal{D}$, and we seek to analytically continue f beyond $\mathcal{D}$. (To carry out analytic continuation in C^m, one may use overlapping polydiscs.) What can be said about this challenge?

Consider, for example, a function $f(z_1, z_2)$ of two complex variables that is analytic within the complete Reinhardt domain whose Reinhardt diagram is displayed in Figure 2.10. Figure 2.11, which shows the logarithmic image of Figure 2.10, illustrates that this domain is not logarithmically convex. However, it becomes logarithmically convex if the region corrresponding to the area below the dashed curved line segment (which is the inverse image of the dashed straight line segment in Figure 2.11) is added to the region corresponding to the shaded area. Indeed, this is the minimum territory that must be added to make the full domain logarithmically convex. A domain augmented in this way is called the *logarithmically convex hull* of the original domain.

It can be shown that *any* function $f(z_1, z_2)$ that is analytic in the complete Reinhardt domain described by the Reinhardt diagram of Figure 2.10 must have an analytic continuation into the minimal logarithmically convex extension obtained by adding to the Reinhardt diagram the area below the dashed curved line. Of course, there may be some functions that can be analytically continued even further. However, there will be other functions that cannot.

Suppose a domain has the property that there exists a function that is analytic in the domain, but this function cannot be analytically continued beyond the domain. Such a domain is called a *domain of holomorphy*. The determination of domains of holomorphy is a major topic in the Theory of Analytic Functions of Several Complex Variables. An important result, for our purposes, is that any logarithmically convex complete Reinhardt domain is a domain of holomorphy.

Often one begins with some particular domain, and would like to know the smallest domain of holomorphy that contains this domain. Such a domain is said to be an *envelope of holomorphy* for the original domain. This question is important, because *any* function that is analytic in the original domain must automatically be analytic in its envelope of holomorphy.

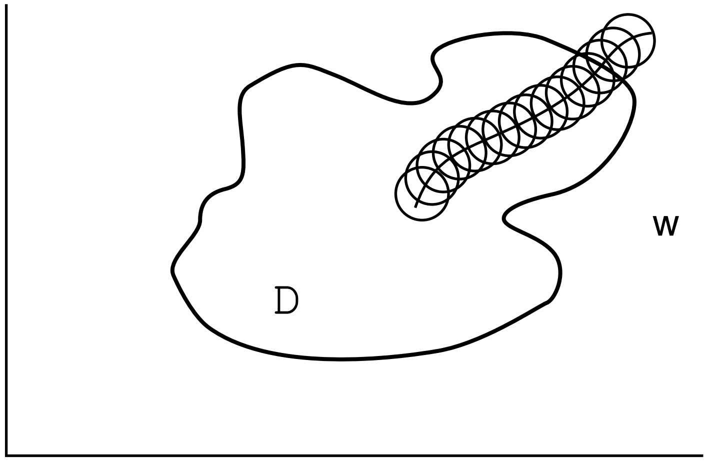


Figure 38.2.9: Analytic continuation along a path out of a domain $\mathcal{D}$ in the complex w plane by making successive related Taylor expansions in overlapping disks along the path.

Roughly speaking, for functions of many complex variables, the known existence of some analyticity often implies the existence of more analyticity.

We conclude that the complete Reinhardt domain described by the Reinhardt diagram of Figure 2.10 is not a domain of holomorphy. By contrast, the minimal logarithmically convex extension of this domain, which is a logarithmically convex complete Reinhardt domain, is a domain of holomorphy. Moreover, it is the envelope of holomorphy for the original domain.

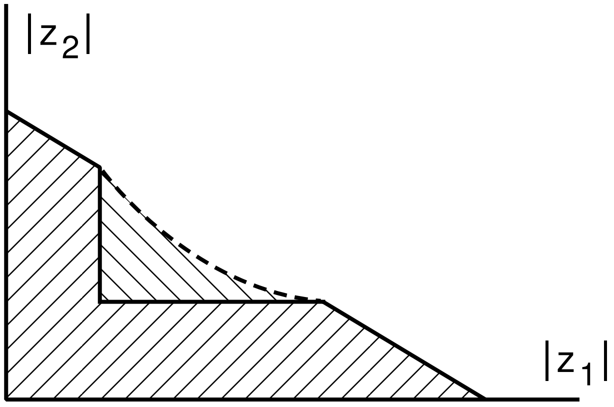


Figure 38.2.10: Reinhardt diagram for a complete Reinhardt domain that is not logarithmically convex. The dashed *curved* line segment is the inverse image of the dashed *straight* line segment in Figure 2.11. The domain becomes logarithmically convex if the region corresponding to the area below the dashed line is annexed to that corresponding to the shaded area.

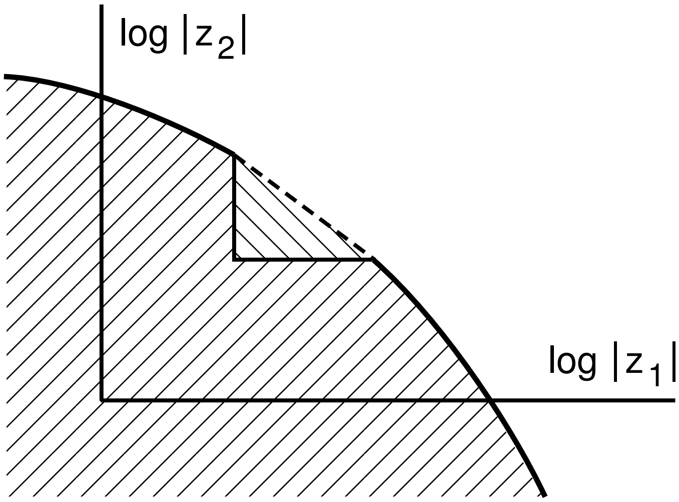


Figure 38.2.11: The logarithmic image of the shaded region of Figure 2.10. Augmenting the image by adding the area below the dashed *straight* line segment makes the image convex.

Exercises

38.2.1. Consider the function

$$f(z) = 1/[1 - (z_1 + z_2)]. \tag{38.2.39}$$

Expand f in a double Taylor series about the origin, and determine the convergence set of this series. Verify that the set is logarithmically convex.

38.2.2. Repeat Exercise 2.1 above for the function

$$f(z) = 1/(1 - z_1 z_2). \tag{38.2.40}$$

38.2.3. Repeat Exercise 2.1 above for the function

$$f(z) = 1/[1 - (z_1 + z_2 + z_1 z_2)]. \tag{38.2.41}$$

38.2.4. Find the ϕ_2, ϕ_3 values that minimize (2.28) by drawing suitable pictures (in the complex plane) of the sets

$$[R_2 \exp(i\phi_2) - 1]^2, \tag{38.2.42}$$

$$[R_3 \exp(i\phi_3)]^2, \tag{38.2.43}$$

$$[R_2 \exp(i\phi_2) - 1]^2 + [R_3 \exp(i\phi_3)]^2. \tag{38.2.44}$$

Verify (2.29).

38.2.5. Verify (2.33) through (2.36). Write a computer program to evaluate R_2 as given by (2.33).

38.2.6. Find the first few coefficients $a_{jk}(x_3)$ as given by (2.30) and (2.38). Replace x_3 by the complex variable z_3, and determine the domain of analyticity for the $a_{jk}(z_3)$.

38.2.7. Consider the map, described by (1.4.21) and (1.4.22), that arises from the monomial Hamiltonian (1.4.20). Suppose this map is expanded in a double Taylor series about the origin. Determine the convergence set for this series, and verify that it is logarithmically convex. Discuss the analytic properties of the Hénon map given by (1.2.23).

38.2.8. Exercise about *real environment.*

38.3 Convergence of Homogeneous Polynomial Series

We have seen in Section 31.2 that, to work with multiple Taylor series in a precise way, it is necessary to order their terms. In this section we will explore the effect of ordering and then *grouping* various terms together during the process of summation. In particular, we will study the effect of grouping terms of like degree together to form homogeneous polynomials, and then summing the resulting series of homogeneous polynomials. For us such series are of particular interest because, as we have seen in earlier chapters, they arise naturally in the Lie algebraic treatment of maps.

As an illustration of the concept of grouping terms, consider the simple series (2.6). Let n_0, n_1, n_2, $n_3 \cdots$ be a set of increasing integers with $n_0 = 0$. Suppose we write the series (2.6) in the form

$$f(z) = \sum_{n=0}^{\infty} b_n(z) = \sum_{n=0}^{n_1-1} b_n(z) + \sum_{n=n_1}^{n_2-1} b_n(z) + \sum_{n=n_2}^{n_3-1} b_n(z) + \cdots = \sum_{\ell=0}^{\infty} \sum_{n=n_\ell}^{n_{\ell+1}-1} b_n(z) = \sum_{\ell=0}^{\infty} c_\ell(z)$$

with

$$c_\ell(z) = \sum_{n=n_\ell}^{n_{\ell+1}-1} b_n(z). \tag{38.3.1}$$

We have carried out group sums of the kind $(b_{n_\ell} + b_{n_\ell+1} + b_{n_\ell+2} + \cdots + b_{n_{\ell+1}-1})$ to produce the quantities c_ℓ, and then formed an infinite sum over the c_ℓ.

What can be said about the convergence of the sum $\sum c_\ell$? According to Definition 1.3, to answer this question we must examine the sequence of partial sums. However inspection reveals that, by construction, the sequence of partial sums of $\sum c_\ell$ is a *subsequence* of the sequence of partial sums of $\sum b_n$. Now we know that every subsequence of a convergent sequence is also convergent; indeed, it must converge to the same limit that the full sequence converges to. Moreover, a subsequence may be convergent even if the full sequence is not. We conclude that the grouping of terms can never spoil the convergence of a series, and it may even help in the sense that it may yield a finite result for what would otherwise have been a divergent series. In the case that it helps, and in the context of multiple Taylor series, we would like to show that the result obtained by arranging and grouping is equivalent to that obtained by analytic continuation out of the original domain of convergence of the series.

Let us apply grouping to the ordering of Figure 2.1b. Suppose the successive groups consist of the terms corresponding to the long sides of the successive triangles. Evidently, the terms corresponding to the long side of any given triangle all have degree $(j + k)$, and their sum is a homogeneous polynomial of degree $(j+k)$. Thus, with this ordering and grouping, the sum (2.6) can be rewritten in the form

$$f(z) = \sum_{n=0}^{\infty} a_{j(n),k(n)} z_1^{j(n)} z_2^{k(n)} = \sum_{\ell=0}^{\infty} P_\ell(z) \tag{38.3.2}$$

where $P_\ell(z)$ is the homogeneous polynomial of total degree ℓ

$$P_\ell(z) = \sum_{j+k=\ell} a_{j,k} z_1^j z_2^k. \tag{38.3.3}$$

We have ordered and grouped a double Taylor series into what we will call a *homogeneous polynomial* series. Evidently, an analogous ordering and grouping can be carried out for any multiple Taylor series. Moreover, we know from the results of Sections 31.1 and 31.2, and the arguments just made, that the homogeneous polynomial series must converge in the same set as the original Taylor series, and must converge to the same result. Indeed, we have from (3.3) the inequality

$$|P_\ell(z)| \leq \sum_{j+k=\ell} |a_{j,k} z_1^j z_2^k|. \tag{38.3.4}$$

Consequently, the homogeneous polynomial series $\sum P_\ell$ converges absolutely whenever the Taylor series converges absolutely. Finally, the homogeneous polynomial series may even converge *outside* the convergence set of the underlying Taylor series.

How can we find the convergence set of the homogenous polynomial series? Remarkably, we will see that this task is *easier* than finding the convergence set of the underlying Taylor series.

Suppose λ is some complex number. By the symbol λz we mean the collection of variables $(\lambda z_1, \lambda z_2, \cdots \lambda z_m)$. Also, $|z|$ will denote the magnitude of z defined by the rule

$$|z|^2 = \sum_{\ell=1}^{m} |z_\ell|^2 = \sum_{\ell=1}^{m} (x_\ell^2 + y_\ell^2) \tag{38.3.5}$$

with

$$z_\ell = x_\ell + i y_\ell. \tag{38.3.6}$$

[However, for exponents we continue to use (2.13).] Consider the function $g(\lambda, z)$ defined by the relation

$$g(\lambda, z) = f(\lambda z) \tag{38.3.7}$$

where $f(z)$ is analytic about the origin and therefore has a convergent expansion of the form (2.11). Combining (2.11) and (3.7) gives the expansion

$$\begin{aligned} g(\lambda, z) = f(\lambda z) &= \sum a_{n_1,n_2,\cdots n_m} (\lambda z_1)^{n_1} (\lambda z_2)^{n_2} \cdots (\lambda z_m)^{n_m} \\ &= \sum \lambda^{|n|} a_{n_1,n_2,\cdots n_m} z_1^{n_1} z_2^{n_2} \cdots z_m^{n_m}. \end{aligned} \tag{38.3.8}$$

For any fixed z, this series will converge for sufficiently small λ. To see this, make some simple estimates. Let $R_1, R_2, \cdots R_m$ be some set of (nonzero) conjugate radii as described in Theorem 2.2, and let R be the smallest of them,

$$R = \min R_\ell. \tag{38.3.9}$$

Next define R' by

$$R' = R - \epsilon \tag{38.3.10}$$

where ϵ is a small positive number. Then from (2.18) we get the bound

$$|a_{n_1,n_2,\cdots n_m}| \le K(R')^{-|n|}. \tag{38.3.11}$$

Also, from (3.5) we have the bounds

$$|z_\ell| \le |z|, \tag{38.3.12}$$

$$|z_1^{n_1} z_2^{n_2} \cdots z_m^{n_m}| \le |z|^{|n|}. \tag{38.3.13}$$

It follows that the series (3.8) has the comparison series

$$\text{comparison series } = \sum K(|\lambda||z|/R')^{|n|} = K\sum_{\ell=0}^{\infty}(|\lambda||z|/R')^\ell \sum_{|n|=\ell} 1. \tag{38.3.14}$$

A moment's reflection shows that the second sum in (3.14) is $N(\ell, m)$, the number of monomials of degree ℓ in m variables,

$$\sum_{|n|=\ell} 1 = N(\ell, m) = (\ell + m - 1)!/[\ell!(m-1)!]. \tag{38.3.15}$$

See (7.3.36). Moreover, for fixed m the quantity $N(\ell, m)$ is a *polynomial* in ℓ of degree $(m-1)$, and for large ℓ has the behavior

$$N(\ell, m) \sim \ell^{m-1}/(m-1)!. \tag{38.3.16}$$

It follows that the comparison series (3.14), and therefore also the series (3.8), converge provided λ is small enough to satisfy the inequality

$$|\lambda| < R'/|z|. \tag{38.3.17}$$

We have seen that the series (3.8) for $g(\lambda, z)$ converges absolutely for sufficiently small λ and z. It follows from the discussion of Section 31.2 that g is an analytic function, in the vicinity of the origin, of all the $(m+1)$ complex variables $z_1, z_2, \cdots z_m$ *and* λ. Moreover, the terms in the series for g can be arranged and grouped to take the form

$$g(\lambda, z) = \sum_{\ell=0}^{\infty} \lambda^\ell \sum_{|n|=\ell} a_{n_1,n_2,\cdots n_m} z_1^{n_1} z_2^{n_2} \cdots z_m^{n_m} = \sum_{\ell=0}^{\infty} \lambda^\ell P_\ell(z) \tag{38.3.18}$$

where $P_\ell(z)$ is the homogeneous polynomial

$$P_\ell(z) = \sum_{|n|=\ell} a_{n_1,n_2,\cdots n_m} z_1^{n_1} z_2^{n_2} \cdots z_m^{n_m}. \tag{38.3.19}$$

Correspondingly, for f itself we have the homogeneous polynomial expansion

$$f(z) = g(1, z) = \sum_{\ell=0}^{\infty} P_\ell(z). \tag{38.3.20}$$

For fixed z, regard (3.18) as a Taylor series in λ with Taylor coefficients $P_\ell(z)$. By essentially the same arguments we have just endured, these coefficients have the bound

$$|P_\ell(z)| \le KN(\ell, m)(|z|/R')^\ell. \tag{38.3.21}$$

Consequently, as expected, this series will converge and produce an analytic function of λ at least within the disc (3.17). It follows from Theorem 1.11 that the $P_\ell(z)$ are given by the integrals

$$P_\ell(z) = (1/2\pi i) \oint d\lambda g(\lambda, z)/\lambda^{\ell+1} = (1/2\pi i) \oint d\lambda f(\lambda z)/\lambda^{\ell+1} \tag{38.3.22}$$

for any contour about the origin in the λ plane for which (3.17) holds. Continue to hold z fixed. Let $\lambda_c(z)$ be the singularity of $g(\lambda, z) = f(\lambda z)$ in the λ plane that is *closest* to the origin. Since $f(\lambda z)$ is always analytic in the disc (3.17), we know that $\lambda_c(z)$ is always nonzero. Define a radius R'' by the rule

$$R'' = \rho|\lambda_c(z)| \tag{38.3.23}$$

where ρ is any number slightly less than but arbitrarily near 1,

$$\rho \simeq 1 \text{ but } \rho < 1. \tag{38.3.24}$$

Then the λ contour in (3.22) can be expanded to give the relation

$$P_\ell(z) = (1/2\pi i) \oint_{|\lambda|=R''} d\lambda f(\lambda z)/\lambda^{\ell+1} \tag{38.3.25}$$

and the bound

$$|P_\ell(z)| \le M/(R'')^\ell \tag{38.3.26}$$

where

$$M = \max |f(\lambda z)| \text{ for } |\lambda| = R''. \tag{38.3.27}$$

We conclude that the homogeneous polynomial expansion (3.20) converges absolutely and uniformly provided $\lambda_c(z)$ satisfies the relation

$$|\lambda_c(z)| \ge \rho' > 1. \tag{38.3.28}$$

Conversely, the series (3.18) must diverge for any λ that satisfies

$$|\lambda| > |\lambda_c(z)|. \tag{38.3.29}$$

For if it converged, then (by Theorem 1.9 and the discussion surrounding it) $g(\lambda, z)$ would be analytic in a disc having a radius larger than $|\lambda_c|$, and could not be singular at $\lambda = \lambda_c$. Finally we note that the calculations involved in the relations (3.22) through (3.27), and the divergence criterion associated with (3.29), hold for general z.

We now have the ingredients for determining the domain of convergence of a homogeneous polynomial expansion. The recipe is this:

1. Let $\mathcal{S}^{2m-1}$ be the unit sphere in C^m defined by the condition

$$|z| = 1. \tag{38.3.30}$$

2. For any point $\hat{z} \in \mathcal{S}^{2m-1}$ let $\lambda_c(\hat{z})$ be the singularity of $f(\lambda\hat{z})$ in the λ plane that is closest to the origin.

3. Define the positive number $\sigma(\hat{z})$ by the rule

$$\sigma(\hat{z}) = |\lambda_c(\hat{z})|. \tag{38.3.31}$$

4. Let $\zeta(\hat{z})$ be the *ray* that goes from the origin to the point $\sigma(\hat{z})\hat{z}$,

$$\zeta(\hat{z}) = \text{ set of all points } r\sigma\hat{z} \text{ with } r \in [0, 1]. \tag{38.3.32}$$

5. Let $\mathcal{H}$ be the union of all rays $\zeta(\hat{z})$ for all points $\hat{z} \in \mathcal{S}^{2m-1}$,

$$\mathcal{H} = \bigcup_{\hat{z} \in \mathcal{S}^{2m-1}} \zeta(\hat{z}). \tag{38.3.33}$$

Then the convergence set of the homogeneous polynomial expansion (3.20) is $\mathcal{H}$. Specifically, the homogeneous polynomial expansion converges absolutely for all z in the interior of $\mathcal{H}$. Moreover, if z is any point in the exterior of $\mathcal{H}$, the terms $P_\ell(z)$ are unbounded for increasing ℓ, and hence the homogeneous polynomial expansion diverges at all exterior points.

We note that this recipe is considerably simpler, particularly in the case of many complex variables, than the analogous recipe given in Section 31.2 for finding conjugate radii of Taylor series. Moreover, it has the advantage that it can be applied, if desired, using only *real* points $\hat{x}$ in $\mathcal{S}^{2m-1}$. This feature is of interest because, in the context of maps, we often need to know only about the convergence of series when all variables are real. Finally, we observe that by construction the *interior* of the convergence set $\mathcal{H}$ has the property that if z is in $\mathcal{H}$, then so is τz where τ is any complex number satisfying $|\tau| \leq 1$. See Exercise 3.4. A domain having this property is called a *complete circular* domain. Thus, the natural domain of convergence for a homogeneous polynomial series is a complete circular domain.

The statements just made about convergence and divergence are easily proved. First, given any $z \neq 0$, there is a unique ray that goes from the origin to z, and this ray (extended if necessary) intersects $\mathcal{S}^{2m-1}$ in the point $\hat{z}$ given by

$$\hat{z} = z/|z|. \tag{38.3.34}$$

Now suppose that z is in the *interior* of $\mathcal{H}$. Then z can be written in the form

$$z = r\sigma(\hat{z})\hat{z} \text{ with } 0 < r \leq \rho'' < \rho < 1. \tag{38.3.35}$$

Here ρ is the same number that appears in (3.23). For $P_\ell(z)$ we find, by using (3.23), (3.26), (3.31), and (3.35), the result

$$\begin{aligned} |P_\ell(z)| &= |P_\ell(r\sigma\hat{z})| = |(r\sigma)^\ell P_\ell(\hat{z})| \\ &\leq |(\rho''\sigma)^\ell P_\ell(\hat{z})| = (\rho''/\rho)^\ell |(\rho\sigma)^\ell P_\ell(\hat{z})| \\ &= (\rho''/\rho)^\ell |(\rho|\lambda_c(\hat{z})|)^\ell P_\ell(\hat{z})| \\ &= (\rho''/\rho)^\ell |(R'')^\ell P_\ell(\hat{z})| \leq (\rho''/\rho)^\ell M. \end{aligned} \tag{38.3.36}$$

However, from (3.35) we have the inequality

$$(\rho''/\rho) < 1. \tag{38.3.37}$$

It follows that (3.20) has a geometric series as a comparison series, and hence it converges absolutely when z is an interior point in $\mathcal{H}$.

To complete the proof, suppose that z is in the *exterior* of $\mathcal{H}$. Then z can be written in the form

$$z = r\sigma(\hat{z})\hat{z} \text{ with } r > 1. \tag{38.3.38}$$

For $P_\ell(z)$ we find the result

$$P_\ell(z) = P_\ell(r\sigma\hat{z}) = (r\sigma)^\ell P_\ell(\hat{z}) \tag{38.3.39}$$

and (3.20) becomes

$$\sum_{\ell=0}^{\infty} P_\ell(z) = \sum_{\ell=0}^{\infty} (r\sigma)^\ell P_\ell(\hat{z}), \tag{38.3.40}$$

which is a series of the form (3.18) with

$$\lambda = r\sigma. \tag{38.3.41}$$

Since $r > 1$, we see from (3.31) and (3.41) that

$$|\lambda| = |r\sigma| = r|\lambda_c| > |\lambda_c|, \tag{38.3.42}$$

and conclude from (3.29) that the series (3.40) is divergent. Because (3.40) is a divergent Taylor series, the terms $P_\ell(z)$ that comprise it must be unbounded. See (3.39) and Theorem 1.9.

It remains to be shown that if the homogeneous polynomial series converges outside the convergence set of the underlying Taylor series, then it provides an analytic continuation of the function specified by the Taylor series. The reader has the pleasure of working out a proof in Exercise 3.5.

Example 3.1: Consider again the function $\psi(\boldsymbol{r})$ given by (2.23). Suppose the series (2.24) is grouped into homogeneous polynomials,

$$\psi(x_1, x_2, x_3) = \sum_{jk\ell} a_{jk\ell} x_1^j x_2^k x_3^\ell = \sum_{m=0}^{\infty} \sum_{j+k+\ell=m} a_{jk\ell} x_1^j x_2^k x_3^\ell = \sum_{m=0}^{\infty} P_m(x_1, x_2, x_3) \tag{38.3.43}$$

where

$$P_m(x_1,x_2,x_3)=\sum_{j+k+\ell=m} a_{jk\ell}x_1^j x_2^k x_3^\ell. \tag{38.3.44}$$

Let us determine the domain of convergence of this homogeneous polynomial series in the (real) x_1,x_2,x_3 3-dimensional space. We parameterize points $\hat{x}_1,\hat{x}_2,\hat{x}_3\in\mathcal{S}^3$ by writing

$$\hat{x}_2=\sin\theta, \tag{38.3.45}$$

$$\hat{x}_1=\cos\theta\cos\phi, \tag{38.3.46}$$

$$\hat{x}_3=\cos\theta\sin\phi. \tag{38.3.47}$$

From (2.26) we see that singularities in λ satisfy the equation

$$(\lambda\cos\theta\cos\phi)^2+(\lambda\sin\theta-1)^2+(\lambda\cos\theta\sin\phi)^2=0.$$

This equation has the solutions

$$\lambda=\sin\theta\pm i\cos\theta \tag{38.3.48}$$

from which we find that

$$|\lambda_c|=1. \tag{38.3.49}$$

It follows that the homogeneous polynomial series (3.43) for $\psi(x_1,x_2,x_3)$ converges in the *unit ball* about the origin,

$$0\le x_1^2+x_2^2+x_3^2\le 1. \tag{38.3.50}$$

Note that this set includes points that lie outside the convergence set described by the Reinhardt diagram of Figure 2.7.

Example 3.2: Suppose the series (2.30) is grouped into homogeneous polynomials,

$$\psi(x_1,x_2;x_3)=\sum_{jk} a_{jk}(x_3)x_1^j x_2^k=\sum_{\ell=0}^{\infty}\sum_{j+k=\ell} a_{jk}(x_3)x_1^j x_2^k=\sum_{\ell=0}^{\infty}P_\ell(x_1,x_2;x_3) \tag{38.3.51}$$

where

$$P_\ell(x_1,x_2;x_3)=\sum_{j+k=\ell} a_{jk}(x_3)x_1^j x_2^k. \tag{38.3.52}$$

For fixed (real) x_3, let us determine the convergence set of this homogeneous polynomial series in the (real) x_1, x_2 plane. We parameterize points $\hat{x}_1,\hat{x}_2\in\mathcal{S}^2$ by writing the relations

$$\hat{x}_1=\cos\phi, \tag{38.3.53}$$

$$\hat{x}_2=\sin\phi. \tag{38.3.54}$$

Then, from(2.32), we see that singularities in λ satisfy the equation

$$(\lambda\cos\phi)^2+(\lambda\sin\phi-1)^2+x_3^2=0. \tag{38.3.55}$$

This equation has the solutions

$$\lambda=\sin\phi\pm i[\cos^2\phi+x_3^2]^{1/2} \tag{38.3.56}$$

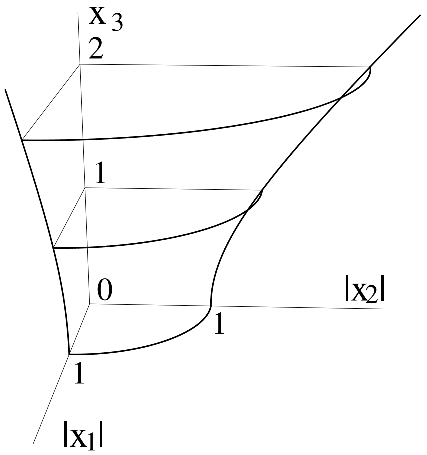


Figure 38.3.1: Real x_1, x_2 convergence sets for the homogeneous polynomial series (3.51) for various values of x_3. Together they form a hyperbola of revolution. Sets are not shown for negative values of x_3 since the sets for $\pm x_3$ are identical.

from which we find that

$$|\lambda_c| = (1 + x_3^2)^{1/2}. \tag{38.3.57}$$

It follows that the homogeneous polynomial series (3.51) for $\psi(x_1, x_2; x_3)$ converges in a *disc* about the origin in the x_1,x_2 plane with radius $(1+x_3^2)^{1/2}$. Figure 3.1 shows several of these discs for various values of x_3. Together they form the hyperbola of revolution

$$x_1^2 + x_2^2 = 1 + x_3^2. \tag{38.3.58}$$

Note that these sets include points that lie outside the convergence sets described by the Reinhardt diagrams of Figure 2.8.

Exercises

38.3.1. Find the convergence set in the real x_1, x_2 plane for the homogeneous polynomial expansion about the origin of the function f given by (2.39).

38.3.2. Repeat Exercise 3.1 for the f given by (2.40).

38.3.3. Repeat Exercise 3.1 for the f given by (2.41).

38.3.4. Verify Equations (3.45) through (3.50).

38.3.5. Verify Equations (3.53) through (3.58).

38.3.6. Verify that the interior of the convergence set $\mathcal{H}$ constructed following the steps (3.30) through (3.33) is a complete circular domain.

38.3.7. Show that if the homogeneous polynomial series converges outside the convergence set of the underlying Taylor series, then it provides an analytic continuation of the function specified by the Taylor series.

38.3.8. Consider the map, described by (1.4.21) and (1.4.22), that arises from the monomial Hamiltonian (1.4.20). See Exercise 2.7. Suppose this map is expanded in a homogeneous polynomial series about the origin. Find the convergence set in the real q, p plane for this expansion.

38.3.9. Using the methods of this section, determine the domain of analyticity and the convergence set for the monopole doublet $\psi(x, y, z)$ given by (13.11.3).

38.4 Application to Potentials and Fields

38.5 Application to Taylor Maps: The Anharmonic Oscillator

38.6 Application to Taylor Maps: The Pendulum

38.7 Convergence of the BCH Series

38.8 Convergence of Lie Transformations and the Factored Product Representation

Bibliography

Taylor Series and Complex Variables

[1] P. Dienes, *The Taylor Series*, Oxford (1931).

[2] K. Knopp, *Theory and Application of Infinite Series*, Blackie (1951).

[3] K. Knopp, *Theory of Functions, Parts I and II*, Dover Publications, New York (1945).

[4] I.I. Hirschman, *Infinite Series*, Holt, Rinehardt and Winston (1962).

[5] W. Rudin, *Principles of Mathematical Analysis*, McGraw-Hill (1976).

[6] R. Remmert, *Theory of Complex Functions*, Springer-Verlag (1991).

[7] R. Remmert, *Classical Topics in Complex Function Theory*, Springer-Verlag (1998).

[8] E.C. Titchmarsh, *The Theory of Functions*, Oxford (1960).

[9] L.V. Ahlfors, *Complex Analysis*, McGraw Hill (1979).

[10] G. Springer, *Introduction to Riemann Surfaces*, Addison-Wesley (1957).

[11] H. Weyl, *The Concept of a Riemann Surface*, Dover (2009).

[12] E. Hille, *Analytic Function Theory, Vol. I and II*, Ginn and Company (1962).

[13] C.L. Siegel, *Topics in Complex Function Theory, Vols. I-III*, Wiley-Interscience (New York, 1971).

[14] V.M. Kadets and M.I. Kadets, *Rearrangement of Series in Banach Spaces*, American Mathematical Society (1991).

[15] E.T. Whittaker and G.N. Watson, *A Course of Modern Analysis*, Cambridge (1952).

[16] E.T. Copson, *Theory of Functions of a Complex Variable*, Oxford University Press (1935).

[17] J. Stalker, *Complex analysis, fundamentals of the classical theory of functions*, Birkhauser (1998).

[18] R.M. Range, *Holomorphic Functions and Integral Representations in Several Complex Variables*, Springer-Verlag (1986).

[19] B.V. Shabat, *Introduction to Complex Analysis, Part II: Functions of Several Variables*, Translations of Mathematical Monographs, Vol. 110, American Mathematical Society (1992).

[20] W. Kaplan, *Functions of Several Complex Variables*, Ann-Arbor (1964).

[21] W. Kaplan, *Introduction to Analytic Functions*, Addison-Wesley (1966).

[22] A.S. Wightman, Analytic functions of several complex variables, p. 227 in *Relations de dispersion et particules élémentaires*, C. De Witt and R. Omnes, eds., Hermann, Paris (1960).

[23] B.A. Fuks, *Theory of Analytic Functions of Several Complex Variables*, Translations of Mathematical Monographs, Vol. 8, American Mathematical Society (1963).

[24] B.A. Fuks, *Functions of a Complex Variable and Some of Their Applications, Vol. I*, Pergamon Press Addison-Wesley (1964).

[25] H. Cartan, *Elementary Theory of Analytic Functions of One or Several Complex Variables*, Addison-Wesley (1973) and Dover (1995).

[26] S. Bochner and W.T. Martin, *Several Complex Variables*, Princeton (1948).

[27] R.C. Gunning and H. Rossi, *Analytic Functions of Several Complex Variables*, Prentice-Hall (1965).

[28] R.C. Gunning, *Introduction to Holomorphic Functions of Several Variables, Vols. I to III*, Wadsworth (1990).

[29] S.G. Krantz, *Complex Analysis: The Geometric Viewpoint*, Mathematical Association of America (1990).

[30] S.G. Krantz, *Function Theory of Several Complex Variables*, Second Edition, American Mathematical Society (2001).

[31] E. Chirka, P. Dolbeault, G. Khenkin, and A. Vitushkin *Introduction to Complex Analysis*, Springer (1997).

[32] L. Hormander, *An Introduction to Complex Analysis in Several Variables*, North Holland (1989).

[33] Kiyosi Ito, edit., *Encyclopedic Dictionary of Mathematics*, Second Edition, MIT Press (1993).

[34] *Encyclopaedia of Mathematical Sciences*, A.G. Vitushkin, et al., eds., Springer-Verlag. Volumes 7-10, 54, 69, and 74 of this series comprise the books *Several Complex Variables I-VII*, which contain many chapters written by various authors.

[35] J.P. D'Angelo, *Several Complex Variables and the Geometry of Real Hypersurfaces*, CRC Press (1993).

[36] W. Ebeling, *Functions of Several Complex Variables and Their Singularities*, American Mathematical Society (2007).

[37] V. Vladimirov, *Methods of the Theory of Functions of Many Complex Variables*, M.I.T. Press (1966).

[38] M. Jarnicki and P. Pflug, *First Steps in Several Complex Variables: Reinhardt Domains*, European Mathematical Society (2008).

[39] J. Taylor, *Several Complex Variables with Connections to Algebraic Geometry and Lie Groups*, American Mathematical Society (2002).

[40] J. Taylor, *Several Complex Variables with Connections to Algebraic Geometry and Lie Groups*, American Mathematical Society (2002).

[41] V. Scheidemann, *Introduction to Complex Analysis in Several Variables*, Birkhäuser Verlag (2005).

[42] Y. Ilyashenko and S. Yakovenko, *Lectures on Analytic Differential Equations*, American Mathematical Society (2008).

[43] H. Zoladek, *The Monodromy Group*, Birkhäser Verlag (2006).

The Anharmonic Oscillator and the Pendulum

Chapter 39

Truncated Power Series Algebra

39.1 Introduction

With regard to the simplest transcendental function, the exponential function, it has been said that

> *God created e^x but never heard of a polynomial.*

In this chapter we will be studying polynomials in multiple variables. For mortals, polynomials are a major portal to the transcendent.

Anyone familiar with the rudiments of Computer Science recognizes that the proper choice of algorithm can make an enormous difference in both computational speed and storage requirements. Thus, of several different algorithms which execute the same computational task, some may perform much better than others. Recall, for example, the problem of sorting n items: the number of operations required for the simple *bubble sort* method scales as n^2, while that for the more sophisticated *quick sort* or *heap sort* scales as only $n \log_2 n$. As a second example, the operation count for the ordinary discrete Fourier transform scales as n^2 (where n is the number of data points) while that for the celebrated fast (discrete) Fourier transform again scales as $n \log_2 n$.

The purpose of this chapter is to describe how truncated power series (finite sums of monomials in several variables) can be manipulated by computer. We will explore various methods for labeling and storing monomials together with their relation to efficient algorithms for executing various polynomial operations. These operations, which we refer to as Truncated Power Series Algebra (TPSA), include addition, multiplication, differentiation, Poisson bracketing, the commutation of vector fields, and the composition of functions. The emphasis here will be on computational speed, storage requirements, and program flexibility. Section 32.2 describes how monomials may be labeled and stored. Subsequent sections describe how truncated power series composed of these monomials can be added, multiplied, and otherwise manipulated.

Since TPSA has important applications to the study of dynamical systems, our discussion will often be couched in those terms — sometimes with a particular emphasis on applications to accelerator physics. However, we stress that much of the material presented here applies generically to any use of TPSA.

Before discussing possible schemes for storing and manipulating polynomials, we introduce some useful terminology and definitions: A typical monomial in d variables may be written in the form

$$z_1^{j_1} z_2^{j_2} \cdots z_d^{j_d}, \tag{39.1.1}$$

where the exponents j_k are a set of non-negative integers. We shall sometimes write the d-tuple of exponents $(j_1, \cdots, j_d)$ simply as a vector j, and similarly abbreviate the corresponding monomial by writing

$$z_1^{j_1} \cdots z_d^{j_d} = z^j. \tag{39.1.2}$$

Indeed, for the sake of brevity we shall sometimes refer to the exponent vector j as "the monomial j". For the degree of this monomial, we introduce the notation

$$|j| = j_1 + j_2 + \cdots + j_d. \tag{39.1.3}$$

Recall from Section 7.3 that $N(m, d)$, the number of monomials of degree m in d variables, is given by the binomial coefficient

$$N(m, d) = \left(\begin{array}{c} m + d - 1 \\ m \end{array} \right) = \frac{(m + d - 1)!}{m!(d - 1)!}. \tag{39.1.4}$$

Various values of $N(m, d)$ are listed in Table 7.3.1. Also $S(m, d)$, the number of monomials of degrees 1 through m in d variables, is given by the relation

$$S(m, d) = \left(\begin{array}{c} m + d \\ m \end{array} \right) - 1 = \frac{(m + d)!}{m!d!} - 1. \tag{39.1.5}$$

See Section 7.9. Various values of $S(m, d)$ are listed in Table 7.9.1. Finally, for some calculations it is also useful to employ the quantity $S_0(m, d)$, the number of monomials of degrees 0 through m in d variables. It is given by the relation

$$S_0(m, d) = S(m, d) + 1 = \left(\begin{array}{c} m + d \\ m \end{array} \right) = \left(\begin{array}{c} m + d \\ d \end{array} \right) = \frac{(m + d)!}{m!d!}. \tag{39.1.6}$$

39.2 Monomial Indexing

Any program that manipulates polynomials must have a scheme for labeling and storing the coefficients of the basis monomials. In this section we will describe some ways in which this can be done.

39.2.1 An Obvious but Memory Intensive Method

One very obvious such scheme uses a multi-dimensional array indexed by the monomial exponents. For purposes of illustration, consider the six-variable case. Then, using a phase-space notation, we have monomials of the form

$$z^j = X^{j_1} P_x^{j_2} Y^{j_3} P_y^{j_4} \tau^{j_5} P_\tau^{j_6}. \tag{39.2.1}$$

Suppose, for example, we are interested in the case of homogeneous polynomials of degrees 0 through 12. Then each j_k lies in the interval $[0,12]$, we have a $13 \times 13 \times \cdots$ (six factors) array, and such an array requires $13^6 \simeq 4.8 \times 10^6$ storage locations. By contrast, the entry for $S(12,6)$ in Tablel 7.9.1 shows that in principle only 18,564 ($= 18,563 + 1$) locations should be required. In general, if we use an $(m+1) \times (m+1) \cdots$ (d factors) array to store the coefficients of monomials of degree 0 through m in d variables, we shall need $(m+1)^d$ storage locations. This number is much larger than $S_0(m,d)$. As a consequence, it is desirable — even essential — to consider other possible schemes for labeling and storing monomials.

39.2.2 Polynomial Grading

Implicit in our discussion so far is the assumption that the objects of interest really are polynomials of degrees 0 through m. More explicitly, we assume that we are interested in *grading* polynomials according to their *total* degree. See Section 8.9. For phase-space variables this assumption seems natural, because we expect that the possible excursions from a design trajectory (in suitable scaled coordinates) could be of comparable size in any direction. It may be less natural (and perhaps a different treatment is called for) if we wish simultaneously to make expansions in various parameter variables. Indeed, in this latter case it might be better to have a scheme where the orders of the phase-space variables and the parameter variables could be set independently. One would then have polynomials in the phase-space variables whose coefficients are either numbers or polynomials in the parameter variables.

A remark about nomenclature: Consider the set of all (zero or positive) integers $j_1 \cdots j_d$ that obey the condition

$$|j| \leq m \tag{39.2.2}$$

for a fixed value of m. They form a collection of $S_0(m,d)$ points in d-dimensional space. Some authors refer to this set of points as a *pyramid.* The reader is invited to sketch these points in the cases $d = 2$ and $d = 3$ to see why the name is apt. A set of values assigned to points on a pyramid is referred to as a pyramidal data structure. Finally, some authors refer to sets of points of the kind described in Subsection 32.2.1 as (possible high dimensional) *boxes* or *cubes.* See Appendix S.

39.2.3 Monomial Ordering

Because a well-defined ordering facilitates the systematic implementation of polynomial algebra on a computer, we will continue this section by describing the concept of monomial orderings.

A *monomial ordering* is a relation $>$ on the set of all monomials (exponents) $\alpha = (\alpha_1, \alpha_2 \cdots)$ that satisfies the following three conditions:

1. The relation $>$ is a *total ordering*, meaning that for any two monomials α and β, exactly one of the following statements holds true:

$$\alpha > \beta \ , \quad \alpha = \beta \ , \quad \beta > \alpha. \tag{39.2.3}$$

This condition allows us to arrange the terms of a polynomial in an unambiguous way.

2. If $\alpha > \beta$, then $(\alpha+\gamma) > (\beta+\gamma)$. To see the value of this condition, note that multiplying both z^α and z^β by z^γ yields the results $z^{\alpha+\gamma}$ and $z^{\beta+\gamma}$. Now consider a polynomial whose terms are ordered using the relation $>$. If the above condition holds, then multiplying this polynomial (term-by-term) by z^γ will not alter the arrangement of the terms.

3. The relation $>$ is a *well-ordering*, meaning that every strictly decreasing sequence $\alpha > \beta > \gamma > \cdots$ must eventually terminate. This condition facilitiates proofs that various polynomial algorithms terminate after a finite number of steps.

Suppose we take the $S_0(m,d)$ monomials of degrees 0 through m in d variables, and list them sequentially in some fashion. Next we try to declare that this list constitutes a monomial ordering. For example, as we go down the list, we might declare that each successive monomial is less than all its predecessors. This declaration is consistent with the requirements 1 and 3. However, it generally violates requirement 2. Consequently, we will distinguish between *orderings* and *arrangements*. By an arrangement we will mean any sequential listing, whereas an ordering will mean a monomial ordering as defined above.

The following examples illustrate some of the commonly used monomial orderings. Here we will somes write $z^\alpha > z^\beta$ if $\alpha > \beta$.

Example 1. Lexicographic Order (lex). Let $\alpha = (\alpha_1, \cdots, \alpha_d)$, $\beta = (\beta_1, \cdots, \beta_d)$. We say $\alpha >_{\rm lex} \beta$ if in the vector difference $\alpha - \beta$ the *left-most* nonzero entry is *positive*. (Note that this construction is identical to that used to order weights in Section 5.8.)

For lexicographic ordering we have the relations

$$(1,0,\cdots,0) >_{\rm lex} (0,1,\cdots,0) >_{\rm lex} \cdots >_{\rm lex} (0,0,\cdots,1), \tag{39.2.4}$$

or, to use a more familiar notation,

$$z_1 >_{\rm lex} z_2 >_{\rm lex} \cdots >_{\rm lex} z_d. \tag{39.2.5}$$

Therefore the variables z_i themselves are arranged in descending order as their subscript label increases. Moreover, it is easily checked that

$$(z_i)^m >_{\rm lex} (z_j)^n \text{ whenever } i < j, \tag{39.2.6}$$

independent of the powers m, n (but assuming $m > 0$).

Unfortunately, this latter property can be inconvenient for problems where the total degree of monomials is important. (A common such problem in accelerator physics is the construction of a factorized Lie map.) However, there is a simple remedy to this situation: first order monomials by total degree, then apply lex ordering only to monomials of the same total degree. A formal definition of this scheme is given in the next example.

Example 2. Graded Lexicographic Order (glex, sometimes called grlex). We say $\alpha >_{\rm glex} \beta$ if either $|\alpha| > |\beta|$, or $|\alpha| = |\beta|$ and $\alpha >_{\rm lex} \beta$.

For monomials of degree one,

$$z_1 >_{\rm glex} z_2 >_{\rm glex} \cdots >_{\rm glex} z_d, \tag{39.2.7}$$

just as for lex order. For quadratic monomials,

$$z_1^2 >_{\text{glex}} z_1z_2 >_{\text{glex}} z_1z_3 >_{\text{glex}} \cdots >_{\text{glex}} z_2^2 >_{\text{glex}} z_2z_3 >_{\text{glex}} \cdots >_{\text{glex}} z_d^2. \tag{39.2.8}$$

Example 3. Graded Reverse Lexicographic Order (grevlex). In this ordering we say $\alpha >_{\text{grevlex}} \beta$ if either $|\alpha| > |\beta|$, or $|\alpha| = |\beta|$ and the *right-most* nonzero entry in the vector difference $\alpha - \beta$ is *negative.*

The grevlex order may seem less intuitive than glex, but it does have certain advantages. As with glex, the variables themselves are ordered:

$$z_1 >_{\text{grevlex}} z_2 >_{\text{grevlex}} \cdots >_{\text{grevlex}} z_d. \tag{39.2.9}$$

For quadratic monomials, by contrast,

$$z_1^2 >_{\text{grevlex}} z_1z_2 >_{\text{grevlex}} z_2^2 >_{\text{grevlex}} z_1z_3 >_{\text{grevlex}} \cdots >_{\text{grevlex}} z_d^2. \tag{39.2.10}$$

Thus as one scans monomials of a fixed degree in descending grevlex order, one encounters later variables only after earlier ones have been "exhausted".

39.2.4 Labeling Based on Ordering

We will now describe how monomial ordering can be used to assign a label or index to each monomial.[1] Basically, the idea is to list the monomials in some sequence, and then label the monomials by where they occur in the list.

Consider, as a case of special interest, monomials in 6 variables as in (2.1). We have found it useful to list them in the arrangement shown in Table 2.1 below. Note that the monomials are graded: First the monomial of degree 0 appears, next those of degree 1, then those of degree 2, etc. Now let us, for the moment, regard each exponent $j = (j_1 \cdots j_6)$ as some six-digit number having the digits (reading from left to right) j_1 through j_6. Then observe that within each group of monomials of fixed degree these six-digit numbers appear in descending order as one reads down a column. For example, the monomials of degree 1 have the ordering $100000 > 010000 > 001000 > 000100 > 000010 > 000001$. Similarly, the monomials of degree 2 have the ordering $200000 > 110000 > 101000 > 100100 > \cdots$, etc. That is, for a given degree, the monomials appear in descending lex order. We will refer to this arrangement of monomials as *modified glex sequencing.*

Finally, each monomial has been given an *index*, starting with 0 for the monomial of degree 0, that increases by 1 for each successive monomial.[2] Thus, as the index increases, we first encounter the monomial of degree 0, then those of degree 1, then those of degree 2, etc. Furthermore, as the index *increases* within each set of monomials of a given degree, we encounter monomials in *descending* lex order. Note, as is easily verified by comparison of Tables 7.9.1 and 2.1, that there is the relation

$$\text{index } (00000m) = S(m, 6). \tag{39.2.11}$$

[1] Equivalently, we will assign a label to each point in a pyramid.

[2] This is the indexing scheme used in the program *MaryLie.* MaryLie is a program for charged-particle beam transport based on Lie algebraic methods. It is named in honor of Queen Henrietta Maria, patron of the English colony that was to become the state of Maryland, and Sophus Lie.

We remark that one might consider a true glex ordering in which the monomials of a given degree would occur in *ascending* lexicographic order as the index increased. According to (2.6) and (2.7), however, such an indexing scheme would list temporal variables first, and this order is not convenient for applications to accelerator physics.

There is, though, another arrangement that is satisfactory, and perhaps even superior: One could list the monomials by degree as before, and within each degree arrange them in descending reverse lexicographical (revlex) order. This could be called *modified grevlex sequencing*. Table 2.2 illustrates, for the case of 6 variables, an indexing scheme based on this procedure. Note that monomials with temporal variables now occur at the end of each monomial set of a given degree.

Table 39.2.1: Modified glex sequencing, a possible glex related indexing scheme for monomials in 6 variables.

Index	Exponents of		
	X P_x	Y P_y	τ P_τ
0	0 0	0 0	0 0
1	1 0	0 0	0 0
2	0 1	0 0	0 0
3	0 0	1 0	0 0
4	0 0	0 1	0 0
5	0 0	0 0	1 0
6	0 0	0 0	0 1
7	2 0	0 0	0 0
8	1 1	0 0	0 0
9	1 0	1 0	0 0
10	1 0	0 1	0 0
11	1 0	0 0	1 0
12	1 0	0 0	0 1
13	0 2	0 0	0 0
14	0 1	1 0	0 0
15	0 1	0 1	0 0
16	0 1	0 0	1 0
17	0 1	0 0	0 1
18	0 0	2 0	0 0
19	0 0	1 1	0 0
20	0 0	1 0	1 0
21	0 0	1 0	0 1
22	0 0	0 2	0 0
23	0 0	0 1	1 0
24	0 0	0 1	0 1
25	0 0	0 0	2 0
26	0 0	0 0	1 1
27	0 0	0 0	0 2
28	3 0	0 0	0 0
29	2 1	0 0	0 0
30	2 0	1 0	0 0
31	2 0	0 1	0 0
32	2 0	0 0	1 0
33	2 0	0 0	0 1
34	1 2	0 0	0 0
35	1 1	1 0	0 0
36	1 1	0 1	0 0
37	1 1	0 0	1 0
38	1 1	0 0	0 1
⋮			
77	0 0	0 1	2 0
78	0 0	0 1	1 1
79	0 0	0 1	0 2
80	0 0	0 0	3 0
81	0 0	0 0	2 1
82	0 0	0 0	1 2
83	0 0	0 0	0 3
⋮			

Table 39.2.2: A possible grevlex related indexing scheme for monomials in 6 variables.

Index	Exponents					
0	0	0	0	0	0	0
1	1	0	0	0	0	0
2	0	1	0	0	0	0
3	0	0	1	0	0	0
4	0	0	0	1	0	0
5	0	0	0	0	1	0
6	0	0	0	0	0	1
7	2	0	0	0	0	0
8	1	1	0	0	0	0
9	0	2	0	0	0	0
10	1	0	1	0	0	0
11	0	1	1	0	0	0
12	0	0	2	0	0	0
13	1	0	0	1	0	0
14	0	1	0	1	0	0
15	0	0	1	1	0	0
16	0	0	0	2	0	0
17	1	0	0	0	1	0
18	0	1	0	0	1	0
19	0	0	1	0	1	0
20	0	0	0	1	1	0
21	0	0	0	0	2	0
22	1	0	0	0	0	1
23	0	1	0	0	0	1
24	0	0	1	0	0	1
25	0	0	0	1	0	1
26	0	0	0	0	1	1
27	0	0	0	0	0	2
28	3	0	0	0	0	0
29	2	1	0	0	0	0
30	1	2	0	0	0	0
31	0	3	0	0	0	0
32	2	0	1	0	0	0
33	1	1	1	0	0	0
34	0	2	1	0	0	0
35	1	0	2	0	0	0
36	0	1	2	0	0	0
37	0	0	3	0	0	0
38	2	0	0	1	0	0
⋮						
77	0	0	0	0	2	1
78	1	0	0	0	0	2
79	0	1	0	0	0	2
80	0	0	1	0	0	2
81	0	0	0	1	0	2
82	0	0	0	0	1	2
83	0	0	0	0	0	3
⋮						

39.2.5 Formulas for Lowest and Highest Indices

For future reference, we note that any indexing scheme based on a graded ordering has the property that all the indices associated with monomials of a fixed degree are contiguous and lie within a fixed range. For example, reference to Tables 2.1 or 2.2 shows that (in the case of 6 variables) monomials of degree 2 begin at index 7 and end at index 27, and those of degree 3 begin at index 28 and end at index 83. Let *itop(ideg)* be the highest (top) index for monomials of degree *ideg*, and let *ibot(ideg)* be the lowest (bottom) index. Then we have the relations

$$itop(ideg) = S(ideg, d), \tag{39.2.12}$$

$$ibot(ideg) = itop(ideg - 1) + 1 = S[(ideg - 1), d] + 1. \tag{39.2.13}$$

Here d is the number of variables. Table 2.3 lists values of *ibot* and *itop* for the case of 6 variables ($d = 6$).

Table 39.2.3: Lowest and highest indices for monomials of degree *ideg* in 6 variables.

ideg	*ibot*	*itop*
0	0	0
1	1	6
2	7	27
3	28	83
4	84	209
5	210	461
6	462	923
7	924	1715
8	1716	3002
9	3003	5004
10	5005	8007
11	8008	12375
12	12736	18563

39.2.6 The Giorgilli Formula

By construction, use of Table 2.1 (or Table 2.2) assigns a unique index i to each monomial exponent j. Moreover, i takes on every possible (non-negative) integer value. Therefore, there is a invertible function $i(j)$ that provides a 1-to-1 mapping between the integers and the exponent vectors j. To proceed further, it would be very useful to have an explicit formula for $i(j)$. Such a formula, which we will call the *Giorgilli* formula, exists. We will illustrate it for the case of 6 variables with monomials indexed as in Table 2.1. In this case the exponent vectors have the form $j = (j_1, \cdots, j_6)$. We begin by defining the integers

$$n(\ell; j_1, \cdots, j_6) = \ell - 1 + \sum_{k=0}^{\ell-1} j_{6-k} \tag{39.2.14}$$

for $\ell \in \{1, 2, \cdots 6\}$. Then to the general monomial z^j we assign the index

$$i(j) = i(j_1, \cdots j_6) = \sum_{\ell=1}^{6} \text{ Binomial } [n(\ell; j_1, \cdots j_6), \ell]. \tag{39.2.15}$$

Here the quantities

$$\text{Binomial } [n, \ell] = \left(\begin{array}{c} n \\ \ell \end{array} \right) = \left\{ \begin{array}{ccc} \frac{n!}{\ell!(n-\ell)!} & , & 0 \leq \ell \leq n \\ 0 & , & \text{otherwise} \end{array} \right\} \tag{39.2.16}$$

denote the usual binomial coefficients.

39.2.7 Finding the Required Binomial Coefficients

At this point something needs to be said about what binomial coefficients are actually required and how they can be computed. Note that the formula (2.15) can be written in the form

$$i(j_1, \cdots, j_6) = \text{ Binomial } [n(1; j_1, \cdots, j_6), 1] + \sum_{\ell=2}^{6} \text{ Binomial } [n(\ell; j_1, \cdots j_6), \ell]. \quad (39.2.17)$$

From (2.14) we obtain

$$n(1; j_1, \cdots j_6) = j_6, \quad (39.2.18)$$

and we know that

$$\text{Binomial } [j_6, 1] = j_6. \quad (39.2.19)$$

Thus we may write (2.15) as

$$i(j_1, \cdots, j_6) = j_6 + \sum_{\ell=2}^{6} \text{ Binomial } [n(\ell; j_1, \cdots j_6), \ell]. \quad (39.2.20)$$

Now let *maxdeg* be the maximum degree of the polynomials being stored. Then we have the inequality

$$\sum_{k=0}^{\ell-1} j_{6-k} \leq maxdeg. \quad (39.2.21)$$

Consequently, according to (2.14), $n(\ell; j_1, \cdots, j_6)$ must lie in the range

$$n \in [\ell - 1, \ell - 1 + maxdeg]. \quad (39.2.22)$$

We therefore need only those binomial coefficients Binomial $[n, \ell]$ with $\ell \in [2, 6]$ and, for each ℓ, values of n lying in the range (2.22).

As an example, Exhibit 2.1 shows a program that computes and stores the required binomial coefficients for the case $maxdeg = 6$. It uses the recursion relation (7.3.56), and needs to be executed only once.

Exhibit 32.2.1: A program to compute and store binomial coefficients.

```
      subroutine binom5
c
c     computes a table of the binomial coefficients
c
      implicit double precision (a-h,o-z)
      integer bin5(24,20)
      common /bin5/ bin5
      save/bin5/
      do 1 i=1,20
      bin5(i,1)=i
      do 1 k=2,20
      if (i-k) 2,3,4
    2 bin5(i,k)=0
      go to 1
    3 bin5(i,k)=1
      go to 1
    4 ip=i-1
      kp=k-1
      bin5(i,k)=bin5(ip,kp)+bin5(ip,k)
    1 continue
      return
      end
```

39.2.8 Computation of the Index i Given the Exponent Array j

We will prove eventually that the formula (2.15) does indeed produce the indexing scheme of Table 2.1. Before doing so we will exhibit a computer program that computes $i(j)$. As an example, Exhibit 2.2 shows a computer program that computes $i(j)$ using (2.20) and stored binomial coefficients, and assuming $maxdeg = 6$. For efficiency, the required binomial coefficients have been hard-wired in, and arranged in a convenient order, with the use of a *data* statement. Alternatively, they could have been computed and rearranged in advance, using a variant of the program shown in Exhibit 2.1, and then stored in a common block for use in the program of Exhibit 2.2.

Note that the program requires various binomial lookups and 10 integer adds to compute an index for the case of 6 variables. In the general case of d variables, computing an index requires $2(d-1)$ such adds. The program is therefore reasonably fast.

Exhibit 32.2.2: A program to compute the index $i(j)$ using (2.20) and stored binomial coefficients.

```
      subroutine ndex(j,ind)
c
c  This subroutine calculates the MaryLie index ind, given j1 through j6
c  which are stored in the array j, based on the Giorgilli formula.
c  The use of rearrangement in this algorithm is due to Liam Healy.
c  AJD 6/20/95
c
      integer j(6)
c  cord = cumulative order: sum of exponents from ib-th to 6th.
      integer cord
c  obin = binomial coefficients rearranged to speed up calculation
c         obin(m,i)=Binomial[(m+6-i),(7-i)]
      integer obin(0:6,5)
      save obin
      data obin
     & /0,1,7,28,84,210,462,
     &  0,1,6,21,56,126,252,
     &  0,1,5,15,35, 70,126,
     &  0,1,4,10,20, 35, 56,
     &  0,1,3, 6,10, 15, 21/
c
c  calculate the index
      ind=j(6)
      cord=ind
      do 100 ib=5,1,-1
        cord=cord+j(ib)
  100   ind=ind+obin(cord,ib)
c
      return
      end
```

39.2.9 Preparing a Look-Up Table for the Exponent Array j Given the Index i

Given any exponent array j, we have seen how to compute a corresponding index $i(j)$. There is also the inverse problem: Given an index i, find the exponent array

$$j(i) = \{j_1(i), j_2(i), \cdots j_d(i)\} \tag{39.2.23}$$

that corresponds to this index. From a computational perspective, the most efficient procedure is to prepare a look-up table (a rectangular array) that contains this information.

One way to construct an appropriate look-up table involves finding an algorithm for generating—in the general case of d variables—a modified glex sequence of exponents. With such an algorithm we can produce a look-up table of exponents simply by storing successive d-tuples of exponents as they are generated. In particular, we can initialize the index by setting $i = 0$ for the first exponent vector, $(0, 0, \cdots, 0)$, and then increment i by 1 with each successive exponent vector in the sequence. We now outline a method for generating a modified glex sequence.

As described earlier, one may view the exponents j_1 through j_d that define a particular monomial as the components of a vector $j = (j_1, \cdots j_d)$. Let us refer to any given sequence of such vectors simply as a *list*. Now look at Table 2.1 or recall the definition of glex ordering to see that the list of monomials of a given degree m always begins with the vector

$$j = (m, 0, \cdots 0) \tag{39.2.24}$$

and ends with the vector

$$j = (0, \cdots 0, m). \tag{39.2.25}$$

In addition, the very next element in the list after (2.25) is the vector

$$j = (m + 1, 0, \cdots 0), \tag{39.2.26}$$

which begins the list of monomials of degree $m + 1$. We conclude that it is easy to specify the monomials that begin and end the degree m portion of a modified glex sequence, and to make the transition from the last monomial of degree m to the first one of degree $m + 1$.

We now seek a rule that converts any given j vector, for some monomial of degree m, into the j vector for the next monomial in the list. If the j vector has the form (2.25), then obviously the next j vector has the form (2.26). Otherwise, it is evident that some of the entries in j must be increased and some decreased in such a way as to keep the total degree constant. Moreover, to achieve a modified glex sequence, the entries on the "left end" of j (those j_k with smaller k) should be decreased as little as possible, and the entries on the "right end" of j (those j_k with larger k) should be increased as much as possible. A careful examination of the entries in Table 2.1 shows that one may convert from any j vector not of the form (2.25) to the next j vector via the following sequence of steps:

- Store the value of j_d as $icarry$, and then set $j_d = 0$.
- Test the j_k from right to left to find the right-most non-zero j_k. Let ℓnzj be the subscript for this j_k. (Here ℓnzj is a mnemonic for "last non-zero j".) Thus j_ℓ, with $\ell = \ell nzj$, is the last non-zero component of j.
- Decrease j_ℓ by 1, set $j_{\ell+1} = (1 + icarry)$, and leave intact all other entries of j.

The result of these steps is the next j in the list.

For example, consider the j in Table 2.1 having index 33, $j = (200001)$. In this case we have $icarry = 1$, and setting $j_6 = 0$ yields the vector (200000). We then find that the right-most non-zero entry is j_1; hence $\ell nzj = 1$ and (with $\ell = \ell nzj = 1$) $j_\ell = j_1 = 2$. Decreasing j_1 by 1 and replacing $j_{\ell+1} = j_2$ by $(1 + icarry)$, we find the new vector (120000). Examination of Table 2.1 shows that this vector has index 34, as desired. Readers are invited to check other cases in Table 2.1 to satisfy themselves that this procedure works in general.

Exhibit 2.3 shows a routine that carries out the algorithm just described in the case of 6 variables and assuming $maxdeg = 4$. This routine has the further feature [resulting from the initialization of ℓ (ℓnzj) and the use of suitable 'if' statements] that it also automatically makes the transition from the last monomial of degree m to the first one of degree $m + 1$; and it does so by use of the same algorithm just described for generating successive j vectors elsewhere in the list. That is, the same algorithm also produces the transition from (2.25) to (2.26). Readers are also invited to check that this procedure works as claimed.

Exhibit 32.2.3: A program to produce a look-up table for the exponents $j(i)$.

```
      subroutine jtable
c
c  This program creates the look-up table jtbl based on a method of Liam Healy:
c        ind = monomial index and imax = maximum value of ind.
c        ipsv = phase space variable and id = number of phase space variables.
c        For example when id=6, ipsv= 1...id corresponds to X...P_t.
c        jtbl(ipsv,ind) is the exponent of phase space variable
c        'ipsv' (1 to id)  for monomial index 'ind' (1 to imax).
c        For example, when id=6, monomial number 109 is X*P_X*P_X*P_t.
c        Consequently, jtbl(1 to 6,109)=1,2,0,0,0,1.
c
      parameter (imax = 209, id=6)
      dimension jtbl(id,imax)
c  j = array of exponents
      dimension j(id)
c  initialize exponents
      data j/id*0/
c  icarry = temporarily stored value of j(id).
c  lnzj = last non-zero j
c
c  Sequentially create exponent table jtbl
c
      do 150 ind=1,imax
c  set quantities
        icarry=j(id)
        j(id)=0
        lnzj=0
c  search for last nonzero j
        do 100 ipsv=1,id-1
          if (j(ipsv).gt.0) lnzj=ipsv
  100   continue
c  find next set of exponents
        if (lnzj.gt.0) j(lnzj)=j(lnzj)-1
        j(lnzj+1)=1+icarry
c  store exponents in jtbl
        do 120 ipsv=1,id
          jtbl(ipsv,ind)=j(ipsv)
  120   continue
  150 continue
c
c write out table
      do 70 i=1,imax
      write(6,500) i,
     & jtbl(1,i),jtbl(2,i),jtbl(3,i),
     & jtbl(4,i),jtbl(5,i),jtbl(6,i)
  500 format (1h ,i4,2x,3(i2,1x,i2,2x))
   70 continue
c
      end
```

39.2.10 Verification of the Giorgilli Formula

The last task for this section, as promised, is to show that the formula (2.15) does indeed produce modified glex indexing. For this purpose, the reader is invited to examine Table 2.4 which displays a modified glex sequence for the simple case of 3 variables through terms of degree 4.

Consider the $N(m,d)$ monomials of degree m in d variables. We may view each of these monomials $(z_1^{j_1}z_2^{j_2}\cdots z_d^{j_d})$ as a product of two constituent monomials:

$$z_1^{j_1}z_2^{j_2}\cdots z_d^{j_d} = z_1^{j_1} \times z_2^{j_2}\cdots z_d^{j_d}. \tag{39.2.27}$$

Thus, for example, monomials of degree two in three variables comprise three distinct groups (see Table 2.4):

1. z_1^2 times a monomial of degree zero in the two variables z_2 and z_3;

2. z_1^1 times monomials of degree one in the two other variables;

3. z_1^0 times monomials of degree two in the two other variables.

In general, of course, one may write all the monomials of order m in d variables as products between the monomial $z_1^{j_1}$—with $j_1 \in \{0,1,\ldots,m\}$—and the monomials of order $m-j_1$ in the remaining $d-1$ variables. The reader may observe that listing the monomials in a modified glex sequence—as in Table 2.4—makes clear the structure just described.

Consider now only those monomials of fixed degree m, and examine their exponents as listed in the modified glex sequence. The reader should note that because the exponents (of fixed degree) are listed in *descending* lexicographic order, eliminating the left-most column—the exponents j_1—will leave behind $d-1$ columns which contain exactly the modified glex sequence for the monomials in $d-1$ variables of degree m and smaller. Thus, for example, the transformation displayed in Figure 2.1 shows explicitly how this happens for degree-three monomials in three variables: removing the left column leaves behind the listing, in a modified glex sequence, of all monomials in two variables of degree zero through three.

As a consequence of the observation just described, the index of a given monomial can be determined by using a simple counting procedure, which we illustrate with the following example:

- Look at Table 2.4 and select an exponent, say $j=(1,0,2)$, whose index is to be determined. This exponent j represents a monomial of *degree* 3 ($|j| = j_1+j_2+j_3 = 1+0+2 = 3$) in 3 variables.

Table 39.2.4: Modified glex sequence for j in 3 variables.

Index	j_1	j_2	j_3
1	1	0	0
2	0	1	0
3	0	0	1
4	2	0	0
5	1	1	0
6	1	0	1
7	0	2	0
8	0	1	1
9	0	0	2
10	3	0	0
11	2	1	0
12	2	0	1
13	1	2	0
14	1	1	1
15	1	0	2
16	0	3	0
17	0	2	1
18	0	1	2
19	0	0	3
20	4	0	0
21	3	1	0
22	3	0	1
23	2	2	0
24	2	1	1
25	2	0	2
26	1	3	0
27	1	2	1
28	1	1	2
29	1	0	3
30	0	4	0
31	0	3	1
32	0	2	2
33	0	1	3
34	0	0	4

Figure 39.2.1: Sample extraction of a two-column array from a three-column array.

$$\begin{array}{ccc} 3 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & 0 & 1 \\ 1 & 2 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 2 \\ 0 & 3 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{array} \quad \mapsto \quad \begin{array}{cc} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 2 & 0 \\ 1 & 1 \\ 0 & 2 \\ 3 & 0 \\ 2 & 1 \\ 1 & 2 \\ 0 & 3 \end{array}$$

- Dropping the first entry from j yields a reduced exponent $j' = (0, 2)$, which represents a monomial of *degree* 2 ($|j'| = j_2 + j_3 = 0 + 2 = 2$) in 2 variables.

- Dropping the first two entries from j yields $j'' = (2)$, which represents a monomial of *degree* 2 ($|j''| = j_3 = 2$) in 1 variable.

- Record the *degrees* of j, j', and j'', as described in the previous steps. In the present case we obtain the numbers $|j| = 3$, $|j'| = 2$, and $|j''| = 2$, respectively. Based on these degrees, construct a "path" through Table 2.4, as illustrated in Figure 2.2: Begin the path at the top and proceed down the j_1 column until you reach exponents of degree $|j| = 3$. Then shift over one column and proceed down the j_2's until you reach exponents of degree $|j'| = 2$. Finally, shift over to the last column and proceed down the j_3's until you reach exponents of degree $|j''| = 2$.

- Now determine the index by the evident procedure of simply adding together the "lengths" of the vertical portions of the path just constructed, and then adding 1. The lengths are given by the relations

$$\begin{aligned} \text{length along } j_1 \text{ column} &= S(2, 3), && (39.2.28) \\ \text{length along } j_2 \text{ column} &= S_0(1, 2), && (39.2.29) \\ \text{length along } j_3 \text{ column} &= S_0(1, 1), && (39.2.30) \end{aligned}$$

and hence the index of the monomial $j = (1, 0, 2)$ is given by

$$i(j) = i(1, 0, 2) = S(2, 3) + S_0(1, 2) + S_0(1, 1) + 1. \qquad (39.2.31)$$

Using (1.6), we may write (2.31) in the more pleasing form

$$i(1, 0, 2) = S_0(2, 3) + S_0(1, 2) + S_0(1, 1) = 10 + 3 + 2 = 15, \qquad (39.2.32)$$

in agreement with the index given in Table 2.4 (or Figure 2.2).

Figure 39.2.2: Path to the exponent $j = (1, 0, 2)$ down the modified glex sequence in 3 variables.

Index	j_1	j_2	j_3	
1	1	0	0	
2	0	1	0	
3	0	0	1	
4	2	0	0	
5	1	1	0	$S(2,3) = 9 = S_0(2,3) - 1$
6	1	0	1	
7	0	2	0	
8	0	1	1	
9	0	0	2	
10	3	0	0	
11	2	1	0	$S_0(1,2) = 3$
12	2	0	1	
13	1	2	0	$S_0(1,1) = 2$
14	1	1	1	
15	1	0	2	1
16	0	3	0	
17	0	2	1	
18	0	1	2	
19	0	0	3	
20	4	0	0	
21	3	1	0	
22	3	0	1	
23	2	2	0	
24	2	1	1	
25	2	0	2	
26	1	3	0	
27	1	2	1	
28	1	1	2	
29	1	0	3	
30	0	4	0	
31	0	3	1	
32	0	2	2	
33	0	1	3	
34	0	0	4	

By generalizing the procedure of the example just described, we can now state a procedure for determining the index $i(j)$ for any monomial whose exponent list is given by $j = (j_1, j_2, \ldots, j_d)$.

1. For each $\nu \in \{1, \ldots, d\}$ define m_ν as the degree of the monomial obtained by dropping from the exponent list j the first $(\nu - 1)$ entries:
$$m_\nu = \sum_{k=\nu}^{d} j_k. \tag{39.2.33}$$
2. Then define i_ν as the total number of monomials in $d - (\nu - 1)$ variables which have degree *less* than m_ν:
$$i_\nu = S_0(m_\nu - 1, d - \nu + 1). \tag{39.2.34}$$
3. The index $i(j)$ is then given by the formula
$$i(j_1, \ldots, j_d) = \sum_{\nu=1}^{d} i_\nu. \tag{39.2.35}$$

Finally, we must demonstrate that the prescription just given for determining the index is indeed equivalent to the formula (2.15). To see this, note first that
$$m_\nu = \sum_{k=\nu}^{d} j_k = \sum_{k=0}^{d-\nu} j_{d-k}. \tag{39.2.36}$$
Then recall the definition (2.14) and simply compute:
$$\begin{aligned} i(j_1, \ldots, j_d) &= \sum_{\nu=1}^{d} i_\nu = \sum_{\nu=1}^{d} S_0(m_\nu - 1, d - \nu + 1) = \sum_{\nu=1}^{d} \begin{pmatrix} m_\nu + d - \nu \\ d - \nu + 1 \end{pmatrix} \\ &= \sum_{\nu=1}^{d} \begin{pmatrix} (\sum_{k=0}^{d-\nu} j_{d-k}) + d - \nu \\ d - \nu + 1 \end{pmatrix} = \sum_{\ell=1}^{d} \begin{pmatrix} (\sum_{k=0}^{\ell-1} j_{d-k}) + \ell - 1 \\ \ell \end{pmatrix} \\ &= \sum_{\ell=1}^{d} \begin{pmatrix} n(\ell; j) \\ \ell \end{pmatrix}. \end{aligned} \tag{39.2.37}$$
This result confirms that the formula (2.15) does indeed return the desired index.

Exercises

39.2.1. Verify that one can easily convert between the glex and grevlex orderings by a “double reversal” procedure: among each set of monomials of a given degree first reverse the order of the variables, then reverse the order of the monomials. For example, consider the monomials of degree 2 in 3 variables. Under glex ordering we have
$$z_1^2 > z_1 z_2 > z_1 z_3 > z_2^2 > z_2 z_3 > z_3^2. \tag{39.2.38}$$

Now reverse the order of the variables by making the replacement $z_1,\ z_2,\ z_3 \to z_3,\ z_2,\ z_1$, and replace $>$ by $<$. Then (2.38) becomes

$$z_3^2 < z_2z_3 < z_1z_3 < z_2^2 < z_1z_2 < z_1^2, \tag{39.2.39}$$

or, equivalently,

$$z_1^2 > z_1z_2 > z_2^2 > z_1z_3 > z_2z_3 > z_3^2. \tag{39.2.40}$$

Upon comparing (2.40) and (2.10), we see that (2.40) is in grevlex order.

39.3 Scalar Multiplication and Polynomial Addition

The use of any indexing scheme optimizes the operations of scalar multiplication and polynomial addition. Let $M_i(z)$ denote the generic monomial

$$M_i(z) = z^{j(i)}, \tag{39.3.1}$$

where it is assumed that the exponent j vectors (arrays) are indexed by an index i. Suppose f is any truncated power series. Then, since the M_i form a basis, there is a unique decomposition of the form

$$f = \sum_i f^i M_i \tag{39.3.2}$$

where the f^i are known coefficients. Indeed, the function f is stored by storing each coefficient f^i at the i^{th} location in some array.

Now suppose h is some other function that is related to f by scalar multiplication:

$$h = af \tag{39.3.3}$$

where a is some scalar. Then h has the decomposition

$$h = \sum_i h^i M_i \tag{39.3.4}$$

with the coefficients h^i given by the relation

$$h^i = af^i. \tag{39.3.5}$$

Thus, when any indexing scheme is employed, multiplication of a function by a scalar is equivalent to scalar multiplication of a vector.

Next suppose that f and g are any two polynomials, and we wish to compute the sum

$$h = f + g. \tag{39.3.6}$$

Then we have unique decompositions

$$f = \sum_i f^i M_i, \tag{39.3.7}$$

$$g = \sum_i g^i M_i, \tag{39.3.8}$$

$$h = \sum_i h^i M_i, \tag{39.3.9}$$

where the f^i and g^i are known coefficients, and the h^i are to be determined by (3.6). It follows, again from the fact that the M_i form a basis, that we have the relation

$$h^i = f^i + g^i. \tag{39.3.10}$$

Thus, when any indexing scheme is employed, polynomial addition is equivalent to the simple process of vector addition.

39.4 Polynomial Multiplication

Multiplication of polynomials is more complicated than addition. As before, define basis monomials $M_i(z)$ by (3.1). Also, suppose the polynomials f and g have the decompositions

$$f = \sum_i f^i M_i, \tag{39.4.1}$$

$$g = \sum_k g^k M_k, \tag{39.4.2}$$

and we wish to compute the product

$$h = fg. \tag{39.4.3}$$

The polynomial h will have the decomposition

$$h = \sum_\ell h^\ell M_\ell, \tag{39.4.4}$$

and the problem is two determine the h^ℓ from the relation

$$\begin{aligned} h &= \sum_\ell h^\ell M_\ell = fg = \sum_i f^i M_i \sum_k g^k M_k \\ &= \sum_{i,k} f^i g^k M_i M_k. \end{aligned} \tag{39.4.5}$$

We see that the basic problem consists of computing the products $M_i M_k$.

There are at least 3 ways to solve this problem in the context of indexing:

1. Given the indices i and k, find (say by table look-up) the corresponding exponents $j(i)$ and $j(k)$. Add these exponents as vectors to get the resultant "sum" exponent vector j^s,

   $$j^s = j(i) + j(k). \tag{39.4.6}$$

   Here we have reckoned with the obvious relation

   $$z^{j(i)} z^{j(k)} = z^{j(i)+j(k)} = z^{j^s}. \tag{39.4.7}$$

Next find the index ℓ corresponding to j^s. If the monomials have been indexed using modified glex sequencing, one can evaluate the Giorgilli formula (2.15) for this purpose to find

$$\ell = i(j^s). \tag{39.4.8}$$

Finally, increment h^ℓ by the quantity $f^i g^k$. (Here we have assumed that all the h^ℓ were initially set to zero.)

2. Given the indices i and k, find directly from a specially prepared look-up table the corresponding value of ℓ. Then increment h^ℓ by the quantity $f^i g^k$.

3. Given the index ℓ, use specially prepared *look-back* tables to find all indices i and k such that

$$M_i M_k = M_\ell. \tag{39.4.9}$$

Then increment h^ℓ by all the products $f^i g^k$.

All these methods will be discussed and compared in subsequent sections. At this point we simply remark that the computation of the products $M_i M_k$ is facilitated by the use of a *graded* indexing scheme. Since we are limiting our computations and their results to polynomials of degree less than or equal to m, we know that many of the products $M_i M_k$ need not be computed because their results are monomials of degree larger than m. The identification of such unneeded products is easy in any graded indexing scheme, because the degree of a product is the sum of the degrees of the factors.

39.5 Look-Up Tables

Let f and g be two polynomials. They can be decomposed into sums of homogeneous polynomials, and hence can be written in the form

$$f = f_0 + f_1 + f_2 + f_3 + f_4 + f_5 + f_6 + \cdots , \tag{39.5.1}$$

$$g = g_0 + g_1 + g_2 + g_3 + g_4 + g_5 + g_6 + \cdots . \tag{39.5.2}$$

Here f_m and g_m denote homogeneous polynomials of degree m. Correspondingly, the product fg of any two polynomials can also be organized into terms of common degree. Doing so gives the result

$$\begin{aligned} fg &= (f_0g_0)_0 + (f_0g_1 + g_0f_1)_1 + (f_0g_2 + g_0f_2 + f_1g_1)_2 \\ &+ (f_0g_3 + g_0f_3 + f_1g_2 + g_1f_2)_3 + (f_0g_4 + g_0f_4 + f_1g_3 + g_1f_3 + f_2g_2)_4 \\ &+ (f_0g_5 + g_0f_5 + f_1g_4 + g_1f_4 + f_2g_3 + g_2f_3)_5 \\ &+ (f_0g_6 + g_0f_6 + f_1g_5 + g_1f_5 + f_2g_4 + g_2f_4 + f_3g_3)_6 + \cdots . \end{aligned} \tag{39.5.3}$$

Here the terms appearing in fg have been collected according to their degrees using parentheses, and the parentheses have been given a subscript indicating the degree of the terms enclosed.

There is no particular problem in carrying out multiplications of the form f_0g_m and g_0f_m (since this operation is equivalent to scalar multiplication if the entries in g_m and f_m are viewed as the components of a vector). Multiplications of the form $f_m g_n$ and $g_m f_n$ with $m, n > 0$ are more complicated to execute.

Suppose we rearrange the computationally intensive terms (those not of the form $f_0 g_m$ and $g_0 f_m$) in (5.3) as shown below:

$$\begin{aligned}
&\text{computationally intensive terms } = \\
&(f_1g_1) + (f_1g_2 + f_1g_3 + f_1g_4 + f_1g_5 + \cdots) + (g_1f_2 + g_1f_3 + g_1f_4 + g_1f_5 + \cdots) \\
&+(f_2g_2) + (f_2g_3 + f_2g_4 + \cdots) + (g_2f_3 + g_2f_4 + \cdots) \\
&+(f_3g_3) + \cdots . \qquad (39.5.4)
\end{aligned}$$

If these multiplications were to be performed with the aid of a look-up table, what would this table look like, and how big would it be?

As a simple but instructive example, consider the case of polynomials through degree 4 in 2 variables ($m = 4$ and $d = 2$). Table 5.1 below shows, using phase-space notation, the monomials for this case listed in a modified glex sequence. We wish to multiply polynomials composed of these monomials, but only retain terms through degree 4.

Table 39.5.1: Modified glex sequence when $m = 4$ and $d = 2$.

i	j_1	j_2	$\vert j\vert$	monomial
1	1	0	1	X
2	0	1		P_x
3	2	0	2	X^2
4	1	1		XP_x
5	0	2		P_x^2
6	3	0	3	X^3
7	2	1		X^2P_x
8	1	2		XP_x^2
9	0	3		P_x^3
10	4	0	4	X^4
11	3	1		X^3P_x
12	2	2		$X^2P_x^2$
13	1	3		XP_x^3
14	0	4		P_x^4

Now consider the result of multiplying a monomial in f with index if and a monomial in g with index ig. The result will be some monomial in h with index ih. (Here, to simplify notation, we denote the relevant indices by if, ig, and ih rather than the i, k, ℓ of the previous Section.) Table 5.2 shows a multiplication table for this process. It gives, for each value of if and ig, the value of ih corresponding to the product monomial. Entries with an asterisk "*" correspond to monomials having degree greater than 4. See Table 5.1. These entries fall outside our interest. Consider, for example, the relation

$$X^2 \times XP_x = X^3P_x. \qquad (39.5.5)$$

Inspection of Table 5.1 shows that the monomials X^2 and XP_x have the indices $if = 3$ and $ig = 4$, respectively; and their product X^3P_x has the index $ih = 11$. Correspondingly, the entry in Table 5.2 for $if = 3$ and $ig = 4$ has the value $ih = 11$.

Table 39.5.2: Multiplication table (when $m = 4$ and $d = 2$) giving values of ih.

$if\backslash ig$	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	3	4	6	7	8	10	11	12	13	*	*	*	*	*
2	4	5	7	8	9	11	12	13	14	*	*	*	*	*
3	6	7	10	11	12	*	*	*	*	*	*	*	*	*
4	7	8	11	12	13	*	*	*	*	*	*	*	*	*
5	8	9	12	13	14	*	*	*	*	*	*	*	*	*
6	10	11	*	*	*	*	*	*	*	*	*	*	*	*
7	11	12	*	*	*	*	*	*	*	*	*	*	*	*
8	12	13	*	*	*	*	*	*	*	*	*	*	*	*
9	13	14	*	*	*	*	*	*	*	*	*	*	*	*
10	*	*	*	*	*	*	*	*	*	*	*	*	*	*
11	*	*	*	*	*	*	*	*	*	*	*	*	*	*
12	*	*	*	*	*	*	*	*	*	*	*	*	*	*
13	*	*	*	*	*	*	*	*	*	*	*	*	*	*
14	*	*	*	*	*	*	*	*	*	*	*	*	*	*

Note that Table 5.2 has the form of a *symmetric* matrix with 5 diagonal entries and 18 entries above the diagonal. This symmetry results from the commutativity of (ordinary) multiplication. Let $\{m, d\}$ denote the number of locations (look-up table size) required to store a multiplication table when this symmetry is taken into account. Then we have the result $\{4, 2\} = (5 + 18) = 23$. Table 5.3 lists values of $\{m, d\}$ for various values of m and d.

As will be seen in later sections, it may sometimes be advantageous to sacrifice the savings in storage associated with symmetry in order to gain speed. We should therefore also calculate the storage required when symmetry is ignored. Let $\{m, d\}^{ns}$ denote the number of locations required to store the *full* multiplication table when symmetry is not taken into account. (Here the superscript ns denotes *no symmetry.*) Then we have the result $\{4, 2\}^{ns} = (5 + 18 + 18) = 41$. Table 5.4 lists values of $\{m, d\}^{ns}$ for various values of m and d.

We close this section with a description of how the dimensions $\{m, d\}$ and $\{m, d\}^{ns}$ can be computed in a general case. Suppose, for example, we wish to know the dimensionality of the look-up table associated with the terms displayed in (5.4). (This is equivalent to retaining terms through degree $m = 6$.) Let us also consider the case of 6 variables, $d = 6$. Thus, we need to find $\{6, 6\}$ and $\{6, 6\}^{ns}$.

We begin with the case where symmetry is taken into account. Since the portion of the look-up table associated with (f_1g_1) is 6×6 and symmetric, we have the result

$$\{f_1g_1\} = [(6 \times 6) - 6)]/2 + 6 = (6 \times 7)/2 = 21. \tag{39.5.6}$$

Here we use the notation { } to denote the dimensionality of the associated portion of the look-up table. Similarly, we find the results

Table 39.5.3: Multiplication look-up table size for polynomials of degree 1 through m in various numbers of variables.

m	$\{m,2\}$	$\{m,3\}$	$\{m,4\}$	$\{m,5\}$	$\{m,6\}$	$\{m,7\}$	$\{m,8\}$	$\{m,9\}$
2	3	6	10	15	21	28	36	45
3	9	24	50	90	147	224	324	450
4	23	75	185	385	714	1218	1950	2970
5	45	180	525	1260	2646	5040	8910	14850
6	82	388	1309	3570	8400	17724	34386	62403
7	134	748	2905	8960	23520	54768	116226	229020
8	210	1354	5975	20655	60087	153615	355113	757185
9	310	2300	11475	44250	142065	397320	997425	2295150
10	445	3746	20941	89501	315546	961576	2612753	6470178
11	615	5852	36489	172116	663894	2197272	6444009	17131686
12	833	8869	61270	317366	1333976	4779320	15086409	42955185

m	$\{m,10\}$	$\{m,11\}$	$\{m,12\}$	$\{m,13\}$	$\{m,14\}$	$\{m,15\}$
2	55	66	78	91	105	120
3	605	792	1014	1274	1575	1920
4	4345	6149	8463	11375	14980	19380
5	23595	36036	53235	76440	107100	146880
6	107250	176176	278551	426244	634032	920040
7	424710	748748	1264627	2058784	3246320	4977600
8	1510795	2851563	5134090	8875802	14811930	23963370
9	4915625	9912760	18990530	34807500	61386150	104652000

m	$\{m,16\}$	$\{m,17\}$	$\{m,18\}$	$\{m,19\}$	$\{m,20\}$	$\{m,21\}$
2	136	153	171	190	210	231
3	2312	2754	3249	3800	4410	5082
4	24684	31008	38475	47215	57365	69069
5	197676	261630	341145	438900	557865	701316
6	1306212	1818813	2488962	3353196	4454065	5840758

$$\{f_2g_2\} = (21 \times 22)/2 = 231, \tag{39.5.7}$$

$$\{f_3g_3\} = (56 \times 57)/2 = 1596. \tag{39.5.8}$$

See Table 7.3.1. Next consider the portion of the look-up table required for $(f_1g_2 + f_1g_3 + f_1g_4 + f_1g_5)$. We write its dimensionality in the form

$$\begin{aligned}\{f_1g_2 + f_1g_3 + f_1g_4 + f_1g_5\} &= \{f_1\} \times \{g_2 + g_3 + g_4 + g_5\} \\ &= 6 \times (21 + 56 + 126 + 252) = 2730. \end{aligned} \tag{39.5.9}$$

Table 39.5.4: Multiplication look-up table size for polynomials of degree 1 through m in various numbers of variables when symmetry is not exploited.

m	$\{m,2\}^{ns}$	$\{m,3\}^{ns}$	$\{m,4\}^{ns}$	$\{m,5\}^{ns}$	$\{m,6\}^{ns}$	$\{m,7\}^{ns}$	$\{m,8\}^{ns}$	$\{m,9\}^{ns}$
2	4	9	16	25	36	49	64	81
3	16	45	96	175	288	441	640	891
4	41	141	356	750	1401	2401	3856	5886
5	85	351	1036	2500	5265	10045	17776	29646
6	155	757	2584	7085	16717	35329	68608	124587
7	259	1477	5776	17865	46957	109417	232288	457821
8	406	2674	11881	41185	119965	306901	709732	1513656
9	606	4566	22881	88375	283921	794311	1994356	4589586
10	870	7437	41757	178751	630631	1922361	5224220	12938355
11	1210	11649	72853	343981	1327327	4393753	12886732	34261371
12	1639	17655	122331	634271	2667029	9556925	30169816	85905366

m	$\{m,10\}^{ns}$	$\{m,11\}^{ns}$	$\{m,12\}^{ns}$	$\{m,13\}^{ns}$	$\{m,14\}^{ns}$	$\{m,15\}^{ns}$
2	100	121	144	169	196	225
3	1200	1573	2016	2535	3136	3825
4	8625	12221	16836	22646	29841	38625
5	47125	71995	106380	152776	214081	293625
6	214215	351989	556648	851929	1267385	1839265
7	849135	1497133	2528800	4117009	6491961	9954385
8	3020590	5701762	10266361	17749225	29620801	47922865
9	9830250	19824156	37979241	69612621	122769241	209300125

m	$\{m,16\}^{ns}$	$\{m,17\}^{ns}$	$\{m,18\}^{ns}$	$\{m,19\}^{ns}$	$\{m,20\}^{ns}$	$\{m,21\}^{ns}$
2	256	289	324	361	400	441
3	4608	5491	6480	7581	8800	10143
4	49216	61846	76761	94221	114500	137886
5	395200	523090	682101	877591	1115500	1402380
6	2611456	3636487	4976595	6704853	8906360	11679493

Here we use the notation $\{f_m\}$ and $\{g_m + \cdots + g_n\}$ to denote the dimensionality of the spaces of polynomials of degree m and degrees m through n, respectively. Similarly, we find the results

$$\begin{aligned}\{f_2g_3 + f_2g_4\} &= \{f_2\} \times \{g_3 + g_4\} \\ &= 21 \times (56 + 126) = 3822. \end{aligned} \tag{39.5.10}$$

Finally we note that, by symmetry, the evaluation of $(f_1g_2 + f_1g_3 + f_1g_4 + f_1g_5)$ and $(g_1f_2 + g_1f_3 + g_1f_4 + g_1f_5)$ can be done with the same look-up information, and similarly for $(f_2g_3 + f_2g_4)$ and $(g_2f_3 + g_2f_4)$, etc. Thus, we have covered all possibilities for polynomials of degree 1 through 6. We find that the total dimension of the look-up table for such polynomials (in 6 variables) is given by the sum

$$\{6, 6\} = 21 + 231 + 1596 + 2730 + 3822 = 8400. \tag{39.5.11}$$

Consider next the case where symmetry is ignored. Then (for $d = 6$) we have the results

$$\{f_1g_1\}^{ns} = 6 \times 6 = 36, \tag{39.5.12}$$

$$\{f_2g_2\}^{ns} = 21 \times 21 = 441, \tag{39.5.13}$$

$$\{f_3g_3\}^{ns} = 56 \times 56 = 3136. \tag{39.5.14}$$

The quantities (5.9) and (5.10) remain unchanged, but equal numbers of storage locations must now be allocated for their reversed counterparts. Consequently, we find that the total number of storage locations required through degree 6 (in 6 variables) when symmetry is not exploited is given by the relation

$$\{6, 6\}^{ns} = 36 + 441 + 3136 + 2 \times 2730 + 2 \times 3822 = 16,717. \tag{39.5.15}$$

We see that this number is slightly less than twice (5.11).

It is remarkable how much the decision to retain only terms through degree m (e.g. ignore the * terms in Table 5.2) affects the size of the multiplication look-up table. Consider the case of polynomials of degree 1 through 12 in 6 variables. According to Table 7.9.1 there are

$$S(12, 6) = 18,563 \tag{39.5.16}$$

basis monomials in this case. Thus one might naively expect that

$$18,563 \times 18,563 \simeq 3.4 \times 10^8 \tag{39.5.17}$$

storage locations would be required for a multiplication look-up table. This number is prohibitively large. However, examination of Table 5.4 gives the much smaller result

$$\{12, 6\}^{ns} = 2,667,029. \tag{39.5.18}$$

According to (2.10) and (5.16) the largest index in the case $m = 12$ and $d = 6$ is 18,563. We note that 2^{15} = 32,768. Thus, the indices could all be stored in 2 byte entries, and the storage required by (5.18) would be approximately 5.3 megabytes.

Exercises

39.5.1.

39.6 Scripts

Suppose we wish to carry out some calculation that potentially involves a great deal of index manipulation, and suppose we have some simple-minded method for doing so. What we can then do is carry out this simple-minded method, while at the same time making and keeping a *record* of the *results* of the *index manipulation*. We will call this record a *script*. Then, any time we wish to repeat the calculation, we can bypass the need for index manipulation by use of the script.

As an example of this approach, consider the problem of multiplying two polynomials f and g to find the product

$$h = fg. \tag{39.6.1}$$

For simplicity, we consider only the computationally intensive terms (5.4). Exhibit 6.1 shows a simple-minded procedure for doing so.

Exhibit 32.6.1: Simple-minded program for polynomial multiplication.

```
c  Loop over degree ifdeg and indices if and ig.
      do 10 ifdeg=1,maxdeg-1
      do 20 if=ibot(ifdeg),itop(ifdeg)
      do 30 ig=1,itop(maxdeg-ifdeg)
c  Look up exponent vectors corresponding to if and ig and add them.
      do l=1,6
      jf(l)=jtbl(l,if)
      jg(l)=jtbl(l,ig)
      jsum(l)=jf(l)+jg(l)
      end do
c  Find the index of vector sum.
      call ndex(jsum,ih)
c  Carry out multiplication, and store result in h(ih).
      h(ih)=h(ih)+f(if)*g(ig)
   30 continue
   20 continue
   10 continue
c
      return
      end
```

What is shown is a fragment of a FORTRAN program with various parameter and dimension statements omitted. The procedure employs a triple loop that goes through all relevant degrees in f and all relevant index pairs if and ig. Suppose *maxdeg* is the maximum degree of the monomials we wish to retain. Then, in the factor f, we only need work with monomials whose degree *ifdeg* lies in the range $1 \le ifdeg \le maxdeg - 1$. Note that exponents (and hence degrees) add under the operation of multiplication; and recall that we are only considering computationally intensive terms so that in both the factors f, g only terms of degree 1 and higher appear. Correspondingly, when working with terms of degree *ifdeg* in the factor f, the only terms in g that are required are those whose degree *igdeg* lies in the range $1 \leq igdeg \leq maxdeg - ifdeg$. All these considerations are implemented conveniently, as shown, with the use of the arrays *ibot* and *itop*. See Section 32.2.5.

For each *if*, *ig* pair the following operations are performed:

1. Look up the corresponding exponents jf and jg from a previously prepared and stored table *jtbℓ*.

2. Add these exponents to find the resulting exponent.

3. Compute the index *ih* corresponding to this exponent.

4. Carry out the multiplication of the monomial coefficients from f and g, and increment the monomial coefficient of h corresponding to the index *ih* found in step 3 by the resulting product.

Evidently this procedure requires numerous look-ups (step 1), numerous additions (step 2), and numerous calls to an index computation routine (in this case the subroutine *ndex*, see Exhibit 2.2).

Exhibit 6.2 shows the same routine except that step 4 above is replaced by storage of the relevant index triplets if, ig, and ih. They are stored sequentially in terms of a “counting” index *ic*. The result of running this routine is the arrays (look-up tables) *iftbℓ*, *igtbℓ*, and *ihtbℓ*. The size of each of these arrays is $\{m, d\}^{ns}$ with $m = maxdeg$ and (in this example) $d = 6$. The array *icmin*, whose purpose will be described later, is also filled.

```
Exhibit 32.6.2: Program for preparation of script for polynomial multiplication.
This program prepares the tables iftbl, igtbl, ihtbl, and records their size,
icmax.  It also fills the array icmin.

c  Initialize counter.
      ic=0
c  Loop over degree ifdeg and indices if and ig.
      do 10 ifdeg=1,maxdeg-1
      do 20 if=ibot(ifdeg),itop(ifdeg)
      do 30 ig=1,itop(maxdeg-ifdeg)
c  Look up exponent vectors corresponding to if and ig and add them.
      do l=1,6
      jf(l)=jtbl(l,if)
      jg(l)=jtbl(l,ig)
      jsum(l)=jf(l)+jg(l)
      end do
c  Find the index of vector sum.
      call ndex(jsum,ih)
c  Update counter and fill tables.
      ic=ic+1
      iftbl(ic)=if
      igtbl(ic)=ig
      ihtbl(ic)=ih
c  Fill icmin.
      if(ig .eq. 1) icmin(if)=ic
  30  continue
  20  continue
  10  continue
c  Set icmax
      icmax=ic
c
      return
      end
```

The arrays *iftbℓ*, *igtbℓ*, and *ihtbℓ* can now be used as a script to carry out multiplication. Exhibit 6.3 shows how. We see, by comparing Exhibits 6.1 and 6.3, that the triple loop has been replaced by a single loop, and all index and exponent manipulations and calculations are replaced by table look up. Indeed, the routine consists entirely of the actual calculation to be performed (step 4 above) and table look up. Evidently this program should run far faster than that of Exhibit 6.1. The price paid for this speed is the storage required for the tables *iftbℓ*, *igtbℓ*, and *ihtbℓ*.

```
Exhibit 32.6.3: Program for polynomial multiplication using a script.

      do 10 ic=1,icmax
      if=iftbl(ic)
      ig=igtbl(ic)
      ih=ihtbl(ic)
  10  h(ih)=h(ih)+f(if)*g(ig)
      end

An even more compact program for polynomial multiplication using a script.

      do 10 ic=1,icmax
  10  h(ihtbl(ic))=h(ihtbl(ic))+f(iftbl(ic))*g(igtbl(ic))
      end
```

Now that we understand the basic idea of how a script works, we should seek to optimize the procedure. In particular, the tables *iftbℓ* and *igtbℓ* are not actually necessary. This is good because their size, $\{m, d\}^{ns}$, can be quite large.

Here the arrays *ibot* and *itop* can again be utilized. Consider the program fragment shown in Exhibit 6.4. Evidently, by construction, this code will produce the *same* set of *if* and *ig* values, and in the *same order*, as those stored in the routine of Exhibit 6.2 and used in the routine of Exhibit 6.3. Thus, the need for the tables *iftbℓ* and *igtbℓ* has been eliminated.

```
Exhibit 32.6.4: Program for producing if and ig pairs using less storage and
fewer look ups.

      do 10 ifdeg=1,maxdeg-1
      do 20 if=ibot(ifdeg),itop(ifdeg)
      do 30 ig=1,itop(maxdeg-ifdeg)
      .
      .
      .
   30 continue
   20 continue
   10 continue
      end
```

We are now ready to reap the fruit of our deliberations. We know that the routine of Exhibit 6.4 produces the same if and ig values, and in the same order, as those stored in the routine of Exhibit 6.2 and used in the multiplication routine of Exhibit 6.3. Also, we see from Exhibit 6.2 that the pointer ic is incremented by 1 each time an if, ig pair is produced. As a result of these facts, we may write a multiplication routine equivalent to that of Exhibit 6.3 using the loop structure of Exhibit 6.4. This multiplication routine is shown in Exhibit 6.5 below. We see that this routine uses the table $ihtb\ell$, as does the routine of Exhibit 6.3, but does not use the tables $iftb\ell$ and $igtb\ell$. As noted earlier, this is a considerable savings in storage since these tables may be large.

Exhibit 32.6.5: Program for polynomial multiplication using only one index look-up table.

```
      ic=1
      do 10 ifdeg=1,maxdeg-1
      do 20 if=ibot(ifdeg),itop(ifdeg)
      do 30 ig=1,itop(maxdeg-ifdeg)
      h(ihtbl(ic))=h(ihtbl(ic))+f(if)*g(ig)
      ic=ic+1
   30 continue
   20 continue
   10 continue
      end
```

There is one last possible improvement to be considered. Suppose one of the factors in (3.1), say f, is sparse. This is often the case in problems of practical interest. In this case, the program shown in Exhibit 6.5 can be improved to exploit sparseness in f. To see how this may be done, consider the contents of the arrays $iftb\ell$, $igtb\ell$, and $ihtb\ell$. Table 6.1, for example, shows the contents of these arrays in the case $m = maxdeg = 4$ and $d = 2$. Inspection of Table 6.1 shows that for each value of if there is a minimum value of ic for which $iftb\ell(ic) = if$. Evidently, for a given value of if, the minimum value of ic occurs when $ig = 1$. Suppose these values are put in an array which we will call $icmin(if)$. For example, Table 6.2 shows such an array in the case $m = 4$ and $d = 2$. Finally, inspection of the code in Exhibit 6.2 shows how the array $icmin$ can be constructed in general, and in particular for the case $d = 6$.

Table 39.6.1: Contents of the arrays $iftb\ell$, $igtb\ell$, and $ihtb\ell$ in the case $m = 4$ and $d = 2$.

ic	$iftb\ell$	$igtb\ell$	$ihtb\ell$
1	1	1	3
2		2	4
3		3	6
4		4	7
5		5	8
6		6	10
7		7	11
8		8	12
9		9	13
10	2	1	4
11		2	5
12		3	7
13		4	8
14		5	9
15		6	11
16		7	12
17		8	13
18		9	14
19	3	1	6
20		2	7
21		3	10
22		4	11
23		5	12
24	4	1	7
25		2	8
26		3	11
27		4	12
28		5	13
29	5	1	8
30		2	9
31		3	12
32		4	13
33		5	14
34	6	1	10
35		2	11

ic	*iftbℓ*	*igtbℓ*	*ihtbℓ*
36	7	1	11
37		2	12
38	8	1	12
39		2	13
40	9	1	13
41		2	14

Table 39.6.2: Contents of the array *icmin* in the case $m = 4$ and $d = 2$.

if	*icmin*
1	1
2	10
3	19
4	24
5	29
6	34
7	36
8	38
9	40

Now consider the program shown in Exhibit 6.6 below. It tests the factor $f(if)$ before going into the do loop over ig. This whole loop is skipped if $f(if)$ is zero. If this loop were ever skipped in the program of Exhibit 6.5, the value of ic would be upset because it is incremented within this loop. However, if ic is properly set each time the program goes into this loop, which is what use of the array *icmin* does, then ic always has the proper value even if this loop has possibly been skipped for some earlier values of if.

Exhibit 32.6.6: Program for polynomial multiplication using only one index look-up table and designed to exploit possible sparseness in the factor f.

```
      do 10 ifdeg=1,maxdeg-1
      do 20 if=ibot(ifdeg),itop(ifdeg)
      if(f(if) .ne. 0.d0) then
      ic=icmin(if)
      do 30 ig=1,itop(maxdeg-ifdeg)
      h(ihtbl(ic))=h(ihtbl(ic))+f(if)*g(ig)
      ic=ic+1
   30 continue
      endif
   20 continue
   10 continue
      end
```

Let us briefly compare the theoretical speeds of the programs shown in Exhibits 6.5 and 6.6. We see that the program in Exhibit 6.6 makes $ifmax = itop(maxdeg - 1) =$

$S(maxdeg-1,d)$ “if” tests. See (2.20). When an if test is passed, there is the additional burden of an $icmin(if)$ look up. However, when an if test fails, there is a savings of $igmax = itop(maxdeg - ifdeg)$ additions and multiplies as well as the overhead and other operations associated with the ig loop. Since $igmax$ can often be large, and multiplications are slow, we conclude (providing the if test burden is not too large) that the program of Exhibit 6.6 should be significantly faster than that of Exhibit 6.5 if f is sparse, and only slightly slower if f is dense.

In many applications both f and g are known to be homogeneous. Homogeneity may be viewed as a kind of sparseness, and this sparseness can also be exploited to produce a still faster multiplication routine. Exhibit 6.7 shows such a routine. Note that it is now necessary to offset the counter ic by the amount *icoff* to take into account the fact that the ig loop now begins at $ig = ibot(igdeg)$ rather than $ig = 1$.

Exhibit 32.6.7: Program for polynomial multiplication using only one index look-up table and designed to exploit the fact that f and g are homogeneous of degrees ifdeg and igdeg, respectively. It is also designed to exploit possible additional sparseness in the factor f.

```
      icoff=itop(igdeg-1)
      do 10 if=ibot(ifdeg),itop(ifdeg)
      if(f(if) . ne. 0.d0) then
      ic=icmin(if)+icoff
      do 20 ig=ibot(igdeg),itop(igdeg)
      h(ihtbl(ic))=h(ihtbl(ic))+f(if)*g(ig)
      ic=ic+1
   20 continue
      endif
   10 continue
      end
```

Suppose it is known in advance which entries in f are nonzero. Then it is possible to eliminate the if statements in programs like those in Exhibits 6.6 and 6.7 with a possible improvement in computational speed—particularly in the case of vector or pipe-lined computer architecture for which the if test burden is relatively large. If (exactly) k entries in f are known to be *nonzero*, then we can set up an array $nzf(i)$ such that the values $nzf(i)$ with $i \in [1,k]$ are the indices for the nonzero entries in f. With this array in hand we can, for example, reformulate the routine of Exhibit 6.7 as shown in Exhibit 6.8. Here we assume, as in Exhibit 6.7, that f is homogeneous of degree $ifdeg$. We remark that by the introduction of still further arrays it is possible to exploit known sparseness in both f and g.

Exhibit 32.6.8: Program for polynomial multiplication using only one look-up table, known sparseness in f, and homogeneity in f and g.

```
      icoff=itop(igdeg-1)
      do 10 i=1,k
      if=nzf(i)
      ic=icmin(if)+icoff
      do 20 ig=ibot(igdeg),itop(igdeg)
      h(ihtbl(ic))=h(ihtbl(ic))+f(if)*g(ig)
```

```
   ic=ic+1
20 continue
10 continue
   end
```

Note that the look-up table *ihtbℓ* in Exhibits 6.5 through 6.8 has size $\{m,d\}^{ns}$. We close this section by commenting that it is also possible to write multiplication routines that employ a look-up table having the minimum size $\{m,d\}$. That is, it is possible to write a routine that exploits the commutative symmetry of multiplication. However, we have not found a way of doing so that simultaneously exploits sparseness.

Exercises

39.6.1.

39.7 Look-Back Tables

Consider the array $ihtb\ell(ic)$ of ih values described in Section 6. See, for example, the right column of Table 6.1. For each value of ic ranging from 1 through $icmax = \{m,d\}^{ns}$, there is a corresponding value of ih, and this value lies in the range $(d+1)$ through $S(m,d)$. (Note that there are no zero or first order monomials under consideration since we are only worried about computationally intensive terms.) Since $\{m,d\}^{ns} > S(m,d)$, there are generally many ic values that yield a given value of ih. Indeed, in (6.1) there are many factors $f(if)$ and $g(ig)$ that contribute to a given $h(ih)$.

Next look at the program in Exhibit 6.3. We see that ic runs *successively* through the values 1,2,$\cdots icmax$. However, we also see that the same result would be achieved if ic ran through the values 1,2,$\cdots icmax$ in *any* order. That is, the outcome of following the script is *independent* of the order in which its instructions are executed. Suppose the array $ihtb\ell(ic)$ is rearranged so that the entries are listed in order of increasing ih. At the same time we rearrange the arrays $iftb\ell$ and $igtb\ell$. Finally, we set up a new ic index that again runs successively through the values 1,2,$\cdots icmax$. For example, Table 7.1 below shows the result of this rearrangement applied to Table 6.1.

Examine Table 7.1. We see that for each value of ih in the column $ihtb\ell$ there are corresponding values of if and ig in the columns $iftb\ell$ and $igtb\ell$, respectively. These if, ig pairs are the indices for the monomials that are *factors* of the monomial labelled by ih. Thus, the arrays $iftb\ell$ and $igtb\ell$ (after the rearrangement just described) provide what we have called look-back tables. That is, given some ih, we can look back using these tables to find the if, ig pairs that produced this ih. We note that each table has $\{m,d\}^{ns}$ entries.

These look-back tables can be used to construct a program for multiplication. To do this, we make some observations. We see that for each value of ih in the column $ihtb\ell$ there is a minimum (bottom) value of the variable new ic, call it $icbot$, and a maximum (top) value, call it $ictop$. Use these observations to construct two arrays: $icbot(ih)$ and $ictop(ih)$. For example, Table 7.2 shows the contents of these arrays in the case $m=4$ and $d=2$.

Also, we see that ih has the minimum value $ihmin$ given by the relation

$$ihmin = d + 1, \tag{39.7.1}$$

and a maximum value $ihmax$ given by

$$ihmax = S(m, d) \tag{39.7.2}$$

with, in this case, $m = maxdeg = 4$ and $d = 2$.

Table 39.7.1: The result of rearranging Table 6.1 in order of increasing ih.

new ic	old ic	$iftb\ell$	$igtb\ell$	$ihtb\ell$
1	1	1	1	3
2	2	1	2	4
3	10	2	1	
4	11	2	2	5
5	3	1	3	6
6	19	3	1	
7	4	1	4	7
8	12	2	3	
9	20	3	2	
10	24	4	1	
11	5	1	5	8
12	13	2	4	
13	25	4	2	
14	29	5	1	
15	14	2	5	9
16	30	5	2	
17	6	1	6	10
18	21	3	3	
19	34	6	1	
20	7	1	7	11
21	15	2	6	
22	22	3	4	
23	26	4	3	
24	35	6	2	
25	36	7	1	
26	8	1	8	12
27	16	2	7	
28	23	3	5	
29	27	4	4	
30	31	5	3	
31	37	7	2	
32	38	8	1	
33	9	1	9	13
34	17	2	8	
35	28	4	5	

new ic	old ic	$iftb\ell$	$igtb\ell$	$ihtb\ell$
36	32	5	4	
37	39	8	2	
38	40	9	1	
39	18	2	9	14
40	33	5	5	
41	41	9	2	

Table 39.7.2: The arrays *icbot* and *ictop* in the case $m = 4$ and $d = 2$.

ih	$icbot$	$ictop$
3	1	1
4	2	3
5	4	4
6	5	6
7	7	10
8	11	14
9	15	16
10	17	19
11	20	25
12	26	32
13	33	38
14	39	41

Now look at Exhibit 7.1. It shows a program for polynomial multiplication using the arrays $iftb\ell$ and $igtb\ell$ as look-back tables. (Note that, although we have used the same notation, here the tables $iftb\ell$ and $igtb\ell$ are *rearranged* versions of their original counterparts.) We see that, with the use of the arrays *icbot* and *ictop*, the program ranges over all the proper values of *ic*, and therefore from our previous discussion must give the same result as the program of Exhibit 6.3.

Exhibit 32.7.1: Program for polynomial multiplication using look-back tables.

```
      do 10 ih=ihmin,ihmax
      do 20 ic=icbot(ih),ictop(ih)
      h(ih)=h(ih)+f(iftbl(ic))*g(igtbl(ic))
   20 continue
   10 continue
      end
```

How do the programs shown in Exhibits 6.5 and 7.1 compare? Here are some reasons to believe that (for multiplication) the use of look-up tables is preferable to the use of look-back tables:

1. Examine Table 6.1. Evidently successive values of $iftb\ell(ic)$ and $igtb\ell(ic)$ are *contiguous* as one goes down the list (increments ic). Therefore successive values of $f[iftb\ell(ic)]$ and $g[igtb\ell(ic)]$ are adjacent in memory. Since modern computers are often designed to fetch many adjacent items from memory at once and place them in fast-access cache memory or in on-chip registers in anticipation that they may be needed shortly, one expects there may be relatively few separate calls to slow access memory to find the coefficients in f and g when look-up tables are used. Note, by contrast, that successive values of $ihtb\ell(ic)$ are not contiguous. Therefore there is the penalty that the results of multiplication have to be scattered into slow access memory. They are, however, not too widely dispersed since if f and g are of given degrees (as in Exhibit 6.7), then all the entries in h will at least be of the same fixed degree, and therefore (in a graded indexing scheme) stored fairly close together. Moreover, if monomial ordering is used, then for each fixed if and all successive ig, the relevant ih values also occur in increasing order. We conclude that, with the use of look-up tables, access to the coefficients in both f and g may be fast, and access to the coefficients in h may be somewhat slow. Now examine Table 7.1. Here the successive values of $ihtb\ell(ic)$ are contiguous. However, those of $iftb\ell(ic)$ and $igtb\ell(ic)$ are not. Indeed, the entries in f and g that need to be accessed are not even of fixed degree. We conclude that, with the use of look-back tables, access to the coefficients in *both* f and g may be quite slow, and *only* access to the coefficients in h may be fast. Therefore, the use of look-up tables is likely to yield faster code.

2. It appears that the look-back method requires the storage of two tables of dimension $\{m, d\}^{ns}$, while the look-up method involves the storage of only one. As we have seen, these tables may be large. However, this objection may not be as serious as it sounds. Examination of Table 7.1 shows that, for each fixed value of $ihtb\ell$, there is a close relation between the contents of $iftb\ell$ and $igtb\ell$. Indeed, one list is the reverse of the other. Therefore, at the expense of some slightly more complicated logic and a few additional look ups, it may be possible to work with a single table of dimension $\{m, d\}^{ns}$.

3. The look-back method also requires storage of the arrays *icbot* and *ictop* which are each roughly of size $S(m, d)$. The look-up method requires the additional arrays *ibot* and *itop* which are only of size $(maxdeg + 1)$ with $maxdeg = m$.

4. Finally, the look-up method can be modified to exploit possible sparseness and homogeneity as shown in Exhibits 6.6 through 6.8. This does not seem to be possible for the look-back method.

Strictly speaking, some of the consideration listed above apply only to the case of computation with one processor. Suppose one has several processors available in some form of large-scale parallel architecture. Then one might assign the computation of various $h(ih)$ values to various processors, all to be computed in parallel. In this case, each computation would use look-back tables, and the use of look-back tables might be preferable to the use of look-up tables.

Let us pause to reflect. Look back over our discussion so far in this section, and the content of the previous two sections. After some thought, we see that what we have learned is that the idea of a script is a unifying concept, and that the use of a look-up table and the use of look-back tables are simply alternate ways of going through the script. Indeed, if we look at Table 6.1 and regard the indices if and ig as entries in a two-component number (if, ig) with most significant digit if and least significant digit ig, then we see that these numbers are arranged in increasing size as we go down the list. Alternatively, if we look at Table 7.1, we see that the list has been "graded" according to the value of ih; and the items having a given ih are again arranged in increasing size based on the two-component numbers (if, ig).

We close this section by observing that look-back tables can sometimes be employed to optimal advantage for other calculations. Their use in the calculation of Poisson brackets is described in Section 8. Here we describe their use in the evaluation of polynomials. Suppose $f(z)$ is a polynomial written in the form

$$f(z) = \sum_i f_i M_i(z), \tag{39.7.3}$$

where the f_i are a given set of coefficients, and we wish to know the value of f at the point $z = w$. Here, as before in Section 32.3, we have used the notation

$$M_i(z) = z^{j(i)}. \tag{39.7.4}$$

The coefficients f_i may be viewed as the entries in a *vector* of dimension $S_0(m, d)$. Define another such vector with entries γ_i given by the relation

$$\gamma_i = M_i(w). \tag{39.7.5}$$

Then the value $f(w)$ can be written in the form

$$f(w) = \sum_i f_i \gamma_i. \tag{39.7.6}$$

We see that (7.6) can be viewed as a *vector dot product* and, providing we know the entries γ_i, this dot product can be computed very efficiently by computers having vector or pipe-lined architecture.

But how can be find the γ_i in an efficient manner? If we use indexing based on a modified *glex* sequence, we have the results

$$\gamma_0 = 1, \tag{39.7.7}$$

$$\gamma_i = w_i \text{ for } i = 1, 2, \cdots d. \tag{39.7.8}$$

Moreover, we claim there are two look-back tables $i1(i)$ and $i2(i)$ such that the remaining γ_i can found from a recursion relation of the form

$$\gamma_i = [\gamma_{i1(i)}][\gamma_{i2(i)}]\ ,\ i = d + 1, \cdots S(m, d). \tag{39.7.9}$$

Thus, it is possible to evaluate the γ_i by carrying out only $[S(m, d) - d]$ multiplications.

To see how to construct the tables $i1(i)$ and $i2(i)$, we observe that each $M_i(z)$ of degree $n = |j(i)|$, and assuming $n \geq 2$, can be factored as the product of a *first* degree monomial with index $i1(i)$ and another monomial of degree $n - 1$ having index $i2(i)$:

$$M_i(z) = [M_{i1(i)}(z)][M_{i2(i)}(z)]. \tag{39.7.10}$$

Exhibit 7.2 shows a routine written to find two tables $i1(i)$ and $i2(i)$ having this property. Inspection of the routine shows that it works as follows: Find and examine the exponent $j(i)$. Proceeding from the left, let $j_k(i)$ be the first nonzero entry in $j(i)$. By construction k must lie in the range $1 \leq k \leq d$, and we have the result

$$z^{j(i)} = z_1^{j_1} z_2^{j_2} \cdots z_d^{j_d} = z_k z^{j'(i)} \tag{39.7.11}$$

where the exponent array $j'(i)$ has the entries

$$j'_\ell(i) = j_\ell(i) - 1 \text{ when } \ell = k, \tag{39.7.12}$$

$$j'_\ell(i) = j_\ell(i) \text{ when } \ell \neq k. \tag{39.7.13}$$

Now define the table entries $i1(i)$ and $i2(i)$ by the rules

$$i1(i) = \text{ index for } z_k = k, \tag{39.7.14}$$

$$i2(i) = \text{ index for the exponent } j'(i). \tag{39.7.15}$$

Table 7.3 displays the result of running this routine for the case $m = 3$ and $d = 6$. The entries in this table should be compared with those in Table 2.1.

Exhibit 7.3 shows a routine for computing the γ_i based on the relations (7.7) through (7.9). Look at Table 7.3. We see that the results of the algorithm of Exhibit 7.2 have the pleasing feature that, for the most part, the values of $i2(i)$ are contiguous for successive values of i. Inspection of the routine of Exhibit 7.3 shows that it requires the values $\gamma_{i1(i)}$ and $\gamma_{i2(i)}$ for successive values of i. Since the addresses are usually contiguous, it is very likely that the values of the required $\gamma_{i1(i)}$ and $\gamma_{i2(i)}$ will either be in fast-access cache or in registers when needed, and consequently this routine should be very fast.

Exhibit 32.7.2: Program for factoring monomials. It produces arrays i1(i) and i2(i) such that monomial(i)=monomial(i1(i))*monomial(i2(i)) and the monomials monomial(i1(i)) in the first factor are all of degree one.

```
      do 10 i=1,imax
      do 20 k=1,6
         j(k)=jtbl(k,i)
   20 continue
      do 30 k=1,6
         if (j(k) .ne. 0) then
             j(k)=j(k)-1
             il(i)=k
             call ndex (j,ij)
             i2(i)=ij
             go to 40
         endif
   30 continue
   40 continue
   10 continue
      end
```

Exhibit 32.7.3: Program for building vector of monomial values using look-back tables i1(i) and i2(i).

```
       gam(0)=1.d0
      do 10 i=1,id
         gam(i)=w(i)
   10 continue
      do 20 i=id+1,imax
         gam(i)=gam(i1(i))*gam(i2(i))
   20 continue
      end
```

Table 39.7.3: The arrays $i1(i)$ and $i2(i)$ in the case $m = 3$ and $d = 6$.

i	$i1$	$i2$	i	$i1$	$i2$
7	1	1	28	1	7
8	1	2	29	1	8
9	1	3	30	1	9
10	1	4	31	1	10
11	1	5	32	1	11
12	1	6	33	1	12
13	2	2	34	1	13
14	2	3	35	1	14
15	2	4	36	1	15
16	2	5	37	1	16
17	2	6	38	1	17
18	3	3	⋮		
19	3	4	77	4	25
20	3	5	78	4	26
21	3	6	79	4	27
22	4	4	80	5	25
23	4	5	81	5	26
24	4	6	82	5	27
25	5	5	83	6	27
26	5	6			
27	6	6	⋮		

Exercises

39.7.1.

39.8 Poisson Bracketing

When computing Poisson brackets of homogeneous polynomials, there are three natural cases to consider. First, there are brackets of the form $[f_1, g_1]$. They are trivial to compute in view of (1.7.10). Next in order of complexity are brackets of the form $[f_m, z_a]$ with $m \geq 2$. They will be referrred to as *single-variable* Poisson brackets, and are an essential ingredient in the computation of $\mathcal{M}z_a$ as in (7.1.1). With the aid of (7.6.10), they can be evaluated by the formula

$$: f_m : z_a = [f_m, z_a] = -\partial f_m / \partial z_a^*. \tag{39.8.1}$$

Finally, there are brackets of the form $[f_m, g_n]$ with $m, n \geq 2$. They will be called *general* Poisson brackets.

The purpose of this section is to describe efficient algorithms for the computation of single-variable and general Poisson brackets.

We begin with the case of single-variable Poisson brackets. That is, we wish to compute

$$g =: f : z_a. \tag{39.8.2}$$

For the discussion of Poisson brackets it is convenient to order the q's and p's in conjugate pairs and to define z^j (for the case of a 6-dimensional phase-space) in analogy with (2.1),

$$z^j = q_1^{j_1} p_1^{j_2} q_2^{j_3} p_2^{j_4} q_3^{j_5} p_3^{j_6}. \tag{39.8.3}$$

Then, for example and in agreement with (8.1), we find the result

$$: z^j : q_1 = [z^j, q_1] = -\partial z^j/\partial p_1 = -j_2 q_1^{j_1} p_1^{j_2 - 1} q_2^{j_3} p_2^{j_4} q_3^{j_5} p_3^{j_6}. \tag{39.8.4}$$

Evidently every monomial single-variable Poisson bracket $[z^j, z_a]$ results in at most one term, and many produce none. This circumstance can be exploited using look-back tables. Suppose the monomial z^k with

$$z^k = q_1^{k_1} p_1^{k_2} q_2^{k_3} p_2^{k_4} q_3^{k_5} p_3^{k_6} \tag{39.8.5}$$

produces the monomial z^j as a result of the multiplication

$$z^k z_b = z^j. \tag{39.8.6}$$

Then we have the result

$$\partial z^j/\partial z_b = j_b z^k \tag{39.8.7}$$

with j_b guaranteed to be nonzero. Let $i = i(k)$ be the index of the monomial with exponent k. Then, corresponding to the relation (8.6), we can define a single-variable multiplication table $msv(i, b)$ by the rule

$$msv(i, b) = \text{ index of monomial with exponent } j. \tag{39.8.8}$$

Exhibit 8.1 below shows a program that prepares a script for single-variable Poisson bracketing. It makes the single-variable multiplication table $msv(i, b)$ as well as the tables $coef(i, b)$ and $scoef(i, b)$, which contain various coefficients such as the $(-j_2)$ that appears in (8.4). From their construction it is evident that all these tables have the modest dimension $itop(maxdeg - 1) \times 6$. [Actually the table $coef(i, b)$ is not needed for single-variable Poisson bracketing, but is useful for multplication-based general Poisson bracketing. See Exhibit 8.5.]

```
Exhibit 32.8.1: Program to produce script for single-variable
Poisson bracket routine.

      subroutine svpbs
c
      data icon /2,1,4,3,6,5/
      data sign /1.d0,-1.d0,1.d0,-1.d0,1.d0,-1.d0/
c
```

```
c Loop over phase-space variables z_k
c
      do 10 k=1,6
      izc=icon(k)
c
c Loop over ic
c
      do 20 ic=ibot(1), itop(maxdeg-1)
c
c find and store exponents
c
      do m=1,6
      jl(m)=jtbl(m,ic)
      end do
      jltizc=jl(izc)
c
c fill tables
c
      jl(izc)=jltizc+1
      ifac=jl(izc)
      coef(ic,k)=dfloat(ifac)
      scoef(ic,k)=sign(k)*coef(ic,k)
      if=ndex(jl)
      msv(ic,k)=if
c
c restore exponent
c

      jl(izc)=jltizc
c
  20  continue
  10  continue
c
      return
      end
```

Exhibit 8.2 shows the actual single-variable Poisson bracketing routine that uses the script prepared by the program of Exhibit 8.1. It works directly with the index ig for each monomial in the result g, and uses the table msv to look back to find the index $if = msv(ig, k)$ of the monomial in f that produced this result. Because each monomial in f contributes to at most one term in g, no attempt has been made to exploit possible sparseness in f. To test in advance of their use the various $f(if)$ to see if they vanished would result in significant computational overhead with little associated reward. Finally, we note that, due to the use of monomial ordering, successive if values appear in an increasing sequence. The required $f(if)$ values are therefore likely to be in cache or in on-chip registers.

Exhibit 32.8.2: Program for single variable Poisson bracket.

```
      subroutine svpb(f,ideg,k,g)
c
c This subroutine finds the single variable Poisson
```

```
c bracket g=:f:z_k.  Here f is homogeneous of degree ideg.
c
      do ig=ibot(ideg-1),itop(ideg-1)
      g(ig)=scoef(ig,k)*f(msv(ig,k))
      end do
c
      return
      end
```

We next turn to the general Poisson bracket case. We wish to compute $h = [f, g]$ when both f and g are of degree two and higher. For a 6-dimensional phase space, the Poisson bracket of any two monomials is given by the standard rule

$$[z^j, z^k] = \sum_{i=1}^{3} (\partial z^j/\partial q_i)(\partial z^k/\partial p_i) - (\partial z^j/\partial p_i)(\partial z^k/\partial q_i). \tag{39.8.9}$$

At first count it might appear that a typical monomial Poisson bracket could contain 6 distinct terms. In fact, there are at most 3 distinct terms. To verify this assertion, consider the $i = 1$ terms on the right side of (8.9). They give the result

$$\begin{aligned}
&(\partial z^j/\partial q_1)(\partial z^k/\partial p_1) - (\partial z^j/\partial p_1)(\partial z^k/\partial q_1)\\
&= (j_1 z_1^{j_1-1} z_2^{j_2} z_3^{j_3} z_4^{j_4} \cdots)(k_2 z_1^{k_1} z_2^{k_2-1} z_3^{k_3} z_4^{k_4} \cdots)\\
&\quad -(j_2 z_1^{j_1} z_2^{j_2-1} z_3^{j_3} z_4^{j_4} \cdots)(k_1 z_1^{k_1-1} z_2^{k_2} z_3^{k_3} z_4^{k_4} \cdots)\\
&= (j_1 k_2 - j_2 k_1)(z_1^{j_1+k_1-1} z_2^{j_2+k_2-1} z_3^{j_3+k_3} z_4^{j_4+k_4} \cdots).
\end{aligned} \tag{39.8.10}$$

Evidently in the Poisson bracket result there is at most one monomial term for each value of i in the sum (8.9), and there is none whenever (for odd ℓ) the coefficient $(j_\ell k_{\ell+1} - j_{\ell+1} k_\ell)$ vanishes.

Exhibit 8.3 below shows a program that prepares a script for general Poisson bracketing.

Exhibit 32.8.3: Program to produce script for general Poisson bracket routine.

```
      subroutine pbsc
c
c
c set counters
c
      ic1=0
      ic2=0
c
c loop over degrees
c
      do 10 ifdeg=2,maxdeg
      maxgdeg=maxdeg+2-ifdeg
      do 20 igdeg=2,maxgdeg
c
c loop over if and ig
c
      do 30 if=ibot(ifdeg),itop(ifdeg)
      do 40 ig=ibot(igdeg),itop(igdeg)
```

```
c
c find and store exponents and their sums
c
      do k=1,6
      jf(k)=jtbl(k,if)
      jg(k)=jtbl(k,ig)
      js(k)=jf(k)+jg(k)
      end do
c
c increment ic2 counter
c
      ic2=ic2+1
c
c find and count possible ih indices and coefficients,
c and set up tables
c
      it=0
      do 50 k=5,1,-2
      iz=k
      izc=k+1
      ival=jf(iz)*jg(izc)-jf(izc)*jg(iz)
      if(ival .eq. 0) go to 50
c
c compute exponents from exponent sums,
c and compute index
c
      jsizt=js(iz)
      jsizct=js(izc)
      js(iz)=jsizt-1
      js(izc)=jsizct-1
      index=ndex(js)
c
c restore exponent sums
c
      js(iz)=jsizt
      js(izc)=jsizct
c
c increment it and ic1 counters,
c and store coefficients and indices
c
      it=it+1
      ic1=ic1+1
      pbcoef(ic1)=dfloat(ival)
      ih(ic1)=index
c
   50 continue
c
c store the number of terms
c
      nt(ic2)=it
c
   40 continue
c
c set reset tables just after leaving ig loop
```

```
c
      irst1(if,igdeg)=ic1
      irst2(if,igdeg)=ic2
c

   30 continue
   20 continue
   10 continue
c
c record maximum required storage
c
      maxic1=ic1
      maxic2=ic2
c
      return
      end
```

The sizes of the arrays produced by this script can be quite large. The array *nt* is indexed by the integer variable *ic2*, and has dimension *maxic2*. The arrays *pbcoef* and *ih* are indexed by the integer variable *ic1*, and have dimension *maxic1*. The values of *maxic1* and *maxic2* are listed in Table 8.1 for various values of *maxdeg*, the maximum degree of the polynomials one is considering.

The computation of *maxic2* is elementary. The variable *ic2* labels all possible *if*, *ig* pairs that can potentially occur in a Poisson bracket calculation, and *maxic2* is the total number of such pairs. For example, if $maxdeg = 3$, we have to consider Poisson bracket terms of the form $[f_2, f_2]$, $[f_2, f_3]$, and $[f_3, f_2]$. Inspection of Table 7.3.1 shows that (for a 6-dimensional phase space) there are 21 f_2 basis monomials and 56 f_3 basis monomials. Thus, in this case we expect the result

$$\mathit{maxic2} = 21 \times 21 + 21 \times 56 + 56 \times 21 = 441 + 1176 + 1176 = 2793,$$

in agreement with the corresponding entry in Table 8.1. We note that the quantites $\mathit{maxic2}(maxdeg)$ for Lie (Poisson bracket) multiplication are much larger than their counterparts $\{maxdeg, 6\}^{ns}$ for ordinary multiplication. Compare Tables 5.4 and 8.1. This increased size results from the (-2) term in (7.6.16), which does not occur for ordinary multiplication. As a simple example, many of the absent terms denoted by an asterisk "*" in Table 5.2 for ordinary multiplication would not be absent for Poisson bracket multiplication.

The computation of *maxic1* is more complicated, and is most easily done simply by counting as in Exhibit 8.3. As described earlier, each possible *if*, *ig* monomial pair is labeled by a value of *ic2*. The quantity *nt(ic2)* is the number of monomial terms that result from Poisson bracketing the monomial pair with label *ic2*. We have already seen that this number (including the possibility of a vanishing bracket) ranges from 0 to 3. The quantity *maxic1* is the sum over *nt*,

$$\mathit{maxic1} = \sum_{\mathit{ic2}=1}^{\mathit{maxic2}} \mathit{nt(ic2)}. \tag{39.8.11}$$

Thus, for a fixed value of *maxdeg*, the quantity *maxic1* is the number of nonzero monomials that can occur in a Poisson bracket calculation including repetitions.

Inspection of Table 8.1 shows that *maxic1*, for *maxdeg* ≤ 6, is smaller than or comparable to *maxic2*. Evidently, although the Poisson bracket of each monomial pair could potentially produce as many as 3 monomial terms, most Poisson brackets produce fewer or are in fact zero. By contrast, *maxic1* exceeds *maxic2* for *maxdeg* > 6.

Table 39.8.1: Array sizes *maxic1* and *maxic2* (in the case of 6 phase-space variables) for various values of *maxdeg*.

maxdeg	*maxic1*	*maxic2*	*maxdeg*	*maxic1*	*maxic2*
2	210	441	8	648342	570619
3	1662	2793	9	1514382	1231279
4	7986	11221	10	3320694	2518565
5	29622	35917	11	6902358	4923317
6	92802	99421	12	13701822	9254645
7	257190	247933			

Exhibit 8.4 shows the actual general Poisson bracket routine that uses the script prepared by the program of Exhibit 8.3. It is designed to exploit possible sparseness in f and known homogeneity in both f and g. The array *pbcoef(ic1)* contains precomputed (and nonzero) coefficients of the form $(j_\ell k_{\ell+1} - j_{\ell+1} k_\ell)$, and the look-up table *ih(ict1)* specifies where contributions to the various terms in h are to be placed. For each if,*ig* pair, the *ih* values for successive *ict1* values are arranged in increasing order in the hope of rapid memory access. Thanks to the contents of the array *nt(ic2)*, only potentially nonzero terms are computed. The arrays *irst1* and *irst2* set the counters *ic1* and *ic2* to take into account skipped terms due to sparseness and homogeneity.

Exhibit 32.8.4: Script-driven program for general Poisson bracket.

```
      subroutine pb(f,ifdeg,g,igdeg,h)
c
c This subroutine computes the Poisson bracket
c h=[f,g].  The input polynomials f and g are
c homogeneous of degrees ifdeg and igdeg, respectively.
c
c clear h
c
      do i=1,imax
      h(i)=0.d0
      end do
c
c find offsets
c
      ic1=irst1(itop(ifdeg-1),igdeg)
      ic2=irst2(itop(ifdeg-1),igdeg)
c
c loop over if and ig
c
      do 10 if=ibot(ifdeg),itop(ifdeg)
      if(f(if) .eq. 0.d0) then
      ic1=irst1(if,igdeg)
      ic2=irst2(if,igdeg)
      go to 10
      end if
      do 20 ig=ibot(igdeg),itop(igdeg)
      if(g(ig) .eq. 0.d0) then
      ic1=ic1+nt(ic2)
      ic2=ic2+1
      go to 20
      end if
      if(nt(ic2) .ne. 0) then
      prod=f(if)*g(ig)
      do 30 it=1,nt(ic2)
      h(ih(ic1)) = h(ih(ic1)) + pbcoef(ic1)*prod
      ic1=ic1+1
   30 continue
      end if
      ic2=ic2+1
   20 continue
   10 continue
c
      return
      end
```

As remarked earlier, Table 8.1 shows that the script arrays required to drive the routine of Exhibit 8.4 can be quite large for vaues of *maxdeg* beyond 9. This may not be an issue as memory becomes ever more plentiful. However, it is worth remarking that there is an alternate approach that requires much less memory but, of course, is computationally slower. What one can do, as in the case of single-variable Poisson brackets, is use look-back tables to find the various terms in $\partial f/\partial z_a$ and $\partial g/\partial z_b$, and then use look-up tables, as in Exhibit 6.8, to carry out the multiplications $(\partial f/\partial z_a)(\partial g/\partial z_b)$. Exhibit 8.5 below shows a general Poisson bracket routine that uses this procedure.

Exhibit 32.8.5: General Poisson bracket program based on multiplication.

```
      subroutine pb(f,ifdeg,g,igdeg,h)
c
      data icon /2,1,4,3,6,5/
c
      icoff=ibot(igdeg-1)
      do 10 iz=1,6
      izc=icon(iz)
c
      do 20 ifl=ibot(ifdeg-1),itop(ifdeg-1)
      if=msv(ifl,iz)
      if(f(if) .ne. 0.d0) then
      ic=icmin(ifl)+icoff
      do 30 igl=ibot(igdeg-1),itop(igdeg-1)
      ig=msv(igl,izc)
      if(g(ig).ne. 0.d0) then
      fac=scoef(ifl,iz)*coef(igl,izc)
      h(ihtbl(ic))=h(ihtbl(ic))+fac*f(if)*g(ig)
      end if
      ic=ic+1
  30  continue
      end if
  20  continue
  10  continue
c
      return
      end
```

39.9 Linear Map Action

Let f be any function of the $2n$ phase-space variables z, and let M be any $2n \times 2n$ matrix. Then we define a *transformed* function g by the rule

$$g(z) = f(Mz). \tag{39.9.1}$$

One may view (9.1) as the action of the linear (usually but not necessarily symplectic) map respresented by the matrix M on the function f. See (8.4.23) and (10.4.36) where actions of this type arise in the concatenation and computation of maps. The purpose of this section is to describe an efficient algorithm for carrying out the operation (9.1).

To achieve efficiency, it is useful to employ a precomputed list of variables $jvblist(iv, ind)$. Here ind is a monomial index, and for an nth-order monomial there will be n non-zero variable numbers. For example, for the indexing scheme of Table 2.1, the monomial with index 7 is $X^2 = XX$. Correspondingly, the variable list (array) *jvblist* will have the entries

$$jvblist(iv = 1 \text{ to } 2, 7) = 1, 1. \tag{39.9.2}$$

As a second example, the monomial with index 19 is YP_y. Correspondingly *jvblist* will have the entries

$$jvblist(iv = 1 \text{ to } 2, 19) = 3, 4. \tag{39.9.3}$$

For a third example, the monomial with index 77 is $P_y\tau^2 = P_y\tau\tau$. Correspondingly *jvblist* will have the entries

$$jvblist(iv = 1 \text{ to } 3, 77) = 4, 5, 5. \tag{39.9.4}$$

Exhibit 9.1 below shows a program that produces *jvblist*. It is a modification of the program in Exhibit 2.3 so that both *jtbl* and *jvblist* are created simultaneously.

Exhibit 32.9.1: A program to produce both jtbl and jvblist based on a method of Liam Healy.

```
       subroutine tables

c        ind = monomial index and imax = maximum value of ind.
c        ipsv = phase space variable and id = number of phase space
variables.
c
       parameter (imax = 209, id=6)
```

```
      dimension jtbl(id,imax), jvblist(id,imax)
c  j = array of exponents
      dimension j(id)
c  initialize exponents
      data j/id*0/
c  icarry = temporarily stored value of j(id).
c  lnzj = last non-zero j
c
c  sequentially create exponent table jtbl and the array jvblist
c
      do ind=1,imax
c
c  set quantities
c
        icarry=j(id)
        j(id)=0
        lnzj=0
c
c  search for last nonzero j
c
        do ipsv=1,id-1
          if (j(ipsv).gt.0) lnzj=ipsv
        enddo
c
c  find next set of exponents
c
        if (lnzj.gt.0) j(lnzj)=j(lnzj)-1
        j(lnzj+1)=1+icarry
c
c  store exponents in jtbl
c
        do ipsv=1,id
          jtbl(ipsv,ind)=j(ipsv)
        enddo
c
c create jvblist
c
      iv=1
      do ipsv=1,id
       do k=1,j(ipsv)
        jvblist(iv,ind)=ipsv
        iv=iv+1
       enddo
      enddo
c
      enddo
c
      return
      end
```

With this background information in mind, we are ready to present the algorithm that carries out the operation (9.1). It is shown in Exhibit 9.2 below.

Exhibit 32.9.2: Program for linear map action.

```
      subroutine xform(f,ideg,em,g)
c
c     Transforms a polynomial f of degree ideg by the linear
c     map whose matrix representation is em.  The coefficients
c     of the resultant polynomial f(em*z) are stored in the
c     array g.  Thus g=f(em*z).
c
c initialise g array
c
      do k=1,imax
      g(k) = 0.d0
      end do
c
c loop over monomials of degree ideg
c
      do 100 n=ibot(ideg),itop(ideg)
      if(f(n) .eq. 0.d0) goto 100
c
c clear the array ta
c
      do k=7,itop(ideg)
      ta(k) = 0.d0
      end do
c
c work on the ideg variables in the monomial
c
c treatment of first variable
c
c transform first variable in monomial and place result in ta1
c and ta
c
      jiv = jvblist(1,n)
      do k=1,6
      ta1(k) = em(jiv,k)
      ta(k) = ta1(k)
      end do
c
c loop over remaining variables
c
      do 110 iv=2,ideg
c
c find next variable
c
      jivn = jvblist(iv,n)
c
c if next variable is the same as the previous one,
c build up product in ta
c
      if(jivn .eq. jiv) go to 120
c
c otherwise, if the next variable is different from
c the previous one, transform that variable, place
c result in ta1, and build up product in ta
```

```
c
      jiv = jivn
      do k=1,6
      ta1(k) = em(jiv,k)
      end do
c
  120 continue
c
c build up product
c
      icdeg = iv-1
      do 130 i1=1,6
      if(ta1(i1) .ne. 0.d0) then
      do 140 i2=ibot(icdeg),itop(icdeg)
  140 ta(iprodex(i2,i1)) = ta1(i1)*ta(i2)
      endif
  130 continue
c
  110 continue
c
c accumulate sum in g
c
      do nn=ibot(ideg),itop(ideg)
      g(nn) = g(nn) + f(n)*ta(nn)
      end do
c
  100 continue
c
      return
      end
```

39.10 General Vector Fields

Let $\boldsymbol{f} = (f_1, f_2, \cdots f_d)$ be a collection of d functions of z, and suppose each function f_a is a truncated power series in z. Let $\mathcal{L}_{\boldsymbol{f}}$ be the vector field associated with $\boldsymbol{f}$. See Section 5.3. Using the tools already developed, it is easy to envision how to construct programs that would represent, multiply by scalars, and from linear combinations of such vector fields. We simply store and manipulate the underlying collections of functions in the obvious way.

Beyond these operations, we would also like to apply $\mathcal{L}_{\boldsymbol{f}}$ to a function g where it assumed that g is also a truncated power series. That is, we wish to find the truncated power series for h defined by the equation

$$h = \mathcal{L}_{\boldsymbol{f}}\, g = \sum_a f_a(\partial g/\partial z_a). \tag{39.10.1}$$

Moreover, suppose $\boldsymbol{f}$ and $\boldsymbol{g}$ are two collections of truncated power series. Let $\mathcal{L}_{\boldsymbol{f}}$ and $\mathcal{L}_{\boldsymbol{g}}$ be their associated vector fields. It is easily verified that the commutator of two vector fields is again a vector field. That is, there is a collection of functions $\boldsymbol{h}$ such that

$$\#\mathcal{L}_{\boldsymbol{f}}\#\mathcal{L}_{\boldsymbol{g}} = \{\mathcal{L}_{\boldsymbol{f}}, \mathcal{L}_{\boldsymbol{g}}\} = \mathcal{L}_{\boldsymbol{f}}\mathcal{L}_{\boldsymbol{g}} - \mathcal{L}_{\boldsymbol{g}}\mathcal{L}_{\boldsymbol{f}} = \mathcal{L}_{\boldsymbol{h}}. \tag{39.10.2}$$

Given the collections of truncated power series for $\boldsymbol{f}$ and $\boldsymbol{g}$, we would like to find the collection of truncated power series for $\boldsymbol{h}$.

The purpose of this section is to describe programs for computing h in (10.1) and $\boldsymbol{h}$ in (10.2).

It is evident from (10.1) that the computation of h involves partial differentiation and function multiplication. The same is true for the computation of $\boldsymbol{h}$. From the representation

$$\mathcal{L}_{\boldsymbol{h}} = \sum_b h_b(\partial/\partial z_b) \tag{39.10.3}$$

and (10.2) we find the result

$$h_a = \mathcal{L}_{\boldsymbol{h}} z_a = \mathcal{L}_{\boldsymbol{f}} \mathcal{L}_{\boldsymbol{g}} z_a - \mathcal{L}_{\boldsymbol{g}} \mathcal{L}_{\boldsymbol{f}} z_a = \mathcal{L}_{\boldsymbol{f}} g_a - \mathcal{L}_{\boldsymbol{g}} f_a. \tag{39.10.4}$$

Therefore, the computation of commutators involves the action (10.1) of vector fields on functions which, in turn, again involves partial differentiation and function multiplication.

In principle, it is possible to write scripted programs that would perform all the operations in (10.1) and (10.3) in an optimal way. However, for simplicity, we will only describe a scripted program for partial differentiation. This program can then be used in conjunction with those for multplication and addition to carry out all the required operations.

The operation of partial differentiation is similar to, and in fact simpler than, single-variable Poisson bracketing. See (8.1). Therefore, it is conviently done with the aid of a script and look-back tables. Exhibit 10.1 shows a program that generates a script for partial differentiation. It makes the single-variable multiplication table $msv(ic, k)$ and the table $coef(ic, k)$ that contains the coefficients j_b that occur in (8.7). Finally, Exhibit 10.2 shows the partial differentiation routine that uses the script prepared by the program of Exhibit 10.1.

```
Exhibit 32.10.1: Program to produce script for partial
differentiation routine.

      subroutine pds
c
c Loop over phase-space variables z_k
c
      do 10 k=1,6
      iz=k
c
c Loop over ic
c
      do 20 ic=ibot(1), itop(maxdeg-1)
c
c Find and store exponents
c
      do m=1,6
      jl(m)=jtbl(m,ic)
      end do
      jltiz=jl(iz)
c
c Fill tables
```

```
c
      jl(iz)=jlizt+1
      ifac=jl(iz)
      coef(ic,k)=dfloat(ifac)
      if=ndex(jl)
      msv(ic,k)=if
c
c Restore exponent
c
      jl(iz)=jltiz
c
  20  continue
  10  continue
c
      return
      end
```

Exhibit 32.10.2: Program for partial differentiation.

```
      subroutine pd(f,ideg,k,g)
c
c This subroutine finds the partial derivative
c g=df/dz_k.  Here f is homogeneous of degree ideg.
c
      do ig=ibot(ideg-1), itop(ideg-1)
      g(ig)=coef(ig,k)*f(msv(ig,k))
      end do
c
      return
      end
```

39.11 Expanding Functions of Polynomials

39.12 Differential Algebra

39.13 Other Methods

Bibliography

Computer Science, Indexing Schemes, Polynomial Manipulation, Etc.

[1] A. Giorgilli, “A computer program for integrals of motion”, *Comp. Phys. Comm.* **16**, p. 331 (1979). See also the Web link http://www.mat.unimi.it/users/antonio/ricerca/papers/laplata.pdf.

[2] A. Haro, M. Canadell, J-L. Figueras, A. Luque, and J-M. Mondelo, *The Parameterization Method for Invariant Manifolds: From Rigorous Results to Effective Computations*, Applied Mathematical Sciences Volume 195, Springer (2016).

[3] Alex Haro, “Automatic Differentiation Tools in Computational Dynamical Systems”. See the Web site http://www.maia.ub.es/~alex/ad/adhds.pdf

Combinatorics

[4] D. Cox, J. Little, and D. O’Shea, *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, Springer-Verlag (1992).

Appendix A

Størmer-Cowell and Nyström Integration Methods

The differential equations of classical mechanics often involve only second derivatives with *no* first derivatives present. In this case it is possible to work directly with the second order equations instead of converting them into a first order set of twice the dimensionality. The result can be a saving in computer time and an increase in accuracy. We shall describe methods due to *Størmer* and *Cowell* and *Nyström*. See Chapter 2 for notation.

A.1 Preliminary Derivation of Størmer-Cowell Method

Consider a set of second-order equations of the form

$$\ddot{\boldsymbol{y}}(t) = \boldsymbol{f}(\boldsymbol{y}, t). \tag{A.1.1}$$

[We remark that if a Hamiltonian is of the form $H = p \cdot p/2 + V(q,t)$, it leads to differential equations of the form (1.1).] Then, using arguments similar to those of Chapter 2, we have the integration formulas

$$\nabla^2 \boldsymbol{y}^{n+1} = \nabla^2 D^{-2} \boldsymbol{f}^{n+1}, \tag{A.1.2}$$

$$\nabla^2 \boldsymbol{y}^{n+1} = \nabla^2 (1-\nabla)^{-1} D^{-2} \boldsymbol{f}^{n}. \tag{A.1.3}$$

By expanding (1.2) and (1.3) and using (2.4.13), we may rewrite our results as

$$\boldsymbol{y}^{n+1} = 2\boldsymbol{y}^n - \boldsymbol{y}^{n-1} + h^2[\nabla/\log(1-\nabla)]^2 \boldsymbol{f}^{n+1}, \tag{A.1.4}$$

$$\boldsymbol{y}^{n+1} = 2\boldsymbol{y}^n - \boldsymbol{y}^{n-1} + h^2(1-\nabla)^{-1}[\nabla/\log(1-\nabla)]^2 \boldsymbol{f}^{n}. \tag{A.1.5}$$

As in Chapter 2, we interpret the right sides (1.4) and (1.5) in terms of power series. After truncation we obtain the predictor and corrector formulas

$$\boldsymbol{y}^{n+1} = 2\boldsymbol{y}^n - \boldsymbol{y}^{n-1} + h^2 \sum_{k=0}^{N} \alpha_k \nabla^k \boldsymbol{f}^{n+1}, \qquad (\mathit{corrector}) \tag{A.1.6}$$

$$\boldsymbol{y}^{n+1} = 2\boldsymbol{y}^n - \boldsymbol{y}^{n-1} + h^2 \sum_{k=0}^{N} \beta_k \nabla^k \boldsymbol{f}^n, \qquad (\textit{predictor}) \qquad \text{(A.1.7)}$$

or the expanded versions

$$\boldsymbol{y}^{n+1} = 2\boldsymbol{y}^n - \boldsymbol{y}^{n-1} + h^2 \sum_{k=0}^{N} \overset{\sim N}{\alpha_k} \; \boldsymbol{f}^{n+1-k}, \qquad (\textit{corrector}) \qquad \text{(A.1.8)}$$

$$\boldsymbol{y}^{n+1} = 2\boldsymbol{y}^n - \boldsymbol{y}^{n-1} + h^2 \sum_{k=0}^{N} \overset{\sim N}{\beta_k} \; \boldsymbol{f}^{n-k}. \qquad (\textit{predictor}) \qquad \text{(A.1.9)}$$

The corrector and predictor truncation errors associated with (1.6) and (1.7) may be estimated using arguments similar to the Adams case. The result is

$$\boldsymbol{y}^{n+1}_{\text{true}} - \boldsymbol{y}^{n+1}_{\text{corr}} \approx h^{N+3}\alpha_{N+1}(d^{N+3}\boldsymbol{y}/dt^{N+3})|_{t=t^n}, \qquad \text{(A.1.10)}$$

$$\boldsymbol{y}^{n+1}_{\text{true}} - \boldsymbol{y}^{n+1}_{\text{pred}} \approx h^{N+3}\beta_{N+1}(d^{N+3}\boldsymbol{y}/dt^{N+3})|_{t=t^n}. \qquad \text{(A.1.11)}$$

Note that the predictor and corrector are one order higher accurate than their Adams counterparts. See (2.4.37) and (2.4.38). This increase in accuracy arises from the fact that the original differential equations being integrated are second order with no first derivatives present.

The coefficients α_k, β_k are listed in Table 1 below, and the associated coefficients $\overset{\sim N}{\alpha_k}, \overset{\sim N}{\beta_k}$ are listed in Tables 2 and 3.

Table 1

k	0	1	2	3	4	5	6	7	8	9
α_k	1	-1	$\frac{1}{12}$	0	$\frac{-1}{240}$	$\frac{-1}{240}$	$\frac{-221}{60480}$	$\frac{-19}{6048}$	$\frac{-9829}{3628800}$	$\frac{-407}{172800}$
β_k	1	0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{19}{240}$	$\frac{3}{40}$	$\frac{863}{12096}$	$\frac{275}{4032}$	$\frac{33953}{518400}$	$\frac{8183}{129600}$
$\lvert\beta_k/\alpha_k\rvert$	1	0	1	∞	19	18	~ 20	~ 22	~ 24	~ 27

Table 2
The Størmer-Cowell Corrector Coefficients $\overset{\sim N}{\alpha_k}$.

$k \quad N$	2	3	4	5	6	7	8	9
0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{19}{240}$	$\frac{18}{240}$	$\frac{4315}{60480}$	$\frac{4125}{60480}$	$\frac{237671}{3628800}$	$\frac{229124}{3628800}$
1	10	10	204	209	53994	55324	3398072	3474995
2	1	1	14	4	-2307	-6297	-653032	-960724
3		0	4	14	7948	14598	1426304	2144252
4			-1	-6	-4827	-11477	-1376650	-2453572
5				1	1578	5568	884504	1961426
6					-221	-1551	-368272	-1086220
7						190	90032	397724
8							-9829	-86752
9								8547

The denominator of each of the coefficients of the first line is to be repeated for all the coefficients of the corresponding column.

Table 3
The Størmer-Cowell Predictor Coefficients $\overset{\sim}{\beta}{}^{N}_{k}$.

k N	2	3	4	5	6	7	8	9
0	$\frac{13}{12}$	$\frac{14}{12}$	$\frac{299}{240}$	$\frac{317}{240}$	$\frac{168398}{120960}$	$\frac{176648}{120960}$	$\frac{5537111}{3628800}$	$\frac{5766235}{3628800}$
1	−2	−5	−176	−266	−185844	−243594	−9209188	−11271304
2	1	4	194	374	317946	491196	21390668	29639132
3		−1	−96	−276	−311704	−600454	−31323196	−50569612
4			19	109	184386	473136	30831050	59700674
5				−18	−60852	−234102	−20331636	−49202260
6					8630	66380	8646188	27892604
7						−8250	−2148868	−10397332
8							237671	2299787
9								−229124

The denominator of each of the coefficients of the first line is to be repeated for all the coefficients of the corresponding column. Note that the entries for $N = 6$ and $N = 7$ are not reduced to lowest terms. Both numerator and denominator should be divided by two.

Exercises

A.1.1. Make a study of the $\overset{\sim}{\alpha}$'s and $\overset{\sim}{\beta}$'s similar to that made in Exercise 2.4.4 for the $\tilde{a}$'s and $\tilde{b}$'s.

A.2 Summed Formulation

In principle the integration formulas (1.8) and (1.9) can be used as they stand. However in practice it has been found that a so called *summed* formulation has better performance with respect to round-off errors. It also reduces the truncation error by an additional factor of h without requiring additional starting values. Because the derivation of the summed formulation is a bit involved, we shall first state the procedure, and then provide the derivation.

A.2.1 Procedure

Suppose we know the values $\boldsymbol{y}^0 \cdots \boldsymbol{y}^N$ and $\boldsymbol{f}^0 \cdots \boldsymbol{f}^N$ from some starting routine such as Runge-Kutta. [Note that to use standard Runge-Kutta, one must first convert the set (1.1) into a first-order set. There are variants of Runge-Kutta due to *Nyström* that work directly with (1.1). See Section 5 at the end of this appendix.] We use the starting values to make some preparatory calculations. Define vectors $\boldsymbol{G}^n$ for $n = -1, 0, \cdots, N$ by the rule

$$\boldsymbol{G}^{-1} = 0, \tag{A.2.1}$$

$$\boldsymbol{G}^n = h \sum_{m=0}^{n} \boldsymbol{f}^m \ \text{ for } n \in [0, N].$$

Next define a vector $\boldsymbol{\sigma}$ using

$$\boldsymbol{\sigma} = h^{-1}(\boldsymbol{y}^N - \boldsymbol{y}^{N-1}) - \sum_{k=0}^{N+1} \overset{\sim}{\alpha}{}_k^{N+1}\, \boldsymbol{G}^{N-k}. \tag{A.2.2}$$

Finally, define vectors $\boldsymbol{g}^n$ for $n = -1, 0, \cdots, N$ by writing

$$\boldsymbol{g}^n = \boldsymbol{G}^n + \boldsymbol{\sigma}. \tag{A.2.3}$$

This completes the preparatory calculations.

The integration routine itself is given by the rules

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h \sum_{k=0}^{N+1} \overset{\sim}{\alpha}{}_k^{N+1}\, \boldsymbol{g}^{n+1-k}, \qquad (\textit{corrector}) \tag{A.2.4}$$

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h \sum_{k=0}^{N+1} \overset{\sim}{\beta}{}_k^{N+1}\, \boldsymbol{g}^{n-k}, \qquad (\textit{predictor}) \tag{A.2.5}$$

$$\boldsymbol{g}^{n+1} = \boldsymbol{g}^n + h\boldsymbol{f}^{n+1}. \tag{A.2.6}$$

Their truncation errors have the estimates

$$\boldsymbol{y}^{n+1}_{\text{true}} - \boldsymbol{y}^{n+1}_{\text{corr}} \approx h^{N+4}\alpha_{N+2}(d^{N+4}\boldsymbol{y}/dt^{N+4})|_{t=t^n}, \tag{A.2.7}$$

$$\boldsymbol{y}^{n+1}_{\text{true}} - \boldsymbol{y}^{n+1}_{\text{pred}} \approx h^{N+3}\beta_{N+1}(d^{N+3}\boldsymbol{y}/dt^{N+3})|_{t=t^n}. \tag{A.2.8}$$

Note that the corrector error is a factor of h smaller than that of the predictor. Whether or not this improvement in accuracy is realized in practice depends upon how many times the corrector is iterated. It can be shown that the simplest sequence *PEC* is insufficient. For this reason, and to check the convergence of successive iterations, it is better to use the sequences *PECEC* or *PECECE*.

A.2.2 Derivation

We now present the derivation for this procedure. We begin by rewriting (1.6) and (1.7) with an upper summation limit of $N + 1$:

$$\nabla^2 \boldsymbol{y}^{n+1} = h^2 \sum_{k=0}^{N+1} \alpha_k \nabla^k \boldsymbol{f}^{n+1}, \qquad (\textit{corrector}) \tag{A.2.9}$$

$$\nabla^2 \boldsymbol{y}^{n+1} = h^2 \sum_{k=0}^{N+1} \beta_k \nabla^k \boldsymbol{f}^{n}. \qquad (\textit{predictor}) \tag{A.2.10}$$

We observe that the vectors $\boldsymbol{g}^n$ obey the rules

$$\boldsymbol{g}^{-1} = \boldsymbol{\sigma}, \tag{A.2.11}$$

$$\boldsymbol{g}^n = h \sum_{m=0}^{n} \boldsymbol{f}^m + \boldsymbol{\sigma} \text{ for } n \geq 0.$$

It is easily checked that

$$\nabla \boldsymbol{g}^n = h\boldsymbol{f}^n \text{ for } n \geq \ 0. \tag{A.2.12}$$

Insert (2.12) into (2.9). The result is

$$\nabla^2 \boldsymbol{y}^{n+1} = h\nabla \sum_{k=0}^{N+1} \alpha_k \nabla^k \boldsymbol{g}^{n+1}. \tag{A.2.13}$$

Suppose we could peel off a ∇ from each side of (2.13). The result would be

$$\nabla \boldsymbol{y}^{n+1} = h \sum_{k=0}^{N+1} \alpha_k \nabla^k \boldsymbol{g}^{n+1}. \tag{A.2.14}$$

Normally this operation is not justified since, by (2.4.5), the two sides of (2.14) could differ by a *constant* vector. However, in our case we assert that the value of $\boldsymbol{\sigma}$ was cleverly defined in (2.2) to insure that (2.14) would be correct. To check this claim, set $n+1 = N$ in (2.14). The result, using (2.3), is

$$\nabla \boldsymbol{y}^N = h\alpha_o \boldsymbol{\sigma} + h \sum_{k=0}^{N+1} \alpha_k \nabla^k \boldsymbol{G}^N. \tag{A.2.15}$$

From Table 1 we find $\alpha_o = 1$, and, after expansion, we see that (2.15) is equivalent to (2.2). Thus (2.14) is correct. Upon expansion it gives the corrector (2.4).

Let us now work on the predictor (2.10). Use of (2.12) gives

$$\nabla^2 \boldsymbol{y}^{n+1} = h\nabla \sum_{k=0}^{N+1} \beta_k \nabla^k \boldsymbol{g}^n. \tag{A.2.16}$$

Again we peel off a ∇ from each side, but this time we must insert a constant vector $\boldsymbol{c}$. The result is

$$\nabla \boldsymbol{y}^{n+1} = h \sum_{k=0}^{N+1} \beta_k \nabla^k \boldsymbol{g}^n + \boldsymbol{c}. \tag{A.2.17}$$

What should $\boldsymbol{c}$ be? We shall see that to within sufficient accuracy it can be set to zero. Assuming this to be the case, the expansion of (2.17) gives the predictor (2.5).

We now estimate the size of $\boldsymbol{c}$. From the analog of (2.4.20) for $\boldsymbol{g}$ we find

$$\boldsymbol{g}^n = (1 - \nabla)\boldsymbol{g}^{n+1} \tag{A.2.18}$$

so that (2.17) can also be written as

$$\nabla \boldsymbol{y}^{n+1} = h(1-\nabla)\sum_{k=0}^{N+1} \beta_k \nabla^k \boldsymbol{g}^{n+1} + \boldsymbol{c}. \tag{A.2.19}$$

Subtract (2.14) from (2.19). The result is

$$\boldsymbol{c} = h[\sum_{k=0}^{N+1} \alpha_k \nabla^k - (1-\nabla)\sum_{k=0}^{N+1} \beta_k \nabla^k]\boldsymbol{g}^{n+1}. \tag{A.2.20}$$

From their definition in (1.4) and (1.5), the coefficients α_k and β_k satisfy the identity

$$(1-z)\sum_0^\infty \beta_k z^k = \sum_0^\infty \alpha_k z^k. \tag{A.2.21}$$

It follows that

$$\sum_{k=0}^{N+1} \alpha_k \nabla^k - (1-\nabla)\sum_{k=0}^{N+1} \beta_k \nabla^k = \beta_{N+1}\nabla^{N+2}. \tag{A.2.22}$$

Using (2.12) and (2.22), the expression for $\boldsymbol{c}$ can be simplified to give

$$\boldsymbol{c} = h^2 \beta_{N+1} \nabla^{N+1} \boldsymbol{f}^{n+1}. \tag{A.2.23}$$

Finally, remembering that $\boldsymbol{f} = \ddot{\boldsymbol{y}}$ and using (2.4.12) we find

$$\boldsymbol{c} \approx h^{N+3}\beta_{N+1}(d^{N+3}\boldsymbol{y}/dt^{N+3})|_{t=t^n}. \tag{A.2.24}$$

Since (2.17) is equivalent to (2.10), we know it has the intrinsic error given in (1.11) with N replaced by $N+1$. This error is of order h smaller than $\boldsymbol{c}$. Thus the error made in dropping $\boldsymbol{c}$ is the dominant predictor error, and (2.8) is correct.

Exercises

A.2.1. Verify (2.21).

A.3 Computation of First Derivative

It often happens that values of $\dot{\boldsymbol{y}}$ are required at various points of a trajectory. For example, from time to time we may need the velocity to compute the energy. If the trajectory is being integrated with the Adams method, the velocity is available at each step. However, with Størmer-Cowell only the coordinates $\boldsymbol{y}^n$ are computed.

This apparent defect can be overcome with numerical differentiation. We observe that by using (2.4.13) the equation

$$\dot{\boldsymbol{y}}^{n+1} = D\boldsymbol{y}^{n+1} \tag{A.3.1}$$

can be written in the form

$$\dot{\boldsymbol{y}}^{n+1} = -h^{-1}\ \log(1-\nabla)\boldsymbol{y}^{n+1}. \tag{A.3.2}$$

To use (3.2) as it stands requires the storage of previous $\boldsymbol{y}$'s. However, we may rewrite it in the form

$$\dot{\boldsymbol{y}}^{n+1} = -h^{-1}[\log(1-\nabla)/\nabla]\nabla\boldsymbol{y}^{n+1}, \tag{A.3.3}$$

or using (2.14),

$$\dot{\boldsymbol{y}}^{n+1} = -[\log(1-\nabla)/\nabla]\sum_{0}^{N+1}\alpha_k\nabla^k\boldsymbol{g}^{n+1}. \tag{A.3.4}$$

From the definition of the coeffients a_k and α_k [see (2.4.23), (1.4), and (1.6)] we learn that

$$-\ [\log(1-z)/z]\sum_{0}^{\infty}\alpha_k z^k = \sum_{0}^{\infty}a_k z^k. \tag{A.3.5}$$

Consequently, we also may write to within sufficient accuracy

$$\dot{\boldsymbol{y}}^{n+1} = \sum_{k=0}^{N+2}a_k\nabla^k\boldsymbol{g}^{n+1}, \tag{A.3.6}$$

or expanding out,

$$\dot{\boldsymbol{y}}^{n+1} = \sum_{k=0}^{N+2}\tilde{a}_k^{N+2}\ \boldsymbol{g}^{n+1-k}. \tag{A.3.7}$$

We conclude that $\dot{\boldsymbol{y}}$ can be computed in terms of the stored $\boldsymbol{g}$'s any time it is required. [If the reader is wondering about the upper summation limit of $N+2$ in (3.6) and (3.7), it is not a misprint. He or she should see Exercise 4.2 at the end of Section A.4.]

Exercises

A.3.1. Verify (3.5).

A.4 Example Program and Numerical Results

We show below, with some associated subroutines, a summed Størmer-Cowell program.

A.4.1 Program

The program is written to solve (2.2.5) with the initial conditions (2.2.6). We have set $N=3$ and $h=1/10$, and the solution is initiated with the Runge-Kutta routine rk3 using a step size of $h/20$.

```
c This is the main program for illustrating a Stormer-Cowell
c method for numerical integration.
c
      implicit double precision (a-h,o-z)
c
c Print heading.
c
      write(6,100)
  100 format
     & (1h ,’time’,4x,’ycomp’,10x,’ydcomp’,10x,’ytrue’,
     & 10x,’ydtrue’,/)
c
c Set up initial conditions and parameters. n is the number of integration
c steps we wish to make.
c
      t=0.d0
      h=.1d0
      n=15
      y=0.d0
      ydot=1.d0
c
      call sc(t,h,n,y,ydot)
c
      end
c
c This is a sixth order Stormer-Cowell integration subroutine.
c
      subroutine sc(t,h,n,y,ydot)
      implicit double precision (a-h,o-z)
      dimension g(5)
c
      write(6,*) ’Starting with Runge-Kutta integration’
c
c Set up initial g values.
c
      g(1)=0.d0
      call evalsc(y,t,f)
      g(2)=h*f
      call prints(t,y,ydot,y1true(t),y2true(t),0)
      do 10 i=2,4
      call rk3(t,h/20.d0,20,y,ydot)
      call evalsc(y,t,f)
      g(i+1)=g(i)+h*f
      if(i .eq. 3) yb=y
      call prints(t,y,ydot,y1true(t),y2true(t),0)
   10 continue
      sigma=(y-yb)/h-(1.d0/240.d0)*
     & (19.d0*g(5)+204.d0*g(4)+14.d0*g(3)+4.d0*g(2))
      do 20 i=1,5
   20 g(i)=g(i)+sigma
      hdiv=h/240.d0
      n=n-3
      tint=t
      write (6,*) ’Continuing with Stormer-Cowell integration’
```

```
c
c Printing and integration loop.
c
      do 100 i=1,n
c
c Predictor step.
c
      t=t+h
      p=y+hdiv*(299.d0*g(5)-176.d0*g(4)+194.d0*g(3)
     & -96.d0*g(2)+19.d0*g(1))
      call evalsc(p,t,f)
      g6=g(5)+h*f
      call dif(g,g6,ydot)
      call prints(t,p,ydot,y1true(t),y2true(t),0)
c
c Corrector steps.
c
      do 50 j=1,3
      c=y+hdiv*(19.d0*g6+204.d0*g(5)+14.d0*g(4)
     & +4.d0*g(3)-1.d0*g(2))
      call evalsc(c,t,f)
      g6=g(5)+h*f
      call dif(g,g6,ydot)
      call prints(t,c,ydot,y1true(t),y2true(t),1)
   50 continue
c
c Update gs
c
      do 60 j=1,4
   60 g(j)=g(j+1)
      g(5)=g6
      y=c
      t=tint+float(i)*h
  100 continue
c
      return
      end

c This subroutine computes ydot from the g values.
c
      subroutine dif(g,g6,ydot)
      implicit double precision (a-h,o-z)
      dimension g(5)
c
      ydot=(1.d0/1440.d0)*(475.d0*g6+1427.d0*g(5)-798.d0*g(4)
     & +482.d0*g(3)-173.d0*g(2)+27.d0*g(1))
c
      return
      end
c
c This subroutine evaluates f, the right side of the
c differential equation for the second order set.
c
      subroutine evalsc(y,t,f)
      implicit double precision (a-h,o-z)
c
      f=2.d0*t-y
c
      return
      end
```

A.4.2 Numerical Results

Below are the results of running this program. The format of the column *ycomp* is the same as that of *y1comp* in Example 2.4.1. The column *ydcomp* contains values of $\dot{y}$ computed using (A.42). We observe that the solution is accurate to essentially eight significant figures.

```
time    ycomp           ydcomp          ytrue           ydtrue

Starting with Runge-Kutta integration
0.0000  0.00000000E+00  0.10000000E+01  0.00000000E+00  0.10000000E+01
0.1000  0.10016658E+00  0.10049958E+01  0.10016658E+00  0.10049958E+01
0.2000  0.20133067E+00  0.10199334E+01  0.20133067E+00  0.10199334E+01
0.3000  0.30447979E+00  0.10446635E+01  0.30447979E+00  0.10446635E+01
Continuing with Stormer-Cowell integration
0.4000  0.41058164E+00  0.10789390E+01  0.41058166E+00  0.10789390E+01
        0.41058166E+00  0.10789390E+01
        0.41058166E+00  0.10789390E+01
        0.41058166E+00  0.10789390E+01
0.5000  0.52057444E+00  0.11224174E+01  0.52057446E+00  0.11224174E+01
        0.52057446E+00  0.11224174E+01
        0.52057446E+00  0.11224174E+01
        0.52057446E+00  0.11224174E+01
0.6000  0.63535750E+00  0.11746644E+01  0.63535753E+00  0.11746644E+01
        0.63535753E+00  0.11746644E+01
        0.63535753E+00  0.11746644E+01
        0.63535753E+00  0.11746644E+01
0.7000  0.75578228E+00  0.12351578E+01  0.75578231E+00  0.12351578E+01
        0.75578232E+00  0.12351578E+01
        0.75578232E+00  0.12351578E+01
        0.75578232E+00  0.12351578E+01
0.8000  0.88264387E+00  0.13032933E+01  0.88264391E+00  0.13032933E+01
        0.88264392E+00  0.13032933E+01
        0.88264392E+00  0.13032933E+01
        0.88264392E+00  0.13032933E+01
0.9000  0.10166730E+01  0.13783900E+01  0.10166731E+01  0.13783900E+01
        0.10166731E+01  0.13783900E+01
        0.10166731E+01  0.13783900E+01
        0.10166731E+01  0.13783900E+01
1.0000  0.11585290E+01  0.14596977E+01  0.11585290E+01  0.14596977E+01
        0.11585290E+01  0.14596977E+01
        0.11585290E+01  0.14596977E+01
        0.11585290E+01  0.14596977E+01
1.1000  0.13087926E+01  0.15464039E+01  0.13087926E+01  0.15464039E+01
        0.13087927E+01  0.15464039E+01
        0.13087927E+01  0.15464039E+01
        0.13087927E+01  0.15464039E+01
```

```
1.2000  0.14679609E+01  0.16376422E+01  0.14679609E+01  0.16376422E+01
        0.14679609E+01  0.16376422E+01
        0.14679609E+01  0.16376422E+01
        0.14679609E+01  0.16376422E+01
1.3000  0.16364418E+01  0.17325012E+01  0.16364418E+01  0.17325012E+01
        0.16364418E+01  0.17325012E+01
        0.16364418E+01  0.17325012E+01
        0.16364418E+01  0.17325012E+01
1.4000  0.18145502E+01  0.18300328E+01  0.18145503E+01  0.18300329E+01
        0.18145503E+01  0.18300328E+01
        0.18145503E+01  0.18300328E+01
        0.18145503E+01  0.18300328E+01
1.5000  0.20025050E+01  0.19292628E+01  0.20025050E+01  0.19292628E+01
        0.20025050E+01  0.19292628E+01
        0.20025050E+01  0.19292628E+01
        0.20025050E+01  0.19292628E+01
```

Exercises

A.4.1. Compare the error estimate (2.7) with the actual error made in the example above.

A.4.2. Check the derivation of (3.7) and find a formula similar to (2.7) for the expected error in $\dot{\boldsymbol{y}}$. Compare with the error in the example above. Suppose the sum in (3.6) were terminated at $N+1$. Show that the error in its expanded form, the analog of (3.7), would then be larger.

A.5 Nyström Runge-Kutta Methods

We close this appendix with a brief description of Nyström Runge-Kutta (NRK) methods. They are analogous to ordinary Runge-Kutta methods, but are designed to work directly with second-order equations of the form (1.1). We will present methods that are analogous to the Runge-Kutta methods RK3 given by (2.3.2), (2.3.3) and RK4 given by (2.3.4), (2.3.5). Unlike Størmer-Cowell methods, we will need integration formulas for both $\boldsymbol{y}$ and $\dot{\boldsymbol{y}}$.

The method that is analogous to RK3 is given by

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\dot{\boldsymbol{y}}^n + (h^2/6)(\boldsymbol{a} + 2\boldsymbol{b}), \tag{A.5.1}$$

$$\dot{\boldsymbol{y}}^{n+1} = \dot{\boldsymbol{y}}^n + (h/6)(\boldsymbol{a} + 4\boldsymbol{b} + \boldsymbol{c}), \tag{A.5.2}$$

where at each step

$$\boldsymbol{a} = \boldsymbol{f}(\boldsymbol{y}^n, t^n), \tag{A.5.3}$$

$$\boldsymbol{b} = \boldsymbol{f}[\boldsymbol{y}^n + (h/2)\dot{\boldsymbol{y}}^n + (h^2/8)\boldsymbol{a}, t^n + h/2], \tag{A.5.4}$$

$$\boldsymbol{c} = \boldsymbol{f}[\boldsymbol{y}^n + h\dot{\boldsymbol{y}}^n + (h^2/2)\boldsymbol{b}, t^n + h]. \tag{A.5.5}$$

This is a three-stage fourth-order method. That is, it is locally correct through order h^4, and makes local errors of order h^5. Note that this method is one order higher in accuracy than its counterpart RK3. Accordingly, we will call it NRK4. This increase in accuracy again arises from the fact that the original differential equations being integrated are second order with no first derivatives present.

The method that is analogous to RK4 is given by

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\dot{\boldsymbol{y}}^n + (h^2/192)(23\boldsymbol{a} + 75\boldsymbol{b} - 27\boldsymbol{c} + 25\boldsymbol{d}), \tag{A.5.6}$$

$$\dot{\boldsymbol{y}}^{n+1} = \dot{\boldsymbol{y}}^n + (h/192)(23\boldsymbol{a} + 125\boldsymbol{b} - 81\boldsymbol{c} + 125\boldsymbol{d}), \tag{A.5.7}$$

where at each step

$$\boldsymbol{a} = \boldsymbol{f}(\boldsymbol{y}^n, t^n), \tag{A.5.8}$$

$$\boldsymbol{b} = \boldsymbol{f}[\boldsymbol{y}^n + (2/5)h\dot{\boldsymbol{y}}^n + (2/25)h^2\boldsymbol{a}, t^n + (2/5)h], \tag{A.5.9}$$

$$\boldsymbol{c} = \boldsymbol{f}[\boldsymbol{y}^n + (2/3)h\dot{\boldsymbol{y}}^n + (2/9)h^2\boldsymbol{a}, t^n + (2/3)h], \tag{A.5.10}$$

$$\boldsymbol{d} = \boldsymbol{f}[\boldsymbol{y}^n + (4/5)h\dot{\boldsymbol{y}}^n + (4/25)h^2(\boldsymbol{a} + \boldsymbol{b}), t^n + (4/5)h]. \tag{A.5.11}$$

This is a four-stage fifth-order method, and is again one order higher in accuracy than its counterpart RK4. Accordingly, we will call it NRK5.

Nyström Runge-Kutta methods can also be described in terms of Butcher tableaux. Let $\bar{b}$, b, and c be s-dimensional vectors with real entries, and let $\bar{a}$ be an $s \times s$ matrix with real entries. Consider stepping formulas of the form

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\dot{\boldsymbol{y}}^n + h^2 \sum_{i=1}^{s} \bar{b}_i \boldsymbol{k}_i, \tag{A.5.12}$$

$$\dot{\boldsymbol{y}}^{n+1} = \dot{\boldsymbol{y}}^n + h \sum_{i=1}^{s} b_i \boldsymbol{k}_i, \tag{A.5.13}$$

where at each step

$$\boldsymbol{k}_i = \boldsymbol{f}(\boldsymbol{y}^n + hc_i\dot{\boldsymbol{y}}^n + h^2 \sum_{j=1}^{s} \bar{a}_{ij}\boldsymbol{k}_j,\ t^n + c_i h). \tag{A.5.14}$$

Evidently the procedures (5.1) through (5.5) and (5.6) through (5.11) are of this form.

In terms of the the notation just introduced, the general problem now is to impose various conditions on the vectors $\bar{b}$, b, and c and the matrix $\bar{a}$ so that the integration method will be of some particular order m, and perhaps have some other desirable properties. For this purpose, it is convenient to arrange the vectors $\bar{b}$, b, and c and the matrix $\bar{a}$ into a tableau (again called a Butcher tableau) of the form

$$\begin{array}{c|ccc} c_1 & \bar{a}_{11} & \cdots & \bar{a}_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & \bar{a}_{s1} & \cdots & \bar{a}_{ss} \\ \hline & & & \\ & \bar{b}_1 & \cdots & \bar{b}_s \\ \hline & b_1 & \cdots & b_s \end{array}\ . \tag{A.5.15}$$

The Butcher tableau for NRK4, the method (5.1) through (5.5), is

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 1/8 & 0 & 0 \\ 1 & 0 & 1/2 & 0 \\ \hline & 1/6 & 2/6 & 0 \\ \hline & 1/6 & 4/6 & 1/6 \end{array} . \tag{A.5.16}$$

The Butcher tableau for NRK5, the method (5.6) through (5.11), is

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 2/5 & 2/25 & 0 & 0 & 0 \\ 2/3 & 2/9 & 0 & 0 & 0 \\ 4/5 & 4/25 & 4/25 & 0 & 0 \\ \hline & 23/192 & 75/192 & -27/192 & 25/192 \\ \hline & 23/192 & 125/192 & -81/192 & 125/192 \end{array} . \tag{A.5.17}$$

At this point we observe that, as in the case of ordinary Runge Kutta, it is sometimes useful to rewrite the relations (5.12) through (5.14) in a somewhat different form. At each step introduce intermediate times t_i and coordinates $\boldsymbol{y}_i$ by the rules

$$t_i = t^n + c_i h, \tag{A.5.18}$$

$$\boldsymbol{y}_i = \boldsymbol{y}^n + hc_i\dot{\boldsymbol{y}}^n + h^2\sum_{j=1}^{s}\bar{a}_{ij}\boldsymbol{k}_j. \tag{A.5.19}$$

With this convention (5.14) can be rewritten in the form

$$\boldsymbol{k}_i = \boldsymbol{f}(\boldsymbol{y}_i, t_i). \tag{A.5.20}$$

Finally we copy (5.12) and (5.13) and place them last,

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\dot{\boldsymbol{y}}^n + h^2\sum_{i=1}^{s}\bar{b}_i\boldsymbol{k}_i, \tag{A.5.21}$$

$$\dot{\boldsymbol{y}}^{n+1} = \dot{\boldsymbol{y}}^n + h\sum_{i=1}^{s}b_i\boldsymbol{k}_i, \tag{A.5.22}$$

Evidently the relations (5.18) through (5.22) are equivalent to the relations (5.12) through (5.14), but in this expanded form it is clear that the $\boldsymbol{k}_i$ are the values of $\boldsymbol{f}$ at the intermediate points t_i, $\boldsymbol{y}_i$.

Finally, we remark that there are Nyström counterparts to embedded Runge-Kutta pairs so that it is possible to develop Nyström procedures with adaptive step-size control. Also, as described in Section 12.2, there are symplectic Nyström-Runge-Kutta procedures. See the books of Hairer *et al.* and Sanz-Serna and Calvo listed in the bibliography at the end of this appendix.

Exercises

A.5.1. Apply one step of the fourth-order Nyström method (5.1) through (5.5) to the differential equation

$$\ddot{y} = t^2 \tag{A.5.23}$$

with the initial conditions

$$y(0) = 0 \text{ and } \dot{y}(0) = 0. \tag{A.5.24}$$

That is, use (5.1) and (5.2) to compute $y(h)$ and $\dot{y}(h)$. Verify that your result has the advertised accuracy. Repeat the calculation for the case $\ddot{y} = t^3$, and show that, as expected, the result is not exact.

Bibliography

[1] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations.* (John Wiley 1962) QA 372.H48. *Error Propagation for Difference Methods.* (John Wiley 1963) QA 431.H44.

[2] J.F. Frankena, "Størmer-Cowell: straight, summed and split. An overview", *J. Computational and Applied Mathematics* **62**, p. 129-154 (1995).

[3] E. Hairer, S. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I. Nonstiff Problems*, Springer (1993).

[4] J.M. Sanz-Serna and M.P. Calvo, *Numerical Hamiltonian Problems*, Chapman and Hall (1994).

Appendix B

Computer Programs for Numerical Integration

In this appendix we list computer programs that are more efficient versions of those described in Chapter 2. The programs are all subroutines, and need to be supplemented by a main calling program and input and output statements of the reader's own design. For the most part, the programs are self-explanatory. All the integration programs call the subroutine *eval* (or *feval*), which computes the vector $\boldsymbol{f}$ appearing in the differential equation $\dot{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{y}, t)$. To change from one set of differential equations to another, it is only necessary to change *eval.* The integration routines themselves are general purpose. One only need specify *ne*, the number of equations to be simultaneously integrated. As listed, the routines are set up to integrate the equation used in the Examples in Chapter 2, *i.e.* $\ddot{x} + x = 2t$.

The routines given in this appendix are suitable for *semi-serious* use. Readers who wish to pursue numerical integration in a serious way may wish to use a canned integration package. Much work has gone into writing some such routines. However, the reader should be sure to understand how the package he or she has selected actually works. Some produce unpleasant surprises. See the references at the end of Chapter 2.

B.1 A 3^{rd} Order Runge-Kutta Routine

B.1.1 Butcher Tableau for $RK3$

The Butcher tableau for $RK3$ is given in (2.3.9).

B.1.2 The Routine $RK3$

```
      subroutine rk3 (h,ns,nf,t,y)
c This is a Runge Kutta routine that makes local errors of order
c h**4.  h is the step size, ns is the number of steps, t is the
c time, and y is the dependent variable array.  Finally, nf is a
c flag to control whatever.
c In the next line, put in after ne the number of equations to be
c integrated.
      parameter (ne=2)
      dimension y(ne),yt(ne),f(ne),a(ne),b(ne),c(ne)
c yt is a temporary storage array and f is ydot.
c a, b, and c are used in integration.
      tint=t
c tint is the initial time.
      do 100 i=1,ns
      call eval (t,y,f)
c eval is a subroutine that evaluates ydot.
      do 10 j=1,ne
   10 a(j)=h*f(j)
      do 20 j=1,ne
   20 yt(j)=y(j)+.5*a(j)
      tt=t+.5*h
c tt is a temporary time
      call eval (tt,yt,f)
      do 30 j=1,ne
   30 b(j)=h*f(j)
      do 40 j=1,ne
   40 yt(j)=y(j)+2.*b(j)-a(j)
      tt=t+h
      call eval (tt,yt,f)
      do 50 j=1,ne
   50 c(j)=h*f(j)
      do 60 j=1,ne
   60 y(j)=y(j)+(a(j)+4.*b(j)+c(j))/6.
      t=tint+float(i)*h
  100 continue
      return
      end
```

Note: The *flag* `nf` is not actually used. It is incorporated to make the program easier to modify.

B.2 A 4^{th} Order Runge-Kutta Routine

B.2.1 Butcher Tableau for $RK4$

The Butcher tableau for $RK4$ is given in (2.3.10).

B.2.2 The Routine $RK4$

```
      subroutine rk4 (h,ns,nf,t,y)
c This is a Runge Kutta routine that makes local errors of order
c h**5.  h is the step size, ns is the number of steps, t is the
c time, and y is the dependent variable array.  Finally, nf is a
c flag to control whatever.
c In the next line, put in after ne the number of equations to be
c integrated.
      parameter (ne=2)
      dimension y(ne),yt(ne),f(ne),a(ne),b(ne),c(ne),d(ne)
c yt is a temporary storage array and f is ydot.
c a, b, c, and d are used in integration.
      tint=t
c tint is the initial time.
      do 100 i=1,ns
      call eval (t,y,f)
c eval is a subroutine that evaluates ydot.
      do 10 j=1,ne
   10 a(j)=h*f(j)
      do 20 j=1,ne
   20 yt(j)=y(j)+.5*a(j)
      tt=t+.5*h
c tt is a temporary time
      call eval (tt,yt,f)
      do 30 j=1,ne
   30 b(j)=h*f(j)
      do 40 j=1,ne
   40 yt(j)=y(j)+.5*b(j)
      call eval (tt,yt,f)
      do 50 j=1,ne
   50 c(j)=h*f(j)
      do 60 j=1,ne
   60 yt(j)=y(j)+c(j)
      tt=t+h
      call eval (tt,yt,f)
      do 70 j=1,ne
   70 d(j)=h*f(j)
      do 80 j=1,ne
   80 y(j)=y(j)+(a(j)+2.*b(j)+2.*c(j)+d(j))/6.
      t=tint+float(i)*h
  100 continue
      return
      end
```

Note: The *flag* `nf` is not actually used. It is incorporated to make the program easier to modify.

B.3 A Subroutine to Compute f

```
      subroutine eval (t,y,f)
c This is a subroutine that evaluates ydot.
      dimension y(*),f(*)
c In the following lines put in the expressions for the f(i).
      f(1)=y(2)
      f(2)=2.*t-y(1)
      return
      end
```

B.4 A Partial Double-Precision Version of $RK3$

```
      subroutine rk3pdp (h,ns,nf,t,y)
c This is a Runge Kutta routine that works in partial double precision
c and makes local errors of order h**4.  h is the step size, ns is the
c number of steps, t is the time, and y is the dependent variable array.
c Finally, nf is a flag to control whatever.
c In the next line, put in after ne the number of equations to be
c integrated.
      parameter (ne=2)
      dimension y(ne),yt(ne),f(ne),a(ne),b(ne),c(ne),yd(ne)
      double precision yd
c yt is a temporary storage array and f is ydot.
c yd is a double precision storage array.
c a, b, and c are used in integration.
      tint=t
c tint is the initial time.
      do 2 j=1,ne
    2 yd(j)=dble(y(j))
c The input array y is transferred into the double precision array yd.
c Beginning of rk3 loop.
      do 100 i=1,ns
      call eval (t,y,f)
c eval is a subroutine that evaluates ydot.
      do 10 j=1,ne
   10 a(j)=h*f(j)
      do 20 k=1,ne
   20 yt(j)=y(j)+.5*a(j)
      tt=t+.5*h
c tt is a temporary time
      call eval (tt,yt,f)
      do 30 j=1,ne
   30 b(j)=h*f(j)
      do 40 j=1,ne
   40 yt(j)=y(j)+2.*b(j)-a(j)
      tt=t+h
      call eval (tt,yt,f)
      do 50 j=1,ne
   50 c(j)=h*f(j)
      do 60 j=1,ne
   60 yd(j)=yd(j)+dble((a(j)+4.*b(j)+c(j))/6.)
c The array yd is incremented in double precision.
      t=tint+float(i)*h
      do 70 j=1,ne
   70 y(j)=sngl(yd(j))
c Preparation of y for transfer out or the next run through the loop.
  100 continue
      return
      end
```

The *error curve* for this routine is exactly the same as that given in Figure (2.3.1) when the step size h is .05 or larger. However, for smaller h the error curve is much better. The error continues to decrease as h^3 until h reaches a value a little less than 10^{-7} and then remains approximately constant as h is decreased further. This is because the only serious round-off error occurs in the statement $y(j) = sngl(yd(j))$, and this error is independent of the number of steps. We see that partial double precision is worthwhile if good accuracy is required.

B.5 A 6^{th} Order 8 Stage Runge-Kutta Routine

This section describes a sixth-order eight-stage Runge-Kutta routine. Note that, according to Table 2.3.1, it should be possible to achieve sixth-order accuracy using seven stages. Therefore the routine in this section, while workable, is not optimal with regard to employing only the minimum number of required stages. On the other hand, again according to Table 2.3.1, there is no eight-stage method that has an order higher than six.

B.5.1 Butcher Tableau for $RK6$

The Butcher tableau for RK6 is

$$\begin{array}{c|cccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\hline
 & 1/6 & 2/6 & 2/6 & 1/6 & * & * & * & *
\end{array} . \tag{B.5.1}$$

B.5.2 The Routine $RK6$

```
      subroutine rk6(h,ns,t,y)
c  Written by Rob Ryne, Spring 1986, based on a routine of
c  J. Milutinovic.
c  For a reference, see page 76 of F. Ceschino and J. Kuntzmann,
c  Numerical Solution of Initial Value Problems, Prentice Hall 1966.
c  This integration routine makes local truncation errors at each
c  step of order h**7.
c  That is, it is locally correct through terms of order h**6.
c  Each step requires 8 function evaluations: The method has
c  8 stages.
c
      implicit double precision (a-h,o-z)
c
      parameter (ne=2)
      dimension y(ne),yt(ne),f(ne),a(ne),b(ne),c(ne),d(ne),
     # e(ne),g(ne),o(ne),p(ne)
c
      tint=t
      do 200 i=1,ns
      call feval(t,y,f)
      do 10 j=1,ne
   10 a(j)=h*f(j)
      do 20 j=1,ne
   20 yt(j)=y(j)+a(j)/9.d+0
      tt=t+h/9.d+0
      call feval(tt,yt,f)
      do 30 j=1,ne
   30 b(j)=h*f(j)
      do 40 j=1,ne
   40 yt(j)=y(j) + (a(j) + 3.d+0*b(j))/24.d+0
      tt=t+h/6.d+0
      call feval(tt,yt,f)
      do 50 j=1,ne
   50 c(j)=h*f(j)
      do 60 j=1,ne
   60 yt(j)=y(j)+(a(j)-3.d+0*b(j)+4.d+0*c(j))/6.d+0
      tt=t+h/3.d+0
      call feval(tt,yt,f)
      do 70 j=1,ne
   70 d(j)=h*f(j)
      do 80 j=1,ne
   80 yt(j)=y(j) + (-5.d+0*a(j) + 27.d+0*b(j) -
     # 24.d+0*c(j) + 6.d+0*d(j))/8.d+0
      tt=t+.5d+0*h
      call feval(tt,yt,f)
      do 90 j=1,ne
   90 e(j)=h*f(j)
      do 100 j=1,ne
  100 yt(j)=y(j) + (221.d+0*a(j) - 981.d+0*b(j) +
     # 867.d+0*c(j)- 102.d+0*d(j) + e(j))/9.d+0
      tt = t+2.d+0*h/3.d+0
      call feval(tt,yt,f)
      do 110 j=1,ne
  110 g(j)=h*f(j)
      do 120 j=1,ne
  120 yt(j) = y(j)+(-183.d+0*a(j)+678.d+0*b(j)-472.d+0*c(j)-
     #  66.d+0*d(j)+80.d+0*e(j) + 3.d+0*g(j))/48.d+0
      tt = t + 5.d+0*h/6.d+0
      call feval(tt,yt,f)
      do 130 j=1,ne
  130 o(j)=h*f(j)
      do 140 j=1,ne
  140 yt(j) = y(j)+(716.d+0*a(j)-2079.d+0*b(j)+1002.d+0*c(j)+
     # 834.d+0*d(j)-454.d+0*e(j)-9.d+0*g(j)+72.d+0*o(j))/82.d+0
      tt = t + h
      call feval(tt,yt,f)
      do 150 j=1,ne
  150 p(j)=h*f(j)
      do 160 j=1,ne
  160 y(j) = y(j)+(41.d+0*a(j)+216.d+0*c(j)+27.d+0*d(j)+
     #  272.d+0*e(j)+27.d+0*g(j)+216.d+0*o(j)+41.d+0*p(j))/840.d+0
      t=tint+i*h
  200 continue
      return
      end
```

Note: This program calls the subroutine `feval`, which is another name for `eval`.

B.6 Embedded Runge-Kutta Pairs

In this section we will describe the construction of Runge-Kutta pairs and provide two examples. Our discussion here is meant to provide background, but not actual code. As in the case of other adaptive codes, much time has been spent by professional mathematicians and numerical analysts writing optimal code for embedded Runge-Kutta procedures. Readers are advised not to try writing such code on their own without first exploring existing programs and without being prepared to expend considerable time and effort.

B.6.1 Preliminaries

Section 2.5.1 sketched the possibility of pairs of Runge-Kutta methods whose orders differ (usually by one) and that, in making one integration step, share many or all intermediate evaluation points. By subtracting the higher-order result from the lower-order result, one can estimate the error in the lower-order result, and adjust the step size accordingly. In this section we will describe two examples of Runge-Kutta pairs. Each example has the feature that both methods of each pair employ the same $\boldsymbol{k}$ vectors. Thus both methods can be carried out simultaneously with little added expense.

Equations (2.3.6) through (2.3.8) described the general Runge-Kutta method and its characterization by an associated Butcher table. If there are two methods that utilize the same $\boldsymbol{k}$ vectors, we may make definitions of the form

$$\boldsymbol{y}^{n+1} = \boldsymbol{y}^n + h\sum_{i=1}^{s} b_i\boldsymbol{k}_i, \quad \text{stepping formula} \tag{B.6.1}$$

$$\hat{\boldsymbol{y}}^{n+1} = \boldsymbol{y}^n + h\sum_{i=1}^{s} \hat{b}_i\boldsymbol{k}_i, \quad \text{error estimator} \tag{B.6.2}$$

where at each step

$$\boldsymbol{k}_i = \boldsymbol{f}(\boldsymbol{y}^n + h\sum_{j=1}^{s} a_{ij}\boldsymbol{k}_j,\ t^n + c_i h). \tag{B.6.3}$$

This pair of methods may be described by an extended Butcher tableau of the form

$$\begin{array}{c|ccc}
c_1 & a_{11} & \cdots & a_{1s} \\
\vdots & \vdots & & \vdots \\
c_s & a_{s1} & \cdots & a_{ss} \\
\hline
 & b_1 & \cdots & b_s \\
\hline
 & \hat{b}_1 & \cdots & \hat{b}_s
\end{array} \tag{B.6.4}$$

As the annotation is intended to indicate, the relation (6.1) is to be used as a stepping formula to propagate the solution, and the relation (6.2) is to be used in conjunction with (6.1) to estimate and control the local error in making a given step.

B.6.2 Fehlberg 4(5) Pair

Fehlberg was the first to propose and develop embedded pairs. One such pair, called Fehlberg 4(5), is that described by the (extended) Butcher tableau below:

Butcher Tableau for Fehlberg 4(5)

$$\begin{array}{c|cccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\
\frac{3}{8} & \frac{3}{32} & \frac{9}{32} & 0 & 0 & 0 & 0 \\
\frac{12}{13} & \frac{1932}{2197} & -\frac{7200}{2197} & \frac{7296}{2197} & 0 & 0 & 0 \\
1 & \frac{439}{216} & -8 & \frac{3680}{513} & -\frac{845}{4104} & 0 & 0 \\
\frac{1}{2} & -\frac{8}{27} & 2 & -\frac{3544}{2565} & \frac{1859}{4104} & -\frac{11}{40} & 0 \\
\hline
 & \frac{25}{216} & 0 & \frac{1408}{2565} & \frac{2197}{4104} & -\frac{1}{5} & 0 \\
\hline
 & \frac{16}{135} & 0 & \frac{6656}{12825} & \frac{28561}{56430} & -\frac{9}{50} & \frac{2}{55}
\end{array} \tag{B.6.5}$$

The procedure has $s = 6$ stages. The stepping formula is locally exact through terms of order h^4. The error estimator is locally exact through terms of order h^5. This procedure is therefore referred to as a 4(5) procedure. Because the stepping formula used to propagate the solution is of order 4, the whole procedure itself is locally exact through terms of order $m = 4$. Note that, according to Table 2.3.1, the highest order a 6-stage method can have is $m = 5$. Although the order 4 is relatively low in view of the number of stages involved, there is some freedom in selecting the entries in the matrix a and the vectors b and $\hat{b}$; and Fehlberg selected them to minimize the size of the order h^5 error terms in the stepping formula.

Error Estimation

Since $\boldsymbol{y}^{n+1}$ is locally exact through order 4 and $\hat{\boldsymbol{y}}^{n+1}$ is locally exact through order 5, we may define a local error vector $\boldsymbol{\Delta}^n$ by the rule

$$\boldsymbol{\Delta}^n = \boldsymbol{y}^{n+1} - \hat{\boldsymbol{y}}^{n+1} = h \sum_{i=1}^{6} d_i \boldsymbol{k}_i \tag{B.6.6}$$

where

$$d_i = b_i - \hat{b}_i. \tag{B.6.7}$$

Note that in this approach only $\boldsymbol{y}^{n+1}$ is computed using (6.1), $\hat{\boldsymbol{y}}^{n+1}$ is not computed, and $\boldsymbol{\Delta}^n$ is computed using the far right side of (6.6). So doing minimizes work and round-off error.

Control of Step Size

We now wish to use $||\boldsymbol{\Delta}^n||$ to control the subsequent step size h_{n+1} or to specify how to repeat, if necessary, the current step with a smaller step size.[1] Exactly how to do so is an art based on considerable experience with various possibilities, and is best left to professional numerical analysts. Typical programs require the user to specify some desired *tolerance* Tol and perhaps some initial step size h_0, and the program automatically adjusts the step size as the integration proceeds based on this information.[2] As is the case with adaptive predictor-corrector and extrapolation methods, the prospective user is advised to first try and understand programs professionally written before attempting to write any of his/her own.

Interpolation-Dense Output

When employing a fixed step size method it is easy to integrate from some initial time t^0 to the final time $t^0 + T$ simply by requiring the relation

$$Nh = T \tag{B.6.8}$$

between the step size h and the interval duration T. Recall Section 2.1. However, when the time step is variable, the time $t^0 + T$ is generally not among the times t^n at which the $\boldsymbol{y}^n$ are computed, and so the quantity $\boldsymbol{y}(t^0 + T)$ is not among the $\boldsymbol{y}^n$.

In the case of a Runge-Kutta method, since it is a single-step method, this problem of finding an accurate $\boldsymbol{y}(t^0 + T)$ is fairly easy to solve. First, over the course of integration, monitor the times t^n and find the first such time, call it t^{n^*}, for which

$$t^{n^*} \geq t^0 + T. \tag{B.6.9}$$

Next, define a step size h^* by the rule

$$h^* = t^0 + T - t^{n^*}. \tag{B.6.10}$$

Note that $h^* \leq 0$. Finally, execute one step of the Rung-Kutta method in use with the time step h^* and the initial condition $\boldsymbol{y}^{n^*}$. So doing determines $\boldsymbol{y}(t^0 + T)$ to the accuracy of the integration method. In effect, this method integrates backward from the time t^{n^*} to the time $t^0 + T$.

There are some situations, for example when graphical output is needed, in which one desires an accurate and efficient method for finding $\boldsymbol{y}(t^n + \theta h_n)$ for any $\theta \in [0, 1]$. There are procedures that prepare, at each integration step, polynomials in θ for this purpose, and these procedures utilize the $\boldsymbol{k}$ vectors computed in the course of a Runge-Kutta step. See, for example, the book of Hairer, Nørsett, and Wanner cited at the end of this appendix.

[1] For computational efficiency, in this application it is convenient to define the vector norm $|| * ||$ to be the component moduli sum norm (3.7.20).

[2] There are also procedures for determining h_0 automatically so that only Tol need be specified.

B.6.3 Dormand-Prince 5(4) Pair

Inspired by the work of Fehlberg, Dormand and Prince and others developed procedures for which the stepping formula is higher order than the error estimator, and the stepping formula is optimized to minimize its still higher-order error terms.[3] One procedure of Dormand and Prince is specified by the following Butcher tableau:

Butcher Tableau for Dormand-Prince 5(4)

$$\begin{array}{c|ccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{3}{10} & \frac{3}{40} & \frac{9}{40} & 0 & 0 & 0 & 0 & 0 \\
\frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & 0 & 0 & 0 & 0 \\
\frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & 0 & 0 & 0 \\
1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} & 0 & 0 \\
1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\
\hline
 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\
\hline
 & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40}
\end{array} \tag{B.6.11}$$

The procedure has 7 stages. But the method for the stepping formula is FSAL, and therefore only requires the work of a 6-stage method after the first integration step. See the material at the end of Subsection 2.3.4. The stepping formula is locally exact through terms of order h^5. The error estimator is locally exact through terms of order h^4. This method is therefore referred to as a 5(4) method.[4] Because the stepping formula is of order 5, and is used to propagate the solution, the whole procedure itself is locally exact through terms of order $m = 5$. Note that, according to Table 2.3.1, the highest order a 6-stage method can have is $m = 5$. Thus the order is optimum in view of the effective number of stages involved. Moreover, there is still some freedom in selecting the entries in the matrix a and the vectors b and $\hat{b}$. As mentioned at the beginning of this subsection, Dormand and Prince selected them to minimize the size of the order h^6 error terms in the stepping formula.[5]

[3] Procedures that use the higher-order formula for stepping are called *local extrapolation* procedures. Note that here the word *extrapolation* has a different meaning than in Section 2.6.

[4] Some authors use the notation (4)5. However, whatever the notation, it is always the order of the stepping formula that is not in parentheses, and the order of the error estimator that is in parentheses.

[5] Why choose, for the stepping formula, a seven-stage FSAL method rather than a six-stage non-FSAL method? Both choices could yield a fifth-order stepping formula and a fourth-order error estimator. Dor-

Error Estimation

Since $\boldsymbol{y}^{n+1}$ is locally exact through order 5 and $\hat{\boldsymbol{y}}^{n+1}$ is locally exact through order 4, we may now define a local error vector $\boldsymbol{\Delta}^n$ by the rule

$$\boldsymbol{\Delta}^n = \hat{\boldsymbol{y}}^{n+1} - \boldsymbol{y}^{n+1} = -h \sum_{i=1}^{7} d_i \boldsymbol{k}_i. \tag{B.6.12}$$

Note that in carrying out the sum (6.12) the $i = 2$ term may be omitted since, according to (6.11), both b_2 and $\hat{b}_2$ vanish, and therefore $d_2 = 0$.

Control of Step Size

We again wish to use $||\boldsymbol{\Delta}^n||$ to control the step size. In this case it should be understood that what is now being controlled is the error in the lower order error estimator with the hope that, since it is of one order higher, the error in the stepping formula will be even better controlled. Again, there are several possible procedures, and again the prospective user is advised to first try and understand programs professionally written before attempting to write any of his/her own.

Interpolation-Dense Output

The same methods and considerations apply here as in Subsection 6.2.

mand and Prince found that by so doing they were better able to choose the entries in the matrix a and the vectors b and $\hat{b}$ to minimize the order h^6 error terms in the stepping formula.

B.7 A 5^{th} Order PECEC Adams Routine

```
      subroutine adams5 (h,ns,nf,t,y)
c This is an N=4 PECEC Adams routine that is locally correct
c through terms of order h**5 and makes local errors of
c order h**6.
c h is the step size, and ns is the number of steps.  t is the time, and
c y is the dependent variable array.  nf is a flag that controls the mode
c of entry.  If nf = 'start', the trajectory is started with Runge Kutta.
c If nf = 'cont', the solution is continued using previous "f" values.
c ns must exceed 4 when Adams is called with nf = 'start'.
c In the next line, put in after ne the number of equations to be
c integrated.
      parameter (ne=2)
      dimension y(ne),yp(ne),yc(ne),f1(ne),f2(ne),f3(ne),
     & f4(ne),f5(ne),f6(ne)
c yp and yc are corrector arrays.  f1 through f6 form the array of
c "f" values.
      dimension a(5),am(5),b(5),bm(5)
c a,am and b,bm are coefficients used in the corrector and predictor,
c respectively.
      data (a(i), i=1,5) /-19.,106.,-264.,646.,251./
      data (b(i), i=1,5) /251.,-1274.,2616.,-2774.,1901./
      save am,bm,f1,f2,f3,f4,f5
      nsa=ns
c nsa is the number of steps to be made by Adams.
      if (nf .eq. 'cont') go to 20
c When nf = 'cont', the integration has already been started earlier,
c and "f" values are assumed to exist.  Otherwise, start with Runge Kutta.
c Set up the initial f array using Runge Kutta.
      call eval (t,y,f1)
c eval is a subroutine that evaluates ydot.
      call rk3 (h/5.,5,0,t,y)
      call eval (t,y,f2)
      call rk3 (h/5.,5,0,t,y)
      call eval (t,y,f3)
      call rk3 (h/5.,5,0,t,y)
      call eval (t,y,f4)
      call rk3 (h/5.,5,0,t,y)
      call eval (t,y,f5)
c Now go into the finite difference procedure.
      nsa=ns-4
c nsa is the number of steps to be made by Adams.  If the integration
c began with Runge Kutta, Adams has 4 fewer steps to make.
      hdiv=h/720.
      do 10 i=1,5
      am(i)=hdiv*a(i)
   10 bm(i)=hdiv*b(i)
c am and bm are used in the corrector and predictor.
   20 tint=t
c tint is the initial time for Adams.
      do 100 i=1,nsa
c Begin with predictor.
```

```
      do 30 j=1,ne
   30 yp(j)=y(j)+bm(1)*f1(j)+bm(2)*f2(j)+bm(3)*f3(j)
     & +bm(4)*f4(j)+bm(5)*f5(j)
c First evaluation.
      call eval (t+h,yp,f6)
c First use of corrector.  Here we use yp as a storage array.
      do 40 j=1,ne
      yp(j)=y(j)+am(1)*f2(j)+am(2)*f3(j)+am(3)*f4(j)
     & +am(4)*f5(j)
   40 yc(j)=yp(j)+am(5)*f6(j)
c Second evaluation.
      call eval (t+h,yc,f6)
c Second use of corrector.
      do 50 j=1,ne
   50 y(j)=yp(j)+am(5)*f6(j)
c Update table of f values.
      do 60 j=1,ne
      f1(j)=f2(j)
      f2(j)=f3(j)
      f3(j)=f4(j)
      f4(j)=f5(j)
   60 f5(j)=f6(j)
      t=tint+float(i)*h
  100 continue
      return
      end
```

Note: Here the *flag* `nf` controls the mode of entry.

B.8 A 10^{th} Order PECEC Adams Routine

```
      subroutine adams10(h,ns,nf,t,y)
c  Written by Rob Ryne, Spring 1986, based on a routine of Alex Dragt.
c  This N=9 Adams integration routine makes local truncation errors
c  at each step of order h**11.  That is, it is locally correct through
c  order h**10.  Due to round off errors, its true precision is
c  realized only when more than 64 bits are used.
c  Warning: because this is a high-order method, the step size must be
c  correspondingly small to achieve stability.  For example, for the simple
c  harmonic oscillator with unit frequency (xdoubleprime+x=0), at least
c  50 steps per oscillation are require to safely achieve stability and
c  for the error analysis based on finite-difference considerations
c  to be relevant.
      implicit double precision (a-h,o-z)
c
      character*6 nf
      parameter (ne=2)
c
      dimension y(ne),yp(ne),yc(ne),f1(ne),f2(ne),f3(ne),f4(ne),
     # f5(ne),f6(ne),f7(ne),f8(ne),f9(ne),f10(ne),f11(ne)
c
      dimension a(10),am(10),b(10),bm(10)
c
      data (a(i),i=1,10)/57281.d0,-583435.d0,2687864.d0,
     # -7394032.d0,13510082.d0,-17283646.d0,16002320.d0,
     # -11271304.d0,9449717.d0,2082753.d0/
      data (b(i),i=1,10)/-2082753.d0,20884811.d0,-94307320.d0,
     # 252618224.d0,-444772162.d0,538363838.d0,-454661776.d0,
     # 265932680.d0,-104995189.d0,30277247.d0/
c
      nsa=ns
      if (nf.eq.'cont') go to 20
c
c  rk start
      iqt=5
      qt=float(iqt)
      hqt=h/qt
      call feval(t,y,f1)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f2)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f3)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f4)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f5)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f6)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f7)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f8)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f9)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f10)
      nsa=ns-9
      hdiv=h/7257600.0d+00
       do 10 i=1,10
       am(i)=hdiv*a(i)
  10  bm(i)=hdiv*b(i)
c
c  Adams routine
c
  20  tint=t
      do 100 i=1,nsa
      do 30 j=1,ne
      yp(j)=y(j)+bm(1)*f1(j)+bm(2)*f2(j)+bm(3)*f3(j)
     # +bm(4)*f4(j)+bm(5)*f5(j)+bm(6)*f6(j)+bm(7)*f7(j)
     # +bm(8)*f8(j)+bm(9)*f9(j)+bm(10)*f10(j)
   30 continue
      call feval(t+h,yp,f11)
      do 40 j=1,ne
      yp(j)=y(j)+am(1)*f2(j)+am(2)*f3(j)+am(3)*f4(j)+am(4)*f5(j)
     # +am(5)*f6(j)+am(6)*f7(j)+am(7)*f8(j)+am(8)*f9(j)+am(9)*f10(j)
  40  yc(j)=yp(j)+am(10)*f11(j)
  41  call feval(t+h,yc,f11)
      do 50 j=1,ne
   50 y(j)=yp(j)+am(10)*f11(j)
      do 60 j=1,ne
      f1(j)=f2(j)
      f2(j)=f3(j)
      f3(j)=f4(j)
      f4(j)=f5(j)
      f5(j)=f6(j)
      f6(j)=f7(j)
      f7(j)=f8(j)
      f8(j)=f9(j)
      f9(j)=f10(j)
   60 f10(j)=f11(j)
      t=tint+i*h
  100 continue
      return
      end
```

Notes: This program calls the subroutine `feval`, which is another name for `eval`. Here the *flag* `nf` controls the mode of entry.

Bibliography

[1] E. Hairer, S. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I. Nonstiff Problems*, Springer (1993).

[2] J.C. Butcher, *The numerical analysis of ordinary differential equations: Runge-Kutta and general linear methods*, John Wiley, (1987).

[3] J.C. Butcher, *Numerical Methods for Ordinary Differential Equations*, First Edition, John Wiley (2003).

[4] J.C. Butcher, *Numerical Methods for Ordinary Differential Equations*, Second Edition, John Wiley (2008). http://www.math.auckland.ac.nz/~butcher/ODE-book-2008/.

[5] J. Dormand and P. Prince, "A family of embedded Runge-Kutta formulae", *J. Comp. Appl. Math.* **6**, 19-26 (1980).

Appendix C

Baker-Campbell-Hausdorff and Zassenhaus Formulas, Bases, and Paths

The purpose of this appendix is to describe the Lie-algebraic results of Henry Frederick Baker (1866-1956), John Edward Campbell (1862-1924), and Felix Hausdorff (1868-1942), and the related results of Hans Zassenhaus (1912-1991). We also discuss differentiating the exponential function, bases for Lie algebras, and paths in Lie groups and Lie algebras.

C.1 Differentiating the Exponential Function

C.2 The Baker-Campbell-Hausdorff Formula

C.3 The Baker-Campbell-Hausdorff Series

$$
\begin{aligned}
\log(e^y e^x) = x + y &- \frac{1}{2}[x,y] + \frac{1}{12}[x,[x,y]] + \frac{1}{12}[[x,y],y] - \frac{1}{24}[x,[[x,y],y]] \\
&- \frac{1}{720}[x,[x,[x,[x,y]]]] + \frac{1}{180}[x,[x,[[x,y],y]]] + \frac{1}{180}[x,[[[x,y],y],y]] \\
&+ \frac{1}{120}[[x,y],[[x,y],y]] + \frac{1}{360}[[x,[x,y]],[x,y]] - \frac{1}{720}[[[[x,y],y],y]y] \\
&+ \frac{1}{1440}[x,[x,[x,[[x,y],y]]]] - \frac{1}{360}[x,[x,[[[x,y],y],y]]] - \frac{1}{240}[x,[[x,y],[[x,y],y]]] \\
&- \frac{1}{720}[x,[[x,[x,y]],[x,y]]] + \frac{1}{1440}[x,[[[[x,y],y],y],y]] + \frac{1}{30240}[x,[x,[x,[x,[x,[x,y]]]]]] \\
&- \frac{1}{5040}[x,[x,[x,[x,[[x,y],y]]]]] + \frac{1}{3780}[x,[x,[x,[[[x,y],y],y]]]] \\
&+ \frac{1}{1680}[x,[x,[[x,y],[[x,y],y]]]] + \frac{1}{10080}[x,[x,[[x,[x,y]],[x,y]]]] \\
&+ \frac{1}{3780}[x,[x,[[[[x,y],y],y],y]]] + \frac{13}{15120}[x,[[x,y],[[[x,y],y],y]]]
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1}{1260}[x,[[x,[[x,y],y]],[x,y]]] - \frac{1}{5040}[x,[[[[[x,y],y],y],y],y]] \\
&+ \frac{1}{1260}[[x,y],[[x,y],[[x,y],y]]] - \frac{1}{2016}[[x,y],[[[[x,y],y],y],y]] \\
&+ \frac{1}{2016}[[x,[x,y]],[x,[[x,y],y]]] - \frac{1}{5040}[[[x,y],y],[[[x,y],y],y]] \\
&+ \frac{1}{10080}[[x,[x,[x,y]]],[x,[x,y]]] + \frac{1}{10080}[[x,[[x,y],y]],[[x,y],y]] \\
&- \frac{1}{1512}[[x,[[[x,y],y],y]],[x,y]] - \frac{1}{5040}[[[x,[x,y]],[x,y]],[x,y]] \\
&+ \frac{1}{30240}[[[[[[x,y],y],y],y],y],y] - \frac{1}{60480}[x,[x,[x,[x,[x,[[x,y],y]]]]]] \\
&+ \frac{1}{10080}[x,[x,[x,[x,[[[x,y],y],y]]]]] + \frac{1}{20160}[x,[x,[x,[[x,y],[[x,y],y]]]]] \\
&+ \frac{1}{15120}[x,[x,[x,[[x,[x,y]],[x,y]]]]] - \frac{23}{120960}[x,[x,[x,[[[[x,y],y],y],y]]]] \\
&- \frac{13}{30240}[x,[x,[[x,y],[[[x,y],y],y]]]] - \frac{1}{2520}[x,[x,[[x,[[x,y],y]],[x,y]]]] \\
&+ \frac{1}{10080}[x,[x,[[[[[x,y],y],y],y],y]]] - \frac{1}{2520}[x,[[x,y],[[x,y],[[x,y],y]]]] \\
&+ \frac{1}{4032}[x,[[x,y],[[[[x,y],y],y],y]]] - \frac{1}{4032}[x,[[x,[x,y]],[x,[[x,y],y]]]] \\
&+ \frac{1}{10080}[x,[[[x,y],y],[[[x,y],y],y]]] - \frac{1}{20160}[x,[[x,[x,[x,y]]],[x,[x,y]]]] \\
&- \frac{1}{20160}[x,[[x,[[x,y],y]],[[x,y],y]]] + \frac{1}{3024}[x,[[x,[[[x,y],y],y]],[x,y]]] \\
&+ \frac{1}{10080}[x,[[[x,[x,y]],[x,y]],[x,y]]] - \frac{1}{60480}[x,[[[[[[x,y],y],y],y],y],y]] \\
&- \frac{1}{1209600}[x,[x,[x,[x,[x,[x,[x,[x,y]]]]]]]] + \frac{1}{151200}[x,[x,[x,[x,[x,[x,[[x,y],y]]]]]]] \\
&- \frac{1}{56700}[x,[x,[x,[x,[x,[[[x,y],y],y]]]]]] - \frac{1}{43200}[x,[x,[x,[x,[[x,y],[[x,y],y]]]]]] \\
&- \frac{1}{100800}[x,[x,[x,[x,[[x,[x,y]],[x,y]]]]]] + \frac{1}{75600}[x,[x,[x,[x,[[[[x,y],y],y],y]]]]] \\
&+ \frac{11}{302400}[x,[x,[x,[[x,y],[[[x,y],y],y]]]]] + \frac{1}{75600}[x,[x,[x,[[x,[[x,y],y]],[x,y]]]]] \\
&+ \frac{1}{75600}[x,[x,[x,[[[[[x,y],y],y],y],y]]]] + \frac{1}{100800}[x,[x,[[x,y],[[x,y],[[x,y],y]]]]] \\
&+ \frac{1}{20160}[x,[x,[[x,y],[[[[x,y],y],y],y]]]] + \frac{1}{67200}[x,[x,[[x,[x,y]],[x,[[x,y],y]]]]] \\
&+ \frac{1}{30240}[x,[x,[[[x,y],y],[[[x,y],y],y]]]] + \frac{11}{201600}[x,[x,[[x,[[x,y],y]],[[x,y],y]]]] \\
&+ \frac{11}{151200}[x,[x,[[x,[[[x,y],y],y]],[x,y]]]] + \frac{1}{43200}[x,[x,[[[x,[x,y]],[x,y]],[x,y]]]] \\
&- \frac{1}{56700}[x,[x,[[[[[[x,y],y],y],y],y],y]]] + \frac{1}{6048}[x,[[x,y],[[x,y],[[[x,y],y],y]]]] \\
&- \frac{23}{302400}[x,[[x,y],[[[[[x,y],y],y],y],y]]] + \frac{23}{302400}[x,[[x,[x,y]],[x,[[[x,y],y],y]]]] \\
&+ \frac{1}{37800}[x,[[x,[x,y]],[[x,[x,y]],[x,y]]]] - \frac{11}{120960}[x,[[[x,y],y],[[[[x,y],y],y],y]]]
\end{aligned}
$$

$$+ \frac{1}{25200}[x,[[x,[x,[[x,y],y]]],[x,[x,y]]]] - \frac{1}{15120}[x,[[x,[[[x,y],y],y]],[[x,y],y]]]$$
$$- \frac{1}{20160}[x,[[[x,y],[[x,y],y]],[[x,y],y]]] + \frac{1}{33600}[x,[[x,[[x,y],[[x,y],y]]],[x,y]]]$$
$$- \frac{1}{7560}[x,[[x,[[[[x,y],y],y],y]],[x,y]]] - \frac{17}{100800}[x,[[[x,[[x,y],y]],[x,y]],[x,y]]]$$
$$+ \frac{1}{151200}[x,[[[[[[x,y],y],y],y],y],y],y]] + \frac{1}{10080}[[x,y],[[x,y],[[x,y],[[x,y],y]]]]$$
$$- \frac{1}{7560}[[x,y],[[x,y],[[[[x,y],y],y],y]]] - \frac{1}{15120}[[x,y],\ [[[x,y],y],[[[x,y],y],y]]]$$
$$+ \frac{1}{43200}[[x,y],[[[[[x,y],y],y],y],y],y]] + \frac{1}{25200}[[x,[x,y]],[x,[[x,y],[[x,y],y]]]]$$
$$- \frac{1}{17280}[[x,[x,y]],[x,[[[[x,y],y],y],y]]] - \frac{1}{8400}[[x,[x,y]],[[x,[[x,y],y]],[x,y]]]$$
$$+ \frac{1}{33600}[[[x,y],y],[[[[[x,y],y],y],y],y]] + \frac{1}{43200}[[x,[x,[x,y]]],[x,[x,[[x,y],y]]]]$$
$$+ \frac{1}{60480}[[x,[[x,y],y]],[x,[[[x,y],y],y]]] + \frac{1}{20160}[[x,[[x,y],y]],[[x,y],[[x,y],y]]]$$
$$+ \frac{1}{60480}[[[[x,y],y],y],[[[[x,\ y],y],y],y]] + \frac{1}{302400}[[x,[x,[x,[x,y]]]],[x,[x,[x,y]]]]$$
$$+ \frac{1}{67200}[[x,[x,[[x,y],y]]],[x,[[x,y],y]]] + \frac{1}{90720}[[x,[[[x,y],y],y]],[[[x,y],y],y]]$$
$$- \frac{1}{25200}[[[x,[x,y]],[x,y]],[x,[[x,y],y]]] - \frac{1}{21600}[[x,[x,[[[x,y],y],y]]],[x,[x,y]]]$$
$$- \frac{1}{30240}[[x,[[x,[x,y]],[x,y]]],[x,[x,y]]] + \frac{1}{60480}[[x,[[[[x,y],y],y],y]],[[x,y],y]]$$
$$+ \frac{1}{15120}[[[x,y],[[[x,y],y],y]],[[x,y],y]] - \frac{1}{10080}[[x,[[x,y],[[[x,y],y],y]]],[x,y]]$$
$$+ \frac{1}{21600}[[x,[[[[[x,y],y],y],y],y]],[x,y]] + \frac{1}{20160}[[[,x[[x,y],y]],[[x,y],y]],[x,y]]$$
$$+ \frac{1}{10080}[[[x,[[[x,y],y],y]],[x,y]],[x,y]] + \frac{1}{50400}[[[[x,[x,y]],\ [x,y]],[x,y]],[x,y]]$$
$$- \frac{1}{1209600}[[[[[[[[x,y],y],y],y],y],y],y],y] + \frac{1}{2419200}[x,[x,[x,[x,[x,[x,[x,[[x,y],y]]]]]]]]$$
$$- \frac{1}{302400}[x,[x,[x,[x,[x,[x,[[[x,y],y],y]]]]]]] + \frac{1}{604800}[x,[x,[x,[x,[x,[[x,y],[[x,y],y]]]]]]]$$
$$- \frac{1}{403200}[x,[x,[x,[x,[x,[[x,[x,y]],[x,y]]]]]]] + \frac{37}{3628800}[x,[x,[x,[x,[x,[[[[x,y],y],y],y]]]]]]$$
$$+ \frac{1}{56700}[x,[x,[x,[x,[[x,y],[[[x,y],y],y]]]]]] + \frac{1}{37800}[x,[x,[x,[x,[[x,[[x,y],y]],[x,y]]]]]]$$
$$- \frac{1}{67200}[x,[x,[x,[x,[[[[[x,y],y],y],y],y]]]]] + \frac{17}{604800}[x,[x,[x,[[x,y],\ [[x,y],[[x,y],y]]]]]]$$
$$- \frac{11}{241920}[x,[x,[x,[[x,y],[[[[x,y],y],y],y]]]]] + \frac{1}{75600}[x,[x,[x,[[x,[x,y]],[x,[[x,y],y]]]]]]$$
$$- \frac{1}{40320}[x,[x,[x,[[[x,y],y],[[[x,y],y],y]]]]] + \frac{1}{241920}[x,[x,[x,[[x,[x,[x,y]]],[x,[x,y]]]]]]$$
$$- \frac{1}{43200}[x,[x,[x,[[x,[[x,y],y]],[[x,y],y]]]]] - \frac{29}{453600}[x,[x,[x,[[x,[[[x,y],y],y]],[x,y]]]]]$$
$$- \frac{1}{50400}[x,[x,[x,[[[x,[x,y]],[x,y]],[x,y]]]]] + \frac{37}{3628800}[x,[x,[x,[[[[[[x,y],y],y],y],y],y]]]]$$

$$
\begin{aligned}
&- \frac{1}{12096}[x,[x,[[x,y],[[x,y],[[[x,y],y],y]]]]] + \frac{23}{604800}[x,[x,[[x,y],[[[[[x,y],y],y],y],y]]]]\\
&- \frac{23}{604800}[x,[x,[[x,[x,y]],[x,[[[x,y],y],y]]]]] - \frac{1}{75600}[x,[x,[[x,[x,y]],[[x,[x,y]],[x,y]]]]]\\
&+ \frac{11}{241920}[x,[x,[[[x,y],y],[[[[x,y],y],y],y]]]] - \frac{1}{50400}[x,[x,[[x,[x,[[x,y],y]]],[x,[x,y]]]]]\\
&+ \frac{1}{30240}[x,[x,[[x,[[[x,y],y],y]],[[x,y]y]]]] + \frac{1}{40320}[x,[x,[[[x,y],[[x,y],y]],[[x,y]y]]]]\\
&- \frac{1}{67200}[x,[x,[[x,[[x,y],[[x,y],y]]],[x,y]]]] + \frac{1}{15120}[x,[x,[[x,[[[[x,y],y],y],y]],[x,y]]]]\\
&+ \frac{17}{201600}[x,[x,[[[x,[[x,y],y]],[x,y]],[x,y]]]] - \frac{1}{302400}[x,[x,[[[[[[x,y],y],y],y],y],y],y]]]\\
&- \frac{1}{20160}[x,[[x,y],[[x,y],[[x,y],[[x,y],y]]]]] + \frac{1}{15120}[x,[[x,y],[[x,y],[[[[x,y],y],y],y]]]]\\
&+ \frac{1}{30240}[x,[[x,y],[[[x,y],y],[[[x,y],y],y]]]] - \frac{1}{86400}[x,[[x,y],[[[[[x,y],y],y],y],y],y]]]\\
&- \frac{1}{50400}[x,[[x,[x,y]],[x,[[x,y],[[x,y],y]]]]] + \frac{1}{34560}[x,[[x,[x,y]],[x,[[[[x,y],y],y],y]]]]\\
&+ \frac{1}{16800}[x,[[x,[x,y]],[[x,[x,y]],[x,y]]]] - \frac{1}{67200}[x,[[[x,y],y],[[[[[x,y],y],y],y],y]]]\\
&- \frac{1}{86400}[x,[[x,[x,[x,y]]],[x,[x,[[x,y],y]]]]] - \frac{1}{120960}[x,[[x,[[x,y],y]],[x,[[[x,y],y],y]]]]\\
&- \frac{1}{40320}[x,[[x,[[x,y],y]],[[x,y],[[x,y],y]]]] - \frac{1}{120960}[x,[[[[x,y],y],y],[[[[x,y],y],y],y]]]\\
&- \frac{1}{604800}[x,[[x,[x,[x,[x,y]]]],[x,[x,[x,y]]]]] - \frac{1}{134400}[x,[[x,[x,[[x,y],y]]],[x,[[x,y],y]]]]\\
&- \frac{1}{181440}[x,[[x,[[[x,y],y],y]],[[[x,y],y],y]]] + \frac{1}{50400}[x,[[[x,[x,y]],[x,y]],[x,[[x,y],y]]]]\\
&+ \frac{1}{43200}[x,[[x,[x,[[[x,y],y],y]]],[x,[x,y]]]] + \frac{1}{60480}[x,[[x,[[x,[x,y]],[x,y]]],[x,[x,y]]]]\\
&- \frac{1}{120960}[x,[[x,[[[[x,y],y],y],y]],[[x,y],y]]] - \frac{1}{30240}[x,[[[x,y],[[[x,y],y],y]],[[x,y],y]]]\\
&+ \frac{1}{20160}[x,[[x,[[x,y],[[[x,y],y],y]]],[x,y]]] - \frac{1}{43200}[x,[[x,[[[[[x,y],y],y],y],y]],[x,y]]]\\
&- \frac{1}{40320}[x,[[[x,[[x,y],y]],[[x,y],y]],[x,y]]] - \frac{1}{20160}[x,[[[x,[[[x,y],y],y]],[x,y]],[x,y]]]\\
&- \frac{1}{100800}[x,[[[[x,[x,y]],[x,y]],[x,y]],[x,y]]] + \frac{1}{2419200}[x,[[[[[[[[x,y],y],y],y],y],y],y],y]]\\
&+ \cdots
\end{aligned}
$$

Exercises

C.3.1. According to the BCH theorem the *exponential* function has the remarkable property that the quantity C in the relation

$$\exp(A)\exp(B) = \exp(C) \tag{C.3.1}$$

depends only on elements in the Lie algebra generated by A and B. See Section 3.7.3 and the BCH series in Section C.3 above. The purpose of this exercise is to study the properties of two other functions.

Begin with the truncated exponential function t1exp defined by

$$\mathrm{t1exp}(A) = I + A. \tag{C.3.2}$$

See (4.1.22). Show that

$$\mathrm{t1exp}(A)\mathrm{t1exp}(B) = \mathrm{t1exp}(C) \tag{C.3.3}$$

with

$$C = A + B + AB. \tag{C.3.4}$$

Evidently the degree 1 term $A + B$ is in the Lie algebra generated by A and B, but the degree 2 term AB is not.

Next consider the truncated exponential function t2exp defined by

$$\mathrm{t2exp}(A) = I + A + A^2/2!. \tag{C.3.5}$$

Show that in this case

$$\mathrm{t2exp}(A)\mathrm{t2exp}(B) = \mathrm{t2exp}(C) \tag{C.3.6}$$

with

$$C = A + B + (1/2)(AB - BA) + \cdots . \tag{C.3.7}$$

Evidently the terms explicitly displayed on the right side of (C.3.7) are in the Lie algebra generated by A and B. Show that the remaining terms, which are of degree 3 and higher, are not.

C.3.2. Exercise on BCH like relation for the Cayley function.

C.4 Zassenhaus Formulas

C.5 Bases

C.6 Paths

C.6.1 Paths in the Group Yield Paths in the Lie Algebra

C.6.2 Paths in the Lie Algebra Yield Paths in the Group

C.6.3 Differential Equations

Bibliography

[1] P.V. Koseleff, Formal calculus for Lie methods in Hamiltonian mechanics, Ph.D. Thesis, École Polytechnique (1993).

[2] E.B. Dynkin, "On the representation by means of commutators of the series $\log(e^x e^y)$ for noncommutative x and y", *Mat. Sbornik (N.S.)* **25**(67), 155-162 (1949).

[3] Karl Goldberg, "The Formal Power Series for Log $e^x e^y$", *Duke Journal of Mathematics* **23**, 13 (1956).

[4] Erik Eriksen, "Properties of Higher-Order Commutator Products and the Baker-Hausdorff Formula", *Journal of Mathematical Physics* **9**, 790 (1968).

[5] J.B. Kogut, *Rev. Mod. Phys.* **51**, 4 (1979).

[6] J.A. Oteo, "The Baker-Campbell-Hausdorff formula and nested commutator identities", *J. Math. Phys.* **32**, 419-424 (1991).

[7] A. Bonfiglioli and R. Fulci, *Topics in Noncommutative Algebra: The Theorem of Campbell, Baker, Hausdorff, and Dynkin*, Lecture Notes in Mathematics 2034, Springer (2012).

[8] F. Casas and A Murua, "An efficient algorithm for computing the Baker-Campbell-Hausdorff series and some of its applications", *J. Math. Phys.* **50**, 033513 (2009).

Appendix D

Canonical Transformations

Appendix E

Mathematica Notebooks

Appendix F

Analyticity, Aberration Expansions, and Smoothing

F.1 The Static Case

According to Poincaré's Theorem 1.3.3, trajectories will be analytic functions of the initial conditions in some domain if the right sides of the equations of motion (1.3.4) are analytic. Correspondingly, according to the results of Section 26.2, the Taylor map (7.5.5) will converge in some domain about the origin. For problems of particular interest to us the Hamiltonians, and hence the equations of motion, will involve the scalar and vector potentials ψ and $\boldsymbol{A}$ as in, for example, (1.5.29), (1.6.16), and (1.6.17). Consequently, we are interested in knowing the analytic properties of ψ and $\boldsymbol{A}$.

In the static case these potentials are determined in terms of the charge density $\rho(\boldsymbol{r})$ and the current density $\boldsymbol{j}(\boldsymbol{r})$ by the Poisson equations

$$\nabla^2\psi = -4\pi\rho, \tag{F.1.1}$$

$$\nabla^2\boldsymbol{A} = -4\pi\boldsymbol{j}/c, \tag{F.1.2}$$

which have the solutions

$$\psi(\boldsymbol{r}) = \int d^3\boldsymbol{r}'\rho(\boldsymbol{r}')/\parallel \boldsymbol{r}'-\boldsymbol{r} \parallel, \tag{F.1.3}$$

$$\boldsymbol{A}(\boldsymbol{r}) = (1/c)\int d^3\boldsymbol{r}'\boldsymbol{j}(\boldsymbol{r}')/\parallel \boldsymbol{r}'-\boldsymbol{r} \parallel . \tag{F.1.4}$$

Here the notation $\parallel\ \parallel$ indicates the Euclidean norm,

$$\parallel \boldsymbol{r}'-\boldsymbol{r} \parallel = [(x'-x)^2+(y'-y)^2+(z'-z)^2]^{1/2}. \tag{F.1.5}$$

It follows immediately from (1.1) that if $\psi(\boldsymbol{r})$ is analytic in the components x, y, z of $\boldsymbol{r}$ at some point $\boldsymbol{r}^0$, then $\rho(\boldsymbol{r})$ must also be analytic at $\boldsymbol{r}^0$. The purpose of this appendix is to show the converse: if $\rho(\boldsymbol{r})$ is analytic at some point $\boldsymbol{r}^0$, then $\psi(\boldsymbol{r})$ will also be analytic at $\boldsymbol{r}^0$. We note, with some surprise, that this is a *local* statement. Although, according to (1.3), the value of $\psi(\boldsymbol{r})$ at the point $\boldsymbol{r}$ is determined by the value of $\rho(\boldsymbol{r}')$ at *all* points $\boldsymbol{r}'$, the quantity $\rho(\boldsymbol{r}')$ need not be analytic everywhere, but only at $\boldsymbol{r}^0$, for $\psi(\boldsymbol{r})$ to be analytic at

$\boldsymbol{r}^0$. Finally, since (1.4) is analogous to (1.3), our proof will also show that if all components of $\boldsymbol{j}(\boldsymbol{r})$ are analytic at $\boldsymbol{r}^0$, then all components of $\boldsymbol{A}(\boldsymbol{r})$ will also be analytic at $\boldsymbol{r}^0$.

Before proceeding further, and so as to not raise the reader's expectations too high, we confess that the analyticity we will prove is analyticity in the vicinity of *real* points. That is, at any real point (x^0, y^0, z^0), we will be able to prove analyticity in the complex variables $(x^0 + i\tilde{x}, y^0 + i\tilde{y}, z^0 + i\tilde{z})$ for $\tilde{x}$, $\tilde{y}$, $\tilde{z}$ finite but possibly small.

Why should we be interested in this question? First, we note that if $\rho(\boldsymbol{r})$ is zero in some region, then it is automatically analytic in that region. Therefore, as a particular case, we will find that vacuum solutions to Poisson's equation (solutions to Laplace's equation) must be analytic. This particular case is in fact the most common case since we are usually interested in orbits that remain within evacuated beam pipes. However, in some cases we are interested in the behavior of orbits that pass through regions of nonzero charge and/or current densities. Examples that come to mind include plasma lenses, electron cloud lenses, lithium lenses, beam-beam effects, and space-charge effects. In these cases we may still expect to have convergent aberration expansions provided $\rho(\boldsymbol{r})$ and $\boldsymbol{j}(\boldsymbol{r})$ are analytic in the region *traversed* by all orbits of interest. We emphasize again that no analyticity assumptions need be made about the behavior of $\rho(\boldsymbol{r})$ and $\boldsymbol{j}(\boldsymbol{r})$ in regions not traversed by orbits. From a mathematical perspective, the discussion that follows will be somewhat discursive; but it will have the advantage of obtaining several interesting results along the way. For a more direct approach, see Exercises 17 and 18.

By a suitable translation, and without loss of generality, we may take $\boldsymbol{r}^0$ to be the origin. Then, by the theory of Section 26.2, the assumption that ρ is analytic implies that it has an expansion of the form

$$\rho(\boldsymbol{r}) = \sum_{ijk} c_{ijk} x^i y^j z^k \tag{F.1.6}$$

that converges in some polydisc $\mathcal{D}$ given by inequalities of the form

$$|x| < R_x, \ |y| < R_y, \ |z| < R_z. \tag{F.1.7}$$

Here we are treating x, y, z as three *complex* variables. For further discussion, it is convenient to work within a smaller domain $\mathcal{R}$ contained within $\mathcal{D}$. Let ϵ be some fixed small positive number. Define a quantity R by the rule

$$R = \text{ minimum of } (R_x - \epsilon), (R_y - \epsilon), (R_z - \epsilon). \tag{F.1.8}$$

Then we define $\mathcal{R}$ to be the closed set

$$|x| \le R, \ |y| \le R, \ |z| \le R. \tag{F.1.9}$$

We are now in a position to obtain a bound on the Taylor coefficients c_{ijk}. From the Cauchy formula

$$c_{jk\ell} = \frac{1}{(2\pi i)^3} \oint_{|x|=R} \oint_{|y|=R} \oint_{|z|=R} dxdydz \frac{\rho(x,y,z)}{x^{j+1} y^{k+1} z^{\ell+1}} \tag{F.1.10}$$

we get the result

$$|c_{jk\ell}| \le K R^{-(j+k+\ell)} \tag{F.1.11}$$

where the constant K is defined by the equation

$$K = \max |\rho(x, y, z)| \text{ for } |x| = |y| = |z| = R. \tag{F.1.12}$$

Suppose the terms in the series (1.6) are grouped together according to their total degree. Doing so gives the result

$$\rho(\boldsymbol{r}) = \sum_{D=0}^{\infty} \sum_{\alpha=1}^{N(D,3)} d_{D\alpha} h_D^\alpha(\boldsymbol{r}). \tag{F.1.13}$$

Here $h_D^\alpha(\boldsymbol{r})$ denotes a monomial of the form $x^i y^j z^k$ and of degree D (that is, $i + j + k = D$) and labelled by an index α. From (6.3.36) we know that $N(D, 3)$, the number of monomials of degree D in 3 variables, is given by the relation

$$N(D, 3) = \frac{(D + 2)!}{D!2!} = (D + 2)(D + 1)/2. \tag{F.1.14}$$

Hence, we may label the monomials of degree D in such a way that the index α ranges over the values of 1 to $N(D, 3)$, as indicated in (1.13). We note that that (1.13) is an ordering and grouping of (1.6), and hence is a permissible operation that cannot change its value for $\boldsymbol{r} \in \mathcal{R}$. See Sections 26.1 through 26.3.

We next study the relation between the monomials h_D^α, harmonic polynomials, and spherical harmonics. For the moment let x, y, and z be real. Introduce, in the standard way, spherical coordinates by the relations

$$x = r \sin\theta \cos\phi, \tag{F.1.15}$$

$$y = r \sin\theta \sin\phi, \tag{F.1.16}$$

$$z = r \cos\theta, \tag{F.1.17}$$

with r given by the relation

$$r = (x^2 + y^2 + z^2)^{1/2}. \tag{F.1.18}$$

Now consider the functions $H_\ell^m(\boldsymbol{r})$ defined by the relation

$$H_\ell^m(\boldsymbol{r}) = r^\ell Y_\ell^m(\theta, \phi) \tag{F.1.19}$$

where the Y_ℓ^m are the usual *spherical harmonics*. We observe that the H_ℓ^m are *homogeneous polynomials* of degree ℓ in the variables x, y, z. In fact, for H_ℓ^ℓ we have the relation

$$\begin{aligned} H_\ell^\ell(\boldsymbol{r}) &= r^\ell Y_\ell^\ell(\theta, \phi) = (2\ell + 1)^{1/2} \{4\pi[(2\ell)!]\}^{-1/2} r^\ell P_\ell^\ell(\cos\theta) \\ &= [(-1)^\ell/(2^\ell \ell!)]\{[(2\ell + 1)/(4\pi)][(2\ell)!]\}^{1/2} r^\ell (\sin\theta)^\ell e^{i\ell\phi} \\ &= [(-1)^\ell/(2^\ell \ell!)]\{[(2\ell + 1)/(4\pi)][(2\ell)!]\}^{1/2} (x + iy)^\ell. \end{aligned} \tag{F.1.20}$$

To see that the remaining H_ℓ^m are also homogeneous polynomials of degree ℓ we define, in the usual way, the angular momentum operator $\boldsymbol{\mathcal{L}}$ by the rule

$$\boldsymbol{\mathcal{L}} = -i\boldsymbol{r} \times \boldsymbol{\partial}. \tag{F.1.21}$$

By convention and construction the Y_ℓ^m have the property

$$\mathcal{L}_- Y_\ell^m = [(\ell+m)(\ell-m+1)]^{1/2} Y_\ell^{m-1} \tag{F.1.22}$$

where $\mathcal{L}_-$ is defined by the rule

$$\mathcal{L}_- = \mathcal{L}_x - i\mathcal{L}_y. \tag{F.1.23}$$

It is easily verified that $\boldsymbol{\mathcal{L}}$ commutes with r, and hence we also have the relation

$$\mathcal{L}_- H_\ell^m = [(\ell+m)(\ell-m+1)]^{1/2} H_\ell^{m-1}. \tag{F.1.24}$$

Moreover, we see from (1.21) and (1.23) that $\mathcal{L}_-$ maps homogeneous polynomials into homogeneous polynomials, and leaves their degrees unchanged. It follows that all the H_ℓ^m are homogeneous polynomials of degree ℓ. Finally, we see from (1.19) that the H_ℓ^m satisfy Laplace's equation,

$$\nabla^2 H_\ell^m = 0, \tag{F.1.25}$$

and therefore are entitled to be called *harmonic* polynomials.

Consider now the functions $H_\ell^{ms}(\boldsymbol{r})$ defined by the relations

$$H_\ell^{ms}(\boldsymbol{r}) = r^{2s} r^\ell Y_\ell^m = r^{2s} H_\ell^m(\boldsymbol{r}),\ s = 0, 1, 2, \cdots . \tag{F.1.26}$$

They are evidently homogeneous polynomials in x, y, z of degree D, with D given by the relation

$$D = \ell + 2s. \tag{F.1.27}$$

See (1.18). They are also linearly independent. Let us count their number for fixed D. The cases of even and odd D need to be treated separately. For even D we have the polynomials

$$r^D Y_D^m, r^2 r^{D-2} Y_{D-2}^m, \cdots r^{D-2} r^2 Y_2^m, r^D, \tag{F.1.28}$$

and their total number is

$$\begin{aligned} \text{number} &= [2D+1] + [2(D-2)+1] + \cdots + [2(2)+1] + 1 \\ &= (D+2)(D+1)/2 = N(D,3). \end{aligned} \tag{F.1.29}$$

For odd D we have the polynomials

$$r^D Y_D^m, r^2 r^{D-2} Y_{D-2}^m, \cdots r^{D-1} r Y_1^m, \tag{F.1.30}$$

and their total number is

$$\begin{aligned} \text{number} &= [2D+1] + [2(D-2)+1] + \cdots + [2(1)+1] \\ &= (D+2)(D+1)/2 = N(D,3). \end{aligned} \tag{F.1.31}$$

Comparison of (1.14), (1.29), and (1.31) shows that the number of homogeneous polynomials H_ℓ^{ms} with fixed degree D equals the number of monomials h_D^α with fixed D. Consequently, there exist expansion coefficients a and relations of the form

$$\begin{aligned} x^i y^j z^k &= h_D^\alpha(\boldsymbol{r}) = \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} a_{m\ell s}^{D\alpha} r^{2s} r^\ell Y_\ell^m \\ &= \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} a_{m\ell s}^{D\alpha} H_\ell^{ms}(\boldsymbol{r}). \end{aligned} \tag{F.1.32}$$

With the aid of (1.32), the series (1.13) can also be written in the form

$$\begin{aligned}\rho(\boldsymbol{r}) &= \sum_{D=0}^{\infty}\sum_{\ell+2s=D}\sum_{m=-\ell}^{\ell} b_{m\ell s} r^{2s} r^{\ell} Y_{\ell}^{m} \\ &= \sum_{D=0}^{\infty}\sum_{\ell+2s=D}\sum_{m=-\ell}^{\ell} b_{m\ell s} r^{2s} H_{\ell}^{m}(\boldsymbol{r}). \qquad \text{(F.1.33)}\end{aligned}$$

In this form we see, as a consequence of analyticity, that spherical harmonics Y_ℓ^m always occur in conjunction with powers of r of the form $r^{\ell+2s}$ with $s = 0, 1, 2, \cdots$.

The transition from the Taylor series (1.13) to what we will call the *harmonic* series (1.33) is not simply a different ordering, and therefore questions of convergence have to be examined anew. We begin by finding bounds for the polynomials $H_\ell^m(\boldsymbol{r})$. For this purpose it is convient to use the familiar expansion

$$\frac{1}{\parallel \boldsymbol{r}' - \boldsymbol{r} \parallel} = 4\pi \sum_{\ell,m} (2\ell+1)^{-1} r^{\ell} Y_\ell^m(\Omega)(r')^{-\ell-1}\overline{Y}_\ell^m(\Omega') \text{ for } r < r'. \qquad \text{(F.1.34)}$$

For real angles we note the bound

$$|Y_\ell^m(\Omega)| \leq [(2\ell+1)/4\pi]^{1/2}. \qquad \text{(F.1.35)}$$

It follows that the expansion (1.34) is absolutely convergent for $\boldsymbol{r}$ and $\boldsymbol{r}'$ real and $r < r'$. Now multiply both sides of (134) by $Y_{\ell'}^{m'}(\Omega')$, integrate over Ω', and use the orthogonality of the Y_ℓ^m to get the integral representation

$$H_\ell^m(\boldsymbol{r}) = r^\ell Y_\ell^m(\Omega) = (4\pi)^{-1}(2\ell+1)(r')^{\ell+1}\int d\Omega'\ Y_\ell^m(\Omega')/\parallel \boldsymbol{r}' - \boldsymbol{r} \parallel. \qquad \text{(F.1.36)}$$

As it stands, the representation (1.36) holds for $\boldsymbol{r}$ real and satisfying $r < r'$. We will now analytically continue it to possibly complex $\boldsymbol{r}$ while keeping $\boldsymbol{r}'$ real. There is no difficulty in extending the left side of (1.36) to complex $\boldsymbol{r}$ since it is a polynomial. The extension of the right side is also straight forward. Moreover, when the left and right sides of (1.36) are extended to complex $\boldsymbol{r}$, they will continue to agree. That is, the integral representation (1.36) is also valid for complex $\boldsymbol{r}$. See Exercise 9. Introduce the unit vector

$$\boldsymbol{e}(\Omega') = \boldsymbol{r}'/r' = \boldsymbol{e}_x \sin\theta'\cos\phi' + \boldsymbol{e}_y \sin\theta'\sin\phi' + \boldsymbol{e}_z\cos\theta'. \qquad \text{(F.1.37)}$$

Also introduce real vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ and a complex vector $\boldsymbol{\zeta}$ by the relation

$$\boldsymbol{r}/r' = \boldsymbol{\xi} + i\boldsymbol{\eta} = \boldsymbol{\zeta}. \qquad \text{(F.1.38)}$$

With the aid of these definitions (1.36) can be recast in the form

$$H_\ell^m(\boldsymbol{r}) = (4\pi)^{-1}(2\ell+1)(r')^{\ell}\int d\Omega'\ Y_\ell^m(\Omega')/\parallel \boldsymbol{e}(\Omega') - \boldsymbol{\zeta} \parallel. \qquad \text{(F.1.39)}$$

Let us examine the denominator $\parallel \boldsymbol{e} - \boldsymbol{\zeta} \parallel$. It has the form

$$\parallel \boldsymbol{e} - \boldsymbol{\zeta} \parallel = [(\boldsymbol{e} - \boldsymbol{\zeta})\cdot(\boldsymbol{e} - \boldsymbol{\zeta})]^{1/2}. \qquad \text{(F.1.40)}$$

For the dot product appearing in (1.40) we find the expression

$$\begin{aligned}(\boldsymbol{e}-\boldsymbol{\zeta})\cdot(\boldsymbol{e}-\boldsymbol{\zeta}) &= \boldsymbol{e}\cdot\boldsymbol{e}-2\boldsymbol{e}\cdot\boldsymbol{\zeta}+\boldsymbol{\zeta}\cdot\boldsymbol{\zeta}\\ &= 1-2\boldsymbol{e}\cdot\boldsymbol{\zeta}+\boldsymbol{\zeta}\cdot\boldsymbol{\zeta}=(1-2\boldsymbol{e}\cdot\boldsymbol{\xi}+\boldsymbol{\xi}\cdot\boldsymbol{\xi}-\boldsymbol{\eta}\cdot\boldsymbol{\eta})\\ &+ 2i(\boldsymbol{\xi}\cdot\boldsymbol{\eta}-\boldsymbol{e}\cdot\boldsymbol{\eta}).\end{aligned} \tag{F.1.41}$$

Suppose the vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are restricted in length to satisfy the relations

$$\boldsymbol{\xi}\cdot\boldsymbol{\xi}\leq 1/16,\ \ \boldsymbol{\eta}\cdot\boldsymbol{\eta}\leq 1/16. \tag{F.1.42}$$

Then the quantities appearing in the real part of (1.41) obey the inequalities

$$|\boldsymbol{e}\cdot\boldsymbol{\xi}|\leq\parallel \boldsymbol{e}\parallel\parallel\boldsymbol{\xi}\parallel\leq 1/4, \tag{F.1.43}$$

$$-1/16\leq\boldsymbol{\xi}\cdot\boldsymbol{\xi}-\boldsymbol{\eta}\cdot\boldsymbol{\eta}\leq 1/16, \tag{F.1.44}$$

and the real part itself satisfies the inequality

$$|1-2\boldsymbol{e}\cdot\boldsymbol{\xi}+\boldsymbol{\xi}\cdot\boldsymbol{\xi}-\boldsymbol{\eta}\cdot\boldsymbol{\eta}|\geq 7/16. \tag{F.1.45}$$

From (1.41) and (1.45) it follows that $\parallel \boldsymbol{e}-\boldsymbol{\zeta}\parallel$ satisfies the inequality

$$|[\parallel \boldsymbol{e}-\boldsymbol{\zeta}\parallel]|=|[(\boldsymbol{e}-\boldsymbol{\zeta})\cdot(\boldsymbol{e}-\boldsymbol{\zeta})]^{1/2}|\geq\sqrt{7}/4. \tag{F.1.46}$$

Now use (1.35) and (1.46) in (1.39) to get the bound

$$|H_\ell^m(\boldsymbol{r})|\leq(16\pi/\sqrt{7})[(2\ell+1)/4\pi]^{3/2}(r')^\ell. \tag{F.1.47}$$

Suppose $\boldsymbol{r}$ is in the closed polydisc $\mathcal{R}$ given by (1.9). Then we have the relations

$$\xi_x=\mathrm{Re}(x)/r'\leq R/r',\ \text{etc.}; \tag{F.1.48}$$

$$\eta_x=\mathrm{Im}(x)/r'\leq R/r',\ \text{etc.} \tag{F.1.49}$$

From these relations it follows that

$$\boldsymbol{\xi}\cdot\boldsymbol{\xi}\leq 3(R/r')^2, \tag{F.1.50}$$

$$\boldsymbol{\eta}\cdot\boldsymbol{\eta}\leq 3(R/r')^2. \tag{F.1.51}$$

Finally set r' to the value

$$r'=4\sqrt{3}R. \tag{F.1.52}$$

Then the inequalities (1.42) are satisfied and (1.47) takes the final form

$$|H_\ell^m(\boldsymbol{r})|\leq(16\pi/\sqrt{7})[(2\ell+1)/4\pi]^{3/2}(4\sqrt{3})^\ell R^\ell \text{ for } \boldsymbol{r}\in\mathcal{R}. \tag{F.1.53}$$

Inspection of (1.33) shows that what we really require are bounds on the polynomials $r^{2s}H_\ell^m(\boldsymbol{r})$. From the definition (1.18) and (1.9) we find the result

$$|r^2|=|x^2+y^2+z^2|\leq|x^2|+|y^2|+|z^2|\leq 3R^2. \tag{F.1.54}$$

Consequently we find for the quantity r^{2s} the bound

$$|r^{2s}| = |r^2|^s \leq 3^s R^{2s} \leq (\sqrt{3})^{2s} R^{2s} \leq (4\sqrt{3})^{2s} R^{2s}. \tag{F.1.55}$$

It follows that the polynomials $r^{2s} H_\ell^m(\boldsymbol{r})$ have the bounds

$$|r^{2s} H_\ell^m(\boldsymbol{r})| \leq (16\pi/\sqrt{7})[(2\ell+1)/4\pi]^{3/2}(4\sqrt{3})^{\ell+2s} R^{\ell+2s}. \tag{F.1.56}$$

We next find bounds on the coefficients $b_{m\ell s}$. For each degree D we have the relation

$$\sum_{i+j+k=D} c_{ijk} x^i y^j z^k = \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} b_{m\ell s} r^{\ell+2s} Y_\ell^m(\theta,\phi). \tag{F.1.57}$$

Multiply both sides of (1.57) by $\overline{Y}_{\ell'}^{m'}$ and integrate over solid angle to obtain the relation

$$b_{m'\ell' s'} r^D = r^D \int d\Omega \sum_{i+j+k=D} c_{ijk} (\sin\theta\cos\phi)^i (\sin\theta\sin\phi)^j (\cos\theta)^k \overline{Y}_{\ell'}^{m'}(\theta,\phi) \tag{F.1.58}$$

where s' satisfies the condition

$$\ell' + 2s' = D. \tag{F.1.59}$$

Here we have also used (1.15) through (1.17). It follows from (1.35) and (1.58) that we have the inequality

$$|b_{m'\ell' s'}| \leq 4\pi[(2\ell+1)/4\pi]^{1/2} \sum_{i+j+k=D} |c_{ijk}|. \tag{F.1.60}$$

Now use (1.11), and the fact that the number of terms appearing in the sum (1.60) is $N(D,3)$, to get the result

$$|b_{m'\ell' s'}| \leq 2\pi K[(2\ell+1)/4\pi]^{1/2}(D+2)(D+1)R^{-D}. \tag{F.1.61}$$

Let us use the results (1.56) and (1.61) to examine the convergence of the series (1.33). Suppose $\boldsymbol{r}$ is in a polydisc of the form (1.9) with R replaced by some value R'' yet to be selected. Consider the series

$$\begin{aligned}
&\sum_{D=0}^{\infty} \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} 2\pi K[(2\ell+1)/4\pi]^{1/2}(D+2)(D+1)R^{-D}|r^{2s} H_\ell^m(\boldsymbol{r})| \\
&\leq K(2/\sqrt{7}) \sum_{D=0}^{\infty} \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} (D+2)(D+1)(2\ell+1)^2(4\sqrt{3})^D (R''/R)^D \\
&\leq K(2/\sqrt{7}) \sum_{D=0}^{\infty} (D+2)(D+1)(4\sqrt{3})^D (R''/R)^D \sum_{\ell+2s=D} (2\ell+1)^3.
\end{aligned} \tag{F.1.62}$$

Here we have used (1.56) with R replaced by R''. Evidently the series (1.62) is convergent if R'' satisfies the inequality

$$R'' < (4\sqrt{3})^{-1} R. \tag{F.1.63}$$

We conclude that the series (1.33) converges absolutely in the polydisc

$$|x| \leq R'',\ |y| \leq R'',\ |z| \leq R'' \tag{F.1.64}$$

with R'' given by (1.63).

With these matters concerning the series (1.6) and (1.33) for $\rho(\boldsymbol{r})$ behind us, we turn to the behavior of $\psi(\boldsymbol{r})$. Break up the integration required by (1.3) into two regions by rewriting it in the form

$$\psi(\boldsymbol{r}) = \psi_<(\boldsymbol{r}) + \psi_>(\boldsymbol{r}) \tag{F.1.65}$$

where

$$\psi_<(\boldsymbol{r}) = \int\limits_{\|\boldsymbol{r}'\| \leq R'} d^3\boldsymbol{r}'\rho(\boldsymbol{r}')/\parallel \boldsymbol{r}' - \boldsymbol{r} \parallel, \tag{F.1.66}$$

$$\psi_>(\boldsymbol{r}) = \int\limits_{\|\boldsymbol{r}'\| \geq R'} d^3\boldsymbol{r}'\rho(\boldsymbol{r}')/\parallel \boldsymbol{r}' - \boldsymbol{r} \parallel . \tag{F.1.67}$$

We will examine each of the functions $\psi_<(\boldsymbol{r})$ and $\psi_>(\boldsymbol{r})$ separately.

The behavior of $\psi_>(\boldsymbol{r})$ near the origin is relatively easy to discern. By analysis similar to that of equations (1.40) through (1.46), we see that $(1/\parallel \boldsymbol{r}' - \boldsymbol{r} \parallel)$ with $\parallel \boldsymbol{r}' \parallel \geq R'$ is analytic in the components of $\boldsymbol{r}$ in a small neighborhood of the origin. Consequently, $\psi_>(\boldsymbol{r})$ is also analytic in the components of $\boldsymbol{r}$ in a small neighborhood of the origin, and this conclusion is independent of the nature of $\rho(\boldsymbol{r}')$ for $\parallel \boldsymbol{r}' \parallel \geq R'$ save for some mild distribution theoretic or integrability conditions such as, for example, that the integral (1.67) be absolutely convergent. We also observe that $(1/\parallel \boldsymbol{r}' - \boldsymbol{r} \parallel)$ with $\parallel r' \parallel \geq R'$ is harmonic in the variables $\boldsymbol{r}$ in a small neighborhood of the origin,

$$\nabla^2(1/\parallel \boldsymbol{r}' - \boldsymbol{r} \parallel) = (\partial_x^2 + \partial_y^2 + \partial_z^2)(1/\parallel \boldsymbol{r}' - \boldsymbol{r} \parallel) = 0 \text{ for } \parallel \boldsymbol{r}' \parallel \geq R' \text{ and } \boldsymbol{r} \text{ near } 0. \tag{F.1.68}$$

Consequently, $\psi_>(\boldsymbol{r})$ is also harmonic around the origin,

$$\nabla^2\psi_>(\boldsymbol{r}) = 0 \text{ for } \boldsymbol{r} \text{ near } 0. \tag{F.1.69}$$

Finding the behavior of $\psi_<(\boldsymbol{r})$ near the origin requires more work. We know that an expansion for $\rho(\boldsymbol{r}')$ of the form (1.33) converges for $\boldsymbol{r}'$ in some sufficiently small polydisc. Select the R' in (1.66) and (1.67) such that the region $\boldsymbol{r}'$ real and $\parallel \boldsymbol{r}' \parallel \leq R'$ lies within this polydisc. Then we may use the expansion (1.33) for $\rho(\boldsymbol{r}')$ in the integral (1.66) to compute $\psi_<(\boldsymbol{r})$.

Let us do this one term at a time. Define functions $\mathcal{X}_{m\ell s}(\boldsymbol{r})$ by the rule

$$\mathcal{X}_{m\ell s}(\boldsymbol{r}) = \int\limits_{\|\boldsymbol{r}'\| \leq R'} d^3\boldsymbol{r}'(r')^{2s+\ell}Y_\ell^m(\Omega')/\parallel \boldsymbol{r}' - \boldsymbol{r} \parallel . \tag{F.1.70}$$

Then, $\psi_<(\boldsymbol{r})$ will have the expansion

$$\psi_<(\boldsymbol{r}) = \sum_{D=0}^{\infty} \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} b_{m\ell s}\mathcal{X}_{m\ell s}(\boldsymbol{r}). \tag{F.1.71}$$

The integral (1.70) can be written in the iterated form

$$\mathcal{X}_{m\ell s}(\boldsymbol{r}) = \int_0^{R'} dr'(r')^2(r')^{2s+\ell} \int d\Omega' \; Y_\ell^m(\Omega')/ \parallel \boldsymbol{r'} - \boldsymbol{r} \parallel . \tag{F.1.72}$$

The second integral in (1.72) has the value

$$\int d\Omega' \; Y_\ell^m(\Omega')/ \parallel \boldsymbol{r'} - \boldsymbol{r} \parallel = [4\pi/(2\ell+1)](r_<^\ell/r_>^{\ell+1})Y_\ell^m(\Omega), \tag{F.1.73}$$

where $r_<$ and $r_>$ are defined by the equations

$$r_< = \text{ the lesser of } r, r', \tag{F.1.74}$$

$$r_> = \text{ the greater of } r, r'. \tag{F.1.75}$$

This value can be used in (1.72) to yield the result

$$\mathcal{X}_{m\ell s}(\boldsymbol{r}) = \mathcal{X}^1_{m\ell s}(\boldsymbol{r}) + \mathcal{X}^2_{m\ell s}(\boldsymbol{r}), \tag{F.1.76}$$

where

$$\mathcal{X}^1_{m\ell s}(\boldsymbol{r}) = -4\pi[(2\ell+1)+(2s+2)]^{-1}(2s+2)^{-1}r^{2s+2}r^\ell Y_\ell^m(\theta,\phi), \tag{F.1.77}$$

$$\mathcal{X}^2_{m\ell s}(\boldsymbol{r}) = 4\pi[(2\ell+1)(2s+2)]^{-1}(R')^{2s+2}r^\ell Y_\ell^m(\theta,\phi). \tag{F.1.78}$$

We see that $\mathcal{X}_{m\ell s}(\boldsymbol{r})$ is a linear combination of the functions $r^{2s+2}r^\ell Y_\ell^m$ and $r^\ell Y_\ell^m$. From our earlier work we know that both these functions are polynomials in the components of $\boldsymbol{r}$, and are therefore entire analytic functions of the variables x, y, z.

The next thing to check is that the series (1.71) for $\psi_<(\boldsymbol{r})$ converges. Following the decomposition (1.76), we will write $\psi_<(\boldsymbol{r})$ in the form

$$\psi_<(\boldsymbol{r}) = \psi^1_<(\boldsymbol{r}) + \psi^2_<(\boldsymbol{r}) \tag{F.1.79}$$

where

$$\psi^1_<(\boldsymbol{r}) = -4\pi \sum_{D=0}^{\infty} \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} b_{m\ell s}[(2\ell+1)+(2s+2)]^{-1}(2s+2)^{-1}r^{2s+2}r^\ell Y_\ell^m(\theta,\phi), \tag{F.1.80}$$

$$\psi^2_<(\boldsymbol{r}) = 4\pi \sum_{D=0}^{\infty} \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} b_{m\ell s}[(2\ell+1)(2s+2)]^{-1}(R')^{2s+2}r^\ell Y_\ell^m(\theta,\phi). \tag{F.1.81}$$

It is easily verified, using the bounds (1.56) and (1.61), that both the series $\psi^1_<(\boldsymbol{r})$ and $\psi^2_<(\boldsymbol{r})$ converge, and converge absolutely, for $\boldsymbol{r}$ sufficiently near the origin and R' sufficiently small. Consequently, $\psi_<(\boldsymbol{r})$ itself is analytic in a neighborhood of the origin. We now know that both $\psi_>$ and $\psi_<$ are analytic around the origin, and hence their sum ψ is analytic about the origin, which is what we have wanted to prove.

At this point we make some observations about the functions $\psi^1_<(\boldsymbol{r})$ and $\psi^2_<(\boldsymbol{r})$. We claim that they satisfy the equations

$$\nabla^2\psi^1_<(\boldsymbol{r}) = -4\pi\rho(\boldsymbol{r}), \tag{F.1.82}$$

$$\nabla^2\psi^2_<(\boldsymbol{r}) = 0. \tag{F.1.83}$$

Since the series for $\psi^1_<$ and $\psi^2_<$ converge absolutely, the operator ∇^2 can be taken under the summation signs and allowed to act term by term. The claim (1.83) then follows immediately from (1.25). To verify (1.82) we note that the operator ∇^2 has the spherical decomposition

$$\nabla^2 = r^{-1}\partial_r^2 r - \mathcal{L}^2/r^2, \tag{F.1.84}$$

and the spherical harmonics have the property

$$\mathcal{L}^2 Y_\ell^m = \ell(\ell+1)Y_\ell^m. \tag{F.1.85}$$

It follows that

$$\begin{aligned}\nabla^2(r^{2s+2}r^\ell Y_\ell^m) &= \{[(r^{-1}\partial_r^2 r) - \ell(\ell+1)/r^2]r^{\ell+2s+2}\}Y_\ell^m\\ &= [(\ell+2s+3)(\ell+2s+2) - \ell(\ell+1)]r^{\ell+2s}Y_\ell^m\\ &= [(2\ell+1)+(2s+2)](2s+2)r^{2s}r^\ell Y_\ell^m.\end{aligned} \tag{F.1.86}$$

We see that the ℓ and s dependent multiplicative factors in (1.86) cancel like factors in the denominators appearing in (1.80), and comparison of the resulting expression for $\nabla^2\psi^1_<$ with that given in (1.33) for ρ shows that the assertion (1.82) is also correct.

Exercises

F.1.1. This section has been devoted to the rather laborious task of showing that if $\rho(\boldsymbol{r})$ is analytic in the components of $\boldsymbol{r}$ at some point $\boldsymbol{r}^0$, then the same is true for $\psi(\boldsymbol{r})$. By contrast, the converse is easy to prove. Show that if $\psi(\boldsymbol{r})$ is analytic in the components of $\boldsymbol{r}$ at some point $\boldsymbol{r}^0$, then the same is true for $\rho(\boldsymbol{r})$.

F.1.2. Consider, as examples, three possible forms for $\rho(\boldsymbol{r})$ as shown below. In each case find the corresponding $\psi(\boldsymbol{r})$, and discuss its analytic properties.

$$\begin{aligned}\rho(\boldsymbol{r}) &= \text{ constant for } \parallel \boldsymbol{r} \parallel \leq R,\\ &= 0 \text{ for } \parallel \boldsymbol{r} \parallel > R;\end{aligned} \tag{F.1.87}$$

$$\rho(\boldsymbol{r}) = a\exp(-br^2); \tag{F.1.88}$$

$$\rho(\boldsymbol{r}) = a\exp(-br). \tag{F.1.89}$$

F.1.3. Show that the electron charge density for any hydrogen atom energy eigenstate,

$$\rho(\boldsymbol{r}) = \overline{\mathcal{X}}(\boldsymbol{r})\mathcal{X}(\boldsymbol{r}), \tag{F.1.90}$$

is *not* analytic at the origin. Relate this "singular" behavior to the fact that the energy eigenstate wave function $\mathcal{X}(\boldsymbol{r})$ satisfies the Schroedinger equation

$$-[\hbar^2/(2m)]\nabla^2\mathcal{X} - (e^2/r)\mathcal{X} = E\mathcal{X}. \tag{F.1.91}$$

F.1.4. Suppose that $f(D, \boldsymbol{r})$ is a homogeneous polynomial of degree D in the components of $\boldsymbol{r}$. According to Exercise 1.5.12 it obeys the relation

$$(\boldsymbol{r}\cdot\boldsymbol{\partial})f(D,\boldsymbol{r}) = Df(D,\boldsymbol{r}). \tag{F.1.92}$$

Show that the operators $(\boldsymbol{r}\cdot\boldsymbol{\partial})$ and $\boldsymbol{\mathcal{L}}$ commute. You have provided another demonstration that $\mathcal{L}_-$ maps homogeneous polynomials into themselves and leaves their degrees unchanged.

F.1.5. Verify (1.25) directly for H_ℓ^ℓ as given by (1.20). Show that ∇^2 commutes with $\boldsymbol{\mathcal{L}}$, and hence (1.25) holds for all H_ℓ^m.

F.1.6. Verify the sums (1.29) and (1.31).

F.1.7. Verify the bound (1.35).

F.1.8. Verify (1.36) through (1.53). Verify that H_ℓ^ℓ as given by (1.20) satisfies the bound (1.53).

F.1.9. The conditions (1.42) were chosen to simplify analysis. Show that the analysis can be modified to improve the bound (1.53) and the requirement (1.63) by replacing $(4\sqrt{3})$ by a smaller factor. Verify that H_ℓ^ℓ as given by (1.20) satisfies your improved bound. Hint: Let δ be a real number in the open interval $0 < \delta < 1$. Show that if $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ satisfy the conditions

$$\boldsymbol{\xi}\cdot\boldsymbol{\xi} \le (1-\delta)^2/4,\ \boldsymbol{\eta}\cdot\boldsymbol{\eta} \le (1-\delta)^2/4, \tag{F.1.93}$$

then one has the inequality

$$|[\|\ \boldsymbol{e} - \boldsymbol{\zeta}\ \|]| \ge \delta^{1/2}. \tag{F.1.94}$$

F.1.10. Since H_ℓ^m is a homogeneous polynomial of degree ℓ, show that both sides of (1.39) can be divided by $(r')^\ell$ so that, with the aid of (1.38), it can be rewritten in the form

$$H_\ell^m(\boldsymbol{\zeta}) = (4\pi)^{-1}(2\ell+1)\int d\Omega' Y_\ell^m(\Omega')/\ \|\ \boldsymbol{e}(\Omega') - \boldsymbol{\zeta}\ \|\ . \tag{F.1.95}$$

Verify that the right side of (1.95) is analytic in the components of $\boldsymbol{\zeta}$ for $\boldsymbol{\zeta}$ sufficiently near 0, and that it has a Taylor expansion about 0 that converges absolutely for $\boldsymbol{\zeta}$ in a sufficiently small polydisc about the origin. (Here you may assume that the composition of two analytic functions is again analytic. See Section 38.2.) Show that the coefficients of this Taylor expansion can be determined from a knowledge of the values of the right side of (1.95) when $\boldsymbol{\zeta}$ is *real* and near 0. We know that both sides of (1.95) are equal when $\boldsymbol{\zeta}$ is real and near 0. (Such a set is an example of what is called a *real environment*. See Exericse 38.2.8.) It follows that both sides of (1.95) have identical Taylor coefficients. Consequently, both sides of (1.95) must also be equal when $\boldsymbol{\zeta}$ is complex and in a sufficiently small polydisc about the origin. Show, in fact, that they must be equal in the domain (1.93).

F.1.11. Verify that (1.61) is a consequence of (1.11).

F.1.12. Verify that the series (1.62) is convergent if (1.64) is satisfied.

F.1.13. Verify that if (in the static case) ρ vanishes in some region, then ψ is analytic in this region.

F.1.14. Verify (1.77) and (1.78). Show that the series (1.80) and (1.81) converge.

F.1.15. Verify (1.86) and (1.82).

F.1.16. Our discussion so far of how ψ inherits the analytic properties of ρ has taken a rather circuitous path through the territory of Taylor and harmonic series. Is there an approach that displays the inheritance directly? There is. Consider $\psi_<$ as given by (1.66). Introduce the variable $\boldsymbol{\Delta}$ by the definition

$$\boldsymbol{r}' = \boldsymbol{r} + \boldsymbol{\Delta}, \tag{F.1.96}$$

and show that (1.66) can also be written in the form

$$\psi_<(\boldsymbol{r}) = \int_{\|\boldsymbol{r}+\boldsymbol{\Delta}\|\leq R'} d^3\boldsymbol{\Delta}\ \rho(\boldsymbol{r}+\boldsymbol{\Delta})/ \parallel \boldsymbol{\Delta} \parallel . \tag{F.1.97}$$

Next introduce polar coordinates for $\boldsymbol{\Delta}$ by the relation

$$\boldsymbol{\Delta} = \Delta \boldsymbol{e}(\Omega). \tag{F.1.98}$$

Show that (1.97) can be rewritten in the form

$$\psi_<(\boldsymbol{r}) = \int d\Omega \int_0^{\tilde{\Delta}} d\Delta\ \Delta\rho(\boldsymbol{r}+\boldsymbol{\Delta}), \tag{F.1.99}$$

where $\tilde{\Delta}$ is given by the expression

$$\tilde{\Delta}(\boldsymbol{r}, \Omega, R') = -\boldsymbol{r}\cdot\boldsymbol{e}(\Omega) + \{(R')^2 + [\boldsymbol{r}\cdot\boldsymbol{e}(\Omega)]^2 - r^2\}^{1/2}. \tag{F.1.100}$$

Finally, introduce the variable τ by writing

$$\Delta = \tau\tilde{\Delta}, \tag{F.1.101}$$

and show that (1.99) can be rewritten as

$$\psi_<(\boldsymbol{r}) = \int d\Omega \int_0^1 d\tau\ \tau\tilde{\Delta}^2\rho[\boldsymbol{r} + \tau\tilde{\Delta}\boldsymbol{e}(\Omega)]. \tag{F.1.102}$$

As it stands, (1.102) may be viewed as an integral representation for $\psi_<(\boldsymbol{r})$ that is valid for small real $\boldsymbol{r}$. Now consider making $\boldsymbol{r}$ complex. Verify that $\tilde{\Delta}$ as given by (1.100) is analytic in $\boldsymbol{r}$ for $\boldsymbol{r}$ contained in a sufficiently small polydisc about 0. By hypothesis, $\rho(\boldsymbol{r})$ is also analytic in some such polydisc. Verify that the same is true for the function $\rho[\boldsymbol{r} + \tau\tilde{\boldsymbol{\Delta}}\boldsymbol{e}(\Omega)]$. Finally, show that $\psi_<(\boldsymbol{r})$ as given by (1.102) must be analytic in some such polydisc.

F.2 The Time Dependent Case

We will now sketch how the results obtained so far can be extended to the time dependent case. In the time dependent case the scalar potential satisfies the inhomogeneous wave equation

$$[\nabla^2 - (1/c^2)\partial_t^2]\psi(\boldsymbol{r}, t) = -4\pi\rho(\boldsymbol{r}, t). \tag{F.2.1}$$

Let us assume that ρ, although time dependent, has a *bounded* Fourier spectrum. That is, we assume that $\rho(\boldsymbol{r}, t)$ can be written in the form

$$\rho(\boldsymbol{r}, t) = (1/2\pi)\int_{-\omega_{\max}}^{\omega_{\max}} d\omega\tilde{\rho}(\boldsymbol{r}, \omega)\exp(-i\omega t) \tag{F.2.2}$$

where $\omega_{\max}$ is some finite frequency cutoff. Then $\psi(\boldsymbol{r}, t)$ will also have a bounded Fourier spectrum,

$$\psi(\boldsymbol{r}, t) = (1/2\pi)\int_{-\omega_{\max}}^{\omega_{\max}} d\omega\tilde{\psi}(\boldsymbol{r}, \omega)\exp(-i\omega t). \tag{F.2.3}$$

We know from (2.1) that the Fourier transform $\tilde{\psi}$ satisfies the inhomogeneous Helmholtz equation

$$(\nabla^2 + k^2)\tilde{\psi}(\boldsymbol{r}, \omega) = -4\pi\tilde{\rho}(\boldsymbol{r}, \omega) \tag{F.2.4}$$

where

$$k = \omega/c. \tag{F.2.5}$$

With the assumption of an outgoing wave boundary condition, equation (2.4) has the solution

$$\tilde{\psi}(\boldsymbol{r}, \omega) = \int d^3\boldsymbol{r}'\tilde{\rho}(\boldsymbol{r}', \omega)[\exp(ik \parallel \boldsymbol{r}' - \boldsymbol{r} \parallel)]/ \parallel \boldsymbol{r}' - \boldsymbol{r} \parallel . \tag{F.2.6}$$

We now assume that $\tilde{\rho}(\boldsymbol{r}, \omega)$, for all ω in the closed interval $[-\omega_{\max}, \omega_{\max}]$, is analytic in the components of $\boldsymbol{r}$ at some point $\boldsymbol{r}^0$. Then it can be shown that $\tilde{\psi}(\boldsymbol{r}, \omega)$ will also be analytic at $\boldsymbol{r}^0$. It follows from (2.3) that $\psi(\boldsymbol{r}, t)$ will also be analytic in the components of $\boldsymbol{r}$ at $\boldsymbol{r}^0$ for all t. Finally, again from (2.3), we see that $\psi(\boldsymbol{r}, t)$ will also be an analytic function of t for all t.

We now outline the proof of this assertion. As before, we take $\boldsymbol{r}^0$ to be the origin and break up the region of integration in (2.6) to write

$$\tilde{\psi}(\boldsymbol{r}, \omega) = \tilde{\psi}_<(\boldsymbol{r}, \omega) + \tilde{\psi}_>(\boldsymbol{r}, \omega) \tag{F.2.7}$$

where

$$\tilde{\psi}_<(\boldsymbol{r}, \omega) = \int\limits_{\|\boldsymbol{r}'\| \leq R'} d^3\boldsymbol{r}'\tilde{\rho}(\boldsymbol{r}', \omega)[\exp(ik \parallel \boldsymbol{r}' - \boldsymbol{r} \parallel)]/ \parallel \boldsymbol{r}' - \boldsymbol{r} \parallel, \tag{F.2.8}$$

$$\tilde{\psi}_>(\boldsymbol{r}, \omega) = \int\limits_{\|\boldsymbol{r}'\| \geq R'} d^3\boldsymbol{r}'\tilde{\rho}(\boldsymbol{r}', \omega)[\exp(ik \parallel \boldsymbol{r}' - \boldsymbol{r} \parallel)]/ \parallel \boldsymbol{r}' - \boldsymbol{r} \parallel . \tag{F.2.9}$$

It follows, from arguments similar to those made earlier, that $\tilde{\psi}_>(\boldsymbol{r}, \omega)$ is analytic in the components of $\boldsymbol{r}$ in a small neighborhood of the origin, and satisfies the Helmholtz equation

$$[\nabla^2 + k^2]\tilde{\psi}_>(\boldsymbol{r}, \omega) = 0 \text{ for } \boldsymbol{r} \text{ near } 0. \tag{F.2.10}$$

To study the behavior of $\tilde{\psi}_<$ we assume, as before, that R' is sufficiently small that $\tilde{\rho}(\boldsymbol{r}', \omega)$ has a convergent expansion of the form (1.33) where the coefficients $b_{m\ell s}(\omega)$ are now ω dependent. Again consider each term at a time and, in analogy to (1.78), examine the integrals

$$\tilde{\mathscr{X}}_{m\ell s}(\boldsymbol{r}, \omega) = \int_0^{R'} dr'(r')^2(r')^{2s+\ell} \int d\Omega' Y_\ell^m(\Omega')[\exp(ik \parallel \boldsymbol{r}' - \boldsymbol{r} \parallel)]/ \parallel \boldsymbol{r}' - \boldsymbol{r} \parallel . \quad \text{(F.2.11)}$$

The second integral in (2.11) has the value

$$\int d\Omega' Y_\ell^m(\Omega')[\exp(ik \parallel \boldsymbol{r}' - \boldsymbol{r} \parallel)]/ \parallel \boldsymbol{r}' - \boldsymbol{r} \parallel = 4\pi ik j_\ell(kr_<)h_\ell^1(kr_>)Y_\ell^m(\Omega). \quad \text{(F.2.12)}$$

Consequently our problem is reduced to studying the behavior of the integral

$$\int_0^{R'} dr'(r')^{\ell+2s+2} 4\pi ik j_\ell(kr_<)h_\ell^1(kr_>). \quad \text{(F.2.13)}$$

It can be shown that this integral has the same analytic behavior as the related integral

$$\int_0^{R'} dr'(r')^{\ell+2s+2}[4\pi/(2\ell+1)]r_<^\ell/r_>^{\ell+1} \quad \text{(F.2.14)}$$

for the analogous time independent case. In particular, $\tilde{\mathscr{X}}_{m\ell s}$ can be written in the form

$$\tilde{\mathscr{X}}_{m\ell s}(\boldsymbol{r}, \omega) = \tilde{\mathscr{X}}^1_{m\ell s}(\boldsymbol{r}, \omega) + \tilde{\mathscr{X}}^2_{m\ell s}(\boldsymbol{r}, \omega) \quad \text{(F.2.15)}$$

where both $\tilde{\mathscr{X}}^1$ and $\tilde{\mathscr{X}}^2$ are entire analytic functions of the variables x, y, z. Correspondingly, $\tilde{\psi}_<(\boldsymbol{r}, \omega)$ can be written in the form

$$\tilde{\psi}_<(\boldsymbol{r}, \omega) = \tilde{\psi}^1_<(\boldsymbol{r}, \omega) + \tilde{\psi}^2_<(\boldsymbol{r}, \omega) \quad \text{(F.2.16)}$$

where both $\tilde{\psi}^1_<$ and $\tilde{\psi}^2_<$ are analytic in a neighborhood about the origin, and satisfy the equations

$$(\nabla^2 + k^2)\tilde{\psi}^1_<(\boldsymbol{r}, \omega) = -4\pi\tilde{\rho}(\boldsymbol{r}, \omega), \quad \text{(F.2.17)}$$

$$(\nabla^2 + k^2)\tilde{\psi}^2_<(\boldsymbol{r}, \omega) = 0. \quad \text{(F.2.18)}$$

Exercises

F.2.1. Review Exercise 1.13. Now consider the time dependent case. Show that the frequency cutoff in (2.2) is essential to the argument. Show that if all frequencies are allowed, then the effect of singularities in ρ can *propagate*. That is, a singularity in ρ at some point $\boldsymbol{r}'$, and some time, can produce a singularity in ψ at some other point $\boldsymbol{r}$ at some later time even though ρ may be analytic at $\boldsymbol{r}$.

F.2.2. Evaluate the integral (2.13), and complete the analysis of the time dependent case.

F.2.3. Extend the method of Exercise 1.16 to the time-dependent case (2.6.

F.3 Smoothing Properties of the Laplacian Kernel

In Section 22.2 we encountered the relation

$$\boldsymbol{H}(\boldsymbol{r}) = [1/(4\pi)] \int_V d^3\boldsymbol{r}' \, \boldsymbol{F}(\boldsymbol{r}')G(\boldsymbol{r}, \boldsymbol{r}'). \tag{F.3.1}$$

See (22.2.92). It follows from the work at the beginning of this appendix that if $\boldsymbol{F}(\boldsymbol{r}')$ is analytic, then $\boldsymbol{H}(\boldsymbol{r})$ will also be analytic. Now we will assume only that $\boldsymbol{F}(\boldsymbol{r}')$ is smooth, and then study what can be said about the properties of $\boldsymbol{H}(\boldsymbol{r})$. By *smooth*, we mean that ....

Bibliography

[1] J.D. Jackson, *Classical Electrodynamics*, John Wiley (1999).

[2] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. II, Chapt. IV, Interscience (1962).

[3] E.H. Lieb and M. Loss, *Analysis*, American Mathematical Society (1997). Suppose $\rho(\boldsymbol{r})$ is not analytic but does have n derivatives. Then, roughly speaking, it can be shown that $\psi(\boldsymbol{r})$ will have $n+2$ derivatives. For precise results, see Chapt. 10 of this book.

[4] A.R. Edmonds, *Angular Momentum in Quantum Mechanics*, Princeton University Press (1960).

[5] F. Treves, *Basic Linear Partial Differential Equations*, Academic Press (1975).

[6] N.M. Giunter, *Potential Theory and its Application to Basic Problems in Mathematical Physics*, F. Ungar, New York (1967).

[7] A. Erdelyi et al., edit., *Higher Transcendental Functions*, Bateman Manuscript Project, Vol. 2, McGraw-Hill (1953).

[8] The method of Exercise 1.16 is due to Robert Warnock. I am grateful to him for this and many other helpful comments.

Appendix G

Invariant Scalar Products

Appendix H

Harmonic Functions

Section 13.2 provided cylindrical harmonic expansions for the harmonic function ψ. This appendix does the same for the gradients of ψ. It also studies the *range* of the transverse gradient operators when acting on the space of harmonic functions. In particular, given a harmonic function χ, it shows that there exists a harmonic function ψ such that $\partial_x \psi = \chi$ or a ψ such that $\partial_y \psi = \chi$. Finally, it provides representations for harmonic functions in two variables.

H.1 Representation of Gradients

We know that the harmonic function $\psi(x, y, z)$ has the representation (13.2.37). We would like to find similar representations for $\partial_x \psi(x, y, z)$, $\partial_y \psi(x, y, z)$, and $\partial_z \psi(x, y, z)$ which, of course, are also harmonic functions.

H.1.1 Low-Order Results

We will begin by finding low-order results, and then work out results to all orders. Refer to (13.2.37) to write ψ in the form

$$\psi = \psi_0 + \psi_c + \psi_s \tag{H.1.1}$$

where

$$\psi_0(x, y, z) = \sum_{\ell=0}^{\infty} (-1)^\ell \frac{1}{2^{2\ell} \ell! \ell!} C_0^{[2\ell]}(z) \rho^{2\ell}, \tag{H.1.2}$$

$$\begin{aligned}\psi_c(x, y, z) &= \sum_{m=1}^{\infty} \cos(m\phi) \sum_{\ell=0}^{\infty} (-1)^\ell \frac{m!}{2^{2\ell} \ell! (\ell + m)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell+m} \\ &= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^\ell \frac{m!}{2^{2\ell} \ell! (\ell + m)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell} \Re(x + iy)^m, \end{aligned}\tag{H.1.3}$$

$$\begin{aligned}\psi_s(x,y,z) &= \sum_{m=1}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell+m}\\ &= \sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)^m. \qquad \text{(H.1.4)}\end{aligned}$$

Let us expand ψ through terms of third order. We can then differentiate this expansion to find the first few terms in the expansions of its gradients. Through terms of degree 3, we find for the constituents of ψ the expansions

$$\psi_0 = C_0^{[0]}(z) - (1/4)(x^2+y^2)C_0^{[2]}(z) + \cdots , \qquad \text{(H.1.5)}$$

$$\begin{aligned}\psi_c &= \Re(x+iy)C_{1,c}^{[0]}(z) + \Re(x+iy)^2C_{2,c}^{[0]}(z) + \Re(x+iy)^3C_{3,c}^{[0]}(z)\\ &-(1/8)(x^2+y^2)\Re(x+iy)C_{1,c}^{[2]}(z) + \cdots\\ &= xC_{1,c}^{[0]}(z) + (x^2-y^2)C_{2,c}^{[0]}(z) + (x^3-3xy^2)C_{3,c}^{[0]}(z) - (1/8)(x^2+y^2)xC_{1,c}^{[2]}(z) + \cdots ,\\ &= xC_{1,c}^{[0]}(z) + (x^2-y^2)C_{2,c}^{[0]}(z)\\ &+x^3[C_{3,c}^{[0]}(z) - (1/8)C_{1,c}^{[2]}(z)] - xy^2[3C_{3,c}^{[0]}(z) + (1/8)C_{1,c}^{[2]}(z)] + \cdots ,\end{aligned} \qquad \text{(H.1.6)}$$

$$\begin{aligned}\psi_s &= \Im(x+iy)C_{1,s}^{[0]}(z) + \Im(x+iy)^2C_{2,s}^{[0]}(z) + \Im(x+iy)^3C_{3,s}^{[0]}(z)\\ &-(1/8)(x^2+y^2)\Im(x+iy)C_{1,s}^{[2]}(z) + \cdots\\ &= yC_{1,s}^{[0]}(z) + 2xyC_{2,s}^{[0]}(z) + (-y^3+3x^2y)C_{3,s}^{[0]}(z) - (1/8)(x^2+y^2)yC_{1,s}^{[2]}(z) + \cdots ,\\ &= yC_{1,s}^{[0]}(z) + 2xyC_{2,s}^{[0]}(z)\\ &-y^3[C_{3,s}^{[0]}(z) + (1/8)C_{1,s}^{[2]}(z)] + x^2y[3C_{3,s}^{[0]}(z) - (1/8)C_{1,s}^{[2]}(z)] + \cdots .\end{aligned} \qquad \text{(H.1.7)}$$

Differentiating these expansions, and retaining terms through second order, give the results

$$\partial_z\psi_0 = C_0^{[1]}(z) - (1/4)(x^2+y^2)C_0^{[3]}(z) + \cdots , \qquad \text{(H.1.8)}$$

$$\partial_z\psi_c = xC_{1,c}^{[1]}(z) + (x^2-y^2)C_{2,c}^{[1]}(z) + \cdots , \qquad \text{(H.1.9)}$$

$$\partial_z\psi_s = yC_{1,s}^{[1]}(z) + 2xyC_{2,s}^{[1]}(z) + \cdots , \qquad \text{(H.1.10)}$$

$$\partial_x\psi_0 = -(1/2)xC_0^{[2]}(z) + \cdots , \qquad \text{(H.1.11)}$$

$$\begin{aligned}\partial_x\psi_c \;=\; & C_{1,c}^{[0]}(z) + 2xC_{2,c}^{[0]}(z) + 3x^2[C_{3,c}^{[0]}(z) - (1/8)C_{1,c}^{[2]}(z)]\\ & -y^2[3C_{3,c}^{[0]}(z) + (1/8)C_{1,c}^{[2]}(z)] + \cdots , \end{aligned} \qquad \text{(H.1.12)}$$

$$\partial_x\psi_s = 2yC_{2,s}^{[0]}(z) + 2xy[3C_{3,s}^{[0]}(z) - (1/8)C_{1,s}^{[2]}(z)] + \cdots , \qquad \text{(H.1.13)}$$

$$\partial_y\psi_0 = -(1/2)yC_0^{[2]}(z) + \cdots\ , \tag{H.1.14}$$

$$\partial_y\psi_c = -2yC_{2,c}^{[0]}(z) - 2xy[3C_{3,c}^{[0]}(z) + (1/8)C_{1,c}^{[2]}(z)] + \cdots\ , \tag{H.1.15}$$

$$\begin{aligned}\partial_y\psi_s \;=\;& C_{1,s}^{[0]}(z) + 2xC_{2,s}^{[0]}(z) - 3y^2[C_{3,s}^{[0]}(z) + (1/8)C_{1,s}^{[2]}(z)]\\ &+x^2[3C_{3,s}^{[0]}(z) - (1/8)C_{1,s}^{[2]}(z)] + \cdots\ .\end{aligned} \tag{H.1.16}$$

Finally, upon employing the decomposition (1.1), we find the results

$$\begin{aligned}&\partial_z\psi = C_0^{[1]}(z) + xC_{1,c}^{[1]}(z) + yC_{1,s}^{[1]}(z)\\ &+(x^2-y^2)C_{2,c}^{[1]}(z) + 2xyC_{2,s}^{[1]}(z) - (1/4)(x^2+y^2)C_0^{[3]}(z) + \cdots\\ &= C_0^{[1]}(z) + xC_{1,c}^{[1]}(z) + yC_{1,s}^{[1]}(z)\\ &+x^2[C_{2,c}^{[1]}(z) - (1/4)C_0^{[3]}(z)] - y^2[C_{2,c}^{[1]}(z) + (1/4)C_0^{[3]}(z)] + 2xyC_{2,s}^{[1]}(z) + \cdots\ ,\end{aligned} \tag{H.1.17}$$

$$\begin{aligned}&\partial_x\psi = C_{1,c}^{[0]}(z) + x[2C_{2,c}^{[0]}(z) - (1/2)C_0^{[2]}(z)] + 2yC_{2,s}^{[0]}(z)\\ &+3x^2[C_{3,c}^{[0]}(z) - (1/8)C_{1,c}^{[2]}(z)] - y^2[3C_{3,c}^{[0]}(z) + (1/8)C_{1,c}^{[2]}(z)]\\ &+2xy[3C_{3,s}^{[0]}(z) - (1/8)C_{1,s}^{[2]}(z)]\cdots\ ,\end{aligned} \tag{H.1.18}$$

$$\begin{aligned}&\partial_y\psi = C_{1,s}^{[0]}(z) - y[2C_{2,c}^{[0]}(z) + (1/2)C_0^{[2]}(z)] + 2xC_{2,s}^{[0]}(z)\\ &-3y^2[C_{3,s}^{[0]}(z) + (1/8)C_{1,s}^{[2]}(z)] + x^2[3C_{3,s}^{[0]}(z) - (1/8)C_{1,s}^{[2]}(z)]\\ &-2xy[3C_{3,c}^{[0]}(z) + (1/8)C_{1,c}^{[2]}(z)] + \cdots\ .\end{aligned} \tag{H.1.19}$$

H.1.2 Results to All Orders

Let us study the effect of the operators ∂_x, ∂_y, and ∂_z on each of the terms in (1.1). Finding the effect of ∂_z is easy because it acts only on the $C_{m,\alpha}^{[n]}(z)$ to raise the value of n by 1. The effects of ∂_x and ∂_y are more complicated. Observe that ψ_0, ψ_c, and ψ_s are sums over the "basis" functions $\rho^{2\ell}$, $\rho^{2\ell}\Re(x+iy)^m$, and $\rho^{2\ell}\Im(x+iy)^m$, respectively. We will first compute ∂_x and ∂_y of these basis functions; then compute ∂_x and ∂_y of ψ_0, ψ_c, and ψ_s; and finally compute $\partial_x\psi$ and $\partial_y\psi$ and also $\partial_z\psi$.

Transverse Gradients of $\rho^{2\ell}$

Let us begin by calculating ∂_x and ∂_y of $\rho^{2\ell}$. We find that

$$\partial_x\rho^{2\ell} = \partial_x(x^2+y^2)^\ell = \ell(x^2+y^2)^{\ell-1}2x = 2\ell\rho^{2\ell-2}\Re(x+iy) \tag{H.1.20}$$

and

$$\partial_y\rho^{2\ell} = \partial_y(x^2+y^2)^\ell = \ell(x^2+y^2)^{\ell-1}2y = 2\ell\rho^{2\ell-2}\Im(x+iy). \tag{H.1.21}$$

Transverse Gradients of $\rho^{2\ell}\Re(x+iy)^m$ and $\rho^{2\ell}\Im(x+iy)^m$

Determining the transverse gradients of $\rho^{2\ell}\Re(x+iy)^m$ and $\rho^{2\ell}\Im(x+iy)^m$ requires somewhat more work. Note that for these calculations we have $m \geq 1$. We find that

$$\begin{aligned}
&\partial_x[\rho^{2\ell}\Re(x+iy)^m] = \partial_x\{\rho^{2\ell}[(x+iy)^m + (x-iy)^m]/2\} \\
&= [\partial_x\rho^{2\ell}][(x+iy)^m + (x-iy)^m]/2 + \rho^{2\ell}\partial_x[(x+iy)^m + (x-iy)^m]/2 \\
&= 2\ell\rho^{2\ell-2}\Re(x+iy)[(x+iy)^m + (x-iy)^m]/2 + \rho^{2\ell}m[(x+iy)^{m-1} + (x-iy)^{m-1}]/2 \\
&= \ell\rho^{2\ell-2}[(x+iy) + (x-iy)][(x+iy)^m + (x-iy)^m]/2 + m\rho^{2\ell}\Re(x+iy)^{m-1} \\
&= \ell\rho^{2\ell-2}\{[(x+iy)^{m+1} + (x-iy)^{m+1}] + \rho^2[(x+iy)^{m-1} + (x-iy)^{m-1}]\}/2 \\
&+m\rho^{2\ell}\Re(x+iy)^{m-1} \\
&= \ell\rho^{2\ell-2}\Re(x+iy)^{m+1} + (\ell+m)\rho^{2\ell}\Re(x+iy)^{m-1},
\end{aligned} \tag{H.1.22}$$

$$\begin{aligned}
&\partial_x[\rho^{2\ell}\Im(x+iy)^m] = \partial_x\{\rho^{2\ell}[(x+iy)^m - (x-iy)^m]/(2i)\} \\
&= [\partial_x\rho^{2\ell}][(x+iy)^m - (x-iy)^m]/(2i) + \rho^{2\ell}\partial_x[(x+iy)^m - (x-iy)^m]/(2i) \\
&= 2\ell\rho^{2\ell-2}\Re(x+iy)[(x+iy)^m - (x-iy)^m]/(2i) + \rho^{2\ell}m[(x+iy)^{m-1} - (x-iy)^{m-1}]/(2i) \\
&= \ell\rho^{2\ell-2}[(x+iy) + (x-iy)][(x+iy)^m - (x-iy)^m]/(2i) + m\rho^{2\ell}\Im(x+iy)^{m-1} \\
&= \ell\rho^{2\ell-2}\{[(x+iy)^{m+1} - (x-iy)^{m+1}] + \rho^2[(x+iy)^{m-1} - (x-iy)^{m-1}]\}/(2i) \\
&+m\rho^{2\ell}\Im(x+iy)^{m-1} \\
&= \ell\rho^{2\ell-2}\Im(x+iy)^{m+1} + (\ell+m)\rho^{2\ell}\Im(x+iy)^{m-1},
\end{aligned} \tag{H.1.23}$$

$$\begin{aligned}
&\partial_y[\rho^{2\ell}\Re(x+iy)^m] = \partial_y\{\rho^{2\ell}[(x+iy)^m + (x-iy)^m]/2\} \\
&= [\partial_y\rho^{2\ell}][(x+iy)^m + (x-iy)^m]/2 + \rho^{2\ell}\partial_y[(x+iy)^m + (x-iy)^m]/2 \\
&= 2\ell\rho^{2\ell-2}\Im(x+iy)[(x+iy)^m + (x-iy)^m]/2 + \rho^{2\ell}im[(x+iy)^{m-1} - (x-iy)^{m-1}]/2 \\
&= \ell\rho^{2\ell-2}[(x+iy) - (x-iy)][(x+iy)^m + (x-iy)^m]/(2i) - m\rho^{2\ell}\Im(x+iy)^{m-1} \\
&= \ell\rho^{2\ell-2}\{[(x+iy)^{m+1} - (x-iy)^{m+1}] - \rho^2[(x+iy)^{m-1} - (x-iy)^{m-1}]\}/(2i) \\
&-m\rho^{2\ell}\Im(x+iy)^{m-1} \\
&= \ell\rho^{2\ell-2}\Im(x+iy)^{m+1} - (\ell+m)\rho^{2\ell}\Im(x+iy)^{m-1},
\end{aligned} \tag{H.1.24}$$

$$\begin{aligned}
&\partial_y[\rho^{2\ell}\Im(x+iy)^m] = \partial_y\{\rho^{2\ell}[(x+iy)^m - (x-iy)^m]/(2i)\} \\
&= [\partial_y\rho^{2\ell}][(x+iy)^m - (x-iy)^m]/(2i) + \rho^{2\ell}\partial_y[(x+iy)^m - (x-iy)^m]/(2i) \\
&= 2\ell\rho^{2\ell-2}\Im(x+iy)[(x+iy)^m - (x-iy)^m]/(2i) + \rho^{2\ell}im[(x+iy)^{m-1} + (x-iy)^{m-1}]/(2i) \\
&= -\ell\rho^{2\ell-2}[(x+iy) - (x-iy)][(x+iy)^m - (x-iy)^m]/2 + m\rho^{2\ell}\Re(x+iy)^{m-1} \\
&= -\ell\rho^{2\ell-2}\{[(x+iy)^{m+1} + (x-iy)^{m+1}] - \rho^2[(x+iy)^{m-1} + (x-iy)^{m-1}]\}/2 \\
&+m\rho^{2\ell}\Re(x+iy)^{m-1} \\
&= -\ell\rho^{2\ell-2}\Re(x+iy)^{m+1} + (\ell+m)\rho^{2\ell}\Re(x+iy)^{m-1}.
\end{aligned} \tag{H.1.25}$$

Transverse Gradients of ψ_0

We are now prepared to compute the transverse gradients of ψ_0. Based on (1.2), (1.20), and (1.21), we find that

$$\begin{aligned}
\partial_x\psi_0 &= \sum_{\ell=0}^{\infty}(-1)^\ell \frac{2\ell}{2^{2\ell}\ell!\ell!}C_0^{[2\ell]}(z)\rho^{2\ell-2}\Re(x+iy) \\
&= \sum_{\ell=1}^{\infty}(-1)^\ell \frac{2\ell}{2^{2\ell}\ell!\ell!}C_0^{[2\ell]}(z)\rho^{2\ell-2}\Re(x+iy) \\
&= \sum_{\ell=0}^{\infty}(-1)^{\ell+1} \frac{(2\ell+2)}{2^{2\ell+2}(\ell+1)!(\ell+1)!}C_0^{[2\ell+2]}(z)\rho^{2\ell}\Re(x+iy) \\
&= \sum_{\ell=0}^{\infty}(-1)^{\ell+1} \frac{2}{2^{2\ell+2}\ell!(\ell+1)!}C_0^{[2\ell+2]}(z)\rho^{2\ell}\Re(x+iy) \qquad \text{(H.1.26)}
\end{aligned}$$

and

$$\begin{aligned}
\partial_y\psi_0 &= \sum_{\ell=0}^{\infty}(-1)^\ell \frac{2\ell}{2^{2\ell}\ell!\ell!}C_0^{[2\ell]}(z)\rho^{2\ell-2}\Im(x+iy) \\
&= \sum_{\ell=1}^{\infty}(-1)^\ell \frac{2\ell}{2^{2\ell}\ell!\ell!}C_0^{[2\ell]}(z)\rho^{2\ell-2}\Im(x+iy) \\
&= \sum_{\ell=0}^{\infty}(-1)^{\ell+1} \frac{(2\ell+2)}{2^{2\ell+2}(\ell+1)!(\ell+1)!}C_0^{[2\ell+2]}(z)\rho^{2\ell}\Im(x+iy) \\
&= \sum_{\ell=0}^{\infty}(-1)^{\ell+1} \frac{2}{2^{2\ell+2}\ell!(\ell+1)!}C_0^{[2\ell+2]}(z)\rho^{2\ell}\Im(x+iy) \qquad \text{(H.1.27)}
\end{aligned}$$

Transverse Gradients of ψ_c and ψ_s

We are also ready to compute the transverse gradients of ψ_c and ψ_s. These results are more lengthy. Based on (1.3), (1.4), and (1.22) through (1.25), we find the following results.

Result for $\partial_x \psi_c$

$$
\begin{aligned}
\partial_x \psi_c &= \sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)[\ell\rho^{2\ell-2}\Re(x+iy)^{m+1}+(\ell+m)\rho^{2\ell}\Re(x+iy)^{m-1}]\\
&= \sum_{m=1}^{\infty}\sum_{\ell=1}^{\infty}(-1)^{\ell}\frac{m!\ell}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell-2}\Re(x+iy)^{m+1}\\
&+\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!(\ell+m)}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)^{m-1}\\
&= \sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!(\ell+1)}{2^{2\ell+2}(\ell+1)!(\ell+m+1)!}C_{m,c}^{[2\ell+2]}(z)\rho^{2\ell}\Re(x+iy)^{m+1}\\
&+\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!(\ell+m)}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)^{m-1}\\
&= \sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!}{2^{2\ell+2}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+2]}(z)\rho^{2\ell}\Re(x+iy)^{m+1}\\
&+\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m-1)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)^{m-1}\\
&= \sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!}{2^{2\ell+2}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+2]}(z)\rho^{2\ell}\Re(x+iy)^{m+1}\\
&+\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_{1,c}^{[2\ell]}(z)\rho^{2\ell}+\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,c}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)\\
&+\sum_{m=3}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m-1)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)^{m-1}\\
&= \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_{1,c}^{[2\ell]}(z)\rho^{2\ell}+\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,c}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)\\
&+\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!}{2^{2\ell+2}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+2]}(z)\rho^{2\ell}\Re(x+iy)^{m+1}\\
&+\sum_{m=3}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m-1)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)^{m-1}\\
&= \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_{1,c}^{[2\ell]}(z)\rho^{2\ell}+\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,c}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)\\
&+\sum_{m=2}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,c}^{[2\ell]}(z)-[1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\}\rho^{2\ell}\Re(x+iy)^{m}.
\end{aligned}
\tag{H.1.28}
$$

Result for $\partial_x \psi_s$

$$
\begin{aligned}
\partial_x \psi_s &= \sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)[\ell\rho^{2\ell-2}\Im(x+iy)^{m+1}+(\ell+m)\rho^{2\ell}\Im(x+iy)^{m-1}]\\
&=\sum_{m=1}^{\infty}\sum_{\ell=1}^{\infty}(-1)^{\ell}\frac{m!\ell}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell-2}\Im(x+iy)^{m+1}\\
&+\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!(\ell+m)}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)^{m-1}\\
&=\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!(\ell+1)}{2^{2\ell+2}(\ell+1)!(\ell+m+1)!}C_{m,s}^{[2\ell+2]}(z)\rho^{2\ell}\Im(x+iy)^{m+1}\\
&+\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!(\ell+m)}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)^{m-1}\\
&=\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!}{2^{2\ell+2}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+2]}(z)\rho^{2\ell}\Im(x+iy)^{m+1}\\
&+\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m-1)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)^{m-1}\\
&=\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!}{2^{2\ell+2}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+2]}(z)\rho^{2\ell}\Im(x+iy)^{m+1}\\
&+\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,s}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)\\
&+\sum_{m=3}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m-1)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)^{m-1}.\\
&=\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,s}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)\\
&+\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!}{2^{2\ell+2}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+2]}(z)\rho^{2\ell}\Im(x+iy)^{m+1}\\
&+\sum_{m=3}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m-1)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)^{m-1}.\\
&=\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,s}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)\\
&+\sum_{m=2}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,s}^{[2\ell]}(z)-[1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\}\rho^{2\ell}\Im(x+iy)^{m}.
\end{aligned}
\tag{H.1.29}
$$

Result for $\partial_y \psi_c$

$$
\begin{aligned}
&\partial_y \psi_c = \sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)[\ell\rho^{2\ell-2}\Im(x+iy)^{m+1}-(\ell+m)\rho^{2\ell}\Im(x+iy)^{m-1}]\\
&=\sum_{m=1}^{\infty}\sum_{\ell=1}^{\infty}(-1)^{\ell}\frac{m!\ell}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell-2}\Im(x+iy)^{m+1}\\
&-\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!(\ell+m)}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)^{m-1}\\
&=\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!(\ell+1)}{2^{2\ell+2}(\ell+1)!(\ell+m+1)!}C_{m,c}^{[2\ell+2]}(z)\rho^{2\ell}\Im(x+iy)^{m+1}\\
&-\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!(\ell+m)}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)^{m-1}\\
&=\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!}{2^{2\ell+2}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+2]}(z)\rho^{2\ell}\Im(x+iy)^{m+1}\\
&-\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m-1)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)^{m-1}\\
&=\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!}{2^{2\ell+2}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+2]}(z)\rho^{2\ell}\Im(x+iy)^{m+1}\\
&-\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,c}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)\\
&-\sum_{m=3}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m-1)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)^{m-1}\\
&=-\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,c}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)\\
&+\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!}{2^{2\ell+2}\ell!(\ell+m+1)!}C_{m,c}^{[2\ell+2]}(z)\rho^{2\ell}\Im(x+iy)^{m+1}\\
&-\sum_{m=3}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m-1)!}C_{m,c}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)^{m-1}\\
&=-\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,c}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy)\\
&-\sum_{m=2}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,c}^{[2\ell]}(z)+[1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\}\rho^{2\ell}\Im(x+iy)^{m}.
\end{aligned}
\tag{H.1.30}
$$

Result for $\partial_y\psi_s$

$$
\begin{aligned}
\partial_y\psi_s &= \sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)[-\ell\rho^{2\ell-2}\Re(x+iy)^{m+1}+(\ell+m)\rho^{2\ell}\Re(x+iy)^{m-1}]\\
&= -\sum_{m=1}^{\infty}\sum_{\ell=1}^{\infty}(-1)^{\ell}\frac{m!\ell}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell-2}\Re(x+iy)^{m+1}\\
&+\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!(\ell+m)}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)^{m-1}\\
&= -\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!(\ell+1)}{2^{2\ell+2}(\ell+1)!(\ell+m+1)!}C_{m,s}^{[2\ell+2]}(z)\rho^{2\ell}\Re(x+iy)^{m+1}\\
&+\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!(\ell+m)}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)^{m-1}\\
&= -\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!}{2^{2\ell+2}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+2]}(z)\rho^{2\ell}\Re(x+iy)^{m+1}\\
&+\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m-1)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)^{m-1}\\
&= -\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!}{2^{2\ell+2}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+2]}(z)\rho^{2\ell}\Re(x+iy)^{m+1}\\
&+\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_{1,s}^{[2\ell]}(z)\rho^{2\ell}+\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,s}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)\\
&+\sum_{m=3}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m-1)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)^{m-1}\\
&= \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_{1,s}^{[2\ell]}(z)\rho^{2\ell}+\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,s}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)\\
&-\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{m!}{2^{2\ell+2}\ell!(\ell+m+1)!}C_{m,s}^{[2\ell+2]}(z)\rho^{2\ell}\Re(x+iy)^{m+1}\\
&+\sum_{m=3}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m-1)!}C_{m,s}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)^{m-1}\\
&= \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_{1,s}^{[2\ell]}(z)\rho^{2\ell}+\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,s}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy)\\
&+\sum_{m=2}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,s}^{[2\ell]}(z)+[1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\}\rho^{2\ell}\Re(x+iy)^{m}.
\end{aligned}
\tag{H.1.31}
$$

Gradients of ψ

The last step is to put all the previous results together using (1.1). So doing gives the following final results.

Result for $\partial_x\psi$

$$
\begin{aligned}
&\partial_x\psi(x,y,z) = \partial_x\psi_0 + \partial_x\psi_c + \partial_x\psi_s \\
&= \sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{2}{2^{2\ell+2}\ell!(\ell+1)!}C_0^{[2\ell+2]}(z)\rho^{2\ell}\Re(x+iy) \\
&+ \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_{1,c}^{[2\ell]}(z)\rho^{2\ell} + \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_0^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy) \\
&+ \sum_{m=2}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,c}^{[2\ell]}(z) - [1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\}\rho^{2\ell}\Re(x+iy)^m \\
&+ \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,s}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy) \\
&+ \sum_{m=2}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,s}^{[2\ell]}(z) - [1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\}\rho^{2\ell}\Im(x+iy)^m \\
&= \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_{1,c}^{[2\ell]}(z)\rho^{2\ell} \\
&+ \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}[C_{2,c}^{[2\ell]}(z) - (1/4)C_0^{[2\ell+2]}(z)]\rho^{2\ell}\Re(x+iy) \\
&+ \sum_{m=2}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,c}^{[2\ell]}(z) - [1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\}\rho^{2\ell}\Re(x+iy)^m \\
&+ \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,s}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy) \\
&+ \sum_{m=2}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,s}^{[2\ell]}(z) - [1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\}\rho^{2\ell}\Im(x+iy)^m.
\end{aligned}
\tag{H.1.32}
$$

Result for $\partial_y \psi$

$$
\begin{aligned}
&\partial_y\psi(x,y,z) = \partial_y\psi_0 + \partial_y\psi_c + \partial_y\psi_s \\
&= \sum_{\ell=0}^{\infty}(-1)^{\ell+1}\frac{2}{2^{2\ell+2}\ell!(\ell+1)!}C_0^{[2\ell+2]}(z)\rho^{2\ell}\Im(x+iy) \\
&-\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,c}^{[2\ell]}(z)\rho^{2\ell}\Im(x+iy) \\
&-\sum_{m=2}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,c}^{[2\ell]}(z)+[1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\}\rho^{2\ell}\Im(x+iy)^m \\
&+\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_{1,s}^{[2\ell]}(z)\rho^{2\ell}+\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,s}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy) \\
&+\sum_{m=2}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,s}^{[2\ell]}(z)+[1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\}\rho^{2\ell}\Re(x+iy)^m \\
&= \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_{1,s}^{[2\ell]}(z)\rho^{2\ell}+\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,s}^{[2\ell]}(z)\rho^{2\ell}\Re(x+iy) \\
&+\sum_{m=2}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,s}^{[2\ell]}(z)+[1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\}\rho^{2\ell}\Re(x+iy)^m \\
&-\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}[C_{2,c}^{[2\ell]}(z)+(1/4)C_0^{[2\ell+2]}]\rho^{2\ell}\Im(x+iy) \\
&-\sum_{m=2}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,c}^{[2\ell]}(z)+[1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\}\rho^{2\ell}\Im(x+iy)^m.
\end{aligned}
\tag{H.1.33}
$$

Result for $\partial_z \psi$

$$
\begin{aligned}
&\partial_z\psi(x,y,z) = \partial_z\psi_0 + \partial_z\psi_c + \partial_z\psi_s \\
&= \sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_0^{[2\ell+1]}(z)\rho^{2\ell} \\
&+\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,c}^{[2\ell+1]}(z)\rho^{2\ell}\Re(x+iy)^m \\
&+\sum_{m=1}^{\infty}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}C_{m,s}^{[2\ell+1]}(z)\rho^{2\ell}\Im(x+iy)^m.
\end{aligned}
\tag{H.1.34}
$$

Exercises

H.1.1. Verify (1.22) through (1.22) for selected example values of ℓ and m.

H.1.2. Verify (1.26) and (1.27).

H.1.3. Verify (1.28) through (1.31).

H.1.4. Verify (1.32) through (1.34). Show that, when expanded to low order, they yield results identical to (1.17) through (1.19).

H.1.5. Write

$$\boldsymbol{B} = \nabla\psi \tag{H.1.35}$$

to find the cylindrical coordinate results

$$B_\rho = \partial\psi/\partial\rho, \tag{H.1.36}$$

$$B_\phi = (1/\rho)\partial\psi/\partial\phi, \tag{H.1.37}$$

$$B_z = \partial\psi/\partial z. \tag{H.1.38}$$

Also invoke the relations

$$B_x = (\cos\phi)B_\rho - (\sin\phi)B_\phi \tag{H.1.39}$$

$$B_y = (\sin\phi)B_\rho + (\cos\phi)B_\phi. \tag{H.1.40}$$

Use these results to derive (1.32) through (1.34).

H.2 Range of Transverse Gradient Operators

We next enquire about the *range* of the operators ∂_x and ∂_y. Consider ∂_x. Suppose χ is some given (real) harmonic function. Can we find a (real) harmonic ψ such that either

$$\partial_x\psi = \chi \tag{H.2.1}$$

or

$$\partial_y\psi = \chi? \tag{H.2.2}$$

We will verify, by construction, that the answer is *yes*.

H.2.1 Solution of $\partial_x\psi = \chi$

Since χ is assumed (real) harmonic, it has a representation of the form

$$\begin{aligned}\chi(x,y,z) &= \sum_{\ell=0}^{\infty}(-1)^\ell\frac{1}{2^{2\ell}\ell!\ell!}B_0^{[2\ell]}(z)\rho^{2\ell}\\ &+ \sum_{m=1}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{m!}{2^{2\ell}\ell!(\ell+m)!}B_{m,c}^{[2\ell]}(z)\rho^{2\ell+m}\\ &+ \sum_{m=1}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^\ell\frac{m!}{2^{2\ell}\ell!(\ell+m)!}B_{m,s}^{[2\ell]}(z)\rho^{2\ell+m}.\end{aligned} \tag{H.2.3}$$

Evidently, we must compare (1.32) and (2.3). We must try to satisfy the pair of equations

$$\begin{aligned}
&\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_{1,c}^{[2\ell]}(z)\rho^{2\ell}+\\
&+\cos(\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}[C_{2,c}^{[2\ell]}(z)-(1/4)C_{0}^{[2\ell+2]}(z)]\rho^{2\ell+1}\\
&+\sum_{m=2}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,c}^{[2\ell]}(z)-[1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\}\rho^{2\ell+m}\\
&\overset{?}{=}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}B_{0}^{[2\ell]}(z)\rho^{2\ell}+\sum_{m=1}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}B_{m,c}^{[2\ell]}(z)\rho^{2\ell+m}
\end{aligned}\tag{H.2.4}$$

and

$$\begin{aligned}
&\sin(\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,s}^{[2\ell]}(z)\rho^{2\ell+1}\\
&+\sum_{m=2}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,s}^{[2\ell]}(z)-[1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\}\rho^{2\ell+m}\\
&\overset{?}{=}\sum_{m=1}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}B_{m,s}^{[2\ell]}(z)\rho^{2\ell+m}.
\end{aligned}\tag{H.2.5}$$

Here we have used (13.2.7) and (13.2.8).

Let us first work on question (2.4). Suppose we set

$$C_{1,c}^{[0]}(z)=B_{0}^{[0]}(z).\tag{H.2.6}$$

Then (2.4) becomes the question

$$\begin{aligned}
&+\cos(\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}[C_{2,c}^{[2\ell]}(z)-(1/4)C_{0}^{[2\ell+2]}(z)]\rho^{2\ell+1}\\
&+\sum_{m=2}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,c}^{[2\ell]}(z)-[1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\}\rho^{2\ell+m}\\
&\overset{?}{=}\sum_{m=1}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}B_{m,c}^{[2\ell]}(z)\rho^{2\ell+m}.
\end{aligned}\tag{H.2.7}$$

Note that the right side of (2.7) can be written in the form

$$\begin{aligned}&\sum_{m=1}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}B_{m,c}^{[2\ell]}(z)\rho^{2\ell+m}\\&=\cos(\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!(\ell+1)!}B_{1,c}^{[2\ell]}(z)\rho^{2\ell+1}\\&+\sum_{m=2}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}B_{m,c}^{[2\ell]}(z)\rho^{2\ell+m}.\end{aligned} \tag{H.2.8}$$

Therefore, upon equating like terms, we see that (2.7) is equivalent to the two questions

$$2[C_{2,c}^{[2\ell]}(z)-(1/4)C_{0}^{[2\ell+2]}(z)]\overset{?}{=}B_{1,c}^{[2\ell]}(z) \tag{H.2.9}$$

and

$$(m+1)C_{m+1,c}^{[2\ell]}(z)-[1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)=B_{m,c}^{[2\ell]}(z)?\text{ for }m\geq 2. \tag{H.2.10}$$

We will solve (2.9) by making the stipulation

$$C_{0}^{[0]}=0, \tag{H.2.11}$$

in which case (2.9) has the solution

$$C_{2,c}^{[0]}(z)=(1/2)B_{1,c}^{[0]}(z). \tag{H.2.12}$$

Let us next rewrite (2.10) in the recursive form

$$C_{m+1,c}^{[2\ell]}(z)=[1/(m+1)]B_{m,c}^{[2\ell]}(z)+\{1/[(4m)(m+1)]\}C_{m-1,c}^{[2\ell+2]}(z)?\text{ for }m\geq 2. \tag{H.2.13}$$

In view of (2.6) and (2.12), this is a well-defined recursion relation. For example, putting $m=2$ in (2.13) gives the result

$$C_{3,c}^{[0]}(z)=(1/3)B_{2,c}^{[0]}(z)+\{1/[(8)(3)]\}C_{1,c}^{[2]}(z), \tag{H.2.14}$$

which, using (2.6), can be rewritten as

$$C_{3,c}^{[0]}(z)=(1/3)B_{2,c}^{[0]}(z)+(1/24)B_{0}^{[2]}(z). \tag{H.2.15}$$

Now put $m=3$. Doing so gives the result

$$C_{4,c}^{[0]}(z)=(1/4)B_{3,c}^{[0]}(z)+\{1/[(12)(4)]\}C_{2,c}^{[2]}(z), \tag{H.2.16}$$

which, using (2.12), can be rewritten as

$$C_{4,c}^{[0]}(z)=(1/4)B_{3,c}^{[0]}(z)+(1/96)B_{1,c}^{[2]}(z). \tag{H.2.17}$$

Finally, put $m = 4$, The result is

$$C^{[0]}_{5,c}(z) = (1/5)B^{[0]}_{4,c}(z) + \{1/[(16)(5)]\}C^{[2]}_{3,c}(z), \tag{H.2.18}$$

which, using (2.15), can be rewritten as

$$C^{[0]}_{5,c}(z) = (1/5)B^{[0]}_{4,c}(z) + (1/240)B^{[2]}_{2,c}(z) + (1/1920)B^{[4]}_{0}(z). \tag{H.2.19}$$

The pattern is now clear. All the $C^{[0]}_{m,c}(z)$ are determined in terms of the $B^{[n]}_{m,c}(z)$, and (2.4) is satisfied.

Move on to look at (2.5), which can also be written in the form

$$\begin{aligned}&\sin(\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C^{[2\ell]}_{2,s}(z)\rho^{2\ell+1}\\ &+\sum_{m=2}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C^{[2\ell]}_{m+1,s}(z)-[1/(4m)]C^{[2\ell+2]}_{m-1,s}(z)\}\rho^{2\ell+m}\\ &\overset{?}{=}\sin(\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!(\ell+1)!}B^{[2\ell]}_{1,s}(z)\rho^{2\ell+1}\\ &+\sum_{m=2}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}B^{[2\ell]}_{m,s}(z)\rho^{2\ell+m}.\end{aligned} \tag{H.2.20}$$

Upon equating like terms in (2.20) we see that it is equivalent to the two relations

$$2C^{[2\ell]}_{2,s}(z)\overset{?}{=}B^{[2\ell]}_{1,s}(z) \tag{H.2.21}$$

and

$$(m+1)C^{[2\ell]}_{m+1,s}(z) - [1/(4m)]C^{[2\ell+2]}_{m-1,s}(z)\overset{?}{=}B^{[2\ell]}_{m,s}(z) \text{ for } m \geq 2. \tag{H.2.22}$$

The relation (2.21) has the solution

$$C^{[0]}_{2,s}(z) = (1/2)B^{[0]}_{1,s}(z), \tag{H.2.23}$$

and (2.22) can be written in the recursive form

$$C^{[2\ell]}_{m+1,s}(z)\overset{?}{=}[1/(m+1)]B^{[2\ell]}_{m,s}(z) + \{1/[(4m)(m+1)]\}C^{[2\ell+2]}_{m-1,s}(z) \text{ for } m \geq 2. \tag{H.2.24}$$

Now make the stipulation

$$C^{[0]}_{1,s}(z) = 0. \tag{H.2.25}$$

Then, for $m = 2$, we get the result

$$C^{[0]}_{3,s}(z) = (1/3)B^{[0]}_{2,s}(z). \tag{H.2.26}$$

Now put $m = 3$ to get the result

$$C^{[0]}_{4,s}(z) = (1/4)B^{[0]}_{3,s}(z) + \{1/[(12)(4)]\}C^{[2]}_{2,s}(z), \tag{H.2.27}$$

which, when combined with (2.23), gives the result

$$C_{4,s}^{[0]}(z) = (1/4)B_{3,s}^{[0]}(z) + (1/96)B_{1,s}^{[2]}(z). \tag{H.2.28}$$

To continue the calculation, put $m = 4$ to find

$$C_{5,s}^{[0]}(z) = (1/5)B_{4,s}^{[0]}(z) + \{1/[(16)(5)]\}C_{3,s}^{[2]}(z), \tag{H.2.29}$$

which, when combined with (2.26), gives the result

$$C_{5,s}^{[0]}(z) = (1/5)B_{4,s}^{[0]}(z) + (1/240)B_{2,s}^{[2]}(z). \tag{H.2.30}$$

The pattern becomes clear when $m = 5$. In this case we find that

$$C_{6,s}^{[0]}(z) = (1/6)B_{5,s}^{[0]}(z) + \{1/[(20)(6)]\}C_{4,s}^{[2]}(z), \tag{H.2.31}$$

which, when combined with (2.28), gives the result

$$C_{6,s}^{[0]}(z) = (1/6)B_{5,s}^{[0]}(z) + (1/480)B_{3,s}^{[2]}(z) + (1/11520)B_{1,s}^{[4]}(z). \tag{H.2.32}$$

We see that all the $C_{m,s}^{[0]}(z)$ are determined in terms of the $B_{m,s}^{[n]}(z)$, and (2.5) is satisfied. Thus, both (2.4) and (2.5) have been satisfied, and therefore our goal (2.1) has been met.

At this point we should comment on the stipulations (2.11) and (2.25). Evidently the stipulation (2.11) specifies that all terms of the form

$$\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_0^{[2\ell]}(z)\rho^{2\ell} \tag{H.2.33}$$

are omitted from ψ. And the stipulation (2.25) specifies that all terms of the form

$$\sin(\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!(\ell+1)!}C_{1,s}^{[2\ell]}(z)\rho^{2\ell+1} \tag{H.2.34}$$

are omitted from ψ. We have seen that these terms are not needed to meet the goal (2.1), and their omission simplifies the recursion relations that specify ψ in terms of χ.

H.2.2 Solution of $\partial_y\psi = \chi$

We next address the question of whether there is a ψ such that (2.2) is satisfied. By symmetry we know that there must be such a ψ, but it is instructive to work out the details. Evidently we must compare (1.33) and (2.3). We must try to satisfy the pair of equations

$$\begin{aligned}&\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}C_{1,s}^{[2\ell]}(z)\rho^{2\ell} + \cos(\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,s}^{[2\ell]}(z)\rho^{2\ell+1}\\ &+\sum_{m=2}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,s}^{[2\ell]}(z) + [1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\}\rho^{2\ell+m}\\ &\overset{?}{=}\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!\ell!}B_0^{[2\ell]}(z)\rho^{2\ell} + \sum_{m=1}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}B_{m,c}^{[2\ell]}(z)\rho^{2\ell+m}\end{aligned} \tag{H.2.35}$$

and

$$
\begin{aligned}
&-\sin(\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}[C_{2,c}^{[2\ell]}(z)+(1/4)C_{0}^{[2\ell+2]}]\rho^{2\ell+1}\\
&-\sum_{m=2}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,c}^{[2\ell]}(z)+[1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\}\rho^{2\ell+2}\\
&\overset{?}{=}\sum_{m=1}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}B_{m,s}^{[2\ell]}(z)\rho^{2\ell+m}. \qquad \text{(H.2.36)}
\end{aligned}
$$

Let us first work on question (2.35). Begin by setting

$$C_{1,s}^{[0]}(z)=B_{0}^{[0]}(z). \qquad \text{(H.2.37)}$$

When this is done, (2.35) becomes

$$
\begin{aligned}
&\cos(\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}C_{2,s}^{[2\ell]}(z)\rho^{2\ell+1}\\
&+\sum_{m=2}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C_{m+1,s}^{[2\ell]}(z)+[1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\}\rho^{2\ell+m}\\
&\overset{?}{=}\cos(\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!(\ell+1)!}B_{1,c}^{[2\ell]}(z)\rho^{2\ell+1}\\
&+\sum_{m=2}^{\infty}\cos(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}B_{m,c}^{[2\ell]}(z)\rho^{2\ell+m},
\end{aligned}
\qquad \text{(H.2.38)}
$$

which is equivalent to the questions

$$2C_{2,s}^{[0]}(z)\overset{?}{=}B_{1,c}^{[0]}(z) \qquad \text{(H.2.39)}$$

and

$$(m+1)C_{m+1,s}^{[0]}(z)+[1/(4m)]C_{m-1,s}^{[2]}(z)\overset{?}{=}B_{m,c}^{[0]}(z) \text{ for } m\geq 2. \qquad \text{(H.2.40)}$$

The question (2.39) has the answer

$$C_{2,s}^{[0]}(z)=(1/2)B_{1,c}^{[0]}(z), \qquad \text{(H.2.41)}$$

and (2.40) can be written in the recursive form

$$C_{m+1,s}^{[0]}(z)\overset{?}{=}[1/(m+1)]B_{m,c}^{[0]}(z)-\{1/[(4m)(m+1)]\}C_{m-1,s}^{[2]}(z). \qquad \text{(H.2.42)}$$

We see from (2.37) and (2.41) that this recursion relation has a unique solution. Setting $m=2$ in (2.42) gives the result

$$C_{3,s}^{[0]}(z)=(1/3)B_{2,c}^{[0]}(z)-\{1/[(8)(3)]\}C_{1,s}^{[2]}(z), \qquad \text{(H.2.43)}$$

which, in view of (2.37), becomes

$$C^{[0]}_{3,s}(z) = (1/3)B^{[0]}_{2,c}(z) - (1/24)B^{[2]}_{0}(z). \tag{H.2.44}$$

Setting $m = 3$ gives

$$C^{[0]}_{4,s}(z) = (1/4)B^{[0]}_{3,c}(z) - \{1/[(12)(4)]\}C^{[2]}_{2,s}(z), \tag{H.2.45}$$

which, in view of (2.41), becomes

$$C^{[0]}_{4,s}(z) = (1/4)B^{[0]}_{3,c}(z) - (1/96)B^{[2]}_{1,c}(z). \tag{H.2.46}$$

Setting $m = 4$ gives

$$C^{[0]}_{5,s}(z) = (1/5)B^{[0]}_{3,c}(z) - \{1/[(16)(5)]\}C^{[2]}_{3,s}(z), \tag{H.2.47}$$

which, in view of (2.44), becomes

$$C^{[0]}_{5,s}(z) = (1/5)B^{[0]}_{4,c}(z) - (1/96)B^{[2]}_{2,c}(z) - (1/96)B^{[4]}_{0}(z). \tag{H.2.48}$$

Evidently we are able to find all the $C^{[0]}_{m,s}(z)$ in terms of the $B^{[n]}_{m,c}(z)$, and therefore (2.35) can be satisfied.

Now turn to satisfying (2.36), which is equivalent to the question

$$\begin{aligned}
&-\sin(\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{2}{2^{2\ell}\ell!(\ell+1)!}[C^{[2\ell]}_{2,c}(z) + (1/4)C^{[2\ell+2]}_{0}]\rho^{2\ell+1} \\
&-\sum_{m=2}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}\{(m+1)C^{[2\ell]}_{m+1,c}(z) + [1/(4m)]C^{[2\ell+2]}_{m-1,c}(z)\}\rho^{2\ell+2} \\
&\overset{?}{=}\sin(\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!(\ell+1)!}B^{[2\ell]}_{1,s}(z)\rho^{2\ell+1} \\
&+\sum_{m=2}^{\infty}\sin(m\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{m!}{2^{2\ell}\ell!(\ell+m)!}B^{[2\ell]}_{m,s}(z)\rho^{2\ell+m}.
\end{aligned} \tag{H.2.49}$$

Upon equating like terms, we find the questions

$$2[C^{[0]}_{2,c}(z) + (1/4)C^{[2]}_{0}] \overset{?}{=} -B^{[0]}_{1,s}(z) \tag{H.2.50}$$

and

$$(m+1)C^{[0]}_{m+1,c}(z) + [1/(4m)]C^{[2]}_{m-1,c}(z) \overset{?}{=} -B^{[0]}_{m,s}(z). \tag{H.2.51}$$

We again make the stipulation (2.11) so that (2.50) has the answer

$$C^{[0]}_{2,c}(z) = -(1/2)B^{[0]}_{1,s}(z). \tag{H.2.52}$$

Moreover, (2.51) is equivalent to the recursion relation

$$C^{[0]}_{m+1,c}(z) \overset{?}{=} -[1/(m+1)]B^{[0]}_{m,s}(z) - \{1/[(4m)(m+1)]\}C^{[2]}_{m-1,c}(z). \tag{H.2.53}$$

We now add the further stipulation

$$C_{1,c}^{[0]}(z)=0. \tag{H.2.54}$$

In view of (2.52) and (2.54), the recursion relation (2.53) now has a unique solution. Setting $m=2$ and using (2.54) give the result

$$C_{3,c}^{[0]}(z)=-(1/3)B_{2,s}^{[0]}(z). \tag{H.2.55}$$

Next set $m=3$ to find the result

$$C_{4,c}^{[0]}(z)=-(1/4)B_{3,s}^{[0]}(z)-\{1/[(12)(4)]\}C_{2,c}^{[2]}(z), \tag{H.2.56}$$

which, in view of (2.52), becomes the relation

$$C_{4,c}^{[0]}(z)=-(1/4)B_{3,s}^{[0]}(z)+(1/96)B_{1,s}^{[2]}(z). \tag{H.2.57}$$

Now set $m=4$ to find the result

$$C_{5,c}^{[0]}(z)=-(1/5)B_{4,s}^{[0]}(z)-\{1/[(16)(5)]\}C_{3,c}^{[2]}(z), \tag{H.2.58}$$

which, in view of (2.55), becomes the relation

$$C_{5,c}^{[0]}(z)=-(1/5)B_{4,s}^{[0]}(z)+(1/240)B_{2,s}^{[2]}(z). \tag{H.2.59}$$

Set $m=5$ to find the result

$$C_{6,c}^{[0]}(z)=-(1/6)B_{5,s}^{[0]}(z)-\{1/[(20)(6)]\}C_{4,c}^{[2]}(z), \tag{H.2.60}$$

which, in view of (2.57), becomes the relation

$$C_{6,c}^{[0]}(z)=-(1/6)B_{5,s}^{[0]}(z)+(1/480)B_{3,s}^{[2]}(z)-(1/11520)B_{1,s}^{[4]}(z). \tag{H.2.61}$$

Evidently we are able to find all the $C_{m,c}^{[0]}(z)$ in terms of the $B_{m,s}^{[n]}(z)$, and therefore (2.36) can be satisfied. Since both (2.35) and (2.36) have been satisfied, our goal (2.2) has been met.

We note that the stipulation (2.54) specifies that all terms of the form

$$\cos(\phi)\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{1}{2^{2\ell}\ell!(\ell+1)!}C_{1,c}^{[2\ell]}(z)\rho^{2\ell+1} \tag{H.2.62}$$

are omitted from ψ. We have seen that these terms [and those of (2.33)] are not needed to meet the goal (2.2), and their omission simplifies the recursion relations that specify ψ in terms of χ.

H.3 Harmonic Functions in Two Variables and Their Associated Fields

It is also useful to have representations for harmonic functions in two variables. We will consider the two cases of harmonic functions in the variable pairs x, z and y, z.

H.3.1 Harmonic Functions in x, z

Suppose $\psi(x,z)$ is a harmonic function in the two variables x, z. That is, we have the relation

$$[(\partial_x)^2 + (\partial_z)^2]\psi(x,z) = 0. \tag{H.3.1}$$

Decompose ψ into *even* and *odd* parts with respect to its x dependence,

$$\psi = \psi_{ev} + \psi_{od}. \tag{H.3.2}$$

Then, because the operation $x \to -x$ commutes with $[(\partial_x)^2 + (\partial_z)^2]$, each separate part of ψ must be harmonic,

$$[(\partial_x)^2 + (\partial_z)^2]\psi_{ev}(x,z) = 0, \tag{H.3.3}$$

$$[(\partial_x)^2 + (\partial_z)^2]\psi_{od}(x,z) = 0. \tag{H.3.4}$$

Series Representation

For the even part let us make the Ansatz

$$\begin{aligned}\psi_{ev}(x,z) &= \sum_{n=0}^{\infty}(-1)^n[1/(2n)!]x^{2n}E^{[2n]}(z)\\ &= E^{[0]}(z) - (1/2)x^2E^{[2]}(z) + (1/24)x^4E^{[4]}(z) + \cdots ,\end{aligned} \tag{H.3.5}$$

where the functions $E^{[n]}(z)$ and the meaning of the $[n]$ notation are yet to be determined. For such a ψ_{ev} to be harmonic there must be the relation

$$\begin{aligned}0 &= \nabla^2\psi_{ev}(x,z) = \partial_x^2\psi_{ev}(x,z) + \partial_z^2\psi(x,z)\\ &= [\partial_z^2E^{[0]}(z) - E^{[2]}(z)] - (1/2)x^2[\partial_z^2E^{[2]}(z) - E^{[4]}(z)] + \cdots ,\end{aligned} \tag{H.3.6}$$

from which we conclude that

$$E^{[n+2]}(z) = \partial_z^2E^{[n]}(z). \tag{H.3.7}$$

Similarly, for the odd part we make the Ansatz

$$\begin{aligned}\psi_{od}(x,z) &= \sum_{n=0}^{\infty}(-1)^n[1/(2n+1)!]x^{2n+1}O^{[2n]}(z)\\ &= xO^{[0]}(z) - (1/6)x^3O^{[2]}(z) + (1/120)x^5O^{[4]}(z) + \cdots .\end{aligned} \tag{H.3.8}$$

Now the harmonic requirement yields the result

$$\begin{aligned}0 &= \nabla^2\psi_{od}(x,z) = \partial_x^2\psi_{od}(x,z) + \partial_z^2\psi_{od}(x,z)\\ &= x[\partial_z^2O^{[0]}(z) - O^{[2]}(z)] - (1/6)x^3[\partial_z^2O^{[2]}(z) - O^{[4]}(z)] + \cdots ,\end{aligned} \tag{H.3.9}$$

from which we conclude that

$$O^{[n+2]}(z) = \partial_z^2O^{[n]}(z). \tag{H.3.10}$$

We see that in both cases the $[n]$ notation is the usual one. Thus, ψ is specified by the two functions $E^{[0]}(z)$ and $O^{[0]}(z)$. These functions may, in principle, be chosen arbitrarily. Often they are required to go to zero as $|z| \to \infty$.

Let us compute the "fields" associated with ψ_{ev} and ψ_{od}, call them $\boldsymbol{B}^{ev}$ and $\boldsymbol{B}^{od}$. We find the results

$$B_x^{ev} = \partial_x \psi_{ev} = -xE^{[2]}(z) + (1/6)x^3 E^{[4]}(z) - (1/120)x^5 E^{[6]}(z) + \cdots , \tag{H.3.11}$$

$$B_y^{ev} = \partial_y \psi_{ev} = 0, \tag{H.3.12}$$

$$B_z^{ev} = \partial_z \psi_{ev} = E^{[1]}(z) - (1/2)x^2 E^{[3]}(z) + (1/24)x^4 E^{[5]}(z) + \cdots ; \tag{H.3.13}$$

$$B_x^{od} = \partial_x \psi_{od} = O^{[0]}(z) - (1/2)x^2 O^{[2]}(z) + (1/24)x^4 O^{[4]}(z) + \cdots , \tag{H.3.14}$$

$$B_y^{od} = \partial_y \psi_{od} - 0, \tag{H.3.15}$$

$$B_z^{od} = \partial_z \psi_{od} = xO^{[1]}(z) - (1/6)x^3 O^{[3]}(z) + (1/120)x^5 O^{[5]}(z) + \cdots . \tag{H.3.16}$$

Note that, because the fields are the gradients of harmonic functions, the Cartesian components of $\boldsymbol{B}^{ev}$ and $\boldsymbol{B}^{od}$ must also be harmonic functions.

Explicit Construction from Analytic Functions

As is well known, there is an intimate connection between harmonic functions in two variables and analytic functions of a complex variable. This connection facilitates obtaining closed-form expressions for harmonic functions rather than dealing solely with power series. Let w be a complex variable written in the form

$$w = u + iv. \tag{H.3.17}$$

Suppose $f(w)$ is a *real-analytic* function. That is, f is defined, analytic, and real for w on the real axis. Such a function can be extended into the complex plane by analytic continuation. For complex arguments, decompose f into real and imaginary parts by writing

$$f(u + iv) = f_r(u, v) + if_i(u, v). \tag{H.3.18}$$

For f_r and f_i we find, by the chain rule, the result

$$[(\partial_u)^2 + (\partial_v)^2][f_r(u, v) + if_i(u, v)] = f''(u + iv)(1 + i^2) = 0. \tag{H.3.19}$$

Thus, upon equating real and imaginary parts in (H.19), we see that both f_r and f_i are harmonic.

Let us expand f as a power series in the quantity iv. From Taylor's theorem we find the result

$$\begin{aligned} f(u + iv) &= \sum_{n=0}^{\infty} f^{[n]}(u)(iv)^n/n! \\ &= \sum_{n=0}^{\infty} (-1)^n [1/(2n)!] v^{2n} f^{[2n]}(u) \\ &\quad + i \sum_{n=0}^{\infty} (-1)^n [1/(2n+1)!] v^{2n+1} f^{[2n+1]}(u). \end{aligned} \tag{H.3.20}$$

Upon comparing (H.18) and (H.20), we see that there are the relations

$$\begin{aligned} f_r(u,v) &= \sum_{n=0}^{\infty}(-1)^n[1/(2n)!]v^{2n}f^{[2n]}(u) \\ &= f^{[0]}(u)-(1/2)v^2f^{[2]}(u)+(1/24)v^4f^{[4]}(u)+\cdots \end{aligned} \tag{H.3.21}$$

and

$$\begin{aligned} f_i(u,v) &= \sum_{n=0}^{\infty}(-1)^n[1/(2n+1)!]v^{2n+1}f^{[2n+1]}(u) \\ &= vf^{[1]}(u)-(1/6)v^3f^{[3]}(u)+(1/120)v^5f^{[5]}(u)+\cdots . \end{aligned} \tag{H.3.22}$$

Observe the resemblance between the pair (H.5) and (H.8) and the pair (H.21) and (H.22). Upon making the identifications

$$z \leftrightarrow u, \tag{H.3.23}$$

$$x \leftrightarrow v, \tag{H.3.24}$$

$$E^{[0]}(z) \leftrightarrow f^{[0]}(u), \tag{H.3.25}$$

and

$$O^{[0]}(z) \leftrightarrow f^{[1]}(u), \tag{H.3.26}$$

we see that these two pairs are the same. Thus, with these identifications, we have the relations

$$\psi_{ev}(x,z)=f_r(z,x), \tag{H.3.27}$$

$$\psi_{od}(x,z)=f_i(z,x). \tag{H.3.28}$$

Note that the identifications (H.25) and (H.26) require the restrictive relation

$$O^{[0]}=E^{[1]}. \tag{H.3.29}$$

Of course, in general, ψ_{ev} and ψ_{od} need not be the real and imaginary parts of the *same* real-analytic function.

Since f_r and f_i are the real and imaginary parts of the analytic function f, they must satisfy the Cauchy-Riemann relations

$$\partial_z f_r = \partial_x f_i, \tag{H.3.30}$$

$$\partial_x f_r = -\partial_z f_i. \tag{H.3.31}$$

Consequently, if ψ_{ev} and ψ_{od} are the real and imaginary parts of the *same* real-analytic function, we have the relations

$$B_x^{ev}=\partial_x\psi_{ev}=\partial_x f_r=-\partial_z f_i=-\partial_z\psi_{od}=-B_z^{od}, \tag{H.3.32}$$

$$B_z^{ev}=\partial_z\psi_{ev}=\partial_z f_r=\partial_x f_i=\partial_x\psi_{od}=B_x^{od}. \tag{H.3.33}$$

These relations also follow from the representations (H.11) through (H.16) and the relation (H.29).

There is another application of the relation between analytic and harmonic functions that is useful. We will formulate and apply it in the subsection after the next where we discuss harmonic functions in the y,z variables. From that discussion the reader can easily infer the analogous results for harmonic functions in the variables x,z.

H.3.2 Harmonic Functions in y, z

The case of harmonic functions in the variables y, z is analogous to the x, z case. It is only necessary to make the substitution $x \to y$ in the relations found above.

Suppose $\psi(y, z)$ is a harmonic function in the two variables y, z. That is, we have the relation

$$[(\partial_y)^2 + (\partial_z)^2]\psi(y, z) = 0. \tag{H.3.34}$$

Decompose ψ into *even* and *odd* parts with respect to its y dependence,

$$\psi = \psi_{ev} + \psi_{od}. \tag{H.3.35}$$

Then, because the operation $y \to -y$ commutes with $[(\partial_y)^2 + (\partial_z)^2]$, each separate part of ψ must be harmonic,

$$[(\partial_y)^2 + (\partial_z)^2]\psi_{ev}(y, z) = 0, \tag{H.3.36}$$

$$[(\partial_y)^2 + (\partial_z)^2]\psi_{od}(y, z) = 0. \tag{H.3.37}$$

Series Representation

For the even part there is the representation

$$\begin{aligned}\psi_{ev}(y, z) &= \sum_{n=0}^{\infty}(-1)^n[1/(2n)!]y^{2n}E^{[2n]}(z)\\ &= E^{[0]}(z) - (1/2)y^2E^{[2]}(z) + (1/24)y^4E^{[4]}(z) + \cdots .\end{aligned} \tag{H.3.38}$$

For the odd part there is the representation

$$\begin{aligned}\psi_{od}(y, z) &= \sum_{n=0}^{\infty}(-1)^n[1/(2n+1)!]y^{2n+1}O^{[2n]}(z)\\ &= yO^{[0]}(z) - (1/6)y^3O^{[2]}(z) + (1/120)y^5O^{[4]}(z) + \cdots .\end{aligned} \tag{H.3.39}$$

Thus, ψ is specified by the two functions $E^{[0]}(z)$ and $O^{[0]}(z)$. These functions may, in principle, be chosen arbitrarily. Often they are required to go to zero as $|z| \to \infty$.

For the "fields" associated with ψ_{ev} and ψ_{od}, call them $\boldsymbol{B}^{ev}$ and $\boldsymbol{B}^{od}$, there are the results

$$B_x^{ev} = \partial_x\psi_{ev} = 0, \tag{H.3.40}$$

$$B_y^{ev} = \partial_y\psi_{ev} = -yE^{[2]}(z) + (1/6)y^3E^{[4]}(z) - (1/120)y^5E^{[6]}(z) + \cdots , \tag{H.3.41}$$

$$B_z^{ev} = \partial_z\psi_{ev} = E^{[1]}(z) - (1/2)y^2E^{[3]}(z) + (1/24)y^4E^{[5]}(z) + \cdots ; \tag{H.3.42}$$

$$B_x^{od} = \partial_x\psi_{od} = 0, \tag{H.3.43}$$

$$B_y^{od} = \partial_y\psi_{od} = O^{[0]}(z) - (1/2)y^2O^{[2]}(z) + (1/24)y^4O^{[4]}(z) + \cdots , \tag{H.3.44}$$

$$B_z^{od} = \partial_z\psi_{od} = yO^{[1]}(z) - (1/6)y^3O^{[3]}(z) + (1/120)y^5O^{[5]}(z) + \cdots . \tag{H.3.45}$$

Again, because the fields are the gradients of harmonic functions, the Cartesian components of $\boldsymbol{B}^{ev}$ and $\boldsymbol{B}^{od}$ must also be harmonic functions.

Explicit Construction from Analytic Functions

Suppose the functions $\psi_{ev}(y,z)$ and $\psi_{od}(y,z)$ are related to the real and imaginary parts of a real-analytic function f. Make the identifications

$$z \leftrightarrow u, \tag{H.3.46}$$

$$y \leftrightarrow v, \tag{H.3.47}$$

$$E^{[0]}(z) \leftrightarrow f^{[0]}(u), \tag{H.3.48}$$

and

$$O^{[0]}(z) \leftrightarrow f^{[1]}(u). \tag{H.3.49}$$

With these identifications, we have the relations

$$\psi_{ev}(y,z) = f_r(y,x), \tag{H.3.50}$$

$$\psi_{od}(y,z) = f_i(y,x). \tag{H.3.51}$$

The identifications (H.48) and (H.49) again require the restrictive relation

$$O^{[0]} = E^{[1]}. \tag{H.3.52}$$

But, since f_r and f_i are the real and imaginary parts of the analytic function f, they must satisfy the Cauchy-Riemann relations

$$\partial_z f_r = \partial_x f_i, \tag{H.3.53}$$

$$\partial_x f_r = -\partial_z f_i. \tag{H.3.54}$$

Thus, we have the relations

$$B_y^{ev} = -B_z^{od}, \tag{H.3.55}$$

$$B_z^{ev} = B_y^{od}. \tag{H.3.56}$$

These relations also follow from the representations (H.40) through (H.45) and the relation (H.52).

H.3.3 More About $\mathrm{B}^{od}(y,z)$ and Another Application of Analytic Function Theory

To keep the promise made in Subsection H.3.1, we now discuss another application of the relation between analytic and harmonic functions. Consider the field $\boldsymbol{B}^{od}(y,z)$ given by (H.43) through (H.45). It is a candidate magnetic field for an infinitely wide (in x) parallel-faced dipole, see Figures 1.6.1 and 1.6.2, with symmetry about the midplane $y=0$. Given a real-analytic function $f(w)$, define a related function $g(w)$ by the rule

$$g(w) = f^{[1]}(w). \tag{H.3.57}$$

Evidently g will also be real analytic. As before, employ the relation (H.17) and decompose g into real and imaginary parts by writing

$$g(u+iv) = g_r(u,v) + g_i(u,v). \tag{H.3.58}$$

In analogy to what was done for f, expand $g(u+iv)$ as a Taylor series in iv to get the relation

$$\begin{aligned} g(u+iv) &= \sum_{n=0}^{\infty} g^{[n]}(u)(iv)^n/n! \\ &= [g^{[0]}(u) - (1/2)v^2 g^{[2]}(u) + (1/24)v^4 g^{[4]}(u) + \cdots] \\ &\quad +i[vg^{[1]}(u) - (1/6)v^3 g^{[3]}(u) + (1/120)v^5 g^{[5]}(u) + \cdots]. \end{aligned} \tag{H.3.59}$$

We see that

$$g_r(u,v) = g^{[0]}(u) - (1/2)v^2 g^{[2]}(u) + (1/24)v^4 g^{[4]}(u) + \cdots \tag{H.3.60}$$

and

$$g_i(u,v) = vg^{[1]}(u) - (1/6)v^3 g^{[3]}(u) + (1/120)v^5 g^{[5]}(u) + \cdots . \tag{H.3.61}$$

Make the identifications (H.46) and (H.47) and compare the series on the right side of (H.59) with the series (H.44) and (H.45) for B_y^{od} and B_z^{od}. We see that they are analogous if we make the identification

$$O^{[0]} = g^{[0]}. \tag{H.3.62}$$

In particular, we have the relations

$$B_y^{od}(y,z) = g_r(z,y) = g^{[0]}(z) - (1/2)y^2 g^{[2]}(z) + (1/24)y^4 g^{[4]}(z) + \cdots , \tag{H.3.63}$$

$$B_z^{od}(y,z) = g_i(z,y) = yg^{[1]}(z) - (1/6)y^3 g^{[3]}(z) + (1/120)y^5 g^{[5]}(z) + \cdots . \tag{H.3.64}$$

The Cauchy-Riemann relations again hold for g_r and g_i,

$$\partial_u g_r(u,v) = \partial_v g_i(u,v), \tag{H.3.65}$$

$$\partial_v g_r(u,v) = -\partial_u g_i(u,v). \tag{H.3.66}$$

In this case, because of (H.63) and (H.64), they have the consequence

$$\partial_z B_y^{od}(y,z) = \partial_z g_r(z,y) = \partial_y g_i(y,z) = \partial_y B_z^{od}(y,z), \tag{H.3.67}$$

$$\partial_y B_y^{od}(y,z) = \partial_y g_r(z,y) = -\partial_z g_i(y,z) = -\partial_z B_z^{od}(y,z). \tag{H.3.68}$$

These relations can also be verified directly form the representations (H.63) and (H.64).

We already know that $\nabla \times \boldsymbol{B}^{od} = 0$ because $\boldsymbol{B}^{od}$ is the gradient of a scalar field. See (H.43) through (H.45). What can be said about $\nabla \cdot \boldsymbol{B}^{od}$? From the second Cauchy-Riemann relation (H.68) we see that

$$\nabla \cdot \boldsymbol{B}^{od} = \partial_y B_y^{od} + \partial_z B_z^{od} = 0, \tag{H.3.69}$$

as expected and required.

Also observe that, in complex notation, the relations (H.58), (H.63), and (H.64) can be expressed in the compact form

$$B_y^{odd}(y,z) + iB_z^{odd}(y,z) = g(z+iy). \tag{H.3.70}$$

Thus, in this application, the selection of one real-analytic function $g(w)$ specifies both components of the field for an infinitely wide parallel-faced dipole. Indeed, the application is even broader. It would also apply to an infinitely wide parallel-faced wiggler. In this case $g(u)$ would be roughly oscillatory in u. All that is required in both applications is that $g(u)$ vanish as $u \rightarrow \pm\infty$, in which case $\boldsymbol{B}^{od}(y,z)$ will vanish as $z \rightarrow \pm\infty$. Suppose, for example, that we set

$$g(u) = B \text{ bump}(u, c\ell, L) \tag{H.3.71}$$

where bump$(u, c\ell, L)$ is one of the bump functions defined in Section 11.11. These functions are real analytic, and therefore the representation (H.70) can be implemented.

There is one last item to be discussed. Namely, it would be good to have a vector potential $\boldsymbol{A}^{od}$ from which $\boldsymbol{B}^{od}$ could be derived. Consider the following Ansatz,

$$A_x^{od}(x,y,z) = 0, \tag{H.3.72}$$

$$A_y^{od}(x,y,z) = xB_z^{od}(y,z), \tag{H.3.73}$$

$$A_z^{od}(x,y,z) = -xB_y^{od}(y,z). \tag{H.3.74}$$

(Evidently, this vector potential is horizontal free.) It is easily verified that

$$(\nabla \times \boldsymbol{A}^{od})_x = \partial_y A_z^{od} - \partial_z A_y^{od} = x[-\partial_y B_y^{od}(y,z) - \partial_z B_z^{od}(y,z)] = 0 = B_x^{od}(y,z), \tag{H.3.75}$$

$$(\nabla \times \boldsymbol{A}^{od})_y = \partial_z A_x^{od} - \partial_x A_z^{od} = -\partial_x A_z^{od} = B_y^{od}(y,z), \tag{H.3.76}$$

$$(\nabla \times \boldsymbol{A}^{od})_z = \partial_x A_y^{od} - \partial_y A_x^{od} = \partial_x A_y^{od} = B_z^{od}(y,z), \tag{H.3.77}$$

as desired. Also, we find that

$$\nabla \cdot \boldsymbol{A}^{od} = \partial_y A_y^{od} + \partial_z A_z^{od} = x[\partial_y B_z^{od}(y,z) - \partial_z B_y^{od}(y,z)] = 0. \tag{H.3.78}$$

Here we have used the first Cauchy-Riemann relation (H.67). Thus, in addition to being horizontal free, the vector potential $\boldsymbol{A}^{od}(x,y,z)$ is in the Coulomb gauge. Since $\boldsymbol{A}^{od}(x,y,z)$ is in the Coulomb gauge, it follows, as is also readily verified from (H.72) through (H.74), that its Cartesian components are harmonic functions.

Note also, from its definition (H.72) through (H.74), that $\boldsymbol{A}^{od}(x,y,z)$ vanishes as $z \rightarrow \pm\infty$ if $\boldsymbol{B}^{od}(y,z)$ does so. This feature is desirable because we would like canonical and mechanical momenta to be equal in field-free regions. We also note the convenient feature that $A_y^{od}(x,y,z)$ vanishes in the midplane,

$$A_y^{od}(x, y=0, z) = xB_z^{od}(y=0,z) = 0. \tag{H.3.79}$$

See (H.64). Thus, for the design orbit which lies in the midplane, there is no difference between mechanical and canonical for the x and y components of the momentum. Finally, as in the case of the vector potentials found in Exercises 13.3.4 and 13.5.5, the vector potential is primarily in the z direction except in the fringe-field regions.

Bibliography

[1] The Ansatz (H.72) through (H.74) for $\boldsymbol{A}^{od}(x, y, z)$ is due to Peter Walstrom.

Appendix I

Poisson Bracket Relations

I.1 Poisson Brackets

$$z_a J_{aa'} z_{a'} = 0 \tag{I.1.1}$$

$$[f_m, z_a] = (\partial_b f_m) J_{bb'} \partial_{b'} z_a = (\partial_b f_m) J_{bb'} \delta_{b'a} = (\partial_b f_m) J_{ba} \tag{I.1.2}$$

$$\begin{aligned}[f_m, z_a] J_{aa'} z_{a'} &= (\partial_b f_m) J_{ba} J_{aa'} z_{a'} = -(\partial_b f_m) \delta_{ba'} z_{a'} \\ &= -z_b(\partial_b f_m) = -m f_m\end{aligned} \tag{I.1.3}$$

$$[f_m, [f_n, z_a]] = [f_m, (\partial_b f_n)] J_{ba} = (\partial_c f_m) J_{cc'} (\partial_{c'} \partial_b f_n) J_{ba} \tag{I.1.4}$$

$$\begin{aligned}[f_m, [f_n, z_a]] J_{aa'} z_{a'} &= (\partial_c f_m) J_{cc'} (\partial_{c'} \partial_b f_n) J_{ba} J_{aa'} z_{a'} = -(\partial_c f_m) J_{cc'} (\partial_{c'} \partial_b f_n) \delta_{ba'} z_{a'} \\ &= -(\partial_c f_m) J_{cc'} z_b (\partial_{c'} \partial_b f_n)\end{aligned} \tag{I.1.5}$$

$$z_b(\partial_{c'} \partial_b f_n) = \partial_{c'}(z_b \partial_b f_n) - \delta_{c'b}(\partial_b f_n) = (n-1)(\partial_{c'} f_n) \tag{I.1.6}$$

$$[f_m, [f_n, z_a]] J_{aa'} z_{a'} = -(n-1)(\partial_c f_m) J_{cc'} (\partial_{c'} f_n) = -(n-1)[f_m, f_n] \tag{I.1.7}$$

$$\begin{aligned}&[f_\ell, [f_m, [f_n, z_a]]] = [f_\ell, (\partial_c f_m)(\partial_{c'} \partial_b f_n)] J_{cc'} J_{ba} \\ =\ &(\partial_d f_\ell) \{\partial_{d'} [(\partial_c f_m)(\partial_{c'} \partial_b f_n)]\} J_{dd'} J_{cc'} J_{ba} \\ =\ &(\partial_d f_\ell) [(\partial_{d'} \partial_c f_m)(\partial_{c'} \partial_b f_n) + (\partial_c f_m)(\partial_{d'} \partial_{c'} \partial_b f_n)] J_{dd'} J_{cc'} J_{ba}\end{aligned} \tag{I.1.8}$$

$$\begin{aligned}
& [f_\ell,[f_m,[f_n,z_a]]]J_{aa'}z_{a'} \\
= \;& (\partial_d f_\ell)(\partial_{d'}\partial_c f_m)(\partial_{c'}\partial_b f_n)J_{dd'}J_{cc'}J_{ba}J_{aa'}z_{a'} + \\
& (\partial_d f_\ell)(\partial_c f_m)(\partial_{d'}\partial_{c'}\partial_b f_n)J_{dd'}J_{cc'}J_{ba}J_{aa'}z_{a'} \\
= \;& -(\partial_d f_\ell)(\partial_{d'}\partial_c f_m)(\partial_{c'}\partial_b f_n)J_{dd'}J_{cc'}\delta_{ba'}z_{a'} - \\
& (\partial_d f_\ell)(\partial_c f_m)(\partial_{d'}\partial_{c'}\partial_b f_n)J_{dd'}J_{cc'}\delta_{ba'}z_{a'} \\
= \;& -(\partial_d f_\ell)(\partial_{d'}\partial_c f_m)z_b(\partial_{c'}\partial_b f_n)J_{dd'}J_{cc'} - \\
& (\partial_d f_\ell)(\partial_c f_m)z_b(\partial_{d'}\partial_{c'}\partial_b f_n)J_{dd'}J_{cc'}
\end{aligned} \tag{I.1.9}$$

$$\begin{aligned}
& \partial_{d'}\partial_{c'}(z_b\partial_b f_n) = \partial_{d'}[\delta_{c'b}\partial_b f_n + z_b\partial_{c'}\partial_b f_n] \\
= \;& \partial_{d'}[\partial_{c'} f_n + z_b\partial_{c'}\partial_b f_n] \\
= \;& \partial_{d'}\partial_{c'} f_n + \partial_{d'}(z_b\partial_{c'}\partial_b f_n) \\
= \;& \partial_{d'}\partial_{c'} f_n + \delta_{d'b}\partial_{c'}\partial_b f_n + z_b\partial_{d'}\partial_{c'}\partial_b f_n \\
= \;& \partial_{d'}\partial_{c'} f_n + \partial_{c'}\partial_{d'} f_n + z_b\partial_{d'}\partial_{c'}\partial_b f_n \\
= \;& 2\partial_{d'}\partial_{c'} f_n + z_b\partial_{d'}\partial_{c'}\partial_b f_n
\end{aligned} \tag{I.1.10}$$

$$z_b\partial_{d'}\partial_{c'}\partial_b f_n = (n-2)\partial_{d'}\partial_{c'} f_n \tag{I.1.11}$$

$$\begin{aligned}
& [f_\ell,[f_m,[f_n,z_a]]]J_{aa'}z_{a'} \\
= \;& -(\partial_d f_\ell)(\partial_{d'}\partial_c f_m)z_b(\partial_{c'}\partial_b f_n)J_{dd'}J_{cc'} - \\
& (\partial_d f_\ell)(\partial_c f_m)z_b(\partial_{d'}\partial_{c'}\partial_b f_n)J_{dd'}J_{cc'} \\
= \;& -(\partial_d f_\ell)(\partial_{d'}\partial_c f_m)(n-1)(\partial_{c'} f_n)J_{dd'}J_{cc'} - \\
& (\partial_d f_\ell)(\partial_c f_m)(n-2)(\partial_{d'}\partial_{c'} f_n)J_{dd'}J_{cc'} \\
= \;& -(n-1)(\partial_d f_\ell)J_{dd'}(\partial_{d'}\partial_c f_m)J_{cc'}(\partial_{c'} f_n) - \\
& (n-2)(\partial_d f_\ell)(\partial_c f_m)J_{dd'}J_{cc'}(\partial_{d'}\partial_{c'} f_n) \\
= \;& -(n-1)(\partial_d f_\ell)J_{dd'}(\partial_{d'}\partial_c f_m)J_{cc'}(\partial_{c'} f_n) + \\
& (n-2)(\partial_d f_\ell)(\partial_c f_m)J_{dd'}J_{c'c}(\partial_{d'}\partial_{c'} f_n) \\
= \;& -(n-1)(\partial_d f_\ell)J_{dd'}(\partial_{d'}\partial_c f_m)J_{cc'}(\partial_{c'} f_n) + \\
& (n-2)(\partial_d f_\ell)J_{dd'}(\partial_{d'}\partial_{c'} f_n)J_{c'c}(\partial_c f_m) \\
= \;& -(\partial_d f_\ell)J_{dd'}(\partial_{d'}\partial_c f_m)J_{cc'}(\partial_{c'} f_n)
\end{aligned} \tag{I.1.12}$$

$$\begin{aligned}&[f_k,[f_\ell,[f_m,[f_n,z_a]]]]=[f_k,(\partial_d f_\ell)\{(\partial_{d'}\partial_c f_m)(\partial_{c'}\partial_b f_n)+(\partial_c f_m)(\partial_{d'}\partial_{c'}\partial_b f_n)\}]J_{dd'}J_{cc'}J_{ba}\\ =\ &[f_k,(\partial_d f_\ell)(\partial_{d'}\partial_c f_m)(\partial_{c'}\partial_b f_n)]J_{dd'}J_{cc'}J_{ba}+[f_k,(\partial_d f_\ell)(\partial_c f_m)(\partial_{d'}\partial_{c'}\partial_b f_n)]J_{dd'}J_{cc'}J_{ba}\end{aligned} \tag{I.1.13}$$

$$\begin{aligned}&[f_k,(\partial_d f_\ell)(\partial_{d'}\partial_c f_m)(\partial_{c'}\partial_b f_n)]=(\partial_e f_k)J_{ee'}\{\partial_{e'}[(\partial_{d'}\partial_c f_m)(\partial_{c'}\partial_b f_n)]\}\\ =\ &(\partial_e f_k)J_{ee'}[(\partial_{e'}\partial_{d'}\partial_c f_m)(\partial_{c'}\partial_b f_n)+]\end{aligned} \tag{I.1.14}$$

$$\begin{aligned}&[f_k,(\partial_d f_\ell)(\partial_c f_m)(\partial_{d'}\partial_{c'}\partial_b f_n)]\\ =\ &\end{aligned} \tag{I.1.15}$$

I.2 Preparatory Results

$$(z,Jz)=z_aJ_{aa'}z_{a'}=0 \tag{I.2.1}$$

$$(:f_n:z,Jz)=([f_n,z],Jz)=[f_n,z_a]J_{aa'}z_{a'}=-nf_n \tag{I.2.2}$$

$$(:f_m::f_n:z,Jz)=([f_m,[f_n,z]],Jz)=[f_m,[f_n,z_a]]J_{aa'}z_{a'}=-(n-1)[f_m,f_n] \tag{I.2.3}$$

$$\begin{aligned}(:f_\ell::f_m::f_n:z,Jz)&=([f_\ell,[f_m,[f_n,z]]],Jz)=[f_\ell,[f_m,[f_n,z_a]]]J_{aa'}z_{a'}\\ &=(\partial_d f_\ell)J_{dd'}(\partial_{d'}\partial_c f_m)J_{cc'}(\partial_{c'}f_n)\\ &=(\partial f_\ell,JS(f_m)J\partial f_n).\end{aligned} \tag{I.2.4}$$

Here $S(f_m)$ is the Hessian of f_m,

$$S_{ab}(f_m)=\partial_a\partial_b f_m. \tag{I.2.5}$$

$$[f_k,[f_\ell,[f_m,[f_n,z_a]]]]J_{aa'}z_{a'}= \tag{I.2.6}$$

I.3 Application

Suppose $t^i = 0$ and $t^f = 1$ in (1.2.57) and (6.6.57). Suppose also that we confine our interest to the case where only h_3 through h_6 are possibly nonzero. Our task will be to find the Poincaré generating function F_+ corresponding to $\mathcal{M}$.

From (6.6.57) we have the result

$$F(z) = -\sum_m (m-2)h_m(z). \tag{I.3.1}$$

And from (6.6.37) we know that

$$F_+(z) = [F + (Z, Jz)]/2. \tag{I.3.2}$$

Also, by definition,

$$Z = \mathcal{M}z. \tag{I.3.3}$$

Upon combining () through () we conclude that F_+ is given by the relation

$$F_+(z) = (1/2)[(\mathcal{M}z, Jz) - \sum_m (m-2)h_m(z)]. \tag{I.3.4}$$

Our task will be to work out the implications of (). What we will find will be a homogeneous polynomial expansion for F_+.

Evidently the hard part is to find an expansion for $(\mathcal{M}z, Jz)$. Let us write

$$\mathcal{M} = \exp(-:H:) = I - :H: + :H:^2/2! - :H:^3/3! : +H:^4/4! + \cdots \tag{I.3.5}$$

where the terms retained are sufficient to compute F_+ through terms of degree 6. Let $Z^{(n)}$ be the contribution to Z made by the term $:-H:^n/n!$ in $\mathcal{M}$ and let $Y^{(n)}$ be the contribution that it makes to F_+. That is, we make the definitions

$$Z^{(n)} = [:-H:^n/n!]z \tag{I.3.6}$$

and

$$Y^{(n)}(z) = (Z^{(n)}, Jz) = ([:-H:^n/n!]z, Jz). \tag{I.3.7}$$

From the relations listed in Section 23.1 above, we easily find the results

$$Y^{(0)}(z) = (Z^{(0)}, Jz) = (z, Jz) = 0, \tag{I.3.8}$$

$$\begin{aligned} Y^{(1)}(z) &= (Z^{(1)}, Jz) = (:-H:z, Jz) = -\sum_m (:h_m:z, Jz) \\ &= \sum_m m h_m(z). \end{aligned} \tag{I.3.9}$$

It follows from the work done so far that

$$\begin{aligned} F_+(z) &= (1/2)[(\mathcal{M}z, Jz) - \sum_m (m-2)h_m(z)] \\ &= (1/2)\{[\sum_m m h_m(z)] - [\sum_m (m-2)h_m(z)] + \cdots\} = H(z) + \cdots. \end{aligned} \tag{I.3.10}$$

As already stated, we ultimately desire to have an expansion of F_+ in homogeneous polynomials. Let F_+^m denote the term in F_+ that is homogeneous of degree n. Then, based on the work done so far, we have the result

$$F_+^m = h_m + \cdots \,. \tag{I.3.11}$$

The next term we need is $Y^{(2)}(z)$. We have the result

$$Y^{(2)}(z) = (Z^{(2)}, Jz) = (:H:^2 z, Jz)/2!. \tag{I.3.12}$$

The term $:H:^2$ has the expansion

$$\begin{aligned} :H:^2 &= \sum_m \sum_n :h_m::h_n: \\ &= :h_3:^2 + :h_3::h_4: + :h_4::h_3: \\ &+ :h_4:^2 + :h_3::h_5: + :h_5::h_3: + \cdots \end{aligned} \tag{I.3.13}$$

where we have displayed only the terms that will contribute to the F_+^m for $m \leq 6$. Correspondingly, we find for $Y^{(2)}$ the result

$$\begin{aligned} Y^{(2)}(z) &= (1/2!)[(:h_3:^2 z, Jz) \\ &+ (:h_3::h_4: z, Jz) + (:h_4::h_3: z, Jz) \\ &+ (:h_4:^2 z, Jz) \\ &+ (:h_3::h_5: z, Jz) + (:h_5::h_3: z, Jz)] \\ &= (1/2!)([h_3, h_3] + [h_4, h_4] \\ &+ [h_3, h_4] + [h_4, h_3] + [h_3, h_5] + [h_5, h_3]) \\ &= (1/2)([h_3, h_4] + [h_3, h_5]). \end{aligned} \tag{I.3.14}$$

We are now able to conclude that

$$F_+^3 = h_3, \tag{I.3.15}$$

$$F_+^4 = h_4, \tag{I.3.16}$$

$$F_+^5 = h_5 + (1/4)[h_3, h_4] + \cdots \,, \tag{I.3.17}$$

$$F_+^6 = h_6 + (1/4)[h_3, h_5] + \cdots \,. \tag{I.3.18}$$

Let us move on to the term

$$Y^{(3)}(z) = (Z^{(3)}, Jz) = -(:H:^3 z, Jz)/3!. \tag{I.3.19}$$

The term $:H:^3$ has the expansion

$$\begin{aligned} :H:^3 &= \sum_\ell \sum_m \sum_n :h_\ell::h_m::h_n: \\ &= :h_3:^3 + :h_3:^2:h_4: + :h_3::h_4::h_3: + :h_4::h_3:^2 + \cdots \end{aligned} \tag{I.3.20}$$

where we have displayed only the terms that will contribute to the F_+^m for $m \leq 6$. Correspondingly, we find for $Y^{(3)}$ the result

$$\begin{aligned} Y^{(3)}(z) &= (1/2!)[(: h_3 :^3 z, Jz) \\ &+ (: h_3 :^2: h_4 : z, Jz) + (: h_3 :: h_4 :: h_3; z, Jz) + (: h_4 :: h_3 :^2 z, Jz)] \\ &= \end{aligned} \tag{I.3.21}$$

We are now able to conclude that

$$F_+^3 = h_3, \tag{I.3.22}$$

$$F_+^4 = h_4, \tag{I.3.23}$$

$$F_+^5 = h_5 + (1/4)[h_3, h_4] + (\partial h_3, JS(h_3)J\partial h_3) \tag{I.3.24}$$

$$F_+^6 = h_6 + [h_3, h_5] + (\partial h_3, JS(h_3)J\partial h_4) + (\partial h_3, JS(h_4)J\partial h_3) + \cdots . \tag{I.3.25}$$

Finally, as we will see, we need the term

$$Y^{(4)}(z) = (Z^{(4)}, Jz) = -(: H :^4 z, Jz)/3!. \tag{I.3.26}$$

The term $: H :^4$ has the expansion

$$\begin{aligned} : H :^4 &= \sum_k \sum_\ell \sum_m \sum_n : h_k :: h_\ell :: h_m :: h_n : \\ &= : h_3 :^4 + \cdots \end{aligned} \tag{I.3.27}$$

where we have again displayed only the terms that will contribute to the F_+^m for $m \leq 6$. Correspondingly, we find for $Y^{(3)}$ the result

$$Y^{(4)}(z) = (1/2!)[(: h_3 :^4 z, Jz) = \tag{I.3.28}$$

We are now able to conclude that

$$F_+^3 = h_3, \tag{I.3.29}$$

$$F_+^4 = h_4, \tag{I.3.30}$$

$$F_+^5 = h_5 + (1/4)[h_3, h_4] + (\partial h_3, JS(h_3)J\partial h_3), \tag{I.3.31}$$

$$F_+^6 = h_6 + [h_3, h_5] + (\partial h_3, JS(h_3)J\partial h_4) + (\partial h_3, JS(h_4)J\partial h_3) + . \tag{I.3.32}$$

Appendix J

Feigenbaum Cascade Denied/Achieved

Section 1.2.1 described Feigenbaum infinite period doubling cascades and mentioned that, for some maps in some parameter ranges, period doubling cascades begin but do not continue to completion. The purpose of this appendix is to provide a simple example of both incomplete and complete cascades.

J.1 Simple Map and Its Initial Bifurcations

Consider the simple one-dimensional map given by the relation

$$x_{n+1} = \mathcal{M}x_n = f(a, b; x_n) = a + bx_n/[1 + (x_n)^2] \tag{J.1.1}$$

where a and b are parameters. Figure 1.1 shows the curves $y = f(a, b; x)$ for $b = 11.5$ and selected values of a. As is evident from (1.1) and from the figure, these curves are all vertical displacements of eachother.

Also shown is the line $y = x$. Any intersection of the line $y = x$ and the curve $y = f(a, b; x)$ corresponds to a fixed point of $\mathcal{M}$. Observe that for sufficiently negative values of a there is only one intersection, and hence only one fixed point. This is the fixed point whose path is shown as a function of a in the lower left portion of the bifurcation diagram provided by Figure 1.2. Its path extends forever to the left, and has the asymptotic form

$$x_\infty \simeq a \text{ as } a \to -\infty. \tag{J.1.2}$$

Further computation shows that this fixed point is stable.

However, as a is increased slightly beyond -5, Figure 1.1 shows that there are two more intersections of the curve $y = f(a, b; x)$ with the line $y = x$. When these two intersections first occur (when the curve and line are tangent), a pair of fixed points is born together in a blue-sky bifurcation. These are the two fixed points that appear out of the blue at $a \simeq -4.8$ and $x_\infty \simeq +1$ in Figure 1.2, and move along separate paths as a is further increased. One of these fixed points (the one on the lower of the two paths) is unstable, and the other (the one on the very top path) is stable.

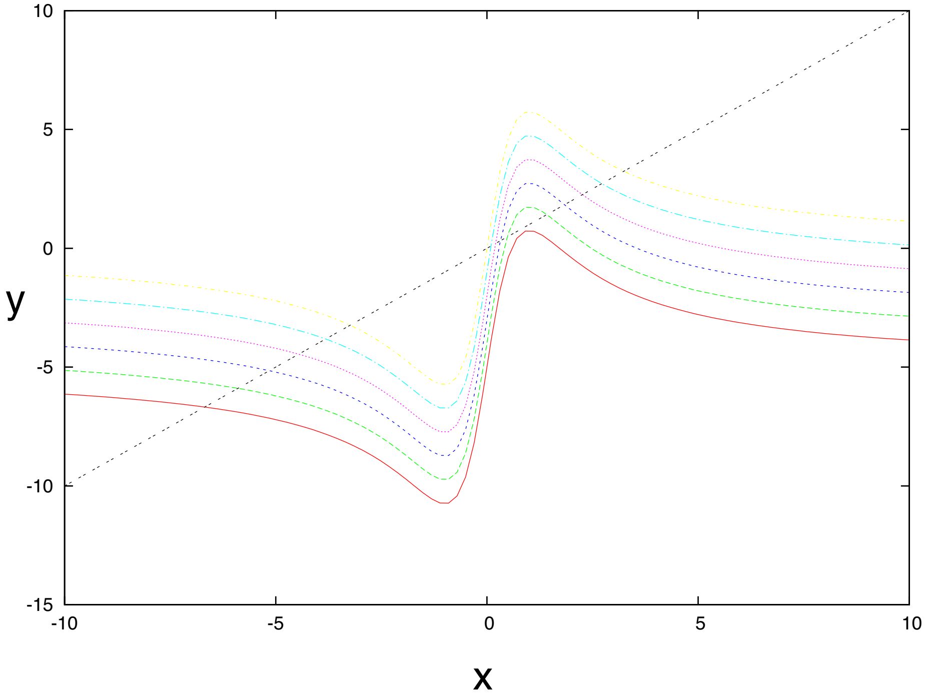


Figure J.1.1: The curves $y = f(a, b; x)$ for $b = 11.5$ and various values of $a \in [-5, 0]$. Also shown is the line $y = x$. Intersections of the line and the curve correspond to fixed points.

J.2 Complete Cascade Denied

As a is increased still further, the fixed point on the very top path begins a period doubling cascade, which we will call the *upper* cascade, at $a \simeq -4.3$ and $x \simeq +1.5$. Again see Figure 1.2. However, as is evident from the figure, the period doubling cascade does not run to completion. Instead it ceases and then begins to undo itself by successive mergers that begin at $a \simeq -3.2$ so that for $a \simeq -.8$ there is again a single fixed point. Its path extends forever to the right, and has the asymptotic form

$$x_\infty \simeq a \text{ as } a \to +\infty. \tag{J.2.1}$$

Further computation shows that this fixed point is stable.

We began our discussion with the fixed point whose path appears in the lower left side of Figure 1.2 and has the asymptotic form (1.2). Let us now follow its history as a is increased. As indicated in Figure 1.2, it too begins a period doubling cascade, which we will call the *lower* cascade, and this cascade begins at $a \simeq +.8$ and $q_\infty \simeq -2.8$. And, like the upper period doubling cascade, it also does not run to completion. Instead it stops and then undoes itself by successive mergers so that for $a \simeq +4.3$ there is again a single fixed point. This fixed point is stable. Inspection of Figure 1.2 shows that, as a is increased still further, this fixed point blue-sky merges with the unstable fixed point that came out of the blue-sky bifurcation at $a \simeq -4.8$ and $q_\infty \simeq +1$ so that they are mutually annihilated at $q_\infty \simeq -1$ when $a \simeq +4.8$.

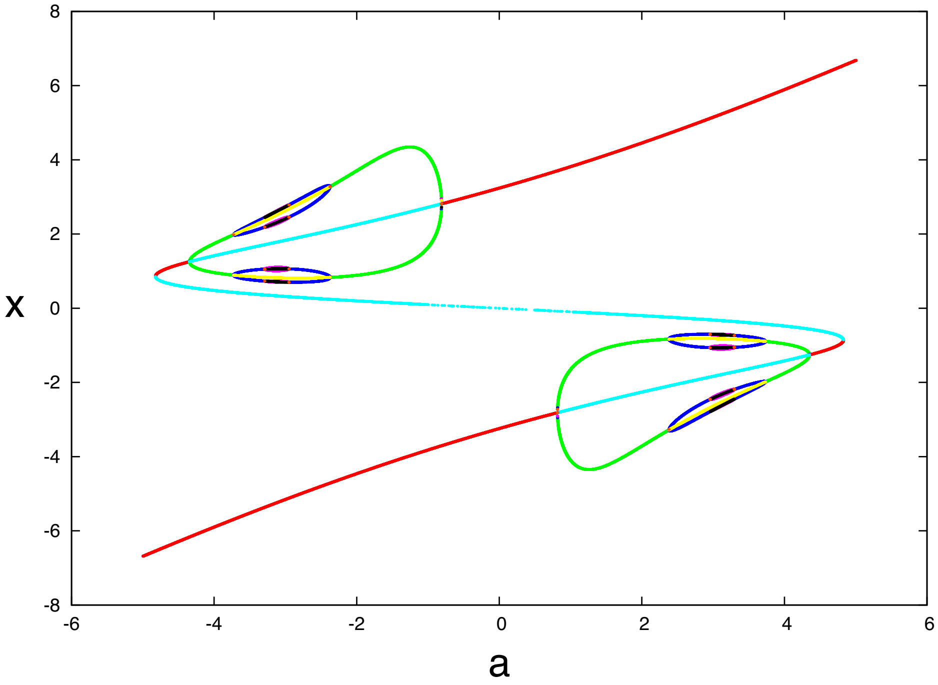


Figure J.1.2: Bifurcation diagram showing x_∞ as a function of a for the map (1.1) with $b = 11.5$ and $a \in [-5, 5]$. For $a = -5$, there is only one fixed point, and it is stable. As a is increased from this value, a blue-sky bifurcation occurs at $x_\infty \simeq +1$ when $a =\simeq -4.8$. Here a pair of fixed points, one stable and one unstable, is born. Now there are three fixed points. The one that bifurcates to larger values of x_∞ is stable, and the one that bifurcates to smaller values of x_∞ is unstable. The original fixed point persists, and remains stable. At $x_\infty \simeq +1$ and $a =\simeq +4.8$ a blue-sky merger occurs where two fixed points, one stable and the other unstable, annihilate. For a values larger than this there is only one fixed point. In between the values $a \simeq -4.8$ and $a \simeq +4.8$ there are two incomplete period-doubling cascades.

Exercises

J.2.1. Figure 2.1 displays intersections between the curve $y = f(a, b; x)$ and the line $y = x$, and we have seen how these intersections are related to the blue-sky bifurcation at $a \simeq -4.8$. Show that the blue-sky merger at $a \simeq +4.8$ can also be understood in terms of intersections between the curve $y = f(a, b; x)$ and the line $y = x$.

J.3 Complete Cascade Achieved

We have seen, from Figure 1.2, that for $b = 11.5$ the period-doubling cascades fail to run to completion. By contrast in Figure 3.1, for which $b = 11.7$, each Feigenbaum cascade runs to completion followed by a region of chaos. Then, it is fascinating to see, each cascade undoes itself by successive mergers as a is further increased until eventually there is again only the one stable fixed point.[1] Note also there that there are two visible windows of stability at $a \simeq -3.3$ and $a \simeq -3.05$. These windows contain stable period-twelve fixed points (as well as numerous unstable fixed points that do not appear because this is a Feigenbaum diagram).

[1] In fact there are dynamical systems for which a large or even infinite number of period doubling cascades followed by inverse cascades occur.

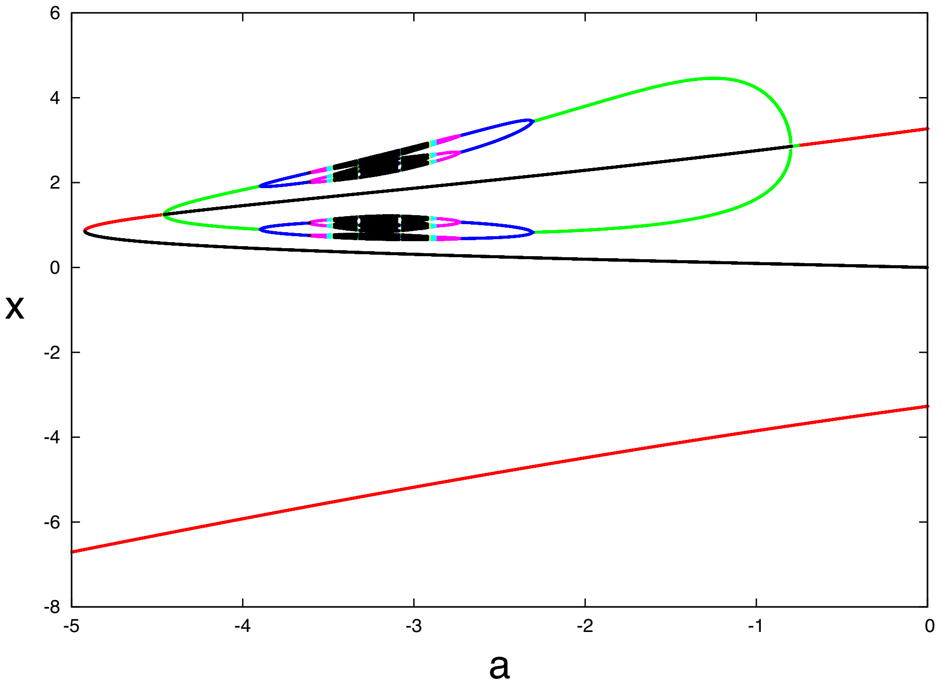


Figure J.3.1: A portion of the Feigenbaum diagram for the map (1.1) with $b = 11.7$. Also shown are the paths of all period-one fixed points, both stable and unstable. The full diagram is similar to that of Figure 1.2 except that both period-doubling cascades now run to completion. Specifically, for the upper cascade shown here, a blue-sky bifurcation again occurs and, as a is further increased, the stable fixed point begins a Feigenbaum perioding doubling cascade that now runs to completion followed by a region of chaos. But then, as a is increased still further, the cascades undoes itself until there is again only a single stable fixed point. The behavior for the lower cascade is analogous.

Bibliography

[1] M. Bier and T. Bountis, “Remerging Feigenbaum trees in dynamical systems”, *Phys. Lett.*, **104A**, 239-244 (1984).

[2] S. Dawson, C. Grebogi, J. Yorke, I. Kan, and H. Koçak, “Antimonotonicity: inevitable reversals of period-doubling cascades”, *Phys. Lett. A*, **162**, 249-254 (1992).

Appendix K

Supplement to Chapter 17

K.1 Computation of Generalized Gradients from Spinning Coil Data

A widely used method to measure the magnetic field in magnets for beam optics relies on spinning coils [*]. By using spinning coils one can achieve very accurate measurements of the angular Fourier components of the magnetic field. In this section, which is restricted to the case of straight elements, we show how it is possible using a short length (i.e. with a length shorter than the region of the magnet where the fields are z-dependent) rectangular spinning coil to recover the full z-dependent profiles of the fields and in particular the profiles of the generalized gradients that are necessary to compute accurately both the linear and nonlinear parts of the transfer map for the magnet.

We consider the case of a rectangular coil rotating in such a way that one side of the coil is always positioned along the magnetic axis of the magnet. The idea is to make repeated measurements of angular field data (integrated over the coil area) by moving the coil along the magnet axis by small steps. The Fourier transforms of the experimental data for each angular harmonic are than calculated, multiplied by a suitable kernel, and then Fourier transformed back to obtain the desired generalized gradients.

For the kind of coil we consider in this section the only relevant component of the magnetic field is B_ϕ because it is the only one generating a flux linked to the coil. It should be mentioned that tangential coils are also used for which the relevant component of the magnetic field is B_ρ. The treatment of that case would follow the same lines as for the kind of coils considered here.

The E.M.F. produced by a rectangular spinning coil (or set of coils, in a realistic setup), with barycenter positioned at z is given by

$$\mathcal{E}(z,t) = -\int_{z-\ell_c}^{z+\ell_c} dz' \int_0^R \frac{dB_\phi}{dt} d\rho, \tag{K.1.1}$$

where $2\ell_c$ is the length of coil and R is its radius. The E.M.F. can be written in terms of a Fourier series in time,

$$\mathcal{E}(z,t) = \sum_{m=0} \mathcal{E}_{m,s}(z)\sin(m\omega t) + \mathcal{E}_{m,c}(z)\cos(m\omega t), \tag{K.1.2}$$

where we assume $\mathcal{E}_{m,s}(z)$ and $\mathcal{E}_{m,c}(z)$ can be experimentally determined over a sufficient number of locations in z in the end and fringe regions where the field varies with z. The angular frequency of the spinning coil is ω.

By using (2.3) we can write B_ϕ as

$$B_\phi = \frac{1}{\rho}\frac{\partial\psi}{\partial\phi} = \sum_{m=1}^{\infty}\int_{-\infty}^{\infty} dk e^{ikz} m \frac{I_m(k\rho)}{\rho}[\hat{b}_m(k)\cos m\phi - \hat{a}_m(k)\sin m\phi]. \tag{K.1.3}$$

By substituting (5.3) into (5.1) with $\phi = \omega t$ we get

$$\mathcal{E}(z,t) = \frac{\omega}{\sqrt{2\pi}}\sum_{m=1}^{\infty}\int_{-\infty}^{\infty} dk e^{ikz}\frac{2m^2\sin k\ell_c}{k}[\hat{b}_m(k)\sin m\omega t + \hat{a}_m(k)\cos m\omega t]. \tag{K.1.4}$$

Then, by comparing (5.2) and (5.4), we find the relations

$$\mathcal{E}_{m,c}(z) = \frac{m^2\omega}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dk e^{ikz}\mathcal{I}_m(kR)\frac{2\sin k\ell_c}{k}\hat{b}_m(k), \tag{K.1.5}$$

$$\mathcal{E}_{m,s}(z) = \frac{m^2\omega}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dk e^{ikz}\mathcal{I}_m(kR)\frac{2\sin k\ell_c}{k}\hat{a}_m(k). \tag{K.1.6}$$

Here we have defined the new function

$$\mathcal{I}_m(kR) = \int_0^R \frac{I_m(k\rho)}{\rho}d\rho = \int_0^{kR}\frac{I_m(x)}{x}dx. \tag{K.1.7}$$

Finally use of (5.5) and (5.6) allows us to write the expression for the generalized gradients,

$$C_{m,\alpha}^{[n]}(z) = \frac{i^n}{2^{m+1}m!m^2\omega}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dk e^{ikz}\frac{k^{m+n+1}\tilde{\mathcal{E}}_{m,\alpha}(k)}{\mathcal{I}_m(kR)\sin k\ell_c}, \tag{K.1.8}$$

where the $\tilde{\mathcal{E}}_{m,\alpha}(k)$ are the Fourier transforms of the experimental data,

$$\tilde{\mathcal{E}}_{m,\alpha}(k) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dz^{-ikz}\mathcal{E}_{m,\alpha}(z). \tag{K.1.9}$$

Notice that because of the asymptotic form of the Bessel function I_m, as $k \to \infty$ the function $\mathcal{I}_m(kR)$ grows exponentially at infinity as $e^{kR}/\sqrt{k}$. Because the function $\mathcal{I}_m(kR)$ is in the denominator of the integrand in (5.8), there is again an effective cut-off in k.

The integral (5.7) for a general m cannot be carried out analytically, but it can easily be reduced to an infinite series either in the Bessel functions or, most conveniently, directly in kR:

$$\mathcal{I}_m(kR) = \frac{1}{kR}\frac{2}{m}\sum_{n=0}^{\infty}(-1)^n(m+2n+1)I_{m+2n+1}(kR), \tag{K.1.10}$$

$$\mathcal{I}_m(kR) = \sum_{\ell=0}\frac{1}{(2\ell+m)\ell!(\ell+m)!}\left(\frac{kR}{2}\right)^{m+2\ell}. \tag{K.1.11}$$

In the particular case $m = 2$ we have

$$\mathcal{I}_2(kR) = \frac{I_1(kR)}{kR} - \frac{1}{2}. \tag{K.1.12}$$

For numerical purposes use of (5.11) may be perfectly adequate (in particular if speed is not an issue). We will see later that we are often interested in values of kR that satisfy the condition $kR < 20$. For such kR values, one can obtain values for $\mathcal{I}_m(kR)$ that are accurate through 15 digits by retaining the first 30 terms in the series.

K.2 Computation of Generalized Gradients from Coil Geometry and Current Data

Bibliography

[1] M. Bassetti and C. Biscari, "Analytic Formulae for Magnetic Multipoles", *Particle Accelerators*, **52**, 221-250 (1996).

Appendix L

Spline Routines

```
!  The following are double precision versions of the subroutines
!  "spline" and "splint" used for 1-D cubic spline interpolation, found in Numerical
!  Recipes pp. 107-110.  Instructions for use:
!
! 1) Call spline(x,y,n,yp1,ypn,y2).
!
!Here x={x_k} is an array of length n containing the x-values on which the function is given, and y is
!array of the same length containing the corresponding function values {f(x_k)}.  Also, yp1 and ypn are
!the first derivatives of the function at the points x_1 and x_n, respectively.  The routine returns an
!array y2 of length n, which contains the second derivatives of the interpolating function at the
!tabulated points x_n.
!
!The subroutine spline is called only once for a given data set, to set up the array y2.
!
! 2) For a given point x at which the interpolating function is desired, call
     splint(xa,ya,y2,n,x,y).
!
!Here xa and ya are the arrays {x_n} and {f(x_n)} as above.  The array y2 is the output from the
!subroutine "spline" above.  Again, n is the number of points in x.  Finally, the double precision numb
!x is the value at which the interpolating function is to be evaluated.  The resulting value f(x) is gi
!as the double precision number y.
!
! C. E. M.  5/27/08

      SUBROUTINE splint(xa,ya,y2a,n,x,y)
      INTEGER n
      double precision x,y,xa(n),y2a(n),ya(n)
      INTEGER k,khi,klo
      double precision a,b,h
      klo=1
      khi=n
1     if (khi-klo.gt.1) then
        k=(khi+klo)/2
        if(xa(k).gt.x)then
          khi=k
        else
          klo=k
        endif
      goto 1
      endif
      h=xa(khi)-xa(klo)
      if (h.eq.0.d0) pause 'bad xa input in splint'
      a=(xa(khi)-x)/h
      b=(x-xa(klo))/h
      y=a*ya(klo)+b*ya(khi)+((a**3-a)*y2a(klo)+(b**3-b)*y2a(khi))*(h**
     *2)/6.d0
      return
      END


      SUBROUTINE spline(x,y,n,yp1,ypn,y2)
      implicit none
      INTEGER n,NMAX
      double precision yp1,ypn,x(n),y(n),y2(n)
      PARAMETER (NMAX=10001)
      INTEGER i,k
      double precision p,qn,sig,un,u(NMAX)
c      if (yp1.gt..99e30) then
c     We set the natural bc with vanishing second derivative.
       if (yp1.gt..99e30) then
        y2(1)=0.d0
        u(1)=0.d0
      else
        y2(1)=-0.5d0
        u(1)=(3.d0/(x(2)-x(1)))*((y(2)-y(1))/(x(2)-x(1))-yp1)
c        u(1)=yp1
      endif
      do 11 i=2,n-1
        sig=(x(i)-x(i-1))/(x(i+1)-x(i-1))
        p=sig*y2(i-1)+2.d0
        y2(i)=(sig-1.d0)/p
        u(i)=(6.d0*((y(i+1)-y(i))/(x(i+
     &        1)-x(i))-(y(i)-y(i-1))/(x(i)-x(i-1)))/(x(i+1)-x(i-1))-sig*
     &        u(i-1))/p
11    continue
      if (ypn.gt..99e30) then
c      We set the natural upper bc with second derivative = 0.
        qn=0.d0
        un=0.d0
      else
        qn=0.5d0
        un=(3.d0/(x(n)-x(n-1)))*(ypn-(y(n)-y(n-1))/(x(n)-x(n-1)))
c         un=ypn
      endif
      y2(n)=(un-qn*u(n-1))/(qn*y2(n-1)+1.d0)
      do 12 k=n-1,1,-1
        y2(k)=y2(k)*y2(k+1)+u(k)
12    continue
      return
      END
```

```
PROGRAM PERSPLINE

C
C       ===============================================================
C        Periodic cubic spline interpolation.
C       ===============================================================
C
        IMPLICIT DOUBLE PRECISION (A-H,O-Z)
        DIMENSION x(801),y(801),y2(801),xa(801),ya(801)

        pi=4.d0*atan(1.d0)
        n=20
open(unit=26,file='output',status='new')
c       Define the data points to be used for interpolation.
        z=0.d0
dz=(2*pi)/(n-2)
do 10 i=1,n
   y(i)=cos(3*z)
           x(i)=z
   ya(i)=y(i)
   xa(i)=x(i)
   z=z+dz
10      continue
call pspline(x,y,n,y2)
c write(*,*) 'Values of M0, MN = ',y2(1),y2(2),y2(3),y2(4)
c write(*,*) 'x=',xa(1),xa(2),xa(3),'...',xa(n)
write(*,*) 'Spline calls completed.'
c       Compute interpolated values.
        z=0.0d0
nmax=801
dz=(2*pi)/(nmax-1)
do 20 j=1,nmax
     call splint(xa,ya,y2,n,z,C1)
   exact=cos(3*z)
           write(26,*) z,C1,exact
   z=z+dz
20      continue
        end



      SUBROUTINE pspline(x,y,n,y2)
c     Takes as input vectors x(n), y(n) defining evaluation of
c     the periodic function at its sampling points.
c     Produces output vectors x - solution to A'x = r
c                             y2 - solution to A'z = u
c     Uses the tridiagonal algorithm for LU decomposition to solve
c     both systems simultaneously.
c     Outputs the vector y2 of second derivatives y'' at sampling points.
      INTEGER n,NMAX
      REAL*8 yp1,ypn,x(n),y(n),y2(n)
      PARAMETER (NMAX=500)
      INTEGER i,k
      REAL*8 p,qn,sig,un,u1(NMAX),u2(NMAX)
      write(*,*) 'Inside spline'
c     Set boundary conditions for lower end.  Here u1 is the intermediate
c     solution for A'x=r in the LU decomposition, and u2 is the
c     intermediate solution of A'z=u in the LU decomposition.
        y2(1)=-1.0d0
        u1(1)=0.0d0
u2(1)=-1.0d0
      do 11 i=2,n-1
        sig=(x(i)-x(i-1))/(x(i+1)-x(i-1))
write(*,*) 'sig =',sig
        p=sig*y2(i-1)+2.0d0
        y2(i)=(sig-1.d0)/p
write(*,*) 'Beta, gamma = ',p,-1.d0*y2(i)
        u1(i)=(6.d0*((y(i+1)-y(i))/(x(i+
     &  1)-x(i))-(y(i)-y(i-1))/(x(i)-x(i-1)))/(x(i+1)-x(i-1))-sig*
     &   u1(i-1))/p
        u2(i)=((sig-1.d0)*u2(i-1))/p
write(*,*) 'u1,u2 = ',u1(i),u2(i)
11    continue
c     Set boundary conditions for upper end.
      x(n)=-1.d0*u1(n-1)/(y2(n-1)+1.d0)
      y2(n)=-1.d0*(1.d0+u2(n-1))/(y2(n-1)+1.d0)
      write(*,*) 'xn, y2n = ',x(n),y2(n)
      do 12 k=n-1,1,-1
        x(k)=y2(k)*x(k+1)+u1(k)
        y2(k)=y2(k)*y2(k+1)+u2(k)
write(*,*) 'x,y2 = ',x(k),y2(k)
12    continue
c     Given the two solutions x(k) and y2(k), we use the Sherman-Morrison
c     formula to construct the solution to the periodic spline system with
c     its off-diagonal terms.  Here 'fact' is the correction to the
c     intermediate solution vector x due to the off-diagonal terms.
        fact = (x(2)+x(n-1))/(1.d0+y2(2)+y2(n-1))
do 10 i=1,n
   y2(i) = x(i) - fact*y2(i)
   write(*,*) 'y2(i) = ',y2(i)
10      continue
      return
      END

      SUBROUTINE splint(xa,ya,y2a,n,x,y)
      implicit double precision(a-h,o-z)
      INTEGER n
      REAL*8 x,y,xa(n),y2a(n),ya(n)
      INTEGER k,khi,klo
      REAL*8 a,b,h
      klo=1
      khi=n
1     if (khi-klo.gt.1) then
        k=(khi+klo)/2
        if(xa(k).gt.x)then
          khi=k
        else
          klo=k
        endif
      goto 1
      endif
      h=xa(khi)-xa(klo)
      if (h.eq.0.0d0) pause 'bad xa input in splint'
      a=(xa(khi)-x)/h
      b=(x-xa(klo))/h
      y=a*ya(klo)+b*ya(khi)+((a**3-a)*y2a(klo)+
     &       (b**3-b)*y2a(khi))*(h**2)/6.d0
      return
      END
```

Appendix M

Routines for Mathieu Separation Constants $a_n(q)$ and $b_n(q)$

```
SUBROUTINE CVA2(KD,M,Q,A)
C
C       ======================================================
C       Purpose: Calculate a specific characteristic value of
C                Mathieu functions
C       Input :  m  --- Order of Mathieu functions
C                q  --- Parameter of Mathieu functions
C                KD --- Case code
C                       KD=1 for cem(x,q)  ( m = 0,2,4,...)
C                       KD=2 for cem(x,q)  ( m = 1,3,5,...)
C                       KD=3 for sem(x,q)  ( m = 1,3,5,...)
C                       KD=4 for sem(x,q)  ( m = 2,4,6,...)
C       Output:  A  --- Characteristic value
C       Routines called:
C             (1) REFINE for finding accurate characteristic
C                 values using an iteration method
C             (2) CV0 for finding initial characteristic
C                 values using polynomial approximation
C             (3) CVQM for computing initial characteristic
C                 values for q  3*m
C             (3) CVQL for computing initial characteristic
C                 values for q  m*m
C       ======================================================
C
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
IF (M.LE.12.OR.Q.LE.3.0*M.OR.Q.GT.M*M) THEN
    CALL CV0(KD,M,Q,A)
    IF (Q.NE.0.0D0) CALL REFINE(KD,M,Q,A,1)
ELSE
   NDIV=10
   DELTA=(M-3.0)*M/NDIV
   IF ((Q-3.0*M).LE.(M*M-Q)) THEN
5             NN=INT((Q-3.0*M)/DELTA)+1
      DELTA=(Q-3.0*M)/NN
      Q1=2.0*M
      CALL CVQM(M,Q1,A1)
      Q2=3.0*M
      CALL CVQM(M,Q2,A2)
      QQ=3.0*M
      DO 10 I=1,NN
 QQ=QQ+DELTA
 A=(A1*Q2-A2*Q1+(A2-A1)*QQ)/(Q2-Q1)
 IFLAG=1
 IF (I.EQ.NN) IFLAG=-1
 CALL REFINE(KD,M,QQ,A,IFLAG)
 Q1=Q2
 Q2=QQ
 A1=A2
 A2=A
10            CONTINUE
      IF (IFLAG.EQ.-10) THEN
 NDIV=NDIV*2
 DELTA=(M-3.0)*M/NDIV
 GO TO 5
      ENDIF
   ELSE
15            NN=INT((M*M-Q)/DELTA)+1
      DELTA=(M*M-Q)/NN
      Q1=M*(M-1.0)
      CALL CVQL(KD,M,Q1,A1)
      Q2=M*M
      CALL CVQL(KD,M,Q2,A2)
      QQ=M*M
      DO 20 I=1,NN
 QQ=QQ-DELTA
 A=(A1*Q2-A2*Q1+(A2-A1)*QQ)/(Q2-Q1)
 IFLAG=1
 IF (I.EQ.NN) IFLAG=-1
 CALL REFINE(KD,M,QQ,A,IFLAG)
 Q1=Q2
 Q2=QQ
 A1=A2
 A2=A
20            CONTINUE
      IF (IFLAG.EQ.-10) THEN
 NDIV=NDIV*2
 DELTA=(M-3.0)*M/NDIV
 GO TO 15
      ENDIF
   ENDIF
ENDIF
RETURN
END


SUBROUTINE REFINE(KD,M,Q,A,IFLAG)
C
C       ======================================================
C       Purpose: calculate the accurate characteristic value
```

```
C                by the secant method
C       Input :  m --- Order of Mathieu functions
C                q --- Parameter of Mathieu functions
C                A --- Initial characteristic value
C       Output:  A --- Refineed characteristic value
C       Routine called:  CVF for computing the value of F for
C                        characteristic equation
C       ========================================================
C
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
EPS=1.0D-14
MJ=10+M
CA=A
DELTA=0.0D0
X0=A
CALL CVF(KD,M,Q,X0,MJ,F0)
X1=1.002*A
CALL CVF(KD,M,Q,X1,MJ,F1)
5       DO 10 IT=1,100
   MJ=MJ+1
   X=X1-(X1-X0)/(1.0D0-F0/F1)
   CALL CVF(KD,M,Q,X,MJ,F)
   IF (ABS(1.0-X1/X).LT.EPS.OR.F.EQ.0.0) GO TO 15
   X0=X1
   F0=F1
   X1=X
10         F1=F
15      A=X
IF (DELTA.GT.0.05) THEN
   A=CA
   IF (IFLAG.LT.0) THEN
      IFLAG=-10
   ENDIF
   RETURN
ENDIF
IF (ABS((A-CA)/CA).GT.0.05) THEN
   X0=CA
   DELTA=DELTA+0.005D0
   CALL CVF(KD,M,Q,X0,MJ,F0)
   X1=(1.0D0+DELTA)*CA
   CALL CVF(KD,M,Q,X1,MJ,F1)
   GO TO 5
ENDIF
RETURN
END


SUBROUTINE CVF(KD,M,Q,A,MJ,F)
C
C       ======================================================
C       Purpose: Compute the value of F for characteristic
C                equation of Mathieu functions
C       Input :  m --- Order of Mathieu functions
C                q --- Parameter of Mathieu functions
```

```
C                A --- Characteristic value
C       Output:  F --- Value of F for characteristic equation
C       ========================================================
C
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
B=A
IC=INT(M/2)
L=0
L0=0
J0=2
JF=IC
IF (KD.EQ.1) L0=2
IF (KD.EQ.1) J0=3
IF (KD.EQ.2.OR.KD.EQ.3) L=1
IF (KD.EQ.4) JF=IC-1
T1=0.0D0
DO 10 J=MJ,IC+1,-1
10         T1=-Q*Q/((2.0D0*J+L)**2-B+T1)
IF (M.LE.2) THEN
   T2=0.0D0
   IF (KD.EQ.1.AND.M.EQ.0) T1=T1+T1
   IF (KD.EQ.1.AND.M.EQ.2) T1=-2.0*Q*Q/(4.0-B+T1)-4.0
   IF (KD.EQ.2.AND.M.EQ.1) T1=T1+Q
   IF (KD.EQ.3.AND.M.EQ.1) T1=T1-Q
ELSE
   IF (KD.EQ.1) T0=4.0D0-B+2.0D0*Q*Q/B
   IF (KD.EQ.2) T0=1.0D0-B+Q
   IF (KD.EQ.3) T0=1.0D0-B-Q
   IF (KD.EQ.4) T0=4.0D0-B
   T2=-Q*Q/T0
   DO 15 J=J0,JF
15            T2=-Q*Q/((2.0D0*J-L-L0)**2-B+T2)
ENDIF
F=(2.0D0*IC+L)**2+T1+T2-B
RETURN
END


SUBROUTINE CV0(KD,M,Q,A0)
C
C       =======================================================
C       Purpose: Compute the initial characteristic value of
C                Mathieu functions for m  12  or q  300 or
C                q  m*m
C       Input :  m  --- Order of Mathieu functions
C                q  --- Parameter of Mathieu functions
C       Output:  A0 --- Characteristic value
C       Routines called:
C             (1) CVQM for computing initial characteristic
C                 value for q  3*m
C             (2) CVQL for computing initial characteristic
C                 value for q  m*m
C       ======================================================
C
```

```
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
Q2=Q*Q
IF (M.EQ.0) THEN
   IF (Q.LE.1.0) THEN
      A0=(((.0036392*Q2-.0125868)*Q2+.0546875)*Q2-.5)*Q2
   ELSE IF (Q.LE.10.0) THEN
      A0=((3.999267D-3*Q-9.638957D-2)*Q-.88297)*Q
     &           +.5542818
   ELSE
      CALL CVQL(KD,M,Q,A0)
   ENDIF
ELSE IF (M.EQ.1) THEN
   IF (Q.LE.1.0.AND.KD.EQ.2) THEN
      A0=(((-6.51E-4*Q-.015625)*Q-.125)*Q+1.0)*Q+1.0
   ELSE IF (Q.LE.1.0.AND.KD.EQ.3) THEN
      A0=(((-6.51E-4*Q+.015625)*Q-.125)*Q-1.0)*Q+1.0
   ELSE IF (Q.LE.10.0.AND. KD.EQ.2) THEN
      A0=(((-4.94603D-4*Q+1.92917D-2)*Q-.3089229)
     &           *Q+1.33372)*Q+.811752
   ELSE IF (Q.LE.10.0.AND.KD.EQ.3) THEN
      A0=((1.971096D-3*Q-5.482465D-2)*Q-1.152218)
     &           *Q+1.10427
   ELSE
      CALL CVQL(KD,M,Q,A0)
   ENDIF
ELSE IF (M.EQ.2) THEN
   IF (Q.LE.1.0.AND.KD.EQ.1) THEN
      A0=(((-.0036391*Q2+.0125888)*Q2-.0551939)*Q2
     &           +.416667)*Q2+4.0
   ELSE IF (Q.LE.1.0.AND.KD.EQ.4) THEN
      A0=(.0003617*Q2-.0833333)*Q2+4.0
   ELSE IF (Q.LE.15.AND.KD.EQ.1) THEN
      A0=(((3.200972D-4*Q-8.667445D-3)*Q
     &           -1.829032D-4)*Q+.9919999)*Q+3.3290504
   ELSE IF (Q.LE.10.0.AND.KD.EQ.4) THEN
      A0=((2.38446D-3*Q-.08725329)*Q-4.732542D-3)
     &           *Q+4.00909
   ELSE
      CALL CVQL(KD,M,Q,A0)
   ENDIF
ELSE IF (M.EQ.3) THEN
   IF (Q.LE.1.0.AND.KD.EQ.2) THEN
      A0=((6.348E-4*Q+.015625)*Q+.0625)*Q2+9.0
   ELSE IF (Q.LE.1.0.AND.KD.EQ.3) THEN
      A0=((6.348E-4*Q-.015625)*Q+.0625)*Q2+9.0
   ELSE IF (Q.LE.20.0.AND.KD.EQ.2) THEN
      A0=(((3.035731D-4*Q-1.453021D-2)*Q
     &           +.19069602)*Q-.1039356)*Q+8.9449274
   ELSE IF (Q.LE.15.0.AND.KD.EQ.3) THEN
      A0=((9.369364D-5*Q-.03569325)*Q+.2689874)*Q
     &           +8.771735
   ELSE
      CALL CVQL(KD,M,Q,A0)
   ENDIF
```

```
ELSE IF (M.EQ.4) THEN
   IF (Q.LE.1.0.AND.KD.EQ.1) THEN
      A0=((-2.1E-6*Q2+5.012E-4)*Q2+.0333333)*Q2+16.0
   ELSE IF (Q.LE.1.0.AND.KD.EQ.4) THEN
      A0=((3.7E-6*Q2-3.669E-4)*Q2+.0333333)*Q2+16.0
   ELSE IF (Q.LE.25.0.AND.KD.EQ.1) THEN
      A0=(((1.076676D-4*Q-7.9684875D-3)*Q
     &           +.17344854)*Q-.5924058)*Q+16.620847
   ELSE IF (Q.LE.20.0.AND.KD.EQ.4) THEN
      A0=((-7.08719D-4*Q+3.8216144D-3)*Q
     &           +.1907493)*Q+15.744
   ELSE
      CALL CVQL(KD,M,Q,A0)
   ENDIF
ELSE IF (M.EQ.5) THEN
   IF (Q.LE.1.0.AND.KD.EQ.2) THEN
      A0=((6.8E-6*Q+1.42E-5)*Q2+.0208333)*Q2+25.0
   ELSE IF (Q.LE.1.0.AND.KD.EQ.3) THEN
      A0=((-6.8E-6*Q+1.42E-5)*Q2+.0208333)*Q2+25.0
   ELSE IF (Q.LE.35.0.AND.KD.EQ.2) THEN
      A0=(((2.238231D-5*Q-2.983416D-3)*Q
     &           +.10706975)*Q-.600205)*Q+25.93515
   ELSE IF (Q.LE.25.0.AND.KD.EQ.3) THEN
      A0=((-7.425364D-4*Q+2.18225D-2)*Q
     &           +4.16399D-2)*Q+24.897
   ELSE
      CALL CVQL(KD,M,Q,A0)
   ENDIF
ELSE IF (M.EQ.6) THEN
   IF (Q.LE.1.0) THEN
      A0=(.4D-6*Q2+.0142857)*Q2+36.0
   ELSE IF (Q.LE.40.0.AND.KD.EQ.1) THEN
      A0=(((-1.66846D-5*Q+4.80263D-4)*Q
     &           +2.53998D-2)*Q-.181233)*Q+36.423
   ELSE IF (Q.LE.35.0.AND.KD.EQ.4) THEN
      A0=((-4.57146D-4*Q+2.16609D-2)*Q-2.349616D-2)*Q
     &           +35.99251
   ELSE
      CALL CVQL(KD,M,Q,A0)
   ENDIF
ELSE IF (M.EQ.7) THEN
   IF (Q.LE.10.0) THEN
      CALL CVQM(M,Q,A0)
   ELSE IF (Q.LE.50.0.AND.KD.EQ.2) THEN
      A0=(((-1.411114D-5*Q+9.730514D-4)*Q
     &           -3.097887D-3)*Q+3.533597D-2)*Q+49.0547
   ELSE IF (Q.LE.40.0.AND.KD.EQ.3) THEN
      A0=((-3.043872D-4*Q+2.05511D-2)*Q
     &           -9.16292D-2)*Q+49.19035
   ELSE
      CALL CVQL(KD,M,Q,A0)
   ENDIF
ELSE IF (M.GE.8) THEN
   IF (Q.LE.3.*M) THEN
```

```
      CALL CVQM(M,Q,A0)
   ELSE IF (Q.GT.M*M) THEN
      CALL CVQL(KD,M,Q,A0)
   ELSE
      IF (M.EQ.8.AND.KD.EQ.1) THEN
 A0=(((8.634308D-6*Q-2.100289D-3)*Q+.169072)*Q
     &              -4.64336)*Q+109.4211
      ELSE IF (M.EQ.8.AND.KD.EQ.4) THEN
 A0=((-6.7842D-5*Q+2.2057D-3)*Q+.48296)*Q+56.59
      ELSE IF (M.EQ.9.AND.KD.EQ.2) THEN
 A0=(((2.906435D-6*Q-1.019893D-3)*Q+.1101965)*Q
     &              -3.821851)*Q+127.6098
      ELSE IF (M.EQ.9.AND.KD.EQ.3) THEN
 A0=((-9.577289D-5*Q+.01043839)*Q+.06588934)*Q
     &              +78.0198
      ELSE IF (M.EQ.10.AND.KD.EQ.1) THEN
 A0=(((5.44927D-7*Q-3.926119D-4)*Q+.0612099)*Q
     &              -2.600805)*Q+138.1923
      ELSE IF (M.EQ.10.AND.KD.EQ.4) THEN
 A0=((-7.660143D-5*Q+.01132506)*Q-.09746023)*Q
     &              +99.29494
      ELSE IF (M.EQ.11.AND.KD.EQ.2) THEN
 A0=(((-5.67615D-7*Q+7.152722D-6)*Q+.01920291)*Q
     &              -1.081583)*Q+140.88
      ELSE IF (M.EQ.11.AND.KD.EQ.3) THEN
 A0=((-6.310551D-5*Q+.0119247)*Q-.2681195)*Q
     &              +123.667
      ELSE IF (M.EQ.12.AND.KD.EQ.1) THEN
 A0=(((-2.38351D-7*Q-2.90139D-5)*Q+.02023088)*Q
     &              -1.289)*Q+171.2723
      ELSE IF (M.EQ.12.AND.KD.EQ.4) THEN
 A0=(((3.08902D-7*Q-1.577869D-4)*Q+.0247911)*Q
     &              -1.05454)*Q+161.471
      ENDIF
   ENDIF
ENDIF
RETURN
END


SUBROUTINE CVQL(KD,M,Q,A0)
C
C       ========================================================
C       Purpose: Compute the characteristic value of Mathieu
C                functions  for q  3m
C       Input :  m  --- Order of Mathieu functions
C                q  --- Parameter of Mathieu functions
C       Output:  A0 --- Initial characteristic value
C       ========================================================
C
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
IF (KD.EQ.1.OR.KD.EQ.2) W=2.0D0*M+1.0D0
IF (KD.EQ.3.OR.KD.EQ.4) W=2.0D0*M-1.0D0
W2=W*W
```

```
W3=W*W2
W4=W2*W2
W6=W2*W4
D1=5.0+34.0/W2+9.0/W4
D2=(33.0+410.0/W2+405.0/W4)/W
D3=(63.0+1260.0/W2+2943.0/W4+486.0/W6)/W2
D4=(527.0+15617.0/W2+69001.0/W4+41607.0/W6)/W3
C1=128.0
P2=Q/W4
P1=DSQRT(P2)
CV1=-2.0*Q+2.0*W*DSQRT(Q)-(W2+1.0)/8.0
CV2=(W+3.0/W)+D1/(32.0*P1)+D2/(8.0*C1*P2)
CV2=CV2+D3/(64.0*C1*P1*P2)+D4/(16.0*C1*C1*P2*P2)
A0=CV1-CV2/(C1*P1)
RETURN
END


SUBROUTINE CVQM(M,Q,A0)
C
C       ======================================================
C       Purpose: Compute the characteristic value of Mathieu
C                functions for q  m*m
C       Input :  m  --- Order of Mathieu functions
C                q  --- Parameter of Mathieu functions
C       Output:  A0 --- Initial characteristic value
C       ======================================================
C
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
HM1=.5*Q/(M*M-1.0)
HM3=.25*HM1**3/(M*M-4.0)
HM5=HM1*HM3*Q/((M*M-1.0)*(M*M-9.0))
A0=M*M+Q*(HM1+(5.0*M*M+7.0)*HM3
     &     +(9.0*M**4+58.0*M*M+29.0)*HM5)
RETURN
END

FUNCTION fac(n)
        IMPLICIT DOUBLE PRECISION (A-H,O-Z)
INTEGER n,J
IF (n.EQ.0) fac = 1.d0
        f=1.d0
DO 11 J=1,n
   f = J*f
11      CONTINUE
        fac = f
RETURN
END
```

Appendix N

Mathieu-Bessel Connection Coefficients

As a consequence of the symmetry properties (13.9.40) through (13.9.43), the Fourier expansions (13.10.144) and (13.10.150) for the $\mathrm{ce}_n(v,q)$ and $\mathrm{se}_n(v,q)$ can be written in the form

$$\mathrm{ce}_{2n}(v,q) = \sum_{m=0}^{\infty} A_{2m}^{2n}(q)\cos(2mv), \tag{N.0.1}$$

$$\mathrm{ce}_{2n+1}(v,q) = \sum_{m=0}^{\infty} A_{2m+1}^{2n+1}(q)\cos[(2m+1)v], \tag{N.0.2}$$

$$\mathrm{se}_{2n+1}(v,q) = \sum_{m=0}^{\infty} B_{2m+1}^{2n+1}(q)\sin[(2m+1)v], \tag{N.0.3}$$

$$\mathrm{se}_{2n+2}(v,q) = \sum_{m=0}^{\infty} B_{2m+2}^{2n+2}(q)\sin[(2m+2)v]. \tag{N.0.4}$$

In all these relations $n = 0, 1, 2, 3, \cdots$. Put another way, the symmetry properties require that the coefficients A_m^n and B_m^n vanish unless both m and n are even or both m and n are odd.

In this appendix we will see that the same symmetry properties hold for the Mathieu-Bessel connection coefficients α_m^n and β_m^n. That is, formulas (13.9.64) and (13.9.65) can be written in the corresponding form

$$\mathrm{Ce}_{2n}(u,q)\ \mathrm{ce}_{2n}(v,q) = \sum_{m=0}^{\infty} \alpha_{2m}^{2n}(k) I_{2m}(k\rho)\cos(2m\phi), \tag{N.0.5}$$

$$\mathrm{Ce}_{2n+1}(u,q)\ \mathrm{ce}_{2n+1}(v,q) = \sum_{m=0}^{\infty} \alpha_{2m+1}^{2n+1}(k) I_{2m+1}(k\rho)\cos[(2m+1)\phi], \tag{N.0.6}$$

$$\mathrm{Se}_{2n+1}(u,q)\ \mathrm{se}_{2n+}(v,q) = \sum_{m=0}^{\infty} \beta_{2m+1}^{2n+1}(k) I_{2m+1}(k\rho)\sin[(2m+1)\phi], \tag{N.0.7}$$

$$\mathrm{Se}_{2n+2}(u,q)\,\mathrm{se}_{2n+2}(v,q)=\sum_{m=0}^{\infty}\beta_{2m+2}^{2n+2}(k)I_{2m+2}(k\rho)\sin[(2m+2)\phi]. \tag{N.0.8}$$

Moreover, there are the relations

$$\alpha_{2m}^{2n}(k)=g_c^{2n}(k)A_{2m}^{2n}(q), \tag{N.0.9}$$

$$\alpha_{2m+1}^{2n+1}(k)=g_c^{2n+1}(k)A_{2m+1}^{2n+1}(q), \tag{N.0.10}$$

$$\beta_{2m+1}^{2n+1}(k)=g_s^{2n+1}(k)B_{2m+1}^{2n+1}(q), \tag{N.0.11}$$

$$\beta_{2m+2}^{2n+2}(k)=g_s^{2n+2}(k)B_{2m+2}^{2n+2}(q), \tag{N.0.12}$$

where

$$g_c^{2n}(k)=[\mathrm{ce}_{2n}(\pi/2,q)\,\mathrm{ce}_{2n}(0,q)]/A_0^{2n}(q), \tag{N.0.13}$$

$$g_c^{2n+1}(k)=-2[\mathrm{ce}'_{2n+1}(\pi/2,q)\,\mathrm{ce}_{2n+1}(0,q)]/[kfA_1^{2n+1}(q)], \tag{N.0.14}$$

$$g_s^{2n+1}(k)=2[\mathrm{se}_{2n+1}(\pi/2,q)\,\mathrm{se}'_{2n+1}(0,q)]/[kfB_1^{2n+1}(q)], \tag{N.0.15}$$

$$g_s^{2n+2}(k)=[\mathrm{se}'_{2n+2}(\pi/2,q)\,\mathrm{se}'_{2n+2}(0,q)]/[qB_2^{2n+2}(q)]. \tag{N.0.16}$$

Here a $'$ denotes d/dv.

Appendix O

Quadratic Forms

O.1 Background

Let L be a real $m \times m$ matrix, let w be a real m-component vector, and let $(*,*)$ denote the usual real inner product. Define a quadratic form $Q(w)$ by the rule

$$Q(w) = (w, Lw). \tag{O.1.1}$$

The matrix L can be uniquely decomposed into symmetric and antisymmetric parts S and A by writing

$$L = S + A \tag{O.1.2}$$

with

$$S = (1/2)(L + L^T) \tag{O.1.3}$$

and

$$A = (1/2)(L - L^T). \tag{O.1.4}$$

Then, since only the symmetric part of L contributes to Q, we may equally well write

$$Q(w) = (w, Sw). \tag{O.1.5}$$

According to standard matrix theory, any real $m \times m$ symmetric matrix S has m real eigenvalues and m associated real eigenvectors that can arranged to form an orthonormal basis. (Note that no assumption needs to be made about the eigenvalues being distinct.) Call the eigenvalues σ_j and the associated orthonormal eigenvectors v_j. Then we may write S in the dyadic form

$$S = \sum_{j=1}^{m} \sigma_j |v_j)(v_j|. \tag{O.1.6}$$

With the aid of this representation for S we find that Q takes the form

$$Q(w) = (w, Sw) = \sum_{j=1}^{m} \sigma_j (w, v_j)(v_j, w) = \sum_{j=1}^{m} \sigma_j (w, v_j)^2. \tag{O.1.7}$$

We see that Q will be positive definite if all $\sigma_j > 0$. Conversely, since the v_j are orthonormal, it is evident that all the σ_j will be positive if Q is positive definite. Similarly, Q will be negative definite if all $\sigma_j < 0$, and conversely. Finally, Q will be indefinite if not all σ_j have the same sign or some are zero.

O.2 Effect of Small Perturbations in the Definite Case

Now suppose, for example, that Q is positive definite and that all the eigenvalues σ_j of S are substantially different from 0. Next suppose that L is slightly perturbed so that S is also slightly perturbed. Thus, we may write that

$$S = S^0 + S^1 \tag{O.2.1}$$

where S^0 is the initial S before perturbation and S^1 is a small symmetric matrix that describes the perturbation. Now the quadratic form Q becomes Q' with

$$Q'(w) = (w, [S^0 + S^1]w) = (w, S^0 w) + (w, S^1 w) = Q(w) + (w, S^1 w). \tag{O.2.2}$$

It is easy to see from (1.7) that

$$Q(w) \geq \sigma_{\min} \sum_{j=1}^{m} (w, v_j)^2 = \sigma_{\min}(w, w) = \sigma_{\min}||w||^2 \tag{O.2.3}$$

where $\sigma_{\min}$ is the smallest eigenvalue of S. Also, we have the estimate

$$|(w, S^1 w)| \leq ||w|| \, ||S^1 w|| \leq ||w|| \, ||S^1|| \, ||w|| = ||S^1|| \, ||w||^2. \tag{O.2.4}$$

It follows that Q' will also be positive definite providing

$$||S^1|| < \sigma_{\min}. \tag{O.2.5}$$

We conclude that if Q is positive definite, it will remain positive definite under small perturbations of L. Similarly, if Q is negative definite, it will remain negative definite under small perturbations of L.

Even more can be said. The *rank* of Q is defined to be the number of nonzero eigenvectors of S, and the *signature* is defined to be the number of positive eigenvalues minus the number of negative eigenvalues. It can be shown that if under a continuous change in S the rank does not change (no eigenvalue passes through the value 0), then the signature also does not change. This result follows from the fact that the eigenvalues of S are continuous functions of the matrix elements of S.

Bibliography

[1] F. Gantmacher, *The Theory of Matrices*, Vols. I and II, Chelsea (1959). See page 309 in Vol. I, which discusses quadratic forms.

Appendix P

Parameterization of the Coset Space $GL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$

P.1 Introduction

Suppose that $M \in GL(2n, \mathbb{R})$ has a symplectic polar decomposition,

$$M = QR \tag{P.1.1}$$

where Q is J-symmetric and R is symplectic.[1] We know that such a decomposition is possible for M sufficiently near the symplectic group and is unique. We know that the ordinary (orthogonal) polar decomposition can be made globally. By contrast, from counter examples, we know that symplectic polar decomposition is not possible globally. We also see that, by construction, J-symmetric matrices Q are related to the cosets $GL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$. We want to find what restrictions must be imposed on M for a symplectic polar decomposition to be possible, and suspect that these restrictions are related to coset structure.

P.2 M Must Have Positive Determinant

From (1.1) we find

$$\det M = (\det Q)(\det R) = \det Q. \tag{P.2.1}$$

According to Lemma 3.6 of Section 4.3 of *Lie Methods*, any J-symmetric matrix Q can be written in the form

$$Q = JA \tag{P.2.2}$$

where A is real and antisymmetric. It follows that

$$\det Q = (\det J)(\det A) = \det A \geq 0. \tag{P.2.3}$$

Here we have used the fact that a real antisymmetric matrix cannot have a negative determinant. It follows from (1.1), (2.1), and (2.3) that, if M is to be nonsingular and have a symplectic polar decomposition, it must have positive determinant,

$$\det M > 0. \tag{P.2.4}$$

[1] We adopt the terminology of Chapter 4 of *Lie Methods* $\cdots$.

Then, from (2.1), we also have

$$\det Q > 0. \tag{P.2.5}$$

P.3 It is Sufficient to Consider $SL(2n,\mathbb{R})/Sp(2n,\mathbb{R})$

Suppose N is any matrix in $GL(2n,\mathbb{R})$ and $\det N > 0$. Define an associated matrix M by the rule

$$M = (\det N)^{-1/(2n)} N. \tag{P.3.1}$$

By construction, M will be in $SL(2n,\mathbb{R})$.

Next assume that M has the symplectic polar decomposition (1.1). Then (1.1) and (3.1) imply that

$$N = (\det N)^{1/(2n)} M = (\det N)^{1/(2n)} QR = Q'R \tag{P.3.2}$$

where

$$Q' = (\det N)^{1/(2n)} Q. \tag{P.3.3}$$

By the lemmas of Section 4.3 of *Lie Methods*, Q' will also be J-symmetric, and therefore N has a symplectic polar decomposition. Thus, it is sufficient to study whether any $M \in SL(2n,\mathbb{R})$ has a symplectic polar decomposition.

P.4 Some Symmetries

Consider the map Σ of $SL(2n,\mathbb{R})$ into itself defined by the rule

$$\Sigma(M) = J(M^T)^{-1}J^{-1}. \tag{P.4.1}$$

We will now explore the properties of Σ.

Suppose M_1 and M_2 are any two $SL(2n,\mathbb{R})$ matrices. Then we find the relation

$$\begin{aligned}\Sigma(M_1M_2) &= J[(M_1M_2)^T]^{-1}J^{-1} = J[M_2^TM_1^T]^{-1}J^{-1} \\ &= J(M_1^T)^{-1}(M_2^T)^{-1}J^{-1} = J(M_1^T)^{-1}J^{-1}J(M_2^T)^{-1}J^{-1} \\ &= \Sigma(M_1)\Sigma(M_2).\end{aligned} \tag{P.4.2}$$

Thus, Σ is a homomorphism.

Next we observe that

$$\Sigma(I) = I \tag{P.4.3}$$

and

$$\Sigma(J) = J(J^T)^{-1}J^{-1} = J(-J)^{-1}J^{-1} = JJJ^{-1} = J. \tag{P.4.4}$$

Similarly, we find

$$\Sigma(J^{-1}) = J^{-1}. \tag{P.4.5}$$

Also, there is the property

$$\Sigma(M^{-1}) = J[(M^{-1})^T]^{-1}J^{-1} = JM^TJ^{-1} = [J(M^T)^{-1}J^{-1}]^{-1} = [\Sigma(M)]^{-1}. \tag{P.4.6}$$

[This result also follows from (4.2) and (4.3).] Consequently, Σ is an isomorphism.

We claim that Σ acts as the identity map on $Sp(2n, \mathbb{R})$. That is, all elements $R \in Sp(2n, \mathbb{R})$ are fixed points of Σ. Indeed, suppose that $R \in Sp(2n, \mathbb{R})$. Then we have the result

$$RJR^T = J, \tag{P.4.7}$$

which is equivalent to the relation

$$J^{-1} = (RJR^T)^{-1} = (R^T)^{-1}J^{-1}R^{-1}, \tag{P.4.8}$$

which in turn is equivalent to the relation

$$R = J(R^T)^{-1}J^{-1} = \Sigma(R). \tag{P.4.9}$$

Note that (4.3) through (4.5) are special cases of (4.9).

Let us next find the action of Σ on any J-symmetric matrix Q. We find the result

$$\Sigma(Q) = J(Q^T)^{-1}J^{-1} = (JQ^TJ^{-1})^{-1} = Q^{-1}. \tag{P.4.10}$$

Note that (2.5) guarantees that Q^{-1} exists.

Upon combining (4.9) and (4.10) we find that the effect of Σ on any matrix M having the factorization (1.1) is given by the relation

$$\Sigma(M) = \Sigma(QR) = \Sigma(Q)\Sigma(R) = Q^{-1}R. \tag{P.4.11}$$

Finally, Σ is an involution. By calculating we find that

$$\begin{aligned} \Sigma^2(M) &= \Sigma[\Sigma(M)] = \Sigma[J(M^T)^{-1}J^{-1}] = \Sigma(J)\Sigma[(M^T)^{-1}]\Sigma(J^{-1}) \\ &= J\Sigma[(M^T)^{-1}]J^{-1} = JJ\{[(M^T)^{-1}]^T\}^{-1}J^{-1}J^{-1} \\ &= JJMJ^{-1}J^{-1} = (-I)M(-I) = M. \end{aligned} \tag{P.4.12}$$

We have found a symmetry for $SL(2n, \mathbb{R})$. We will now see that Σ produces an associated symmetry σ on the Lie algebra $s\ell(2n, \mathbb{R})$. Let B be any element in the Lie algebra $s\ell(2n, \mathbb{R})$. Let σ be the associated induced map in the Lie algebra defined by the relation

$$\Sigma[\exp(B)] = \exp[\sigma(B)]. \tag{P.4.13}$$

By calculation we find

$$\begin{aligned} \Sigma[\exp(B)] &= J\{[\exp(B)]^T\}^{-1}J^{-1} = J\{\exp(B^T)\}^{-1}J^{-1} \\ &= J\exp(-B^T)J^{-1} = \exp(-JB^TJ^{-1}). \end{aligned} \tag{P.4.14}$$

Upon comparing (4.13) and (4.14) in the vicinity of the identity, we conclude that

$$\sigma(B) = -JB^TJ^{-1}. \tag{P.4.15}$$

Let us explore the properties of σ. Any element $B \in s\ell(2n, \mathbb{R})$ can be written uniquely in the form

$$B = JS + JA \tag{P.4.16}$$

where S is symmetric, A is antisymmetric, and JA is traceless. The elements JS are a basis for $sp(2n,\mathbb{R})$, and together with the elements of the form JA form a basis for $s\ell(2n,\mathbb{R})$. Computation gives the results

$$\sigma(JS) = -J(JS)^T J^{-1} = -JS^T J^T J^{-1} = -JS(-J)J^{-1} = JS, \tag{P.4.17}$$

$$\sigma(JA) = -J(JA)^T J^{-1} = -JA^T J^T J^{-1} = JA(-J)J^{-1} = -JA. \tag{P.4.18}$$

We see from (4.17) that $sp(2n,\mathbb{R})$ is invariant under σ. That is, σ acts as the identity on $sp(2n,\mathbb{R})$. This is the local consequence of the global result that Σ acts as the identity on $Sp(2n,\mathbb{R})$. And, since σ is manifestly linear, we have

$$\sigma(B) = \sigma(JS + JA) = \sigma(JS) + \sigma(JA) = JS - JA. \tag{P.4.19}$$

From (4.16) through (4.19) we find that

$$\sigma^2(B) = \sigma[\sigma(B)] = B. \tag{P.4.20}$$

Thus, σ is an involution on $s\ell(2n,\mathbb{R})$.

P.5 Connection between Symmetries and Being J-Symmetric

A matrix Q is called J-symmetric if it satisfies the condition

$$JQ^T J^{-1} = Q, \tag{P.5.1}$$

which is equivalent to the condition

$$\sigma(Q) = -Q. \tag{P.5.2}$$

Suppose we represent Q in the form

$$Q = JX \tag{P.5.3}$$

where the properties of X are yet to be determined. Then we find

$$\sigma(Q) = \sigma(JX) = -J(JX)^T J^{-1} = -JX^T J^T J^{-1} = JX^T. \tag{P.5.4}$$

Thus, in this representation, the J-symmetric condition (5.2) yields the requirement

$$JX^T = -JX, \tag{P.5.5}$$

from which it follows that

$$X^T = -X. \tag{P.5.6}$$

That is,

$$Q = JA' \tag{P.5.7}$$

where A' is antisymmetric.

Suppose we instead represent Q in the form

$$Q = \exp(JX). \tag{P.5.8}$$

Then we find the relation

$$\begin{aligned}\sigma(Q) &= \sigma[\exp(JX)] = -J[\exp(JX)]^T J^{-1} = -J\exp[(JX)^T]J^{-1} \\ &= -J\exp(-X^T J)J^{-1} = -\exp(-JX^T).\end{aligned} \tag{P.5.9}$$

In this context requiring (5.2) produces the relation

$$\exp(JX) = \exp(-JX^T), \tag{P.5.10}$$

which again produces (5.5) and hence (5.6) and consequently

$$Q = \exp(JA) \tag{P.5.11}$$

where A is antisymmetric.

As a side comment, we have discovered, upon comparing (5.7) and (5.11), the relation

$$JA' = \exp(JA), \tag{P.5.12}$$

or,

$$A' = -J\exp(JA), \tag{P.5.13}$$

which maps antisymmetric matrices A into antisymmetric matrices A'. Keeping the first few terms in the power series we find that

$$\begin{aligned}A' &= -J[I + JA + (JA)^2/2! + (JA)^3/3! + \cdots] \\ &= -J + A - J(JA)^2/2! - J(JA)^3/3! + \cdots \\ &= -J + A + AJA/2! + AJAJA/3! + \cdots .\end{aligned} \tag{P.5.14}$$

In this form we see that *any* Taylor series in JA would have the same mapping property.

Finally, let us apply Σ to Q as given by (5.11). We find the result

$$\Sigma(Q) = \Sigma[\exp(JA)] = \exp[\sigma(JA)] = \exp(-JA) = Q^{-1} \tag{P.5.15}$$

as before.

P.6 Relation to Darboux Matrices

According to *Lie Methods* the matrix $N(M)$ is J-symmetric, and we seek a J-symmetric matrix Q such that

$$Q^2 = N(M). \tag{P.6.1}$$

Use the result of Lemma 3.6 of *Lie Methods* to write the representations

$$N(M) = JA' \tag{P.6.2}$$

and

$$Q = JA \tag{P.6.3}$$

where A' and A are antisymmetric. With these representations (6.1) becomes

$$JAJA = JA', \tag{P.6.4}$$

which yields the relation

$$A' = AJA. \tag{P.6.5}$$

Since A is assumed to be antisymmetric, (6.5) can also be written in the form

$$-A' = AJA^T, \tag{P.6.6}$$

which shows that A, if it exists, is a Darboux matrix connecting $-A'$ and J. Thus, the problem of finding Q is equivalent to showing that it is possible to find a Darboux matrix connecting $-A'$ and J that is also antisymmetric.

P.7 Some Observations on $SL(2n,\mathbb{R})/Sp(2n,\mathbb{R})$

By its nature, $SL(2n,\mathbb{R})/Sp(2n,\mathbb{R})$ is a homogeneous space. That is, $SL(2n,\mathbb{R})$ when acting on this space can send any point into any other point. See Section 5.12 for a description of group action on cosets. Between the JS and the JA there are the relations (4.3.2) through (4.3.4). Consequently $SL(2n,\mathbb{R})/Sp(2n,\mathbb{R})$ is a reductive homogeneous space. Since σ is an involution on $s\ell(2n,\mathbb{R})$ that leaves $sp(2n,\mathbb{R})$ invariant, it follows that $SL(2n,\mathbb{R})/Sp(2n,\mathbb{R})$ is a symmetric space. Moreover, at least for the cases $n = 2$ and $n = 3$ and presumably for all n, the elements of the form JA with JA traceless transform irreducibly under the action of $sp(2n,\mathbb{R})$. Therefore $SL(2n,\mathbb{R})/Sp(2n,\mathbb{R})$ is an irreducible symmetric space. For example, in the case $n = 2$ the elements of the form JA with JA traceless carry the irreducible representation $\Gamma(0,1)$ of $sp(4,\mathbb{R})$; and in the case $n = 3$ they carry the irreducible representation $\Gamma(0,1,0)$ of $sp(6,\mathbb{R})$. See the weight diagrams 27.5.4 and 27.8.4 in Chapter 27.

Digesting Goodman Notes

P.8 Action of σ on $s\ell(2n,\mathbb{R})$

Consider the action of σ on $s\ell(2n,\mathbb{R})$. First we see that, for a real symmetric matrix,

$$\sigma(S) = -JS^TJ^{-1} = -JSJ^{-1} = JSJ \tag{P.8.1}$$

so that

$$[\sigma(S)]^T = [JSJ]^T = J^TS^TJ^T = JSJ = \sigma(S). \tag{P.8.2}$$

Therefore, σ maps the space of real symmetric matrices into itself.

Next suppose B and B' are two matrices in $s\ell(2n,\mathbb{R})$. Define their inner product to be

$$(B,B') = \mathrm{tr}(BB'). \tag{P.8.3}$$

Then we find that σ preserves the inner product,

$$\begin{aligned}(\sigma(B),\sigma(B')) &= \mathrm{tr}[(-J)B^TJ^{-1}(-J)(B')^TJ^{-1}] = \mathrm{tr}[JB^T(B')^TJ^{-1}] = \mathrm{tr}[B^T(B')^T] \\ &= \mathrm{tr}[(B'B)^T] = \mathrm{tr}[B'B] = \mathrm{tr}[BB'] = (B,B'). \end{aligned} \tag{P.8.4}$$

Suppose we write for any real symmetric matrix S the decomposition

$$S = S^a + S^c \tag{P.8.5}$$

where S^a anticommutes with J and S^c commutes with J. See Section 3.8 of *Lie Methods*. Then we find that

$$\sigma(S^a) = -JS^aJ^{-1} = S^aJJ^{-1} = S^a, \tag{P.8.6}$$

and

$$\sigma(S^c) = -JS^cJ^{-1} = -S^cJJ^{-1} = -S^c. \tag{P.8.7}$$

We see that σ, which we already know is a linear operator that maps the space of real symmetric matrices into itself, has eigenvalues ±1. This is to be expected because σ is an involution.

As an application of this result, we find that

$$(S^a,S^c) = -(\sigma(S^a),\sigma(S^c)) = -(S^a,S^c) \tag{P.8.8}$$

from which it follows that

$$(S^a,S^c) = 0 \tag{P.8.9}$$

for any S^a, S^c pair.

P.9 Lie Triple System

Let S, S', and S'' be real symmetric matrices. Then we have

$$\{S',S\} = A \tag{P.9.1}$$

where A is an antisymmetric matrix. And,

$$\{S'',\{S',S\}\} = S''' \tag{P.9.2}$$

where S''' is again a symmetric matrix. Thus symmetric matrices comprise a Lie triple system.

Let S^a, $S^{a\prime}$, and $S^{a\prime\prime}$ be real symmetric matrices that anticommute with J. Then we have

$$\{S^{a\prime},S^a\} = A^c \tag{P.9.3}$$

where A^c is an antisymmetric matrix that commutes with J. And,

$$\{S^{a''}, \{S^{a'}, S^a\}\} = S^{a'''} \tag{P.9.4}$$

where $S^{a'''}$ is again a symmetric matrix that anticommutes with J. Thus symmetric matrices that anticommute with J also comprise a Lie triple system.

Similarly, it can be verified that real symmetric matrices of the form S^c comprise a Lie triple system,

$$\{S^{c''}, \{S^{c'}, S^c\}\} = S^{c'''} \tag{P.9.5}$$

where $S^{c'''}$ is again a symmetric matrix that commutes with J.

P.10 A Factorization Theorem (Theorem 1.1 of Goodman)

P.10.1 A Particular Mapping from Real Symmetric Matrices to Positive-Definite Matrices

Define for any S an associated matrix $P(S)$ by the rule

$$P(S) = \exp(S^a)\exp(S^c)\exp(S^a). \tag{P.10.1}$$

Evidently, P is real, symmetric and nonsingular. It is also positive definite because we have

$$\begin{aligned}(v, Pv) &= (v, \exp(S^a)\exp(S^c)\exp(S^a)v) = (\exp(S^a)v, \exp(S^c)\exp(S^a)v)\\ &= ([\exp(S^a)v], \exp(S^c)[\exp(S^a)v]) > 0 \text{ if } v \neq 0\end{aligned} \tag{P.10.2}$$

because $\exp(S^a)$ is symmetric and invertible and $\exp(S^c)$ is positive definite.

We will eventually see that the map (10.1) is invertible. That is, given any real symmetric positive-definite matrix P, there are (unique) matrices S^a and S^c such that (10.1) holds. Thus, (10.1) provides a factorization of any real symmetric positive-definite matrix P.

P.10.2 The Map Is Real Analytic

Evidently, by the nature of the exponential function, $P(S)$ is a real analytic function of S^a and S^c. We claim that $P(S)$ is also an analytic function of S. Note that

$$S^a = (S - J^{-1}SJ)/2 \tag{P.10.3}$$

and

$$S^c = (S + J^{-1}SJ)/2 \tag{P.10.4}$$

Therefore S^a and S^c are analytic functions of S. Since the various exponential functions appearing in (10.1) are analytic functions of their arguments, it follows that $P(S)$ is an analytic function of S.

P.10.3 Trace and Determinant Properties

From (10.3) we find

$$\begin{aligned} \mathrm{tr}(S^a) &= (1/2)[\mathrm{tr}(S) - \mathrm{tr}(J^{-1}SJ)] = (1/2)[\mathrm{tr}(S) - \mathrm{tr}(JJ^{-1}S)] \\ &= (1/2)[\mathrm{tr}(S) - \mathrm{tr}(S)] = 0. \end{aligned} \tag{P.10.5}$$

From (10.4) we find

$$\begin{aligned} \mathrm{tr}(S^c) &= (1/2)[\mathrm{tr}(S) + \mathrm{tr}(J^{-1}SJ)] = (1/2)[\mathrm{tr}(S) + \mathrm{tr}(JJ^{-1}S)] \\ &= (1/2)[\mathrm{tr}(S) + \mathrm{tr}(S)] = \mathrm{tr}(S). \end{aligned} \tag{P.10.6}$$

Now take the determinant of both sides of (10.1). So doing gives the result

$$\det[P(S)] = \exp[\mathrm{tr}(S^a)]\exp[\mathrm{tr}(S^c)]\exp[\mathrm{tr}(S^a)] = \exp[\mathrm{tr}(S)]. \tag{P.10.7}$$

P.10.4 Study of the Inverse of the Map

Conversely, it is claimed that S is an analytic function of P. If true, then, by (10.3) and (10.4), S^a and S^c are also analytic functions of P. Proceed as follows: Since P is real, symmetric, and positive definite, there is a real symmetric matrix Z such that

$$P(S) = \exp(Z) = \exp(S^a)\exp(S^c)\exp(S^a), \tag{P.10.8}$$

and Z will be analytic in P and therefore in S. What we want to do is find S^a and S^c in terms of Z.

P.10.5 Formula for S^a in terms of Z

Begin by finding S^a in terms of Z. Apply Σ to both sides of (10.8) to find the result

$$\Sigma[P(S)] = \Sigma[\exp(Z)] = \Sigma[\exp(S^a)]\Sigma[\exp(S^c)]\Sigma[\exp(S^a)], \tag{P.10.9}$$

from which it follows that

$$\exp[\sigma(Z)] = \exp[\sigma(S^a)]\exp[\sigma(S^c)]\exp[\sigma(S^a)], \tag{P.10.10}$$

from which it follows that

$$\exp[\sigma(Z)] = \exp(S^a)\exp(-S^c)\exp(S^a). \tag{P.10.11}$$

Next take inverses of both sides of (10.8) to find the relation

$$\exp(-Z) = \exp(-S^a)\exp(-S^c)\exp(-S^a), \tag{P.10.12}$$

from which it follows that

$$\exp(-S^c) = \exp(S^a)\exp(-Z)\exp(S^a). \tag{P.10.13}$$

Use (10.13) in (10.11) to get the relation

$$\exp[\sigma(Z)] = \exp(2S^a)\exp(-Z)\exp(2S^a), \tag{P.10.14}$$

which is a relation between Z and S^a.

Next show that, given Z, (10.14) has, in fact, a unique solution S^a. In particular, we want to verify the assertion

$$\exp(2S^a) = \exp(Z/2)\exp(T)\exp(Z/2) \tag{P.10.15}$$

where

$$\exp(2T) = \exp(-Z/2)\exp[\sigma(Z)]\exp(-Z/2). \tag{P.10.16}$$

Note that the right side of (10.16) is real, symmetric, and positive definite. Therefore T and consequently $\exp(T)$ are well defined (real analytic) functions of Z. Correspondingly, the right side of (10.15) is well defined, real, symmetric, and positive definite. Therefore S^a is well defined, and a real analytic function of Z. We also observe that (10.16) can be rewritten in the form

$$\exp[\sigma(Z)] = \exp(Z/2)\exp(2T)\exp(Z/2). \tag{P.10.17}$$

To prove the assertion, take (10.15) and (10.16) to be the definition of S^a. Then, using (10.15), we find that

$$\begin{aligned}&\exp(2S^a)\exp(-Z)\exp(2S^a) =\\ &\exp(Z/2)\exp(T)\exp(Z/2)\exp(-Z)\exp(Z/2)\exp(T)\exp(Z/2) =\\ &\exp(Z/2)\exp(2T)\exp(Z/2) = \exp[\sigma(Z)].\end{aligned} \tag{P.10.18}$$

Here, in the last step, we have also used (10.17). Thus, we see that (10.14) is satisfied.

P.10.6 Uniqueness of Solution for S^a

What about uniqueness? Suppose $\hat{S}^a$ also satisfies (10.14). That is, assume

$$\exp[\sigma(Z)] = \exp(2\hat{S}^a)\exp(-Z)\exp(2\hat{S}^a). \tag{P.10.19}$$

Substitute (10.19) into (10.16) to get

$$\begin{aligned}\exp(2T) &= \exp(-Z/2)\exp[\sigma(Z)]\exp(-Z/2)\\ &= \exp(-Z/2)\{\exp(2\hat{S}^a)\exp(-Z)\exp(2\hat{S}^a)\}\exp(-Z/2)\\ &= \exp(-Z/2)\{\exp(2\hat{S}^a)\exp(-Z/2)\exp(-Z/2)\exp(2\hat{S}^a)\}\exp(-Z/2)\\ &= [\exp(-Z/2)\exp(2\hat{S}^a)\exp(-Z/2)]^2.\end{aligned} \tag{P.10.20}$$

Therefore, by the uniqueness of the positive-definite square root of a positive-definite matrix, we have

$$\exp(T) = \exp(-Z/2)\exp(2\hat{S}^a)\exp(-Z/2), \tag{P.10.21}$$

which can be rewritten in the form

$$\exp(2\hat{S}^a) = \exp(Z/2)\exp(T)\exp(Z/2). \tag{P.10.22}$$

Now compare (10.15) and (10.22) to get

$$\exp(2S^a) = \exp(2\hat{S}^a), \tag{P.10.23}$$

from which it follows that

$$S^a = \hat{S}^a. \tag{P.10.24}$$

P.10.7 Verification of Expected Symmetry for S^a

Also, does the S^a just found satisfy (8.6)? Apply Σ to both sides of (10.14) to get the relation

$$\exp\{\sigma[\sigma(Z)]\} = \exp[2\sigma(S^a)]\exp[-\sigma(Z)]\exp[2\sigma(S^a)], \tag{P.10.25}$$

form which it follows by (4.20) that

$$\exp(Z) = \exp[2\sigma(S^a)]\exp[-\sigma(Z)]\exp[2\sigma(S^a)]. \tag{P.10.26}$$

Rewrite (10.26) in the form

$$\exp[-2\sigma(S^a)]\exp(Z)\exp[-2\sigma(S^a)] = \exp[-\sigma(Z)]. \tag{P.10.27}$$

Now invert both sides of (10.27) to get

$$\exp[\sigma(Z)] = \exp[2\sigma(S^a)]\exp(-Z)\exp[2\sigma(S^a)]. \tag{P.10.28}$$

Upon comparing (10.14) and (10.28) we see that S^a and $\sigma(S^a)$ obey the same equation. Therefore, from the uniqueness of the solution, we have

$$\sigma(S^a) = S^a, \tag{P.10.29}$$

which is (8.6).

P.10.8 Formula for S^c in Terms of Z

The last thing to do is, given Z, find S^c. Look at (10.8). It can be rewritten in the form

$$\exp(S^c) = \exp(-S^a)\exp(Z)\exp(-S^a). \tag{P.10.30}$$

And, since S^a is now known, we may regard (10.30) as a formula for S^c.

P.10.9 Verification of Expected Symmetry for S^c

However, it would be good to check that (8.7) holds. Apply Σ to both sides of (10.30) and manipulate to find

$$\begin{aligned}\exp[\sigma(S^c)] &= \exp[-\sigma(S^a)]\exp[\sigma(Z)]\exp[-\sigma(S^a)]\\ &= \exp(-S^a)\exp[\sigma(Z)]\exp(-S^a)\\ &= \exp(-S^a)\{\exp(2S^a)\exp(-Z)\exp(2S^a)\}\exp(-S^a)\\ &= \exp(S^a)\exp(-Z)\exp(S^a)\\ &= \exp(-S^c),\end{aligned} \tag{P.10.31}$$

from which it follows that

$$\sigma(S^c) = -S^c, \tag{P.10.32}$$

as required by (8.7). Here we used (10.14) and the relation

$$\exp(-S^c) = \exp(S^a)\exp(-Z)\exp(S^a) \tag{P.10.33}$$

which follows from (10.30).

P.10.10 Conclusion

In summary, we have learned that both S^a and S^c are real-analytic functions of Z.

P.10.11 Motivation for Mapping

Suppose P is a real, symmetric, and positive-definite matrix. Use it to define a matrix Q by the relation

$$Q = P^{-1/2}JP^{-1/2}. \tag{P.10.34}$$

(Note that Goodman defines Q by $Q = P^{1/2}JP^{-1/2}$, but presumably this is a misprint.) By calculation we find that

$$Q^T = -P^{-1/2}JP^{-1/2} = -Q. \tag{P.10.35}$$

Evidently Q is real, antisymmetric, and nonsingular. Then $\hat{P}$ given by

$$\hat{P} = Q^TQ \tag{P.10.36}$$

will be real, symmetric, and positive definite. Goodman claims that

$$P = \exp(X)\exp(Y)\exp(X) \tag{P.10.37}$$

with

$$\exp(X) = (P^{1/2}\hat{P}^{1/2}P^{1/2})^{1/2} \tag{P.10.38}$$

and

$$\exp(Y) = \exp(-X)P\exp(-X). \tag{P.10.39}$$

Let us see if this is true. Evidently (10.37) and (10.39) are logically equivalent for any matrices P, X, and Y. So, perhaps we should examine the properties of X and Y.

Evidently X is real analytic in P. Squaring both sides of (10.38) gives

$$\exp(2X) = P^{1/2}\hat{P}^{1/2}P^{1/2}. \tag{P.10.40}$$

Define Z by writing

$$\exp(Z) = P. \tag{P.10.41}$$

Evidently Z is real analytic in P and

$$\exp(Z/2) = P^{1/2}. \tag{P.10.42}$$

With this definition, (10.40) can be rewritten in the form

$$\exp(2X) = \exp(Z/2)\hat{P}^{1/2}\exp(Z/2). \tag{P.10.43}$$

Next, define T by writing

$$\exp(T) = \hat{P}^{1/2}. \tag{P.10.44}$$

Then we have the result

$$\exp(2T) = \hat{P}. \tag{P.10.45}$$

Also, (10.43) can now be written in the form

$$\exp(2X) = \exp(Z/2)\exp(T)\exp(Z/2). \tag{P.10.46}$$

Now work out an expression for $\hat{P}$. From (10.34) through (10.36) we find that

$$\begin{aligned}\hat{P} &= -P^{-1/2}JP^{-1/2}P^{-1/2}JP^{-1/2} = P^{-1/2}JP^{-1}J^{-1}P^{-1/2}\\ &= \exp(-Z/2)J\exp(-Z)J^{-1}\exp(-Z/2) = \exp(-Z/2)\exp(-JZJ^{-1})\exp(-Z/2)\\ &= \exp(-Z/2)\exp[\sigma(Z)]\exp(-Z/2).\end{aligned} \tag{P.10.47}$$

Thus, we get the result

$$\exp(2T) = \exp(-Z/2)\exp[\sigma(Z)]\exp(-Z/2). \tag{P.10.48}$$

We see that (10.46) is the counterpart to (10.22), and (10.48) is the counterpart to (10.16). Therefore we have the relations

$$X = S^a \tag{P.10.49}$$

and

$$Y = S^c. \tag{P.10.50}$$

P.11 Theorem 1.2 of Goodman Due to Mostow

Consider the triplet $\{k \in O(2n, \mathbb{R}), S^c, S^a\}$. Use it to construct $g \in GL(2n, \mathbb{R})$ by the rule

$$g = k\exp(S^c)\exp(S^a). \tag{P.11.1}$$

It is claimed that (11.1) provides an analytic isomorphism between the triplet and $GL(2n, \mathbb{R})$.

Let us pause to do a dimension count. The dimension of S^c plus the dimension of S^a is the dimension of all symetric matrices S. And the dimension of $O(2n,\mathbb{R})$ is the dimension of all antisymmetric matrices A. Taken together, these dimensions add up to the dimension of all $2n \times 2n$ matrices, which is just the dimension of $GL(2n,\mathbb{R})$.

To continue, first we verify that (11.1) is injective. That is, different triplets must yield different elements in $GL(2n,\mathbb{R})$. For suppose that two triplets yield the same element $g \in GL(2n,\mathbb{R})$,

$$g = k_1 \exp(S_1^c)\exp(S_1^a) \tag{P.11.2}$$

and

$$g = k_2 \exp(S_2^c)\exp(S_2^a). \tag{P.11.3}$$

From (11.2) we find

$$g^T g = \exp(S_1^a)\exp(S_1^c)k_1^T k_1 \exp(S_1^c)\exp(S_1^a) = \exp(S_1^a)\exp(2S_1^c)\exp(S_1^a), \tag{P.11.4}$$

and from (11.3) we find

$$g^T g = \exp(S_2^a)\exp(S_2^c)k_2^T k_1 \exp(S_2^c)\exp(S_2^a) = \exp(S_2^a)\exp(2S_2^c)\exp(S_2^a). \tag{P.11.5}$$

Thus, we have the relation

$$\exp(S_1^a)\exp(2S_1^c)\exp(S_1^a) = \exp(S_2^a)\exp(2S_2^c)\exp(S_2^a). \tag{P.11.6}$$

From Theorem 1.1 and (11.6) we conclude that

$$S_1^a = S_2^a \tag{P.11.7}$$

and

$$S_1^c = S_2^c. \tag{P.11.8}$$

Then, from (11.2) and (11.3), se see that

$$k_1 = k_2. \tag{P.11.9}$$

Next, we verify that any $g \in GL(2n,\mathbb{R})$ can be written in the form (11.1). Given any $g \in GL(2n,\mathbb{R})$, define a real symmetric positive-definite matrix P by the rule

$$P = g^T g. \tag{P.11.10}$$

Evidently P is real analytic in g. Therefore, by the factorization theorem, there are unique matrices S^a and S^c, that depend real-analytically on P and therefore real-analytically on g, such that

$$P = \exp(S^a)\exp(2S^c)\exp(S^a). \tag{P.11.11}$$

Now define k by the rule

$$k = (g^T)^{-1}\exp(S^a)\exp(S^c). \tag{P.11.12}$$

Then k is also real-analytic in g. Moreover, we find that

$$k^T = \exp(S^c)\exp(S^a)(g)^{-1}, \tag{P.11.13}$$

from which it follows that

$$\begin{aligned} k^T k &= \exp(S^c)\exp(S^a)(g)^{-1}(g^T)^{-1}\exp(S^a)\exp(S^c) \\ &= \exp(S^c)\exp(S^a)P^{-1}\exp(S^a)\exp(S^c) \\ &= \exp(S^c)\exp(S^a)\exp(-S^a)\exp(-2S^c)\exp(-S^a)\exp(S^a)\exp(S^c) \\ &= \exp(S^c)\exp(-2S^c)\exp(S^c) = I. \qquad \text{(P.11.14)} \end{aligned}$$

Therefore $k \in O(2n, \mathbb{R})$.

Finally, suppose that $g \in SL(2n, \mathbb{R})$. In this case, take the determinant of both sides of (11.1) to get the result

$$1 = \det(g) = \det(k)\exp[\mathrm{tr}(\mathrm{S^c})]\exp[\mathrm{tr}(\mathrm{S^a})] = \det(\mathrm{k})\exp[\mathrm{tr}(\mathrm{S^c})] \qquad \text{(P.11.15)}$$

where we have used (10.5). We know that

$$\det(k) = \pm 1. \qquad \text{(P.11.16)}$$

We see that, in order for (11.15) to be satisfied, we must have the relations

$$\det(k) = +1 \text{ so that } k \in SO(2n, \mathbb{R}) \qquad \text{(P.11.17)}$$

and

$$\mathrm{tr}(S^c) = 0. \qquad \text{(P.11.18)}$$

We conclude the following: Consider the triplet $\{k \in SO(2n, \mathbb{R}), S^c \text{ with } \mathrm{tr}(S^c) = 0, S^a\}$. Use it to construct $g \in SL(2n, \mathbb{R})$ by the rule

$$g = k\exp(S^c)\exp(S^a). \qquad \text{(P.11.19)}$$

Then (11.19) provides an analytic isomorphism between the triplet and $SL(2n, \mathbb{R})$.

P.12 Goodman's Work on Symplectic Polar Decomposition

Consider the group $G = SL(2n, \mathbb{R})$, and its subgroups $H = Sp(2n, \mathbb{R})$ and $K = SO(2n, \mathbb{R})$. We want to study the coset space G/H.

P.12.1 Some More Symmetry Operations

To do so it is useful to introduce some additional symmetry operations on G. The operation Σ has already been defined by (4.1). We will define two more.

Introduce the operation Θ by the rule

$$\Theta(M) = (M^T)^{-1} \qquad \text{(P.12.1)}$$

for any $M \in G$. Evidently this map preserves the condition $\det(M) = 1$, and therefore sends G unto itself. Also we find that

$$\Theta(I) = I, \qquad \text{(P.12.2)}$$

$$\Theta(J) = J, \tag{P.12.3}$$

$$\Theta(M_1M_2) = [(M_1M_2)^T]^{-1} = [(M_2^TM_1^T)]^{-1} = (M_1^T)^{-1}(M_2^T)^{-1} = \Theta(M_1)\Theta(M_2), \tag{P.12.4}$$

$$\Theta[\Theta(M)] = \{[(M^T)^{-1}]^T\}^{-1} = M. \tag{P.12.5}$$

Thus Θ, like Σ, is an isomorphism and an involution.

Next, we discover that Σ and Θ commute. From the definition of Σ we find that

$$\Sigma[\Theta(M)] = J\{[\Theta(M)]^T\}^{-1}J^{-1}. \tag{P.12.6}$$

But, from (12.1),

$$\{[\Theta(M)]^T\}^{-1} = \{[(M^T)^{-1}]^T\}^{-1} = M. \tag{P.12.7}$$

Therefore, we conclude that

$$\Sigma[\Theta(M)] = JMJ^{-1}. \tag{P.12.8}$$

Applying the symmetries in opposite order gives

$$\Theta[\Sigma(M)] = \{[\Sigma(M)]^T\}^{-1}. \tag{P.12.9}$$

But, from (4.1), we have the relations

$$[\Sigma(M)]^T = [J(M^T)^{-1}J^{-1}]^T = JM^{-1}J^{-1} \tag{P.12.10}$$

and

$$\{[\Sigma(M)]^T\}^{-1} = [JM^{-1}J^{-1}]^{-1} = JMJ^{-1}. \tag{P.12.11}$$

It follows that

$$\Theta[\Sigma(M)] = JMJ^{-1}. \tag{P.12.12}$$

We see that the right sides of (12.8) and (12.12) agree, and therefore

$$\Sigma[\Theta(M)] = \Theta[\Sigma(M)] \tag{P.12.13}$$

or, in operator notation,

$$\Sigma\Theta = \Theta\Sigma. \tag{P.12.14}$$

Next define the operation Υ as the product

$$\Upsilon = \Sigma\Theta = \Theta\Sigma. \tag{P.12.15}$$

From (12.8) or (12.2) we find that

$$\Upsilon(M) = JMJ^{-1}. \tag{P.12.16}$$

We see that Υ is also an isomorphism. And, from (12.16), we see that

$$\Upsilon[\Upsilon(M)] = J[JMJ^{-1}]J^{-1} = M, \tag{P.12.17}$$

and therefore Υ is an involution as is expected for the product of commuting involutions.

Let θ and τ be the associated induced maps in the Lie algebra defined by

$$\Theta[\exp(B)] = \exp[\theta(B)], \tag{P.12.18}$$

$$\Upsilon[\exp(B)] = \exp[\tau(B)]. \tag{P.12.19}$$

For (12.18) we find

$$\Theta[\exp(B)] = \{[\exp(B)]^T\}^{-1} = [\exp(B^T)]^{-1} = \exp(-B^T), \tag{P.12.20}$$

and conclude that

$$\theta(B) = -B^T. \tag{P.12.21}$$

For (12.19) we find

$$\Upsilon[\exp(B)] = J\exp(B)J^{-1} = \exp(JBJ^{-1}), \tag{P.12.22}$$

and conclude that

$$\tau(B) = JBJ^{-1}. \tag{P.12.23}$$

Let us check some expected relations: First, we find the results

$$\theta[\theta(B)] = \theta[-B^T] = -[-B^T]^T = B \tag{P.12.24}$$

so that θ, like σ, is an involution, as expected. Second, we find the results

$$\sigma[\theta(B)] = \sigma[-B^T] = -J[-B^T]^T J^{-1} = JBJ^{-1} = \tau(B), \tag{P.12.25}$$

$$\theta[\sigma(B)] = \theta[-JB^TJ^{-1}] = -[-JB^TJ^{-1}]^T = JBJ^{-1} = \tau(B). \tag{P.12.26}$$

Thus, we conclude that

$$\sigma\theta = \theta\sigma = \tau. \tag{P.12.27}$$

It follows that τ is also an involution, as is also obvious from (12.22). Moreover, we note the relations

$$\sigma\tau = \tau\sigma = \theta, \tag{P.12.28}$$

$$\theta\tau = \tau\theta = \sigma. \tag{P.12.29}$$

Finally we should check the effects of θ and τ on scalar products. First we see that, for a real symmetric matrix,

$$\theta(S) = -S^T = -S \tag{P.12.30}$$

so that

$$[\theta(S)]^T = -S^T = -S = \theta(S). \tag{P.12.31}$$

Therefore, θ maps the space of real symmetric matrices into itself.

Next suppose B and B' are two matrices in $s\ell(2n, \mathbb{R})$. As before, define their inner product to be

$$(B, B') = \mathrm{tr}(BB'). \tag{P.12.32}$$

Then we find that θ preserves the inner product,

$$\begin{aligned}(\theta(B), \theta(B')) &= \mathrm{tr}[(-B^T)(-B')^T)] = \mathrm{tr}[B^T(B')^T] = \mathrm{tr}[(B'B)^T] \\ &= \mathrm{tr}[B'B] = \mathrm{tr}[BB'] = (B, B').\end{aligned} \tag{P.12.33}$$

Using (12.25), because σ and θ preserve the inner product, we see that τ also preserves the inner product,

$$(\tau(B), \tau(B')) = (\sigma[\theta(B)], \sigma[\theta(B')]) = ([\theta(B)], [\theta(B')]) = (B, B'). \tag{P.12.34}$$

P.12.2 Fixed-Point Subgroups Associated with Symmetry Operations

In Section 4 we found that the fixed points of Σ comprise the subgroup $Sp(2n,\mathbb{R})$. Now we will find that the fixed points of Θ and Υ also yield subgroups of $GL(2n,\mathbb{R})$.

Suppose g is a fixed point of Θ. Then we find that

$$\Theta(g)=g \;\Leftrightarrow\; (g^T)^{-1}=g \;\Leftrightarrow\; g^Tg=I. \tag{P.12.35}$$

Thus, the fixed points of Θ comprise $SO(2n,\mathbb{R})$. [Here we have already assumed $g\in SL(2n,\mathbb{R})$ so that we know that $\det g=1$. Otherwise the fixed points of Θ comprise $O(2n,\mathbb{R})$.]

Suppose g is a fixed point of Υ. Then we find that

$$\Upsilon(g)=g \;\Leftrightarrow\; JgJ^{-1}=g \;\Leftrightarrow\; Jg=gJ. \tag{P.12.36}$$

Thus, the fixed points of Υ comprise the matrices in $GL(2n,\mathbb{R})$ that commute with J. They also obviously form a group. But what is this group?

Write g in the block form

$$g=\begin{pmatrix} a & b\\ c & d\end{pmatrix}, \tag{P.12.37}$$

where the matrices a,b,c, and d are real and $n\times n$. Then we find the results

$$Jg=\begin{pmatrix} c & d\\ -a & -b\end{pmatrix}, \tag{P.12.38}$$

and

$$gJ=\begin{pmatrix} -b & a\\ -d & c\end{pmatrix}. \tag{P.12.39}$$

Therefore, requiring that g commute with J yields the restrictions

$$c=-b \tag{P.12.40}$$

and

$$d=a. \tag{P.12.41}$$

Thus, g is of the form

$$g=\begin{pmatrix} a & b\\ -b & a\end{pmatrix}. \tag{P.12.42}$$

Next define matrices A and B by the rules

$$A=\begin{pmatrix} a & 0\\ 0 & a\end{pmatrix}, \tag{P.12.43}$$

and

$$B=\begin{pmatrix} b & 0\\ 0 & b\end{pmatrix}. \tag{P.12.44}$$

Then both A and B commute with J, and we also have the relation

$$JB = \begin{pmatrix} 0 & b \\ -b & 0 \end{pmatrix}. \tag{P.12.45}$$

Therefore, we may also write

$$g = A + JB. \tag{P.12.46}$$

Suppose g_1 and g_2 are two matrices that commute with J and we use the representation (12.46) to write

$$g_k = A_k + JB_k \tag{P.12.47}$$

Then, recalling that the A_k and B_k commute with J and that $J^2 = -I$, we find the product relation

$$g_1 g_2 = (A_1 A_2 - B_1 B_2) + J(A_1 B_2 + B_1 A_2). \tag{P.12.48}$$

We see that, in (12.47) and (12.48), the matrix J plays a role analogous to the imaginary number i.

This analogy can be made explicit using the machinery of Section 3.9 of *Lie Methods*. Suppose m is an arbitrary $n \times n$ matrix with possibly complex entries. Evidently it can be written in the form

$$m = a + ib \tag{P.12.49}$$

where a and b are real $n \times n$ matrices. Let us multiply two such matrices together. So doing gives the result

$$m_1 m_2 = (a_1 a_2 - b_1 b_2) + i(a_1 b_2 + b_1 a_2). \tag{P.12.50}$$

Note the resemblance between the pairs (12.46), (12.49) and (12.48), (12.50).

To pursue the analogy further, let W be the unitary and (complex) symplectic matrix

$$W = \frac{1}{\sqrt{2}} \begin{pmatrix} I & iI \\ iI & I \end{pmatrix}. \tag{P.12.51}$$

Here each block in W is $n \times n$. Now, as in Section 3.9 of *Lie Methods*, define an associated $2n \times 2n$ matrix $g(m)$ by the rule

$$g(m) = M(m) = W \begin{pmatrix} m & 0 \\ 0 & \overline{m} \end{pmatrix} W^{-1}. \tag{P.12.52}$$

Then it is easily verified that there are the relations

$$g(I) = I, \tag{P.12.53}$$

$$g(m_1 m_2) = g(m_1) g(m_2), \tag{P.12.54}$$

$$g(m^{-1}) = g^{-1}(m). \tag{P.12.55}$$

Also, if (12.52) is multiplied out explicitly, we find the result

$$g(m) = \begin{pmatrix} \mathrm{Re}(m) & \mathrm{Im}(m) \\ -\mathrm{Im}(m) & \mathrm{Re}(m) \end{pmatrix} = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}. \tag{P.12.56}$$

It follows that $g(m)$ is real for any m.

Matrices of the form (12.49) constitute the group $SL(n,\mathbb{C})$ provided we add the condition

$$\det(m) = 1. \tag{P.12.57}$$

Now take the determinant of both sides of (12.52). Doing so gives the result

$$\begin{aligned} \det(g) &= [\det(W)][\det(m)][\det(\overline{m})][\det(W^{-1})] \\ &= [\det(m)][\det(\overline{m})] = |\det(m)|^2 \geq 0. \end{aligned} \tag{P.12.58}$$

If (12.57) holds, then from (12.58) we also have the condition

$$\det(g) = 1. \tag{P.12.59}$$

From (12.49) through (12.59) we conclude that the set of matrices $g \in SL(2n,\mathbb{R})$ that also commute with J constitutes a group that is isomorphic to $SL(n,\mathbb{C})$. More precisely, the set of matrices $g \in SL(2n,\mathbb{R})$ that also commute with J constitutes a group that is the representation $SL(n,\mathbb{C}) \oplus \overline{SL(n,\mathbb{C})}$ of $SL(n,\mathbb{C})$. If we relax the determinant condition, we conclude that the set of matrices $g \in GL(2n,\mathbb{R},+)$ that also commute with J constitutes a group that is the representation $GL(n,\mathbb{C}) \oplus \overline{GL(n,\mathbb{C})}$ of $GL(n,\mathbb{C})$.

To summarize, let G^{Σ} be the fixed-point group associated with the symmetry Σ. Then we have the result

$$G^{\Sigma} = Sp(2n,\mathbb{R}) = H. \tag{P.12.60}$$

Similarly, we have

$$G^{\Theta} = SO(2n,\mathbb{R}) = K, \tag{P.12.61}$$

and

$$G^{\Upsilon} \cong SL(n,\mathbb{C}). \tag{P.12.62}$$

Also, from Section 3.9 of *Lie Methods*, we know that

$$K \cap H = SO(2n,\mathbb{R}) \cap Sp(2n,\mathbb{R}) \cong U(n) \oplus \overline{U(n)}. \tag{P.12.63}$$

Finally, suppose

$$m = \exp(i\phi)I, \tag{P.12.64}$$

which corresponds to

$$a = \cos(\phi)I \tag{P.12.65}$$

and

$$b = \sin(\phi)I. \tag{P.12.66}$$

Then we find the result

$$g(m) = \begin{pmatrix} \cos(\phi)I & \sin(\phi)I \\ -\sin(\phi)I & \cos(\phi)I \end{pmatrix} = I\cos(\phi) + J\sin(\phi) = \exp(\phi J). \tag{P.12.67}$$

P.13 Decomposition of Lie Algebras

Denote the Lie algebras of $G = SL(2n,\mathbb{R})$, $H = Sp(2n,\mathbb{R})$, and $K = SO(2n,\mathbb{R})$ by the symbols $\mathfrak{g}$, $\mathfrak{h}$, and $\mathfrak{k}$, respectively. The effects of σ, θ, and τ on $\mathfrak{g} = sl(2n,\mathbb{R})$ have already been determined in (4.15), (12.21), and (12.23), respectively. To recapitulate, we find the results

$$\sigma(B) = -JB^TJ^{-1} = B \text{ for } B \in \mathfrak{h} \tag{P.13.1}$$

and

$$\theta(B) = -B^T = B \text{ for } B \in \mathfrak{k}. \tag{P.13.2}$$

That is, $\mathfrak{h}$ is the +1 eigenspace of σ in $\mathfrak{g}$, and $\mathfrak{k}$ (the antisymmetric matrices) is the +1 eigenspace of θ in $\mathfrak{g}$.

Define the subspace $\mathfrak{p}$ by the requirement

$$\mathfrak{p} = \{B \in \mathfrak{g} \mid \theta(B) = -B\}. \tag{P.13.3}$$

That is, $\mathfrak{p}$ consists of the symmetric traceless matrices, and is the -1 eigenspace of θ in $\mathfrak{g}$. Then we have the direct sum decomposition (±1 eigenspaces of θ)

$$\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}, \tag{P.13.4}$$

which is just the familiar statement that any matrix can be uniquely decomposed into antisymmetric and symmetric parts. These parts are also mutually orthogonal relative to the trace form. Indeed, we have

$$(A,S) = \mathrm{tr}(AS) = \mathrm{tr}[(AS)^T] = \mathrm{tr}(S^TA^T) = \mathrm{tr}(-SA) = \mathrm{tr}(-AS) = -(A,S) \tag{P.13.5}$$

and therefore

$$(A,S) = 0. \tag{P.13.6}$$

Here A and S are antisymmetric and symmetric matrices, respectively.

Likewise, define the subspace $\mathfrak{q}$ by the requirement

$$\mathfrak{q} = \{B \in \mathfrak{g} \mid \sigma(B) = -B\}. \tag{P.13.7}$$

Then we have the direct sum decomposition (±1 eigenspaces of σ)

$$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{q}. \tag{P.13.8}$$

We should check that these eigenspaces are also mutually orthogonal relative to the trace form. The elements of $\mathfrak{h}$ are matrices of the form JS where S is symmetric. From (4.18) we know that the elements of $\mathfrak{q}$ are matrices of the form JA where A is antisymmetric. For their inner product we find that

$$\begin{aligned}(JS,JA) &= \mathrm{tr}(JSJA) = \mathrm{tr}[(JSJA)^T] = \mathrm{tr}[A^TJ^TS^TJ^T] \\ &= \mathrm{tr}[-AJSJ] = \mathrm{tr}[-JSJA] = -(JS,JA),\end{aligned} \tag{P.13.9}$$

from which we conclude that

$$(JS,JA) = 0. \tag{P.13.10}$$

Finally, we observe that the relation (13.8) is just the assertion that every traceless matrix can be written as the sum $(JS+JA)$ where A is chosen to make JA traceless. (Note that JS is automatically traceless for any S.)

Next we can refine the decompositions (13.4) and (13.8) using both θ and σ. For example, the decomposition (13.4) used θ to decompose $\mathfrak{g}$ into $\mathfrak{k}+\mathfrak{p}$. We can now further decompose $\mathfrak{k}$ and $\mathfrak{p}$ using σ. Alternatively, the decomposition (13.8) used σ to decompose $\mathfrak{g}$ into $\mathfrak{h}+\mathfrak{q}$. We can now further decompose $\mathfrak{h}$ and $\mathfrak{q}$ using θ. These refinements are possible because θ and σ commute. The eigenspaces of θ are invariant under the action of σ, and the eigenspaces of σ are invariant under the action of θ.

Let s and s' be *sign* variables that take on the values ±1. Define subspaces $\mathfrak{g}_{ss'}$ by the requirements

$$\sigma(B)=sB \tag{P.13.11}$$

and

$$\theta(B)=s'B \tag{P.13.12}$$

for any $B\in\mathfrak{g}_{ss'}$. Then we have the direct sum decomposition

$$\mathfrak{g}=\mathfrak{g}_{1,1}\oplus\mathfrak{g}_{1,-1}\oplus\mathfrak{g}_{-1,1}\oplus\mathfrak{g}_{-1.-1}. \tag{P.13.13}$$

Also, we have the result

$$\tau(B)=ss'B \tag{P.13.14}$$

for any $B\in\mathfrak{g}_{ss'}$.

Let us examine the contents of each subspace $\mathfrak{g}_{ss'}$. Begin with $\mathfrak{g}_{1,1}$. From (4.17) we see that it must be of the form JS. From (12.23) and (13.14) we see that it must be of the form JS^c. Thus we have

$$B\in\mathfrak{g}_{1,1}\ \Leftrightarrow\ B=JS^c. \tag{P.13.15}$$

Similarly, we find that elements in $\mathfrak{g}_{1,-1}$ must be of the form JS^a,

$$B\in\mathfrak{g}_{1,-1}\ \Leftrightarrow\ B=JS^a. \tag{P.13.16}$$

Together $\mathfrak{g}_{1,1}$ and $\mathfrak{g}_{1,-1}$ span $\mathfrak{h}=sp(2n,\mathbb{R})$,

$$\mathfrak{h}=\mathfrak{g}_{1,1}\oplus\mathfrak{g}_{1,-1}. \tag{P.13.17}$$

Next consider $\mathfrak{g}_{-1,1}$. By (4.18) it must be of the form JA. By (13.14) it must be of the form JA^a where A^a is an antisymmetric matrix that anticommutes with J,

$$B\in\mathfrak{g}_{-1,1}\ \Leftrightarrow\ B=JA^a. \tag{P.13.18}$$

Finally, any $B\in\mathfrak{g}_{-1,-1}$ must be of the form JA^c where A^c is an antisymmetric matrix that commutes with J,

$$B\in\mathfrak{g}_{-1,-1}\ \Leftrightarrow\ B=JA^c. \tag{P.13.19}$$

In summary, the claim is that any matrix $B\in sl(2n,\mathbb{R})$ can be uniquely be decomposed as the sum

$$B=JS^c+JS^a+JA^a+JA^c, \tag{P.13.20}$$

which is, in fact, almost obvious upon inspection.

These elements are also mutually orthogonal. We find from (8.9) the result

$$(JS^c, JS^a) = \mathrm{tr}(JS^cJS^a) = \mathrm{tr}(S^cJJS^a) = -\mathrm{tr}(S^cS^a) = -(S^c, S^a) = 0. \tag{P.13.21}$$

From (13.10) we find the results

$$(JS^c, JA^a) = (JS^c, JA^c) = (JS^a, JA^a) = (JS^a, JA^c) = 0. \tag{P.13.22}$$

Lastly, we find

$$(JA^a, JA^c) = \mathrm{tr}(JA^aJA^c) = -\mathrm{tr}(A^aJJA^c) = \mathrm{tr}(A^aA^c) = (A^a, A^c). \tag{P.13.23}$$

But we also find

$$(JA^a, JA^c) = \mathrm{tr}(JA^aJA^c) = \mathrm{tr}(JA^aA^cJ) = \mathrm{tr}(JJA^aA^c) = -(A^a, A^c), \tag{P.13.24}$$

from which we conclude that

$$(JA^a, JA^c) = 0. \tag{P.13.25}$$

These orthogonality proofs just provided are brute force. A more elegant proof, in the style of (8.8) and (8.9), can be given based on (13.1) and (13.2) and the fact that σ and θ preserve the inner product.

Together $\mathfrak{g}_{1,1}$ and $\mathfrak{g}_{-1,1}$ span $\mathfrak{k} = so(2n, \mathbb{R})$,

$$\mathfrak{k} = \mathfrak{g}_{1,1} \oplus \mathfrak{g}_{-1,1}. \tag{P.13.26}$$

Inspection of (3.15) and (3.18) shows that

$$B \in \mathfrak{g}_{1,1} \oplus \mathfrak{g}_{-1,1} \;\Leftrightarrow\; B = JS^c + JA^a. \tag{P.13.27}$$

Matrices of the form ($B = JS^c + JA^a$) are evidently antisymmetric. It is also easily verified that matrices of the form ($JS^a + JA^c$) are symmetric. Since all the matrices appearing in (13.20), when taken together, span $g\ell(2n, \mathbb{R})$, it follows that matrices of the form ($B = JS^c + JA^a$) span the full space of antisymmetric matrices. See also (13.28) and (13.30) below.

It remains to be seen what matrices in (13.20) are traceless. We already know that JS^c and JS^a are traceless, and JA^a is also traceless because it is antisymmetric. The remaining candidate is JA^c. It contains the possibility $JJ = -I$, which is not traceless.

Finally, using the notation of intersecting sets, we may write

$$\mathfrak{g}_{1,1} = \mathfrak{h} \cap \mathfrak{k} = \text{ matrices of the form } JS^c \cong A^c, \tag{P.13.28}$$

$$\mathfrak{g}_{1,-1} = \mathfrak{h} \cap \mathfrak{p} = \text{ matrices of the form } JS^a \cong S^a, \tag{P.13.29}$$

$$\mathfrak{g}_{-1,1} = \mathfrak{q} \cap \mathfrak{k} = \text{ matrices of the form } JA^a \cong A^a, \tag{P.13.30}$$

$$\mathfrak{g}_{-1,-1} = \mathfrak{q} \cap \mathfrak{p} = \text{ matrices of the form } JA^c \cong S^c. \tag{P.13.31}$$

Here we have included the fact that various categories of matrices are are isomorphic. For example, matrices of the form JS^c are evidently antisymmetric, and commute with J. Therefore they are of the form A^c. We also note, from the discussion of the previous paragraph, that all the $\mathfrak{g}_{s,s'}$ are traceless with the possible exception of $\mathfrak{g}_{-1,-1}$.

By using these isomorphisms in (13.20), we find that any matrix $B \in s\ell(2n,\mathbb{R})$ can also be uniquely be decomposed as the sum

$$B = A^c + S^a + A^a + S^c, \tag{P.13.32}$$

which is also obvious upon inspection.

We have already noted in (13.17) that $\mathfrak{g}_{1,1}$ and $\mathfrak{g}_{1,-1}$ together span $\mathfrak{h} = sp(2n,\mathbb{R})$. They, in fact, provide the Cartan decomposition of $sp(2n,\mathbb{R})$.

Finally, Goodman makes the claims

$$\mathfrak{g}_{1,-1} = \mathfrak{h} \cap \mathfrak{p} = \{B \in \mathcal{E} \mid \mathrm{tr}B = 0\}, \tag{P.13.33}$$

$$\mathfrak{g}_{-1,-1} = \mathfrak{q} \cap \mathfrak{p} = \{B \in \mathcal{F} \mid \mathrm{tr}B = 0\}. \tag{P.13.34}$$

But, $\mathcal{E}$ is defined by the requirements

$$\mathcal{E} = \{B \mid B^T = B \text{ and } \sigma(B) = B\}, \tag{P.13.35}$$

which is equivalent to the requirements

$$\mathcal{E} = \{B \mid \theta(B) = -B \text{ and } \sigma(B) = B\}. \tag{P.13.36}$$

In our notation, these requirements simply state that

$$\mathcal{E} = \mathfrak{g}_{1,-1}. \tag{P.13.37}$$

We already know that matrices in $\mathfrak{g}_{1,-1}$ are traceless, and so (13.33) is verified. Moreover, $\mathcal{F}$ is defined by the requirements

$$\mathcal{F} = \{B \mid B^T = B \text{ and } \sigma(B) = -B\}, \tag{P.13.38}$$

which is equivalent to the requirements

$$\mathcal{F} = \{B \mid \theta(B) = -B \text{ and } \sigma(B) = -B\}. \tag{P.13.39}$$

In our notation, these requirements simply state that

$$\mathcal{F} = \mathfrak{g}_{-1,-1}. \tag{P.13.40}$$

We know from (13.19) that in this case there is a matrix that has trace, namely JJ, and therefore in this case (13.34) is also verified provided the JJ case is excluded.

P.14 Preparation for Lemma 2.1 of Goodman

Let S_d consist of all the real *diagonal* symmetric matrices s_d of the form

$$s_d = \begin{pmatrix} d & 0 \\ 0 & d \end{pmatrix} \tag{P.14.1}$$

where

$$d = \operatorname{diag}[d_1, \cdots, d_n] \text{ and } \operatorname{tr}(d) = 0. \tag{P.14.2}$$

(Goodman uses the symbol A for what we call S_d. However, we have already used A to denote antisymmetric matrices.) Then, from Section 12.2, we have the relation

$$\tau(s_d) = s_d. \tag{P.14.3}$$

And, from (12.21), we see that

$$\theta(s_d) = -s_d. \tag{P.14.4}$$

Therefore, from (12.29), we conclude that

$$\sigma(s_d) = -s_d, \tag{P.14.5}$$

and consequently

$$S_d \subset \mathfrak{g}_{-1,-1} = \mathfrak{q} \cap \mathfrak{p}. \tag{P.14.6}$$

Let S_d^+ be the open subset of S_d consisting of the matrices s_d with

$$d_1 > d_2 > \cdots > d_n. \tag{P.14.7}$$

Let S_d^{+c} be the closure of S_d^+. Then we also have the relations

$$S_d^+ \subset \mathfrak{g}_{-1,-1} = \mathfrak{q} \cap \mathfrak{p}, \tag{P.14.8}$$

and

$$S_d^{+c} \subset \mathfrak{g}_{-1,-1} = \mathfrak{q} \cap \mathfrak{p}. \tag{P.14.9}$$

Also, let D^+ be the set of matrices d of the form (14.2) with (14.7) satisfied, and let D^{+c} denote its closure.

P.15 Lemma 2.1 of Goodman

Suppose x is of the form $x = \exp(B)$ where $B \in \mathfrak{g}_{-1,-1}$. Then we know there is a matrix of the form JA^c such that

$$x = \exp(JA^c). \tag{P.15.1}$$

We may also require that JA^c be traceless. By inspection, JA^c is symmetric and commutes with J. Therefore, consistent with (13.31), there is a real traceless symmetric matrix S^c such that

$$JA^c = S^c, \tag{P.15.2}$$

and we have the relation

$$x = \exp(S^c). \tag{P.15.3}$$

It follows from (15.3) that x is real, symmetric, and positive definite.

Let us see what can be said about S^c. Since it commutes with J, by the work of Section 12.2, it must be of the form

$$S^c = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}. \tag{P.15.4}$$

Since S^c is traceless, α must be traceless. Since S^c is symmetric, the matrices α and β must have have the properties

$$\alpha^T = \alpha, \tag{P.15.5}$$

$$\beta^T = -\beta. \tag{P.15.6}$$

Note that, by (15.6), the matrix β is traceless. Define the matrix m by the rule

$$m = \alpha + i\beta. \tag{P.15.7}$$

Then, we have the relation

$$S^c = S^c(m) = M(m) = W \begin{pmatrix} m & 0 \\ 0 & \overline{m} \end{pmatrix} W^{-1}. \tag{P.15.8}$$

We also observe that

$$m^\dagger = \alpha^T - i\beta^T = \alpha + i\beta = m \tag{P.15.9}$$

so that m is Hermitian. Since both α and β are traceless, m is also traceless.

Since m is Hermitian and traceless, there is a unitary matrix v and a matrix $d \in D^{+c}$ such that

$$m = vdv^{-1}. \tag{P.15.10}$$

Since d is real, we also have the relation

$$\overline{m} = \overline{v}d\overline{v^{-1}}. \tag{P.15.11}$$

It follows that S^c has the representation

$$S^c(m) = W \begin{pmatrix} v & 0 \\ 0 & \overline{v} \end{pmatrix} \begin{pmatrix} d & 0 \\ 0 & d \end{pmatrix} \begin{pmatrix} v^{-1} & 0 \\ 0 & \overline{v^{-1}} \end{pmatrix} W^{-1}. \tag{P.15.12}$$

Insert factors of W and W^{-1} into (15.12) to rewrite it in the form

$$S^c(m) = W \begin{pmatrix} v & 0 \\ 0 & \overline{v} \end{pmatrix} W^{-1} W \begin{pmatrix} d & 0 \\ 0 & d \end{pmatrix} W^{-1} W \begin{pmatrix} v^{-1} & 0 \\ 0 & \overline{v^{-1}} \end{pmatrix} W^{-1}. \tag{P.15.13}$$

Since d is real, we see that (15.13) can be rewritten in the form

$$S^c(m) = M(v)M(d)M(v^{-1}). \tag{P.15.14}$$

We also know, since d is real, that we have

$$M(d) = s_d, \tag{P.15.15}$$

with $s_d \in S_d^{+c}$, so that there is also the relation

$$S^c = M(v) s_d M(v^{-1}). \tag{P.15.16}$$

Define the matrix O by the rule

$$O = M(v). \tag{P.15.17}$$

Since v is unitary, O will be real, symplectic, and orthogonal. Thus we have the result

$$S^c = O s_d O^{-1}. \tag{P.15.18}$$

Finally, exponentiate both sides of (15.18). So doing gives the result

$$x = O \exp(s_d) O^{-1} \tag{P.15.19}$$

with

$$O \in SO(2n, \mathbb{R}) \cap Sp(2n, \mathbb{R} \cong U(n) \oplus \overline{U(n)}. \tag{P.15.20}$$

Note that since O is orthogonal, (15.19) can also be written in the form

$$x = O \exp(s_d) O^T, \tag{P.15.21}$$

from which we again see that x is real, symmetric, and positive definite.

Goodman claims that s_d depends analytically on x when the eigenvalues of x, which will be the quantities $\exp(d_j)$, are distinct. We will worry about proving this later. We know from (15.3) that S^c is real analytic in x, and from (15.4) through (15.7) we see that m is analytic in S^c. What remains to be shown is that the eigenvalues of m, with m Hermitian and traceless, are analytic in m under the assumption that the eigenvalues are distinct.

P.16 Preparation for Theorem 2.1 of Goodman

Let u be any $n \times n$ unitary matrix, and consider $M(u)$. From Section 3.9 of *Lie methods* we know that

$$M(u) \in SO(2n, \mathbb{R}) \cap Sp(2n, \mathbb{R}) \cong U(n) \oplus \overline{U(n)}, \tag{P.16.1}$$

and given any $M' \in SO(2n, \mathbb{R}) \cap Sp(2n, \mathbb{R})$ there is a unique $u \in U(n)$ such that

$$M(u) = M'. \tag{P.16.2}$$

Let $L \subset SO(2n, \mathbb{R}) \cap Sp(2n, \mathbb{R})$ be the subgroup of matrices g such that

$$g s_d g^{-1} = s_d \text{ for all } s_d \in S_d. \tag{P.16.3}$$

By definition for any such s_d there is a corresponding d given by (14.1) and (14.2), and we have the relation

$$s_d = M(d). \tag{P.16.4}$$

Also, given any $g \in SO(2n,\mathbb{R}) \cap Sp(2n,\mathbb{R})$, there is a unique $u \in U(n)$ such that

$$M(u) = g. \tag{P.16.5}$$

With these definitions, the relation (16.3) becomes

$$M(u)M(d)[M(u)]^{-1} = M(d) \text{ for all } d. \tag{P.16.6}$$

By the isomorphic property of M this relation is equivalent to the requirement

$$udu^{-1} = d \text{ for all } d. \tag{P.16.7}$$

All such u must be diagonal unitary matrices, and therefore have the explicit form

$$u = \text{diag}[\exp(i\phi_1), \cdots, \exp(i\phi_n)]. \tag{P.16.8}$$

Thus, we have the isomorphism

$$L \cong T^n. \tag{P.16.9}$$

Moreover, since $SO(2n,\mathbb{R})$ has rank n, L is a maximal torus in $SO(2n,\mathbb{R})$. Goodman says that this means that $SO(2n,\mathbb{R})/L$ is the flag manifold for $SO(2n,\mathbb{R})$.

By the definition of L we see from (16.3) that

$$\ell s_d \ell^{-1} = s_d \text{ for all } \ell \in L \text{ and all } s_d \in S_d. \tag{P.16.10}$$

It follows that

$$\ell \exp(s_d)\ell^{-1} = \exp(\ell s_d \ell^{-1}) = \exp(s_d). \tag{P.16.11}$$

Introduce the notation

$$\hat{s}_d = \exp(s_d). \tag{P.16.12}$$

Then, for future use, we have the relation

$$\ell \hat{s}_d = \hat{s}_d \ell \text{ for all } \ell \in L \text{ and all } \hat{s}_d \in \exp(S_d). \tag{P.16.13}$$

Also, we see directly that

$$\ell \in SO(2n,\mathbb{R}) \cap Sp(2n,\mathbb{R}) \tag{P.16.14}$$

since ℓ is of the form $M(u)$ with u given by (16.8).

For insight, let us work out the explicit matrix form of ℓ. We have the relation

$$\ell = M(u) = \begin{pmatrix} \text{Re}(u) & \text{Im}(u) \\ -\text{Im}(u) & \text{Re}(u) \end{pmatrix} = \begin{pmatrix} C & S \\ -S & C \end{pmatrix}. \tag{P.16.15}$$

Here C and S are $n \times n$ diagonal matrices given by the relations

$$C = \begin{pmatrix} \cos(\phi_1) & & & \\ & \cos(\phi_2) & & \\ & & \ddots & \\ & & & \cos(\phi_n) \end{pmatrix}, \tag{P.16.16}$$

$$S = \begin{pmatrix} \sin(\phi_1) & & & \\ & \sin(\phi_2) & & \\ & & \ddots & \\ & & & \sin(\phi_n) \end{pmatrix}. \tag{P.16.17}$$

P.17 Theorem 2.1 of Goodman

Recall that $G = SL(2n, \mathbb{R})$, $H = Sp(2n, \mathbb{R})$, and $K = SO(2n, \mathbb{R})$. Goodman claims that any $g \in G$ has the decomposition

$$g = k\hat{s}_d h \tag{P.17.1}$$

where $k \in K$, $h \in H$, and $\hat{s}_d \in \exp(S_d^{+c})$.

The argument goes as follows. Recall that, according to Theorem 1.2, every $g \in SL(2n, \mathbb{R})$ has the factorization

$$g = k \exp(S^c) \exp(S^a) \tag{P.17.2}$$

with $k \in SO(2n, \mathbb{R})$ and $\mathrm{tr}(S^c) = 0$. Also, we know from (15.3) and (15.20) that there is the representation

$$\exp(S^c) = O \exp(s_d) O^T, \tag{P.17.3}$$

so that we also have the factorization

$$g = kO \exp(s_d) O^T \exp(S^a). \tag{P.17.4}$$

Rewrite this relation in the form

$$g = [kO][\exp(s_d)][O^T \exp(S^a)]. \tag{P.17.5}$$

Note from (13.29) that $S^a \in \mathfrak{g}_{1,-1} = \mathfrak{h} \cap \mathfrak{p}$ and $S^a \cong JS^a$. Therefore,

$$\exp(S^a) \in Sp(2n, \mathbb{R}). \tag{P.17.6}$$

Also O is in both $SO(2n, \mathbb{R})$ and $Sp(2n, \mathbb{R})$. Consequently there are the following group relations,

$$kO \in SO(2n, \mathbb{R}), \tag{P.17.7}$$

$$O^T \exp(S^a) \in Sp(2n, \mathbb{R}). \tag{P.17.8}$$

Define group elements k', $\hat{s}_d$, and h by the rules

$$k' = kO, \tag{P.17.9}$$

$$\hat{s}_d = \exp(s_d), \tag{P.17.10}$$

$$h = O^T \exp(S^a). \tag{P.17.11}$$

Then, we may write

$$g = k'\hat{s}_d h, \tag{P.17.12}$$

which is a factorization of the form (17.1).

Consider the map

$$(K, S_d) \mapsto G \tag{P.17.13}$$

given by the rule

$$g(k, s_d) = k[\exp(s_d)]. \tag{P.17.14}$$

Then, we find that

$$g(k\ell,s_d)=k\ell[\exp(s_d)]=k[\exp(s_d)]\ell \tag{P.17.15}$$

where we have used (16.12). Also we know that

$$\ell\in SO(2n,\mathbb{R})\cap Sp(2n,\mathbb{R}) \tag{P.17.16}$$

and therefore

$$\ell\in Sp(2n,\mathbb{R}). \tag{P.17.17}$$

Thus, we may write (17.15) in the form

$$g(k\ell,s_d)=k\ell[\exp(s_d)]=k[\exp(s_d)]\ell=g(k,s_d)h \tag{P.17.18}$$

with

$$h=\ell \text{ and } h\in Sp(2n,\mathbb{R}). \tag{P.17.19}$$

We see that $g(k\ell,s_d)$ and $g(k,s_d)$ are in the same coset in the coset space G/H,

$$g(k\ell,s_d)\sim g(k,s_d) \bmod Sp(2n,\mathbb{R}). \tag{P.17.20}$$

Also, we know that $k\ell$ and k are in the same coset in the coset space K/L,

$$k\ell\sim k \bmod L. \tag{P.17.21}$$

Therefore (17.14) provides a map

$$(K/L,S_d)\mapsto G/H. \tag{P.17.22}$$

But, by (17.1), we know that every element of G/H can be obtained in this way. Thus, we conjecture that there is a correspondence of the form

$$[K/L]\times S_d^{+c}\leftrightarrow G/H. \tag{P.17.23}$$

Let us check dimensions. We want to check the relation

$$\dim K-\dim L+\dim S_d^{+c}=\dim G-\dim H. \tag{P.17.24}$$

We have the counts

$$\dim K=n(2n-1), \tag{P.17.25}$$

$$\dim L=n, \tag{P.17.26}$$

$$\dim S_d^{+c}=n-1, \tag{P.17.27}$$

$$\dim H=n(2n+1), \tag{P.17.28}$$

$$\dim G=(2n)^2-1. \tag{P.17.29}$$

Therefore we have to check the relation

$$n(2n-1)-n+(n-1)=[(2n)^2-1]-n(2n+1)? \tag{P.17.30}$$

A little algebra shows that both sides of (17.30) simplify to the expression $(2n^2 - n - 1)$ so that the relation does indeed hold.

We still have to check uniqueness. Suppose (k_1, s_{d1}) and (k_2, s_{d2}) are sent to g_1 and g_2 under (17.14),

$$k_1[\exp(s_{d1})] = g_1, \tag{P.17.31}$$

$$k_2[\exp(s_{d2})] = g_2. \tag{P.17.32}$$

Also suppose that

$$g_2 \sim g_1 \bmod H. \tag{P.17.33}$$

Then, we want to show that

$$k_2 \sim k_1 \bmod L \tag{P.17.34}$$

and

$$s_{d2} = s_{d1}. \tag{P.17.35}$$

Suppose (17.33) holds. Then there is an $h \in H$ such that

$$g_2 = g_1 h, \tag{P.17.36}$$

and therefore, from (17.31) and (17.32), there is the relation

$$k_2[\exp(s_{d2})] = k_1[\exp(s_{d1})]h. \tag{P.17.37}$$

To be continued.

Application to Dragt's Symplectic Polar Decompostion

P.18 Search for Counter Examples

In view of the results of the previous section, we will also get elements g' in all possible cosets $GL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$ by the the rule

$$g'(k, s'_d) = ks'_d \tag{P.18.1}$$

where s'_d is of the form (14.1) but d is no longer required to be traceless. Comparison of this rule with (17.14) shows that we may get some overlap by this procedure, but (18.1) appears easier to work with.

The search for counter examples can be begun with the case $k = I$ for which

$$g'(I, s'_d) = Is'_d = s'_d, \tag{P.18.2}$$

and we have found counter examples in the 4 × 4 matrix context. They demonstrate that symplectic polar decomposition of a matrix M is not possible globally even with the restriction $\det(M) > 0$. See Section 4.3.5 of *Lie Methods*.

We can next examine the case $s'_d = I$ for which the g' are matrices of the form

$$g'(k, I) = k. \tag{P.18.3}$$

When k is in the vicinity of the identity, g' will be near the identity, and therefore have a symplectic polar decomposition. But what happens when we consider all $k \in SO(2n,\mathbb{R})$? Exercise 4.3.19 of *Lie Methods* studies a particular one-parameter subgroup of $SO(4,\mathbb{R})$ and shows that for all such elements symplectic polar decomposition is always possible, but the ray$\lambda^2 N(M)$ need not always intersect the unit ball around I. Exercise 4.3.22 shows that all elements of $SO(4,\mathbb{R})$ have symplectic polar decompositions.

Let H be the subgroup consisting of all elements $g \in GL(2n,\mathbb{R},+)$ such that

$$\{g, J\} = 0. \tag{P.18.4}$$

We know that H is isomorphic to $GL(n,\mathbb{C})$. Exercise 4.3.21 shows that all elements of H have symplectic polar decompositions.

We should now think about more general elements of $GL(4,\mathbb{R})$. For example, Exercise 4.3.20 describes a one parameter closed path of elements that includes elements that do and do not have symplectic polar decompositions. It would be interesting to see where the decomposition first fails.

Also, suppose symplectic polar decomposition is possible for some matrix M. What can be said about matrices in the neighborhood of M? Similarly, suppose symplectic polar decomposition is impossible for some matrix M. What can be said then about matrices in the neighborhood of M? What is the nature of the transition from having to not having a symplectic polar decomposition?

Bibliography

[1] The essential contents of this Appendix, including all major insights, were provided in kind correspondence from Professor Roe Goodman.

[2] G. Heckman and H. Schlichtkrull, "Harmonic Analysis and Special Functions on Symmetric Spaces", *Perspectives in Mathematics* **16**, Academic Press (1994).

[3] G. Mostow, "Some New Decomposition Theorems for Semisimple Groups", *Memoirs of the Amer. Math. Soc.* **14**, 31-54 (1955).

[4] G. Hochschild, *The Structure of Lie Groups*, Holden–Day (1965).

Appendix Q

Improving Convergence of Fourier Representation

Q.1 Introduction

Suppose $f(u)$ is a function defined on the interval $u \in [0, 2\pi]$, and suppose f is continuous and has a continuous first derivative. What can be said about a Fourier representation of f over this interval? We observe that, by definition, a Fourier series produces a periodic function, and straight-forward application of Fourier's theorem to f produces a function with period 2π. But, f may not have a periodic extension unless some kind of singularity (say a discontinuity in f or one of its derivatives) is introduced at the points $u = 0, \pm 2\pi, \pm 4\pi, \cdots$. For example, if $f(0) \neq f(2\pi)$, the periodic extension of f cannot be continuous. The net effect of this discontinuity is that in this case the Fourier coefficients of f can fall off no faster than $(1/n)$ for large n. In Section 14.5 we saw that this situation can be improved somewhat by doubling the domain of definition for f and imposing an *evenness* condition on its extension. The net result is a modified Fourier representation over the domain $[-2\pi, 2\pi]$ whose coefficients fall off like $(1/n)^2$. The purpose of this appendix is to describe and apply a further trick that makes it possible to obtain a Fourier-like representation for which the coefficients fall off still faster.

Begin by writing f in the form

$$f(u) = c + [d/(2\pi)]u + g(u) \tag{Q.1.1}$$

where

$$c = f(0) \text{ and } d = f(2\pi) - f(0). \tag{Q.1.2}$$

Then the function $g(u)$ is also defined for $u \in [0, 2\pi]$, is continuous, and has a continuous first derivative. Moreover, it has the property

$$g(0) = g(2\pi) = 0. \tag{Q.1.3}$$

Extend g to the interval $[-2\pi, 0]$ by requiring that g be odd,

$$g(u) = -g(-u) \text{ for } u \in [-2\pi, 0]. \tag{Q.1.4}$$

Then, because of (1.3) and (1.4), the extended g is continuous at $u = 0$. Moreover, we find that

$$g(-2\pi) = -g(2\pi) = 0. \tag{Q.1.5}$$

Finally, from (1.4) we find that

$$g'(u) = g'(-u) \;\; \text{for} \;\; u \in [-2\pi, 0], \tag{Q.1.6}$$

from which it follows that the extended g' is continuous at $u = 0$. Thus, g has now been defined for $u \in [-2\pi, 2\pi]$, and is continuous and has a continuous first derivative in the open interval $u \in (-2\pi, 2\pi)$.

Further extend g to the full interval $u \in (-\infty, \infty)$ by requiring that g be periodic with period 4π,

$$g(u + 4\pi) = g(u). \tag{Q.1.7}$$

We will now see that this extension results in a g that is also continuous and has a continuous first derivative at the points $u = \pm 2\pi$ and their 4π periodic extensions. Thus the net result is that, by these extensions, g has been defined everywhere and is continuous and has a continuous first derivative everywhere. Let us check first the case $u = 2\pi$. From the definitions so far we have the relations

$$g(2\pi + \epsilon) = g(-2\pi + \epsilon) = -g(2\pi - \epsilon). \tag{Q.1.8}$$

In view of (1.3) and (1.8), continuity at $u = 2\pi$ has been established. Also, using periodicity and (1.6), we have the relations

$$g'(2\pi + \epsilon) = g'(-2\pi + \epsilon) = g'(2\pi - \epsilon), \tag{Q.1.9}$$

from which it follows that g has a continuous first derivative at $u = 2\pi$. Similarly, we find that

$$g(-2\pi - \epsilon) = g(2\pi - \epsilon) = -g(-2\pi + \epsilon) \tag{Q.1.10}$$

and

$$g'(-2\pi - \epsilon) = g'(2\pi - \epsilon) = g'(-2\pi + \epsilon), \tag{Q.1.11}$$

from which it follows that g is continuous and has a continuous first derivative at $u = -2\pi$.

We are now ready to invoke the results of Fourier. Since g is 4π periodic, it has an expansion over the interval $u \in [-2\pi, 2\pi]$ of the form

$$g(u) = \sum_{n=0}^{\infty} a_n \cos(nu/2) + \sum_{n=1}^{\infty} b_n \sin(nu/2). \tag{Q.1.12}$$

Since g is odd, all the a_n must vanish, and we are left with

$$g(u) = \sum_{n=1}^{\infty} b_n \sin(nu/2). \tag{Q.1.13}$$

The coefficients b_n are given by the integrals

$$b_n = [1/(2\pi)] \int_{-2\pi}^{2\pi} du \; g(u) \sin(nu/2) = (1/\pi) \int_{0}^{2\pi} du \; g(u) \sin(nu/2). \tag{Q.1.14}$$

Note that the integrals on the far right side of (1.14) depend only on the knowledge of g, and hence f, in the original interval $u \in [0, 2\pi]$. Finally, since g will generally have a discontinuous second derivative at the points $u = 0, \pm 2\pi, \pm 4\pi, \cdots$, the coefficients b_n will generally fall off according to the rule

$$b_n \sim (1/n)^3 \;\text{ as }\; n \to \infty. \tag{Q.1.15}$$

The net result of our efforts is that f has the representation

$$f(u) = c + [d/(2\pi)]u + \sum_{n=1}^{\infty} b_n \sin(nu/2) \tag{Q.1.16}$$

with the coefficients b_n obeying (1.15).

Consider the function $h(v)$ defined by

$$h(v) = f(v + \pi). \tag{Q.1.17}$$

It is defined on the interval $v \in [-\pi, \pi]$. According to (1.16) it has the expansion

$$\begin{aligned} h(v) &= c + [d/(2\pi)](v + \pi) + \sum_{n=1}^{\infty} b_n \sin[n(v + \pi)/2] \\ &= [h(\pi) + h(-\pi)]/2 + \{[h(\pi) - h(-\pi)]/(2\pi)\}v + \sum_{n=1}^{\infty} b_n \sin[n(v + \pi)/2]. \end{aligned} \tag{Q.1.18}$$

with

Q.2 Application

Bibliography

[1] ***

Appendix R

Abstract Lie Group Theory

The purpose of this appendix is to show that the Jacobi identity in a Lie algebra is related to the assumed associativity of group multiplication in the corresponding Lie group. When a Lie group is realized in terms of matrices, the associative condition for group multiplication is automatically satisfied because matrix multiplication is associative. Correspondingly, the Jacobi identity is readily verified for Lie algebras realized in terms of matrices with the Lie product taken to be the matrix commutator. Treating the case of an abstract Lie group requires somewhat more effort.

Bibliography

[1] Eugene Lerman, *Notes on Lie Groups* (2012).
https://faculty.math.illinois.edu/~lerman/519/s12/427notes.pdf
Observe that the equation in Corollary 6.7a should read
$\exp[(t_1 + t_2)X] = [\exp(t_1X)][\exp(t_2X)]$.

Appendix S

Mathematica Realization of TPSA and Taylor Map Computation

S.1 Background

The forward integration method (Section 10.12.4) for computing Taylor maps can be implemented by a code employing the tools of *automatic differentiation* (AD) described by Neidinger [1].[1] In this approach arrays of Taylor coefficients of various functions are referred to as AD variables or *pyramids* since, as will be seen, they have a hyper-pyramidal structure. Generally the first entry in the array will be the value of the function about some expansion point, and the remaining entries will be the higher-order Taylor coefficients about the expansion point and truncated beyond some specified order. Such truncated Taylor expansions are also commonly called *jets*. Recall Section 7.5.

In our application elements in these arrays will be addressed and manipulated with the aid of scalar indices and associated look-up and look-back tables generated at run time. We have also replaced the original APL implementation of Neidinger with a code written in the language of *Mathematica* (Version 6, or 7) [2,3]. Where necessary, for those unfamiliar with the details of *Mathematica*, we will explain the consequences of various *Mathematica* commands. Recall that we wish to integrate equations of the form

$$\dot{z}_a = f_a(\boldsymbol{z}, t), \quad a = 1, m \tag{S.1.1}$$

and their associated complete variational equations. The inputs to the code are the right sides (RS) of (1.1). Other input parameters are the number of variables m, the desired order of the Taylor map p, and the *initial* conditions $(z_a^d)^i$ for the design solution.

Various AD tools for describing and manipulating pyramids are outlined in Section S.2. There we show how pyramid operations are encoded in the case of polynomial RS, as needed, for example, for the Duffing equation. For brevity, we omit the cases of rational, fractional power, and transcendental RS. These cases can also be handled using various methods based on functional identities and known Taylor coefficients, or the differential equations that such

[1] Some authors refer to AD as *truncated power series algebra* (TPSA) since AD algorithms arise from manipulating multivariable truncated power series. Other authors refer to AD as *Differential Algebra* (DA). There is a substantial literature on this subject. See the Web site http://www.autodiff.org/.

functions obey along with certain recursion relations [1]. In Section S.3, based on the work of Section S.2, we in effect obtain and integrate numerically the complete variational equations (10.12.36) in pyramid form, i.e. valid for any map order and any number of variables. Section S.4 treats the specific case of the Duffing equation. A final Section S.5 describes in more detail the relation between integrating equations for pyramids and the complete variational equations.

S.2 AD Tools

This section describes how arithmetic expressions representing $f_a(\boldsymbol{z}, t)$, the right sides of (1.1) where $\boldsymbol{z}$ denotes the dependent variables, are replaced with expressions for arrays (pyramids) of Taylor coefficients. These pyramids in turn constitute the input to our code. Such an ad-hoc replacement, according to the problem at hand, as opposed to operator overloading where the kind of operation depends on the type of its argument, is also the approach taken in [1,4,5].

Let u, v, w be general arithmetic expressions, i.e. scalar-valued functions of $\boldsymbol{z}$. They contain various arithmetic operations such as addition/subtraction ($\pm$), multiplication ($*$), and raising to a power ($\wedge$). (They may also entail the computation of various transcendental functions such as the sine function, etc. However, as stated earlier, for simplicity we will omit these cases.) The arguments of these operations may be a constant, a single variable or multiple variables z_a, or even some other expression. The idea of AD is to redefine the arithmetic operations in such a way (see Definition 1), that all functions u, v, w can be consistently replaced with the arrays of coefficients of their Taylor expansions. For example, by redefining the usual product of numbers ($*$) and introducing the pyramid operation `PROD`, $u * v$ is replaced with `PROD[U,V]`.

We use upper typewriter font for pyramids (`U,V,...`) and for operations on pyramids (`PROD, POW, ...`). Everywhere, equalities written in typewriter fonts have equivalent *Mathematica* expressions. That is, they have associated realizations in *Mathematica* and directly correspond to various operations and commands in *Mathematica*. In effect, our code operates entirely on pyramids. However, as we will see, any pyramid expression contains, as its first entry, its usual arithmetic counterpart.

We begin with a description of our method of monomial labeling. In brief, we list all monomials in a polynomial in some sequence, and label them by where they occur in the list. Next follow Definition 1 and the recipes for encoding operations on pyramids. Subsequently, by using Definition 2, which simply states the rule by which an arithmetic expression is replaced with its pyramid counterpart, we show how a general expression can be encoded by using only the pyramids of a constant and those of the various variables involved.

S.2.1 Labeling Scheme

A monomial $G_{\boldsymbol{j}}(\boldsymbol{z})$ in m variables is of the form

$$G_{\boldsymbol{j}}(\boldsymbol{z}) = (z_1)^{j_1}(z_2)^{j_2}\cdots(z_m)^{j_m}. \tag{S.2.1}$$

Here we have introduced an exponent vector $\boldsymbol{j}$ by the rule

$$\boldsymbol{j} = (j_1, j_2, \cdots j_m). \tag{S.2.2}$$

Evidently $\boldsymbol{j}$ is an m-tuple of non-negative integers. The degree of $G_{\boldsymbol{j}}(\boldsymbol{z})$, denoted by $|\boldsymbol{j}|$, is given by the sum of exponents,

$$|\boldsymbol{j}| = j_1 + j_2 + \cdots + j_m. \tag{S.2.3}$$

The set of all exponents for monomials in m variables with degree less than or equal to p will be denoted by Γ_m^p,

$$\Gamma_m^p = \{\boldsymbol{j} \mid |\boldsymbol{j}| \leq p\}. \tag{S.2.4}$$

According to Section 32.1, this set has $L(m,p)$ entries with $L(m,p)$ given by a binomial coefficient,

$$L(m,p) = S_0(m,p) = \binom{p+m}{p}. \tag{S.2.5}$$

With this notation, a Taylor series expansion (about the origin) of a scalar-valued function u of m variables $\boldsymbol{z} = (z_1, z_2, \ldots z_m)$, truncated beyond terms of degree p, can be written in the form

$$u(\boldsymbol{z}) = \sum_{\boldsymbol{j} \,\in\, \Gamma_m^p} \mathtt{U}(\boldsymbol{j})\; G_{\boldsymbol{j}}(\boldsymbol{z}). \tag{S.2.6}$$

Assuming that m and p are fixed input variables, we will often simply write Γ and L. Here, for now, $\mathtt{U}$ simply denotes an array of numerical coefficients. When employed in code that has symbolic manipulation capabilities, each $\mathtt{U}(\boldsymbol{j})$ may also be a symbolic quantity.

To proceed, what is needed is some way of listing monomials systematically. With such a list, as described in Subsections 32.3.3 and 32.3.4, we may assign a label r to each monomial based on where it appears in the list. We will use a variant of *modified glex sequencing*, the only change being that we will begin the list with the monomial of degree 0. For example, Table 2.1 shows a list of monomials in three variables. As one goes down the list, first the monomial of degree $D = 0$ appears, then the monomials of degree $D = 1$, etc. Within each group of monomials of fixed degree the individual monomials appear in descending lex order. Note that Table 2.1 is similar to Table 32.2.4 except that it begins with the monomial of degree 0. Other possible listings include ascending true glex order in which monomials appear in ascending lex order within each group of degree D, and lex order for the whole monomial list as in [1].

Table S.2.1: A labeling scheme for monomials in three variables.

r	j_1	j_2	j_3	D
1	0	0	0	0
2	1	0	0	1
3	0	1	0	1
4	0	0	1	1
5	2	0	0	2
6	1	1	0	2
7	1	0	1	2
8	0	2	0	2
9	0	1	1	2
10	0	0	2	2
11	3	0	0	3
12	2	1	0	3
13	2	0	1	3
14	1	2	0	3
15	1	1	1	3
16	1	0	2	3
17	0	3	0	3
18	0	2	1	3
19	0	1	2	3
20	0	0	3	3
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
28	1	2	1	4
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

With the aid of the scalar index r the relation (2.6) can be rewritten in the form

$$u(\boldsymbol{z}) = \sum_{r=1}^{L(m,p)} \mathtt{U}(r) G_r(\boldsymbol{z}), \tag{S.2.7}$$

because (by construction and with fixed m) for each positive integer r there is a unique exponent $\boldsymbol{j}(r)$, and for each $\boldsymbol{j}$ there is a unique r. Here $\mathtt{U}$ may be viewed as a vector with entries $\mathtt{U}(\mathtt{r})$, and $G_r(\boldsymbol{z})$ denotes $G_{\boldsymbol{j}(r)}(\boldsymbol{z})$.

Consider, in an m-dimensional space, the points defined by the heads of the vectors $\boldsymbol{j} \in \Gamma_m^p$. See (2.4). Figure 2.1 displays them in the case $m = 3$ and $p = 4$. Evidently they form a grid that lies on the surface and interior of what can be viewed as an m-dimensional *pyramid* in m-dimensional space. At each grid point there is an associated coefficient $\mathtt{U}(\mathtt{r})$.

Because of its association with this pyramidal structure, we will refer to the entire set of coefficients in (2.6) or (2.7) as the *pyramid* $\mathtt{U}$ of $u(\boldsymbol{z})$.

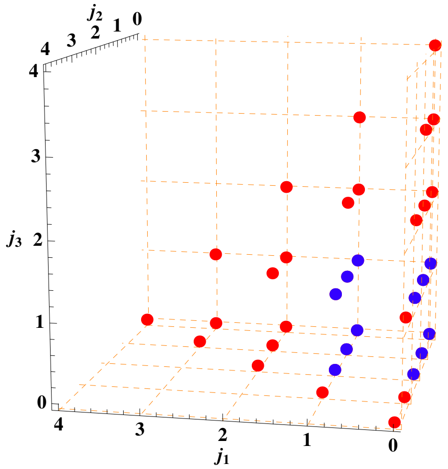


Figure S.2.1: A grid of points representing the set Γ_3^4. For future reference a subset of Γ_3^4, called a *box*, is shown in blue.

S.2.2 Implementation of Labeling Scheme

We have seen that use of modified glex sequencing, for any specified number of variables m, provides a labeling rule such that for each positive integer r there is a unique exponent $\boldsymbol{j}(r)$, and for each $\boldsymbol{j}$ there is a unique r. That is, there is a invertible function $r(\boldsymbol{j})$ that provides a 1-to-1 correspondence between the positive integers and the exponent vectors $\boldsymbol{j}$. To proceed further, it would be useful to have this function and its inverse in more explicit form.

From the work of Subsection 32.2.6, we already know a formula for $r(\boldsymbol{j})$ based on the Giorgilli formula (32.2.15),

$$r(\boldsymbol{j}) = r(j_1, \cdots j_m) = 1 + i(j_1, \cdots j_m). \tag{S.2.8}$$

Below is simple *Mathematica* code that implements this formula (which we call *Gfor*) in the case of three variables, and evaluates it for selected exponents $\boldsymbol{j}$. Observe that these evaluations agree with results in Table 2.1.

```
Gfor[j1_, j2_, j3_] := (
s1 = j3; s2 = 1 + j3 + j2; s3 = 2 + j3 + j2 + j1;
t1 = Binomial[s1, 1]; t2 = Binomial[s2, 2]; t3 = Binomial[s3, 3];
r = 1 + t1 + t2 + t3; r
)
Gfor[0, 0, 0]
Gfor[1, 0, 0]
Gfor[2, 0, 1]
Gfor[1, 2, 1]
1
2
13
28
```

(S.2.9)

For the inverse relation we have found it convenient to introduce a rectangular matrix associated with the set Γ_m^p. By abuse of notation, it will also be called Γ. It has $L(m,p)$ rows and m columns with entries

$$\Gamma_{r,a} = j_a(r). \tag{S.2.10}$$

For example, looking a Table 2.1, we see (when $m = 3$) that $\Gamma_{1,1} = 0$ and $\Gamma_{17,2} = 3$. Indeed, if the first and last columns of Table 2.1 are removed, what remains (when $m = 3$) is the matrix $\Gamma_{r,a}$. In the language of Subsection 32.2.9, Γ is a *look up table* that, given r, produces the associated $\boldsymbol{j}$. In our *Mathematica* implementation Γ is the matrix GAMMA with elements GAMMA[[r, a]].

The matrix GAMMA is constructed using the *Mathematica* code illustrated below,

```
Needs["Combinatorica`"];
m = 3; p = 4;
GAMMA = Compositions[0, m];
Do[GAMMA = Join[GAMMA, Reverse[Compositions[d, m]]], {d, 1, p, 1}];
L = Length[GAMMA]
r = 17; a = 2;
GAMMA[[r]]
GAMMA[[r, a]]
35
{0, 3, 0}
3
```

(S.2.11)

It employs the *Mathematica* commands Compositions, Reverse, and Join.

We will now describe the ingredients of this code and illustrate the function of each:

- The command Needs["Combinatorica‘"]; loads a combinatorial package.

- The command Compositions[i, m] produces, as a list of arrays (a rectangular array), all *compositions* (under addition) of the integer i into m integer parts. Furthermore, the compositions appear in *ascending* lex order. For example, the command Compositions[0, 3] produces the single row

  $$\begin{array}{ccc} 0 & 0 & 0 \end{array} \tag{S.2.12}$$

  As a second example, the command Compositions[1, 3] produces the rectangular array

  $$\begin{array}{ccc} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{array} \tag{S.2.13}$$

  As a third example, the command Compositions[2, 3] produces the rectangular array

  $$\begin{array}{ccc} 0 & 0 & 2 \\ 0 & 1 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 2 & 0 & 0 \end{array} \tag{S.2.14}$$

- The command Reverse acts on the list of arrays, and reverses the order of the list while leaving the arrays intact. For example, the nested sequence of commands Reverse[Compositions[1, 3]] produces the rectangular array

  $$\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \tag{S.2.15}$$

  As a second example, the nested sequence of commands Reverse[Compositions[2, 3]] produces the rectangular array

  $$\begin{array}{ccc} 2 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{array} \tag{S.2.16}$$

  Now the compositions appear in *descending* lex order.

- Look, for example, at Table 2.1. We see that the exponents j_a for the $r = 1$ entry are those appearing in (2.12). Next, exponents for the $r = 2$ through $r = 4$ entries are those appearing in (2.15). Following them, the exponents for the $r = 5$ through $r = 10$ entries, are those appearing in (2.16), etc. Evidently, to produce the exponent list of Table 2.1, what we must do is successively *join* various lists. That is what the *Mathematica* command `Join` accomplishes.

We are now ready to describe how `GAMMA` is constructed:

- The second line in (2.11) sets the values of m and p. They are assigned the values $m = 3$ and $p = 4$ for this example, which will construct `GAMMA` for the case of Table 2.1. The third line in (2.11) initially sets `GAMMA` to a row of m zeroes. The fourth line is a `Do` loop that successively redefines `GAMMA` by generating and joining to it successive descending lex order compositions. The net result is the exponent list of Table 2.1.

- The quantity $L = L(m, p)$ is obtained by applying the *Mathematica* command `Length` to the the rectangular array `GAMMA`.

- The last 6 lines of (2.11) illustrate that L is computed properly and that the command `GAMMA[[r, a]]` accesses the array `GAMMA` in the desired fashion. Specifically, in this example, we find from (2.5) that $L(3, 4) = 35$ in agreement with the *Mathematica* output for L. Moreover, `GAMMA[[17]]` produces the exponent array $\{0, 3, 0\}$, in agreement with the $r = 17$ entry in Table 2.1, and `GAMMA[[17, 2]]` produces $\Gamma_{17,2} = 3$, as expected.

S.2.3 Pyramid Operations: General Procedure

Here we *derive* the pyramid operations in terms of $\boldsymbol{j}$-vectors by using the ordering previously described, and provide scripts to *encode* them in the r-representation (2.7).

Definition 1. Suppose that $w(\boldsymbol{z})$ arises from carrying out various *arithmetic operations* on $u(\boldsymbol{z})$ and $v(\boldsymbol{z})$, and the associated pyramids `U` and `V` are known. The corresponding pyramid operation on `U` and `V` is so defined that it yields the pyramid `W` of $w(\boldsymbol{z})$.

Here we assume that u, v, w are polynomials such as (2.6).

S.2.4 Pyramid Operations: Scalar Multiplication and Addition

We begin with the operations of scalar multiplication and addition, which are easy to define and implement. If

$$w(\boldsymbol{z}) = c\, u(\boldsymbol{z}), \tag{S.2.17}$$

then

$$\mathtt{W}(r) = c\, \mathtt{U}(r), \tag{S.2.18}$$

and we write

$$\mathtt{W} = c\, \mathtt{U}. \tag{S.2.19}$$

If

$$w(\boldsymbol{z}) = u(\boldsymbol{z}) + v(\boldsymbol{z}), \tag{S.2.20}$$

then

$$\mathtt{W}(r) = \mathtt{U}(r) + \mathtt{V}(r), \tag{S.2.21}$$

and we write

$$\mathtt{W} = \mathtt{U} + \mathtt{V}. \tag{S.2.22}$$

In both cases all operations are performed coordinate-wise (as for vectors).

Implementation of scalar multiplication and vector addition is easy in *Mathematica* because, as the example below illustrates, it has built in vector routines. There we define two vectors, multiply them by scalars, and add the resulting vectors.

```
Unprotect[V];
U = {1, 2, 3};
V = {4, 5, 6};
W = .1U + .2V
{.9, 1.2, 1.5}                                   (S.2.23)
```

Since V is a "protected" symbol in the *Mathematica* language, and, for purposes of illustration, we wish to use it as an ordinary vector variable, it must first be unprotected as in line 1 above. The last line shows that the *Mathematica* output is indeed the desired result.

S.2.5 Pyramid Operations: Background for Polynomial Multiplication

The operation of polynomial multiplication is more involved. Now we have the relation

$$w(\boldsymbol{z}) = u(\boldsymbol{z}) * v(\boldsymbol{z}), \tag{S.2.24}$$

and we want to encode

$$\mathtt{W} = \mathtt{PROD}[\mathtt{U}, \mathtt{V}]. \tag{S.2.25}$$

Shown below is *Mathematica* code that implements this operation,

$$\mathtt{PROD}[\mathtt{U\_}, \mathtt{V\_}] := \mathtt{Table}[\mathtt{U}[[\mathtt{B}[[\mathtt{k}]]]] \cdot \mathtt{V}[[\mathtt{Brev}[[\mathtt{k}]]]], \{\mathtt{k}, \mathtt{1}, \mathtt{L}, \mathtt{1}\}]; \tag{S.2.26}$$

Our next task is to describe and explain the ingredients in (2.26).

Let us write $u(\boldsymbol{z})$ in the form (2.6), but with a change of dummy indices, so that it has the representation

$$u(\boldsymbol{z}) = \sum_{\boldsymbol{i} \,\in\, \Gamma_m^p} \mathtt{U}(\boldsymbol{i})\; G_{\boldsymbol{i}}(\boldsymbol{z}). \tag{S.2.27}$$

Similarly, write $v(\boldsymbol{z})$ in the form

$$v(\boldsymbol{z}) = \sum_{\boldsymbol{j} \,\in\, \Gamma_m^p} \mathtt{V}(\boldsymbol{j})\; G_{\boldsymbol{j}}(\boldsymbol{z}). \tag{S.2.28}$$

Then there is the result

$$u(\boldsymbol{z}) * v(\boldsymbol{z}) = \sum_{\boldsymbol{i} \,\in\, \Gamma_m^p} \; \sum_{\boldsymbol{j} \,\in\, \Gamma_m^p} \mathtt{U}(\boldsymbol{i})\mathtt{V}(\boldsymbol{j}) G_{\boldsymbol{i}}(\boldsymbol{z}) * G_{\boldsymbol{j}}(\boldsymbol{z}). \tag{S.2.29}$$

From (2.1) we observe that

$$\begin{aligned} G_{\boldsymbol{i}}(\boldsymbol{z}) * G_{\boldsymbol{j}}(\boldsymbol{z}) &= (z_1)^{i_1}(z_2)^{i_2}\cdots(z_m)^{i_m} * (z_1)^{j_1}(z_2)^{j_2}\cdots(z_m)^{j_m} \\ &= (z_1)^{i_1+j_1}(z_2)^{i_2+j_2}\cdots(z_m)^{i_m+j_m} = G_{\boldsymbol{i}+\boldsymbol{j}}(\boldsymbol{z}). \end{aligned} \tag{S.2.30}$$

Therefore, we may also write

$$u(\boldsymbol{z}) * v(\boldsymbol{z}) = \sum_{\boldsymbol{i}\,\in\,\Gamma_m^p}\ \sum_{\boldsymbol{j}\,\in\,\Gamma_m^p} \mathtt{U}(\boldsymbol{i})\mathtt{V}(\boldsymbol{j})G_{\boldsymbol{i}+\boldsymbol{j}}(\boldsymbol{z}). \tag{S.2.31}$$

Now we see that there are two complications. First, there may be terms on the right side of (2.31) whose degree is higher than p and therefore need not be computed. Second, there are generally many terms on the right side of (2.31) that contribute to a given monomial term in $w(\boldsymbol{z}) = u(\boldsymbol{z}) * v(\boldsymbol{z})$. Suppose we write

$$w(\boldsymbol{z}) = \sum_{\boldsymbol{k}} \mathtt{W}(\boldsymbol{k})\ G_{\boldsymbol{k}}(\boldsymbol{z}). \tag{S.2.32}$$

Upon comparing (2.31) and (2.32) we conclude that there is the multidimensional *Cauchy* product rule

$$\mathtt{W}(\boldsymbol{k}) = \sum_{\boldsymbol{i}+\boldsymbol{j}=\boldsymbol{k}} \mathtt{U}(\boldsymbol{i})\mathtt{V}(\boldsymbol{j}) = \sum_{\boldsymbol{j}\le\boldsymbol{k}} \mathtt{U}(\boldsymbol{k}-\boldsymbol{j})\mathtt{V}(\boldsymbol{j}). \tag{S.2.33}$$

Here, by $\boldsymbol{j} \le \boldsymbol{k}$, we mean that the sum ranges over all $\boldsymbol{j}$ such that $j_a \le k_a$ for all $a \in [1, m]$. That is,

$$\boldsymbol{j} \le \boldsymbol{k} \ \Leftrightarrow\ j_a \le k_a \text{ for all } a \in [1, m]. \tag{S.2.34}$$

Evidently, to implement the relation (2.33) in terms of r labels, we need to describe the exponent relation $\boldsymbol{j} \le \boldsymbol{k}$ in terms of r labels. Suppose $\boldsymbol{k}$ is some exponent vector with label $r(\boldsymbol{k})$ as, for example, in Table 2.1. Introduce the notation

$$k = r(\boldsymbol{k}). \tag{S.2.35}$$

This notation may be somewhat confusing because k is not the norm of the vector $\boldsymbol{k}$, but rather the label associated with $\boldsymbol{k}$. However, this notation is very convenient. Now, given a label k, we can find $\boldsymbol{k}$. Indeed, from (2.10), we have the result

$$k_a = \Gamma_{k,a}. \tag{S.2.36}$$

Having found $\boldsymbol{k}$, we define a set of exponents B_k by the rule

$$B_k = \{\boldsymbol{j} | \boldsymbol{j} \le \boldsymbol{k}\}. \tag{S.2.37}$$

This set of exponents is called the k^{th} *box*. Note that the heads of the vectors $\boldsymbol{j}$ that satisfy (2.37) for some fixed vector $\boldsymbol{k}$ do indeed lie within some hyper-rectangular volume (box). For example (when $m = 3$), suppose $k = 28$. Then we see from Table 2.1 that $\boldsymbol{k} = (1, 2, 1)$. Table 2.2 lists, in modified glex order, all the vectors in B_{28}, i.e. all vectors $\boldsymbol{j}$ such that

Table S.2.2: The vectors in $B_{28} = \{\boldsymbol{j} | \boldsymbol{j} \leq (1,2,1)\}$.

r	j_1	j_2	j_3	D
1	0	0	0	0
2	1	0	0	1
3	0	1	0	1
4	0	0	1	1
6	1	1	0	2
7	1	0	1	2
8	0	2	0	2
9	0	1	1	2
14	1	2	0	3
15	1	1	1	3
18	0	2	1	3
28	1	2	1	4

$\boldsymbol{j} \leq (1,2,1)$. These are the vectors whose heads are shown in blue in Figure 2.1. Finally, with this notation, we can rewrite (2.33) in the form

$$\mathtt{W}(\boldsymbol{k}) = \sum_{\boldsymbol{j} \in B_k} \mathtt{U}(\boldsymbol{k} - \boldsymbol{j}) \mathtt{V}(\boldsymbol{j}). \tag{S.2.38}$$

What can be said about the vectors $(\boldsymbol{k} - \boldsymbol{j})$ as $\boldsymbol{j}$ ranges over B_ℓ? Table 2.3 lists, for example, the vectors $\boldsymbol{j} \in B_{28}$ and the associated vectors $\boldsymbol{i}$ with $\boldsymbol{i} = (\boldsymbol{k} - \boldsymbol{j})$. Also listed are the labels $r(\boldsymbol{j})$ and $r(\boldsymbol{i})$. Compare columns 2,3,4, which specify the $\boldsymbol{j} \in B_{28}$, with columns 5,6,7, which specify the associated $\boldsymbol{i}$ vectors. We see that every vector that appears in the $\boldsymbol{j}$ list also occurs somewhere in the $\boldsymbol{i}$ list, and vice versa. This to be expected because the operation of multiplication is commutative: we can also write (2.33) in the form

$$\mathtt{W}(\boldsymbol{k}) = \sum_{\boldsymbol{j} \in B_k} \mathtt{U}(\boldsymbol{j}) \mathtt{V}(\boldsymbol{k} - \boldsymbol{j}). \tag{S.2.39}$$

We also observe the more remarkable feature that the two lists are *reverses* of each other: running down the $\boldsymbol{j}$ list gives the same vectors as running up the $\boldsymbol{i}$ list, and vice versa. This feature is a consequence of our ordering procedure.

As indicated earlier, what we really want is a version of (2.33) that involves labels instead of exponent vectors. Looking at Table 2.3, we see that this is easily done. We may equally well think of B_k as containing a collection of labels $r(\boldsymbol{j})$, and we may introduce a *reversed* array $Brev_k$ of *complementary* labels $r^c(\boldsymbol{j})$ where

$$r^c(\boldsymbol{j}) = r(\boldsymbol{i}). \tag{S.2.40}$$

That is, for example, B_{28} would consist of the first column of Table 2.3 and $Brev_{28}$ would consist of the last column of Table 2.3. Finally, we have already introduced k as being the label associated with $\boldsymbol{k}$. We these understandings in mind, we may rewrite (2.33) in the label form

$$\mathtt{W}(k) = \sum_{r \in B_k} \mathtt{U}(r^c)\mathtt{V}(r) = \sum_{r \in B_k} \mathtt{U}(r)\mathtt{V}(r^c). \tag{S.2.41}$$

This is the rule $\mathtt{W} = \mathtt{PROD}[\mathtt{U},\mathtt{V}]$ for multiplying pyramids. In the language of Section 32.7, B_k and $Brev_k$, when taken together, provide a *look back table* that, given k, look back to find all monomial pairs with labels r, r^c which produce, when multiplied, the monomial with label k.

Table S.2.3: The vectors $\boldsymbol{j}$ and $\boldsymbol{i} = (\boldsymbol{k} - \boldsymbol{j})$ for $\boldsymbol{j} \in B_{28}$ and $k_a = \Gamma_{28,a}$.

$r(\boldsymbol{j})$	j_1	j_2	j_3	i_1	i_2	i_3	$r(\boldsymbol{i})$
1	0	0	0	1	2	1	28
2	1	0	0	0	2	1	18
3	0	1	0	1	1	1	15
4	0	0	1	1	2	0	14
6	1	1	0	0	1	1	9
7	1	0	1	0	2	0	8
8	0	2	0	1	0	1	7
9	0	1	1	1	1	0	6
14	1	2	0	0	0	1	4
15	1	1	1	0	1	0	3
18	0	2	1	1	0	0	2
28	1	2	1	0	0	0	1

S.2.6 Pyramid Operations: Implementation of Multiplication

The code shown below in (2.42) illustrates how B_k and $Brev_k$ are constructed using *Mathematica.*

```
JSK[list_, K_] :=
Position[Apply[And, Thread[#1<=#2&[#,K]]]& /@ list, True]//Flatten;
B = Table[JSK[GAMMA, GAMMA[[k]]], {k, 1, L}];
Brev = Reverse /@ B;                                        (S.2.42)
```

As before, some explanation is required. The main tasks are to implement the $\boldsymbol{j} \leq \boldsymbol{k}$ operation (2.34) and then to employ this implementation. We will begin by implementing the $\boldsymbol{j} \leq \boldsymbol{k}$ operation. Several steps are required, and each of them is described briefly below:

- When *Mathematica* is presented with a statement of the form $j <= k$, with j and k being *integers*, it replies with the answer True or the answer False. (Here $j <= k$

denotes $j \leq k$.) Two sample *Mathematica* runs are shown below:

```
3 <= 4
True                                        (S.2.43)
```

```
5 <= 4
False                                       (S.2.44)
```

- A *Mathematica* function can be constructed that does the same thing. It takes the form

```
#1 <= #2 & [j, k]                           (S.2.45)
```

Here the symbols #1 and #2 set up two *slots* and the symbol & means the operation to its left is to be regarded as a function and is to be applied to the arguments to its right by inserting the arguments into the slots. Below is a *Mathematica* run illustrating this feature.

```
j = 3; k = 4;
#1 <= #2 & [j, k]
True                                        (S.2.46)
```

Observe that the output of this run agrees with that of (2.43).

- The same operation can be performed on pairs of *arrays* (rather than pairs of numbers) in such a way that corresponding entries from each array are compared, with the output then being an array of True and False answers. This is done using the *Mathematica* command Thread. Below is a *Mathematica* run illustrating this feature.

```
j = {1, 2, 3}; k = {4, 5, 1};
Thread[#1 <= #2 & [j, k]]
{True, True, False}                         (S.2.47)
```

Note that the first two answers in the output array are True because the statements $1 \leq 4$ and $2 \leq 5$ are true. The last answer in the output array is False because the statement $3 \leq 1$ is false.

- Suppose, now, that we are given two arrays $\boldsymbol{j}$ and $\boldsymbol{k}$ and we want to determine if $\boldsymbol{j} \leq \boldsymbol{k}$ in the sense of (2.34). This can be done by *applying* the logical And operation (using the *Mathematica* command Apply) to the True/False output array described above. Below is a *Mathematica* run illustrating this feature.

```
j = {1, 2, 3}; k = {4, 5, 1};
Apply[And, Thread[#1 <= #2 & [j, k]]]
False                                       (S.2.48)
```

Note that the output answer is False because at least one of the entries in the output array in (2.47) is False. The output answer would be True if, and only if, all entries in the output array in (2.47) were True.

- Now that the $\boldsymbol{j} \le \boldsymbol{k}$ operation has been defined for two exponent arrays, we would like to construct a related operator/function, to be called `JSK`. (Here the letter `S` stands for *smaller than or equal to*.) It will depend on the exponent array $\boldsymbol{k}$, and its task will be to search a list of exponent arrays to find those $\boldsymbol{j}$ within it that satisfy $\boldsymbol{j} \le \boldsymbol{k}$. The first step in this direction is to slightly modify the function appearing in (2.48). Below is a *Mathematica* run that specifies this modified function and illustrates that it has the same effect.

```
j = {1, 2, 3}; k = {4, 5, 1};
Apply[And, Thread[#1 <= #2 & [#, k]]] & [j]
False                                                  (S.2.49)
```

  Comparison of the functions in (2.48) and (2.49) reveals that what has been done is to replace the argument j in (2.48) by a slot `#`, then follow the function by the character `&`, and finally add the symbols [j]. What this modification does is to redefine the function in such a way that it acts on what follows the second `&`.

- The next step is to extend the function appearing in (2.49) so that it acts on a list of exponent arrays. To do this, we replace the symbols [j] by the symbols `/@ list`. The symbols `/@` indicate that what stands to their left is to act on what stands to their right, and what stands to their right is a list of exponent arrays. The result of this action will be a list of True/False results with one result for each exponent array in the list. Below is a *Mathematica* run that illustrates how the further modified function acts on lists.

```
k = {4, 5, 1};
ja = {3, 4, 1}; jb = {1, 2, 3}; jc = {1, 2, 1};
list = {ja, jb, jc};
Apply[And, Thread[#1 <= #2 & [#, k]]] & /@ list
{True, False, True}                                    (S.2.50)
```

  Observe that the output answer list is $\{\text{True}, \text{False}, \text{True}\}$ because $\{3, 4, 1\} \le \{4, 5, 1\}$ is True, $\{1, 2, 3\} \le \{4, 5, 1\}$ is False, and $\{1, 2, 1\} \le \{4, 5, 1\}$ is True.

- What we would really like to know is where the True items are in the list, because that will tell us where the $\boldsymbol{j}$ that satisfy $\boldsymbol{j} \le \boldsymbol{k}$ reside. This can be accomplished by use of the *Mathematica* command `Position` in conjunction with the result True. Below is a *Mathematica* run that illustrates how this works.

```
k = {4, 5, 1};
ja = {3, 4, 1}; jb = {1, 2, 3}; jc = {1, 2, 1};
list = {ja, jb, jc};
Position[Apply[And, Thread[#1 <= #2 & [#, k]]] & /@ list, True]
{{1}, {3}}                                             (S.2.51)
```

Note that the output is an array of positions in the list for which $\boldsymbol{j} \leq \boldsymbol{k}$. There is, however, still one defect. Namely, the output array is an array of single-element subarrays, and we would like it to be simply an array of location numbers. This defect can be remedied by appending the *Mathematica* command `Flatten`, preceded by `//`, to the instruction string in (2.51). The *Mathematica* run below illustrates this modification.

```
k = {4,5,1};
ja = {3,4,1}; jb = {1,2,3}; jc = {1,2,1};
list = {ja,jb,jc};
Position[Apply[And,Thread[#1 <= #2 & [#,k]]] & /@ list,True]//Flatten
{1, 3}                                                          (S.2.52)
```

Now the output is a simple array containing the positions in the list for which $\boldsymbol{j} \leq \boldsymbol{k}$.

- The last step is to employ the ingredients in (2.52) to define the operator `JSK[list,k]`. The *Mathematica* run below illustrates how this can be done.

```
k = {4,5,1};
ja = {3,4,1}; jb = {1,2,3}; jc = {1,2,1};
list = {ja,jb,jc};
JSK[list_,k_] :=
Position[Apply[And,Thread[#1 <= #2 & [#,k]]] & /@ list,True]//Flatten;
JSK[list, k]
{1, 3}                                                          (S.2.53)
```

Lines 4 and 5 above define the operator `JSK[list,k]`, line 6 invokes it, and line 7 displays its output, which agrees with the output of (2.52).

- With the operator `JSK[list,k]` in hand, we are prepared to construct tables B and *Brev* that will contain the B_k and the $Brev_k$. The *Mathematica* run below illustrates how this can be done.

```
B = Table[JSK[GAMMA,GAMMA[[k]]],{k,1,L,1}];
Brev = Reverse /@ B;
B[[8]]
Brev[[8]]
B[[28]]
Brev[[28]]
{1,3,8}
{8,3,1}
{1,2,3,4,6,7,8,9,14,15,18,28}
{28,18,15,14,9,8,7,6,4,3,2,1}                        (S.2.54)
```

The first line employs the *Mathematica* command `Table` in combination with an implied Do loop to produce a two-dimensional array `B`. Values of k in the range $[1, L]$ are selected sequentially. For each k value the associated exponent array $\boldsymbol{k}(k) =$ `GAMMA[[k]]` is obtained. The operator `JSK` then searches the full `GAMMA` array to find the list of r values associated with the $\boldsymbol{j} \leq \boldsymbol{k}$. All these r values are listed in a row. Thus, the array `B` consists of list of L rows, of varying width. The rows are labeled by $k \in [1, L]$, and in each row are the r values associated with the $\boldsymbol{j} \leq \boldsymbol{k}$. In the second line the *Mathematica* command `Reverse` is applied to `B` to produce a second array called `Brev`. Its rows are the reverse of those in `B`. For example, as the *Mathematica* run illustrates, `B[[8]]`, which is the 8^{th} row of `B`, contains the list $\{1, 3, 8\}$, and `Brev[[8]]` contains the list $\{8, 3, 1\}$. Inspection of the $r = 8$ monomial in Table 2.1, that with exponents $\{0, 2, 0\}$, shows that it has the monomials with exponents {0,0,0}, {0,1,0}, and {0,2,0} as factors. And further inspection of Table 2.1 shows that the exponents of these factors have the r values $\{1, 3, 8\}$. Similarly `B[[28]]`, which is the 28^{th} row of B, contains the same entries that appear in the first column of Table 2.3. And `Brev[[28]]`, which is the 28^{th} row of *Brev*, contains the same entries that appear in the last column of Table 2.3.

Finally, we need to explain how the arrays B and *Brev* can be employed to carry out polynomial multiplication. This can be done using the *Mathematica* dot product command:

- The exhibit below shows a simple *Mathematica* run that illustrates the use of the dot product command.

```
Unprotect[V];
U = {.1, .2, .3, .4, .5, .6, .7, .8};
V = {1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8};
U.V
u = {1, 3, 5};
v = {6, 4, 2};
U[[u]]
V[[v]]
U[[u]].V[[v]]
5.64
{.1, .3, .5}
{1.6, 1.4, 1.2}
1.18
```

(S.2.55)

As before, `V` must be unprotected. See line 1. The rest of the first part this run (lines 2 through 4) defines two vectors `U` and `V` and then computes their dot product. Note that if we multiply the entries in `U` and `V` pairwise and add, we get the result

$$.1 \times 1.1 + .2 \times 1.2 + \cdots + .8 \times 1.8 = 5.64,$$

which agrees with the *Mathematica* result for $\mathtt{U}\cdot\mathtt{V}$. See line 10.

The second part of this *Mathematica* run, lines 5 through 9, illustrates a powerful feature of the *Mathematica* language. Suppose, as illustrated, we define two arrays `u` and `v` of integers, and use these arrays as *arguments* for the vectors by writing `U[[u]]` and `V[[v]]`. Then *Mathematica* uses the integers in the two arrays `u` and `v` as labels to select the corresponding entries in `U` and `V`, and from these entries it makes new corresponding vectors. In this example, the 1^{st}, 3^{rd}, and 5^{th} entries in `U` are .1, .3, and .5. And the 6^{th}, 4^{th}, and 2^{nd} entries in `V` are 1.6, 1.4, and 1.2. Consequently, we find that

$$\mathtt{U}[[\mathtt{u}]] = \{.1, .3, .5\},$$

$$\mathtt{V}[[\mathtt{v}]] = \{1.6, 1.4, 1.2\},$$

in agreement with lines 11 and 12 of the *Mathematica* results. Correspondingly, we expect that $\mathtt{U}[[\mathtt{u}]]\cdot\mathtt{V}[[\mathtt{v}]]$ will have the value

$$\mathtt{U}[[\mathtt{u}]]\cdot\mathtt{V}[[\mathtt{v}]] = .1\times 1.6 + .3\times 1.4 + .5\times 1.2 = 1.18,$$

in agreement with the last line of the *Mathematica* output.

- Now suppose, as an example, that we set $k = 8$ and use `B[[k]]` and `Brev[[k]]` in place of the arrays `u` and `v`. The *Mathematica* fragment below shows what happens when this is done.

```
k = 8;
B[[k]]
Brev[[k]]
U[[B[[k]]]]
V[[Brev[[k]]]]
U[[B[[k]]]] · V[[Brev[[k]]]]
{1, 3, 8}
{8, 3, 1}
{.1, .3, .8}
{1.8, 1.3, 1.1}
1.45                                   (S.2.56)
```

From (2.54) we see that $\mathtt{B}[[8]] = \{1, 3, 8\}$ and $\mathtt{Brev}[[8]] = \{8, 3, 1\}$ in agreement with lines 7 and 8 of the *Mathematica* output above. Also, the 1^{st}, 3^{rd}, and 8^{th} entries in `U` are .1, .3, and .8. And the 8^{th}, 3^{rd}, and 1^{st} entries in `V` are 1.8, 1.3, and 1.1. Therefore we expect the results

$$\mathtt{U}[[\mathtt{B}[[\mathtt{k}]]]] = \{.1, .3, .8\},$$

$$\mathtt{V}[[\mathtt{Brev}[[\mathtt{k}]]]] = \{1.8, 1.3, 1.1\},$$

$$\mathtt{U}[[\mathtt{B}[[\mathtt{k}]]]]\cdot\mathtt{V}[[\mathtt{Brev}[[\mathtt{k}]]]] = .1\times 1.8 + .3\times 1.3 + .8\times 1.1 = 1.45,$$

in agreement with the last three lines of (2.56).

- Finally, suppose we carry out the operation $\texttt{U[[B[[k]]]]}\cdot\texttt{V[[Brev[[k]]]]}$ for all $k \in [1, L]$ and put the results together in a Table with entries labeled by k. According to (2.41), the result will be the pyramid for the product of the two polynomials whose individual pyramids are `U` and `V`. The *Mathematica* fragment (2.26), which is displayed again below, shows how this can be done to define a *product* function, called `PROD`, that acts on general pyramids `U` and `V`, using the command Table with an implied Do loop over k.

$$\texttt{PROD[U\_, V\_] := Table[U[[B[[k]]]]} \cdot \texttt{V[[Brev[[k]]]], \{k, 1, L, 1\}];}$$

Let us verify that this whole multiplication procedure works for a simple example. For the sake of brevity, we will consider the case of $m = 2$ variables and work through terms of degree $p = 3$. In this case pyramids have the modest length $L(2, 3) = 10$. Table 2.4 provides a labeling scheme for monomials in two variables using our standard modified glex sequencing.

Table S.2.4: A labeling scheme for monomials in two variables.

r	j_1	j_2
1	0	0
2	1	0
3	0	1
4	2	0
5	1	1
6	0	2
7	3	0
8	2	1
9	1	2
10	0	3
.	.	.
.	.	.

Suppose, for example, that u and v are the functions

$$u(\boldsymbol{z}) = 1 + 2z_1 + 3z_2 + 4z_1z_2 \tag{S.2.57}$$

and

$$v(\boldsymbol{z}) = 5 + 6z_1 + 7z_2^2. \tag{S.2.58}$$

From Table 2.4 we find that the corresponding pyramids `U` and `V` are

$$\texttt{U} = \{1, 2, 3, 0, 4, 0, 0, 0, 0, 0\} \tag{S.2.59}$$

and

$$\texttt{V} = \{5, 6, 0, 0, 0, 7, 0, 0, 0, 0\}. \tag{S.2.60}$$

Polynomial multiplication gives the result

$$\begin{aligned} w(\boldsymbol{z}) &= u(\boldsymbol{z}) * v(\boldsymbol{z}) \\ &= 5 + 16z_1 + 15z_2 + 12z_1^2 + 38z_1z_2 + 7z_2^2 + 24z_1^2z_2 + 14z_1z_2^2 + 21z_2^3 + 28z_1z_2^3. \end{aligned} \tag{S.2.61}$$

Correspondingly, through terms of degree 3, the pyramid $\mathtt{W} = \mathtt{PROD}[\mathtt{U}, \mathtt{V}]$ is given by

$$\mathtt{W} = \{5, 16, 15, 12, 38, 7, 0, 24, 14, 21\}. \tag{S.2.62}$$

Below is an execution of a *Mathematica* program illustrating the use of the product function for the polynomials u and v given by (2.57) and (2.58).

```
Clear["Global‘ * "];
Needs["Combinatorica‘"];
m = 2; p = 3;
GAMMA = Compositions[0, m];
Do[GAMMA = Join[GAMMA, Reverse[Compositions[d, m]]], {d, 1, p, 1}];
L = Length[GAMMA]
JSK[list_, k_] :=
Position[Apply[And, Thread[#1 <= #2 & [#, k]]] & /@ list, True]//Flatten;
B = Table[JSK[GAMMA, GAMMA[[r]]], {r, 1, L, 1}];
Brev = Reverse/@ B;
PROD[U_, V_] := Table[U[[B[[k]]]].V[[Brev[[k]]]], {k, 1, L, 1}];
U = {1, 2, 3, 0, 4, 0, 0, 0, 0, 0};
V = {5, 6, 0, 0, 0, 7, 0, 0, 0, 0};
10
PROD[U, V]
{5, 16, 15, 12, 38, 7, 0, 24, 14, 21}                                    (S.2.63)
```

The first 11 lines of the code set up the necessary arrays and define the product function in pyramid form. The next two lines specify the pyramids U and V given in (2.59) and (2.60). The third line from the bottom, which results from the command in line 6, illustrates that indeed $L(2, 3) = 10$. The final two lines show that use of the product function when applied to the pyramids U and V does indeed product the pyramid W given by (2.62).

S.2.7 Pyramid Operations: Implementation of Powers

With operation of multiplication in hand, it is easy to implement the operation of raising a pyramid to a power. The code shown below in (2.64) demonstrates how this can be done.

$$\begin{aligned}&\mathtt{POWER[U\_,0] := C1;}\\&\mathtt{POWER[U\_,1] := U;}\\&\mathtt{POWER[U\_,2] := PROD[U,U];}\\&\mathtt{POWER[U\_,3] := PROD[U,POWER[U,2]];}\\&\ldots\end{aligned}\qquad\text{(S.2.64)}$$

Here C1 is the pyramid for the Taylor series having *one* as its *constant* term and all other terms zero,

$$\mathtt{C1} = \{\mathtt{1,0,0,0},\cdots\}.\qquad\text{(S.2.65)}$$

It can be set up by the *Mathematica* code

$$\mathtt{C1 = Table[KroneckerDelta[k,1],\{k,1,L,1\}];}\qquad\text{(S.2.66)}$$

which employs the Table command, the Kronecker delta function, and an implied Do loop over k. This code should be executed before executing (2.64), but after the value of L has been established.

S.2.8 Replacement Rule and Automatic Differentiation

Definition 2. The transformation $A(\boldsymbol{z}) \rightsquigarrow \mathtt{A}$ means replacement of every real variable z_a in the *arithmetic expression* $A(\boldsymbol{z})$ with an associated pyramid, and of every operation on real variables in $A(\boldsymbol{z})$ with the associated operation on pyramids.

Automatic differentiation is based on the following corollary: if $A(\boldsymbol{z}) \rightsquigarrow \mathtt{A}$, then $\mathtt{A}$ is the pyramid of $A(\boldsymbol{z})$.

For simplicity, we will begin our discussion of the replacement rule with examples involving only a single variable z. In this case monomial labeling, the relation between labels and exponents, is given directly by the simple rules

$$r(j) = 1 + j \text{ and } j(r) = r - 1.\qquad\text{(S.2.67)}$$

See Table 2.5.

As a first example, consider the expression

$$A = 2 + 3(z * z).\qquad\text{(S.2.68)}$$

We have agreed to consider the case $m = 1$. Suppose we also set $p = 2$, in which case $L = 3$. In ascending glex order, see Table 2.5, the pyramid for A is then

$$2 + 3z^2 \rightsquigarrow \mathtt{A} = (2, 0, 3).\qquad\text{(S.2.69)}$$

Now imagine that A was not such a simple polynomial, but some complicated expression. Then the pyramid $\mathtt{A}$ could be generated by computing derivatives of A at $z = 0$ and dividing

Table S.2.5: A labeling scheme for monomials in one variable.

r	j
1	0
2	1
3	2
4	3
.	.
.	.

them by the appropriate factorials. Automatic differentiation offers another way to find $\mathtt{A}$. Assume that all operations in the arithmetic expression A have been encoded according to *Definition 1*. For our example, these are + and $\mathtt{PROD}$. Let $\mathtt{C1}$ and $\mathtt{Z}$ be the pyramids associated with 1 and z,

$$1 \rightsquigarrow \mathtt{C1} = (1, 0, 0), \tag{S.2.70}$$

$$z \rightsquigarrow \mathtt{Z} = (\mathtt{0}, \mathtt{1}, \mathtt{0}). \tag{S.2.71}$$

The quantity $2+3z^2$ results from performing various arithmetic operations on 1 and z. *Definition 1* says that the pyramid of $2+3z^2$ is identical to the pyramid obtained by performing the same operations on the pyramids $\mathtt{C1}$ and $\mathtt{Z}$. That is, suppose we replace 1 and z with their associated pyramids $\mathtt{C1}$ and $\mathtt{Z}$, and also replace $*$ with $\mathtt{PROD}$. Then, upon evaluating $\mathtt{PROD}$, multiplying by the appropriate scalar coefficients, and summing, the result will be the same pyramid $\mathtt{A}$,

$$2\ \mathtt{C1} + 3\ \mathtt{PROD}[\mathtt{Z}, \mathtt{Z}] = \mathtt{A}. \tag{S.2.72}$$

In this way, by knowing only the basic pyramids $\mathtt{C1}$ and $\mathtt{Z}$ (prepared beforehand), one can compute the pyramid of an arbitrary $A(z)$. Finally, in contrast to numerical differentiation, all numerical operations involved are accurate to machine precision. *Mathematica* code that implements (2.72) will be presented shortly in (2.73).

Frequently, if $A(z)$ is some complicated expression, the replacement rule will result in a long chain of nested pyramid operations. At every step in the chain the present pyramid, the pyramid resulting from the previous step, will be combined with some other pyramid to produce a new pyramid. Each such operation has two arguments (the present pyramid and some other pyramid), and *Definition 1* applies to each step in the chain. Upon evaluating all pyramid operations, the final result will be the pyramid of $A(z)$.

By using the replacement operation the above procedure can be represented as:

$$1 \rightsquigarrow \mathtt{C1}, \quad z \rightsquigarrow \mathtt{Z}, \quad A \rightsquigarrow \mathtt{A}.$$

The following general recipe then applies: In order to derive the pyramid associated with some arithmetic expression, apply the $\rightsquigarrow$ rule to all its variables, or parts, and replace all operations with operations on pyramids. Here "apply the $\rightsquigarrow$ rule" to something means replace that something with the associated pyramid. And the term "parts" means subexpressions. *Definition 1* guarantees that the result will be the same pyramid $\mathtt{A}$ no matter how we split the arithmetic expression A into subexpressions. It is only necessary to recognize, in case of using subexpressions, that one pyramid expression should be viewed as a function of another.

For illustration, suppose we regard the A given by (2.68) to be the composition of two functions, $F(z) = 2 + 3z$ and $G(z) = z^2$, so that $A(z) = F(G(z))$. Instead of associating a constant and a single variable with their respective pyramids, let us now associate whole subexpressions. In addition, let us label the pyramid expressions on the right of ⇝ with with some names, F and G:

$$2 + 3z \rightsquigarrow 2\ \mathtt{C1} + 3\ \mathtt{Z} = \mathtt{F[Z]}$$

$$z^2 \rightsquigarrow \mathtt{PROD[Z,Z]} = \mathtt{G[Z]}$$

$$A(z) \rightsquigarrow \mathtt{F[G[Z]]} = \mathtt{A}.$$

We have indicated the explicit dependence on Z. It is important to note that F[Z] is a pyramid *expression* prior to executing any the pyramid operations, i.e it is not yet a pyramid, but is simply the result of formal replacements that follow the association rule.

Mathematica code for the simple example (2.72) is shown below,

```
C1 = {1, 0, 0};
Z = {0, 1, 0};
2 C1 + 3 PROD[Z, Z]
{2, 0, 3}                                    (S.2.73)
```

Note that the result (2.73) agrees with (2.69). This example does not use any nested expressions. We will now illustrate how the same results can be obtained using nested expressions.

We begin by displaying a simple *Mathematica* program/execution, that employs ordinary variables, and uses *Mathematica*'s intrinsic abilities to handle nested expressions. The program/execution is

```
f[z_] := 2 + 3z;
g[z_] := z^2;
f[g[z]]
2 + 3z^2                                     (S.2.74)
```

With *Mathematica* the underscore in z_ indicates that z is a dummy variable name, and the symbols := indicate that f is defined with a delayed assignment. That is what is done in line one above. The same is done in line two for g. Line three requests evaluation of the nested function $f(g(z))$, and the result of this evaluation is displayed in line four. Note that the result agrees with (2.68).

With this background, we are ready to examine a program with analogous nested pyramid operations. The same comments apply regarding the use of underscores and delayed assignments. The program is

```
C1 = {1, 0, 0};
Z = {0, 1, 0};
F[Z_] := 2 C1 + 3 Z;
G[Z_] := PROD[Z, Z];
F[G[Z]]
{2, 0, 3}                                        (S.2.75)
```

Note that line (2.75) agrees with line (2.73), and is consistent with line (2.69).

S.2.9 Taylor Rule

We close this section with an important consequence of the replacement rule and nested operations, which we call the *Taylor* rule. We begin by considering functions of a single variable. Suppose the function $G(x)$ has the special form

$$G(x) = z^d + x \tag{S.2.76}$$

where z^d is some constant. Let F be some other function. Consider the composite (nested) function A defined by

$$A(x) = F(G(x)) = F(z^d + x). \tag{S.2.77}$$

Then, assuming the necessary analyticity and by the chain rule, A evidently has a Taylor expansion in x about the origin of the form

$$\begin{aligned} A &= A(0) + A'(0)x + (1/2)A''(0)x^2 + \cdots \\ &= F(z^d) + F'(z^d)x + (1/2)F''(z^d)x^2 + \cdots . \end{aligned} \tag{S.2.78}$$

We conclude that if we know the Taylor expansion of A about the origin, then we also know the Taylor expansion of F about z^d, and vice versa. Suppose, for example, that

$$F(z) = 1 + 2z + 3z^2 \tag{S.2.79}$$

and

$$z^d = 4. \tag{S.2.80}$$

Then there is the result

$$A(x) = F(G(x)) = F(z^d + x) = 1 + 2(4 + x) + 3(4 + x)^2 = 57 + 26x + 3x^2. \tag{S.2.81}$$

We now show that this same result can be obtained using pyramids. The *Mathematica* fragment below illustrates how this can be done.

```
C1 = {1, 0, 0};
X = {0, 1, 0};
zd = 4;
F[Z_] := 1 C1 + 2 Z + 3 PROD[Z, Z];
G[X_] := zd C1 + X;
F[G[X]]
{57, 26, 3}                                      (S.2.82)
```

Note that (2.82) agrees with (2.81). See also Table 2.5.

Let us also illustrate the Taylor rule in the two-variable case. Let $F(z_1, z_2)$ be some function of two variables. Introduce the functions $G(x_1)$ and $H(x_1)$ having the special forms

$$G(x_1) = z_1^d + x_1, \tag{S.2.83}$$

$$H(x_2) = z_2^d + x_2, \tag{S.2.84}$$

where z_1^d and z_2^d are some constants. Consider the function A defined by

$$A(x_1, x_2) = F(G(x_1), H(x_2)) = F(z_1^d + x_1, z_2^d + x_2). \tag{S.2.85}$$

Then, again assuming the necessary analyticity and by the chain rule, A evidently has a Taylor expansion in x_1 and x_2 about the origin $(0, 0)$ of the form

$$\begin{aligned} A &= A(0,0) + [\partial_1 A(0,0)]x_1 + [\partial_2 A(0,0)]x_2 \\ &\quad +(1/2)[(\partial_1)^2 A(0,0)]x_1^2 + [\partial_1\partial_2 A(0,0)]x_1x_2 + (1/2)[(\partial_2)^2 A(0,0)]x_2^2 + \cdots \\ &= F(z_1^d, z_2^d) + [\partial_1 F(z_1^d, z_2^d)]x_1 + [\partial_2 F(z_1^d, z_2^d)]x_2 \\ &\quad +(1/2)[(\partial_1)^2 F(z_1^d, z_2^d)]x_1^2 + [\partial_1\partial_2 AF(z_1^d, z_2^d)]x_1x_2 + (1/2)[(\partial_2)^2 A(F(z_1^d, z_2^d))]x_2^2 + \cdots \end{aligned} \tag{S.2.86}$$

where

$$\partial_1 = \partial/\partial x_1, \ \ \partial_2 = \partial/\partial x_2 \tag{S.2.87}$$

when acting on A, and

$$\partial_1 = \partial/\partial z_1, \ \ \partial_2 = \partial/\partial z_2 \tag{S.2.88}$$

when acting on F. We conclude that if we know the Taylor expansion of A about the origin $(0, 0)$, then we also know the Taylor expansion of F about (z_1^d, z_2^d), and vice versa.

As a concrete example, suppose that

$$F(z_1, z_2) = 1 + 2z_1 + 3z_2 + 4z_1^2 + 5z_1z_2 + 6z_2^2 \tag{S.2.89}$$

and

$$z_1^d = 7, \ \ z_2^d = 8. \tag{S.2.90}$$

Then, hand calculation shows that $F(G(x_1), H(x_2))$ takes the form

$$\begin{aligned} F(z_1^d + x_1, z_2^d + x_2) &= F(G(x_1), H(x_2)) \\ &= 899 + 98x_1 + 4x_1^2 + 134x_2 + 5x_1x_2 + 6x_2^2. \end{aligned} \tag{S.2.91}$$

Below is a *Mathematica* execution that finds the same result,

$$\begin{array}{l}
\texttt{F}[\texttt{z1}\_,\texttt{z2}\_] := 1 + 2\ \texttt{z1} + 3\ \texttt{z2} + 4\ \texttt{z1}^2 + 5\ \texttt{z1}\ \texttt{z2} + 6\ \texttt{z2}^2\\
\texttt{G}[\texttt{x1}\_] := \texttt{zd1} + \texttt{x1};\\
\texttt{H}[\texttt{x2}\_] := \texttt{zd2} + \texttt{x2};\\
\texttt{zd1} = 7;\\
\texttt{zd2} = 8;\\
\texttt{A} = \texttt{F}[\texttt{G}[\texttt{x1}],\texttt{H}[\texttt{x2}]]\\
\texttt{Expand}[\texttt{A}]\\
1 + 2\ (7 + x1) + 4\ (7 + x1)^2 + 3\ (8 + x2) + 5\ (7 + x1)\ (8 + x2) + 6\ (8 + x2)^2\\
899 + 98\ x1 + 4\ x1^2 + 134\ x2 + 5\ x1\ x2 + 6\ x2^2
\end{array} \tag{S.2.92}$$

The calculation above dealt with the case of a function of two ordinary variables. We now illustrate, for the same example, that there is an analogous result for pyramids. Following the replacement rule, we should make the substitutions

$$z_1^d + x_1 \rightsquigarrow \texttt{zd1 C1} + \texttt{X1}, \tag{S.2.93}$$

$$z_2^d + x_2 \rightsquigarrow \texttt{zd2 C1} + \texttt{X2}, \tag{S.2.94}$$

$$\begin{array}{l}
1 + 2\ z_1 + 3\ z_2 + 4\ z_1^2 + 5\ z_1\ z_2 + 6\ z_2^2 \rightsquigarrow\\
\texttt{C1} + 2\ \texttt{Z1} + 3\ \texttt{Z2} + 4\ \texttt{PROD}[\texttt{Z1},\texttt{Z1}] + 5\ \texttt{PROD}[\texttt{Z1},\texttt{Z2}] + 6\ \texttt{PROD}[\texttt{Z2},\texttt{Z2}].
\end{array} \tag{S.2.95}$$

The *Mathematica* fragment below, executed for the case $m = 2$ and $p = 2$, in which case $L = 6$, illustrates how the analogous result is obtained using pyramids,

$$\begin{array}{l}
\texttt{C1} = \{\texttt{1},\texttt{0},\texttt{0},\texttt{0},\texttt{0},\texttt{0}\};\\
\texttt{X1} = \{\texttt{0},\texttt{1},\texttt{0},\texttt{0},\texttt{0},\texttt{0}\};\\
\texttt{X2} = \{\texttt{0},\texttt{0},\texttt{1},\texttt{0},\texttt{0},\texttt{0}\};\\
\texttt{F}[\texttt{Z1}\_,\texttt{Z2}\_] := \texttt{C1} + 2\ \texttt{Z1} + 3\ \texttt{Z2} + 4\ \texttt{PROD}[\texttt{Z1},\texttt{Z1}] + 5\ \texttt{PROD}[\texttt{Z1},\texttt{Z2}]\\
+6\ \texttt{PROD}[\texttt{Z2},\texttt{Z2}];\\
\texttt{G}[\texttt{X1}\_] := \texttt{z01 C1} + \texttt{X1};\\
\texttt{H}[\texttt{X2}\_] := \texttt{z02 C1} + \texttt{X2};\\
\texttt{zd1} = 7;\\
\texttt{zd2} = 8;\\
\texttt{F}[\texttt{G}[\texttt{X1}],\texttt{H}[\texttt{X2}]]\\
\{899, 98, 134, 4, 5, 6\}
\end{array} \tag{S.2.96}$$

Note that, when use is made of Table 2.4, the last line of (2.96) agrees with (2.91) and the last line of (2.92).

S.3 Numerical Integration and Replacement Rule

S.3.1 Numerical Integration

Consider the set of differential equations (1.1). As described in Chapter 2, a standard procedure for their numerical integration from an initial time $t^i = t^0$ to some final time t^f is to divide the time axis into a large number of steps N, each of small duration h, thereby introducing successive times t^n defined by the relation

$$t^n = t^0 + nh \ \text{ with } \ n = 0, 1, \cdots, N. \tag{S.3.1}$$

By construction, there will also be the relation

$$Nh = t^f - t^i. \tag{S.3.2}$$

The goal is to compute the vectors $\boldsymbol{z}^n$, where

$$\boldsymbol{z}^n = \boldsymbol{z}(t^n), \tag{S.3.3}$$

starting from the vector $\boldsymbol{z}^0$. The vector $\boldsymbol{z}^0$ is assumed given as a set of definite numbers, i.e. the initial conditions at t^0.

If we assume for the solution piece-wise analyticity in t, or at least sufficient differentiability in t (which will be the case if the f_a are piece-wise analytic or at least have sufficient differentiability in t), we may convert the set of differential equations (1.1) into a set of recursion relations for the $\boldsymbol{z}^n$ in such a way that the $\boldsymbol{z}^n$ obtained by solving the recursion relations differ from the true $\boldsymbol{z}^n$ by only small truncation errors of order h^m. (Here m is *not* the number of variables, but rather some fixed integer describing the accuracy of the integration method.) One such procedure, a fourth-order *Runge Kutta* (RK4) method, is the set of marching/recursion rules

$$\boldsymbol{z}^{n+1} = \boldsymbol{z}^n + \frac{1}{6}(\boldsymbol{a} + 2\boldsymbol{b} + 2\boldsymbol{c} + \boldsymbol{d}) \tag{S.3.4}$$

where, at each step,

$$\begin{aligned} \boldsymbol{a} &= h\boldsymbol{f}(\boldsymbol{z}^n, t^n), \\ \boldsymbol{b} &= h\boldsymbol{f}(\boldsymbol{z}^n + \frac{1}{2}\boldsymbol{a}, t^n + \frac{1}{2}h), \\ \boldsymbol{c} &= h\boldsymbol{f}(\boldsymbol{z}^n + \frac{1}{2}\boldsymbol{b}, t^n + \frac{1}{2}h), \\ \boldsymbol{d} &= h\boldsymbol{f}(\boldsymbol{z}^n + \boldsymbol{c}, t^n + h). \end{aligned} \tag{S.3.5}$$

Thanks to the genius of Runge and Kutta, the relations (3.4) and (3.5) have been constructed in such a way that the method is locally (at each step) correct through order h^4, and makes local truncation errors of order h^5. Recall Section 2.3.2

In the case of a single variable, and therefore a single differential equation, the relations (3.4) and (3.5) may be encoded in the *Mathematica* form shown below. Here `Zvar` is the dependent variable, `t` is the time, `Zt` is a temporary variable, `tt` is a temporary time, and ns is the number of integration steps. The program employs a Do loop over i so that the operations (3.4) and (3.5) are carried out ns times.

```
RK4 := (
   t0 = t;
Do[
   Aa = h F[Zvar, t];
   Zt = Zvar + (1/2)Aa;
   tt = t + h/2;
   Bb = h F[Zt, tt];
   Zt = Zvar + (1/2)Bb;
   Cc = h F[Zt, tt];
   Zt = Zvar + Cc;
   tt = t + h;
   Dd = h F[Zt, tt];
   Zvar = Zvar + (1/6)(Aa + 2 Bb + 2 Cc + Dd);
   t = t0 + i h;,
   {i, 1, ns, 1}
   ]
 )
```

(S.3.6)

S.3.2 Replacement Rule, Single Equation/Variable Case

We now make what, for our purposes, is a fundamental observation: The operations that occur in the Runge Kutta recursion rules (3.4) and (3.5) and realized in the code above can be extended to pyramids by application of the replacement rule. In particular, the dependent variable $\boldsymbol{z}$ can be replaced by a pyramid, and the various operations involved in the recursion rules can be replaced by pyramid operations. Indeed if we look at the code above, apart from the evaluation of F, we see that the quantities Zvar, Zt, Aa, Bb, Cc, and Dd can be viewed, if we wish, as pyramids since the only operations involved are scalar multiplication and addition. The only requirement for a pyramidal interpretation of the RK4 *Mathematica* code is that the right side of the differential equation, F[∗, ∗], be defined for pyramids. Finally, we remark that the features that make it possible to interpret the RK4 *Mathematica* code either in terms of ordinary variables or pyramidal variables will hold for *Mathematica* realizations of many other familiar numerical integration methods including other forms of Runge Kutta, predictor-corrector methods, and extrapolation methods.

To make these ideas concrete, and to understand their implications, let us begin with a simple example. Suppose, in the single variable case, that the right side of the differential equation has the simple form

$$f(z,t) = -2tz^2. \tag{S.3.7}$$

The differential equation with this right side can be integrated analytically to yield the solution

$$z(t) = z^0/[1 + z^0(t - t^0)^2]. \tag{S.3.8}$$

In particular, for the case $t^0 = 0$, $z^0 = 1$, and $t = 1$, there is the result

$$z(1) = z^0/[1 + z^0] = 1/2. \tag{S.3.9}$$

Let us also integrate the differential equation with the right side (3.7) numerically. Shown below is the result of running the associated *Mathematica* Runge Kutta code for this case.

```
Clear["Global`*"];
F[Z_, t_] := −2 t Z^2;
h = .1;
ns = 10;
t = 0;
Zvar = 1.;
RK4;
t
Zvar
1.
0.500001
```

(S.3.10)

Note that the last line of (3.10) agrees with (3.9) save for a “1” in the last entry. As expected, and as experimentation shows, this small difference, due to accumulated truncation error, becomes even smaller if `h` is decreased (and correspondingly, `ns` is increased).

Suppose we expand the solution (3.9) about the design initial condition $z^{d0} = 1$ by replacing z^0 by $z^{d0} + x$ and expanding the result in a Taylor series in x about the point x=0. Below is a *Mathematica* run that performs this task.

```
zd0 = 1;
Series[(zd0 + x)/(1 + zd0 + x), {x, 0, 5}]
```

$$\frac{1}{2} + \frac{x}{4} - \frac{x^2}{8} + \frac{x^3}{16} - \frac{x^4}{32} + \frac{x^5}{64} + O[x]^6$$

(S.3.11)

We will now see that the same Taylor series can be obtained by the operation of numerical integration applied to pyramids. The *Mathematica* code below shows, for our example differential equation, the application of numerical integration to pyramids.

```
Clear["Global` * "];
Needs["Combinatorica`"];
m = 1; p = 5;
GAMMA = Compositions[0, m];
Do[GAMMA = Join[GAMMA, Reverse[Compositions[d, m]]], {d, 1, p, 1}];
L = Length[GAMMA];
JSK[list_, k_] :=
Position[Apply[And, Thread[#1 <= #2 & [#, k]]] & /@ list, True]//Flatten;
B = Table[JSK[GAMMA, GAMMA[[r]]], {r, 1, L, 1}];
Brev = Reverse/@ B;
PROD[U_, V_] := Table[U[[B[[k]]]].V[[Brev[[k]]]], {k, 1, L, 1}];
F[Z_, t_] := −2 t PROD[Z, Z];
h = .01;
ns = 100;
t = 0;
zd0 = 1;
C1 = {1, 0, 0, 0, 0, 0};
X = {0, 1, 0, 0, 0, 0};
Zvar = zd0 C1 + X;
RK4;
t
Zvar
1.
{0.5, 0.25, −0.125, 0.0625, −0.03125, 0.015625}                (S.3.12)
```

The first 11 lines of the code set up what should be by now the familiar procedure for labeling and multiplying pyramids. In particular, $m = 1$ because we are dealing with a single variable, and $p = 5$ since we wish to work through fifth order. The line

$$\texttt{F[Z\_, t\_] := -2 t PROD[Z, Z]} \tag{S.3.13}$$

defines F[∗, ∗] for the case of pyramids, and is the result of applying the replacement rule to the right side of f as given by (3.7),

$$-\,2\,t\,z^2 \rightsquigarrow -2\ \texttt{t PROD[Z, Z]}. \tag{S.3.14}$$

Lines 13 through 15 play the same role as lines 3 through 5 in (3.10) except that, in order to improve numerical accuracy, the step size `h` has been decreased and correspondingly the number of steps `ns` has been increased. Lines 16 through 19 now initialize `Zvar` as a pyramid with a constant part `zd0` and first-order monomial part with coefficient 1,

$$\texttt{Zvar = zd0 C1 + X}. \tag{S.3.15}$$

These lines are the pyramid equivalent of line 6 in (3.10). Finally lines 20 through 22 are the same as lines 7 through 9 in (3.10). In particular, the line `RK4` in (3.10) and the line `RK4` in (3.12) refer to exactly the *same* code, namely that in (3.6).

Let us now compare the outputs of (3.10) and (3.12). Comparing the penultimate lines in each we see that the final time $t = 1$ is the same in each case. Comparing the last lines shows that the output `Zvar` for (3.12) is a pyramid whose first entry agrees with the last line of (3.10). Finally, all the entries in the pyramid output agree with the Taylor coefficients in the expansion (3.11). We see, in the case of numerical integration (of a single differential equation), that replacing the dependent variable by a pyramid, with the initial value of the pyramid given by (3.15), produces a Taylor expansion of the final condition in terms of the initial condition.

What accounts for this near miraculous result? It's the Taylor rule described described in Subsection 2.9. We have already learned that to expand some function $F(z)$ about some point z^d we must evaluate $F(z^d + x)$. See (2.77). We know that the final $Zvar$, call it $Zvar^{\rm fin}$, is an analytic function of the initial $Zvar$, call it $Zvar^{\rm in}$, so that we may write

$$Zvar^{\rm fin} = Zvar^{\rm fin}(Zvar^{\rm in}) = g(Zvar^{\rm in}) \tag{S.3.16}$$

where g is the function that results from following the trajectory from $t = t^{\rm in}$ to $t = t^{\rm fin}$. Therefore, by the Taylor rule, to expand $Zvar^{\rm fin}$ about $Zvar^{\rm in} = z^{d0}$, we must evaluate $Zvar^{\rm fin}(z^{d0} + x)$. That, with the aid of pyramids, is what the code (3.12) accomplishes.

S.3.3 Multi Equation/Variable Case

Because of *Mathematica's* built-in provisions for handling arrays, the work of the previous section can easily be extended to the case of several differential equations. Consider, as an example, the two-variable case for which $\boldsymbol{f}$ has the form

$$\begin{aligned} f_1(\boldsymbol{z}, t) &= -z_1^2, \\ f_2(\boldsymbol{z}, t) &= +2z_1z_2. \end{aligned} \tag{S.3.17}$$

The differential equations associated with this $\boldsymbol{f}$ can be solved in closed form to yield, with the understanding that $t^0 = 0$, the solution

$$\begin{aligned} z_1(t) &= z_1^0/(1 + tz_1^0), \\ z_2(t) &= z_2^0(1 + tz_1^0)^2. \end{aligned} \tag{S.3.18}$$

For the final time $t = 1$ we find the result

$$\begin{aligned} z_1(1) &= z_1^0/(1 + z_1^0), \\ z_2(1) &= z_2^0(1 + z_1^0)^2. \end{aligned} \tag{S.3.19}$$

Let us expand the solution (3.19) about the design initial conditions

$$\begin{aligned} z_1^{d0} &= 1, \\ z_2^{d0} &= 2, \end{aligned} \tag{S.3.20}$$

by writing

$$\begin{aligned} z_1^0 &= z_1^{d0} + x_1 = 1 + x_1, \\ z_2^0 &= z_2^{d0} + x_2 = 2 + x_2. \end{aligned} \tag{S.3.21}$$

Doing so gives the results

$$\begin{aligned} z_1(1) &= (1+x_1)/(2+x_1) = (2+x_1-1)/(2+x_1) = 1 - 1/(2+x_1) = \\ &= 1 - (1/2)(1+x_1/2)^{-1} = 1 - (1/2)[1 - x_1/2 + (x_1/2)^2 - (x_1/2)^3 + \cdots \\ &= (1/2) + (1/4)x_1 - (1/8)x_1^2 + (1/16)x_1^3 + \cdots , \end{aligned} \tag{S.3.22}$$

$$\begin{aligned} z_2(1) &= (2+x_2)(2+x_1)^2 \\ &= 8 + 8x_1 + 4x_2 + 2x_1^2 + 4x_1x_2 + x_1^2x_2. \end{aligned} \tag{S.3.23}$$

We will now explore how this same result can be obtained using the replacement rule applied to the operation of numerical integration. As before, we will label individual monomials by an integer r. Recall that Table 2.5 shows our standard modified glex sequencing applied to the case of two variables.

The *Mathematica* code below shows, for our two-variable example differential equation, the application of numerical integration to pyramids. Before describing the code in some detail, we take note of the bottom two lines. When interpreted with the aid of Table 2.4, we see that the penultimate line of (3.24) agrees with (3.22), and the last line of (3.24) nearly agrees with (3.23). The only discrepancy is that for the monomial with label $r = 7$ in the last line of (3.24). In the *Mathematica* output it has the value -1.16563×10^{-7} while, according to (3.23), the true value should be zero. This small discrepancy arises from the truncation error inherent in the RK4 algorithm, and becomes smaller as the step size `h` is decreased (and `ns` is correspondingly increased), or if some more accurate integration algorithm is used. We conclude that, with the use of pyramids, it is also possible in the two-variable case to obtain Taylor expansions of the final conditions in terms of the initial conditions. Indeed, what is involved is again the Taylor rule applied, in this instance, to the case of two variables.

```
Clear["Global‘ * "];
Needs["Combinatorica‘"];
m = 2; p = 3;
GAMMA = Compositions[0, m];
Do[GAMMA = Join[GAMMA, Reverse[Compositions[d, m]]], {d, 1, p, 1}];
L = Length[GAMMA];
JSK[list_, k_] :=
Position[Apply[And, Thread[#1 <= #2 & [#, k]]] & /@ list, True]//Flatten;
B = Table[JSK[GAMMA, GAMMA[[r]]], {r, 1, L, 1}];
Brev = Reverse/@ B;
PROD[U_, V_] := Table[U[[B[[k]]]].V[[Brev[[k]]]], {k, 1, L, 1}];
F[Z_, t_] := {−PROD[Z[[1]], Z[[1]]], 2. PROD[Z[[1]], Z[[2]]]};
h = .01;
ns = 100;
t = 0;
zd0 = {1., 2.};
C1 = Table[KroneckerDelta[k, 1], {k, 1, L, 1}];
X[1] = Table[KroneckerDelta[k, 2], {k, 1, L, 1}];
X[2] = Table[KroneckerDelta[k, 3], {k, 1, L, 1}];
Zvar = {zd0[[1]] C1 + X[1], zd0[[2]] C1 + X[2]};
RK4;
t
Zvar
1.
{{0.5, 0.25, 0., −0.125, 0., 0., 0.0625, 0., 0., 0, },
  {8., 8., 4., 2., 4., 0., −1.16563 × 10^-7, 1., 0., 0.}}                (S.3.24)
```

Let us compare the structures of the routines for the single variable case and multi (two) variable case as illustrated in (3.12) and (3.24). The first difference occurs at line 3 where the number of variables m and the maximum degree p are specified. In (3.24) m is set to 2 because we wish to treat the case of two variables, and p is set to 3 simply to limit the lengths of the output arrays. The next difference occurs in line 12 where the right side `F` of the differential equation is specified. The major feature of the definition of `F` in (3.24) is that it is specified as two pyramids because the right side of the definition has the structure $\{*, *\}$ where each item $*$ is an instruction for computing a pyramid. In particular, the two pyramids are those for the two components of $\boldsymbol{f}$ as given by (3.17) and use of the replacement rule,

$$-z_1^2 \rightsquigarrow -\texttt{PROD[Z[[1]], Z[[1]]]}, \tag{S.3.25}$$

$$2z_1z_2 \rightsquigarrow 2.\,\texttt{PROD[Z[[1]],Z[[2]]]}. \qquad (S.3.26)$$

The next differences occur in lines 16 through 20 of (3.24). In line 16, since specification of the initial conditions now requires two numbers, see (3.20), zd0 is specified as a two-component array. In lines 17 and 18 of (3.12) the pyramids C1 and X are set up explicitly for the case $p = 5$. By contrast, in lines 17 through 19 of (3.24), the pyramids C1, X[1], and X[2] are set up for general p with the aid of the Table command and the Kronecker delta function. Recall (2.66) and observe from Tables 2.1, 2.4, and 2.5 that, no matter what the values of m and p, the constant monomial has the label $r = 1$ and the monomial x_1 has the label $r = 2$. Moreover, as long as $m \geq 2$ and no matter what the value of p, the x_2 monomial has the label $r = 3$. Finally, compare line 19 in (3.12) with line 20 in (3.24), both of which define the initial Zvar. We see that the difference is that in (3.12) Zvar is defined as a single pyramid while in (3.24) it is defined as a pair of pyramids of the form $\{*, *\}$. Most remarkably, all other corresponding lines in (3.12) and (3.24) are the same. In particular, the *same* RK4 code, namely that given by (3.6), is used in the scalar case (3.10), the single pyramid case (3.12), and the two-pyramid case (3.24). This multi-use is possible because of the convenient way in which *Mathematica* handles arrays.

We conclude that the pattern for the multivariable case is now clear. Only the following items need to be specified in an m dependent way:

- The value of m.

- The entries in F with entries entered as an array $\{*, *, \cdots\}$ of m pyramids.

- The design initial condition array zd0.

- The pyramids for C1 and X[1] through X[m].

- The entries for the initial Zvar specified as an array

  {zd0[[1]] C1 + X[1], zd0[[2]] C1 + X[2], $\cdots$, zd0[[m]] C1 + X[m]} of m pyramids.

S.4 Duffing Equation Application

Let us now apply the methods just developed to the case of the Duffing equation with parameter dependence as described by the relations (10.12.133) through (10.12.138). *Mathematica* code for this purpose is shown below. By looking at the final lines that result from executing this code, we see that the final output is an array of the form $\{\{*\}, \{*\}, \{*\}\}$. That is, the final output is an array of three pyramids. This is what we expect, because now we are dealing with three variables. See line 3 of the code, which sets $m = 3$. Also, for convenience of viewing, results are calculated and displayed only through third order as a consequence of setting $p = 3$.

```
Clear["Global‘*"];
Needs["Combinatorica‘"];
m = 3; p = 3;
GAMMA = Compositions[0, m];
Do[GAMMA = Join[GAMMA, Reverse[Compositions[d, m]]], {d, 1, p, 1}];
L = Length[GAMMA];
JSK[list_, k_] :=
Position[Apply[And, Thread[#1 <= #2 & [#, k]]] & /@ list, True]//Flatten;
B = Table[JSK[GAMMA, GAMMA[[r]]], {r, 1, L, 1}];
Brev = Reverse/@ B;
PROD[U_, V_] := Table[U[[B[[k]]]].V[[Brev[[k]]]], {k, 1, L, 1}];
POWER[U_, 2] := PROD[U, U];
POWER[U_, 3] := PROD[U, POWER[U, 2]];
C0 = Table[0, {k, 1, L, 1}];
F[Z_, t_] := {Z[[2]],
-2. beta PROD[Z[[3]], Z[[2]]] - PROD[POWER[Z[[3]], 2], Z[[1]]]-
POWER[Z[[1]], 3] - eps Sin[t] POWER[Z[[3]], 3],
C0};
ns = 100;
t = 0;
h = (2Pi)/ns;
beta = .1; eps = 1.5;
zd0 = {.3, .4, .5};
C1 = Table[KroneckerDelta[k, 1], {k, 1, L, 1}];
X[1] = Table[KroneckerDelta[k, 2], {k, 1, L, 1}];
X[2] = Table[KroneckerDelta[k, 3], {k, 1, L, 1}];
X[3] = Table[KroneckerDelta[k, 4], {k, 1, L, 1}];
Zvar = {zd0[[1]] C1 + X[1], zd0[[2]] C1 + X[2], zd0[[3]] C1 + X[3]};
RK4;
t
Zvar
```

$$
\begin{array}{l}
2\pi \\
\{\{-0.0493158, 0.973942, -0.110494, 5.51271, 3.54684, 3.46678, \\
\quad 11.2762, 2.36463, 1.0985, 23.3332, -1.03541, -3.23761, -12.8064, \\
\quad 4.03421, -23.4342, -17.8967, 1.96148, 5.07403, -36.9009, 25.1379\}, \\
\ \{0.439713, 1.05904, 0.427613, 3.3177, 0.0872459, 0.635397, -3.02822, \\
\quad 1.77416, -4.10115, 3.16981, -2.43002, -5.33643, -7.77038, -6.08476, \\
\quad -0.541465, -21.1672, -1.4091, -9.54326, 14.6334, -39.2312\}, \\
\ \{0.5, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}\}
\end{array}
\tag{S.4.1}
$$

The first unusual fragments in the code are lines 12 and 13, which define functions that implement the calculation of second and third powers of pyramids. Recall Subsection 2.7. The first new fragment is line 14, which defines the pyramid `C0` with the aid of the Table command and an implied Do loop. As a result of executing this code, `C0` is an array of L zeroes. The next three lines, lines 15 through 18, define `F`, which specifies the right sides of equations (10.12.133) through (10.12.135). See (10.12.136) through (10.12.138). The right side of `F` is of the form $\{*, *, *\}$, an array of three pyramids. By looking at (10.12.136) and recalling the replacement rule, we see that the first pyramid should be `Z[[2]]`,

$$z_2 \rightsquigarrow \texttt{Z[[2]]}. \tag{S.4.2}$$

The second pyramid on the right side of `F` is more complicated. It arises by applying the replacement rule to the right side of (10.12.137) to obtain the associated pyramid,

$$
\begin{array}{l}
-\,2\beta z_3 z_2 - z_3^2 z_1 - z_1^3 - \epsilon z_3^3 \sin t \rightsquigarrow \\
-2.\,\texttt{beta}\ \texttt{PROD[Z[[3]], Z[[2]]]} - \texttt{PROD[POWER[Z[[3]], 2], Z[[1]]]} - \\
\texttt{POWER[Z[[1]], 3]} - \texttt{eps}\ \texttt{Sin[t]}\ \texttt{POWER[Z[[3]], 3]}.
\end{array}
\tag{S.4.3}
$$

The third pyramid on the right side of F is simplicity itself. From (10.12.138) we see that this pyramid should be the result of applying the replacement rule to the number 0. Hence, this pyramid is `C0`,

$$0 \rightsquigarrow \texttt{C0} = \{\texttt{0}, \texttt{0}, \cdots, \texttt{0}\}. \tag{S.4.4}$$

The remaining lines of the code require little comment. Line 20 sets the initial time to 0, and line 21 defines h in such a way that the final value of t will be 2π. Line 22 establishes the parameter values $\beta = .1$ and $\epsilon = 1.5$, which are those for Figure 1.4.9. Line 23 specifies that the design initial condition is

$$z_1(0) = z_1^{d0} = .3,\ z_2(0) = z_2^{d0} = .4,\ z_3(0) = z_3^{d0} = .5 = \sigma, \tag{S.4.5}$$

and consequently

$$\omega = 1/\sigma = 2. \tag{S.4.6}$$

See (10.12.104). Also, it follows from (10.12.103) and (10.12.106) that

$$q(0) = \omega Q(0) = \omega z_1(0) = (2)(.3) = .6, \tag{S.4.7}$$

$$q'(0) = \omega^2 \dot{Q}(0) = \omega^2 z_2(0) = (2^2)(.4) = 1.6. \tag{S.4.8}$$

Next, lines 24 through 28 specify that the expansion is to be carried out about the initial conditions (7.124). Finally, line 29 invokes the `RK4` code given by (3.6). That is, as before, *no* modifications are required in the integration code.

A few more comments about the output are appropriate. Line 32 shows that the final time t is indeed 2π, as desired. The remaining output lines display the three pyramids that specify the final value of `Zvar`. From the first entry in each pyramid we see that

$$z_1(2\pi) = -0.0493158, \tag{S.4.9}$$

$$z_2(2\pi) = 0.439713, \tag{S.4.10}$$

$$z_3(2\pi) = .5, \tag{S.4.11}$$

when there are no deviations in the initial conditions. The remaining entries in the pyramids are the coefficients in the Taylor series that describe the changes in the final conditions that occur when changes are made in the initial conditions (including the parameter σ). We are, of course, particularly interested in the first two pyramids. The third pyramid has entries only in the first place and the fourth place, and these entries are the same as those in the third pyramid pyramid for `Zvar` at the start of the integration, namely those in `zd0[3] C1 + X[3]`. The fact that the third pyramid in `Zvar` remains constant is the expected consequence of (10.12.138).

At this point we should also describe how the $\mathcal{M}_8$ employed in Section 22.12 was actually computed. It could have been computed by setting $p = 8$ in (4.1) and specifying a small step size h and a great number of steps ns to insure good accuracy. Of course, when $p = 8$, the pyramids are large. Therefore, one does not usually print them out, but rather writes them to files or sends them directly to other programs for further use.

However, rather than using RK4 in (4.1), we replaced it with an adaptive 4-5th order Runge-Kutta-Fehlberg routine that dynamically adjusts the time step h during the course of integration to achieve a specified local accuracy, and we required that the error at each step be no larger than 10^{-12}. (Recall Subsection 2.1.1.) Like the RK4 routine, the Runge-Kutta-Fehlberg routine, when implemented in *Mathematica*, has the property that it can integrate any number of equations both in scalar variable and pyramid form without any changes in the code.[2]

S.5 Relation to the Complete Variational Equations

At this point it may not be obvious to the reader that the use of pyramids in integration routines to obtain Taylor expansions is the same as integrating the complete variational equations. We now show that the integration of pyramid equations is equivalent to the forward integration of the complete variational equations. For simplicity, we will examine the single variable case with no parameter dependence. The reader who has mastered this case should be able to generalize the results obtained to the general case.

[2] A *Mathematica* version of this code is available from Dobrin Kaltchev (kaltchev@triumf.ca) upon request.

In the single variable case with no parameter dependence (1.1) becomes

$$\dot{z} = f(z,t). \tag{S.5.1}$$

Let $z^d(t)$ be some design solution and introduce a deviation variable ζ by writing

$$z = z^d + \zeta. \tag{S.5.2}$$

Then the equation of motion (5.1) takes the form

$$\dot{z}^d + \dot{\zeta} = f(z^d + \zeta, t). \tag{S.5.3}$$

Also, the relations (10.12.14) and (10.12.15) take the form

$$f(z^d + \zeta, t) = f(z^d, t) + g(z^d, t, \zeta) \tag{S.5.4}$$

where g has an expansion of the form

$$g(z^d, t, \zeta) = \sum_{j=1}^{\infty} g^j(t)\zeta^j. \tag{S.5.5}$$

Finally, (10.12.16) and (10.12.17) become

$$\dot{z}^d = f(z^d, t), \tag{S.5.6}$$

$$\dot{\zeta} = g(z^d, t, \zeta) = \sum_{j=1}^{\infty} g^j(t)\zeta^j, \tag{S.5.7}$$

and (10.12.18) becomes

$$\zeta = \sum_{j=1}^{\infty} h^j(t)(\zeta_i)^j. \tag{S.5.8}$$

Insertion of (5.8) into both sides of (5.7) and equating like powers of ζ_i now yields the set of differential equations

$$\dot{h}^{j''}(t) = \sum_{j=1}^{\infty} g^j(t) U_j^{j''}(h^s) \text{ with } j, j'' \geq 1 \tag{S.5.9}$$

where the (universal) functions $U_j^{j''}(h^s)$ are given by the relations

$$\left(\sum_{j'=1}^{\infty} h^{j'}(\zeta_i)^{j'} \right)^j = \sum_{j''=1}^{\infty} U_j^{j''}(h^s)(\zeta_i)^{j''}. \tag{S.5.10}$$

The equations (5.6) and (5.9) are to be integrated from $t = t^{\text{in}} = t^0$ to $t = t^{\text{fin}}$ with the initial conditions

$$z^d(t^0) = z^{d0}, \tag{S.5.11}$$

$$h^1(t^0) = 1, \tag{S.5.12}$$

$$h^{j''}(t^0) = 0 \text{ for } j'' > 1. \tag{S.5.13}$$

Let us now consider the numerical integration of pyramids. Upon some reflection, we see that the numerical integration of pyramids is equivalent to finding the numerical solution to a differential equation with pyramid arguments. For example, in the single-variable case, let $\texttt{Zvar}(t)$ be the pyramid appearing in the integration process. Then, its integration is equivalent to solving numerically the pyramid differential equation

$$(d/dt)\texttt{Zvar}(t) = \texttt{F}(\texttt{Zvar}, t). \tag{S.5.14}$$

We now work out the consequences of this observation. By the inverse of the replacement rule, we may associate a Taylor series with the pyramid $\texttt{Zvar}(t)$ by writing

$$\texttt{Zvar}(t) \rightsquigarrow c_0(t) + \sum_{j\geq 1} c_j(t)x^j. \tag{S.5.15}$$

By (5.15) it is intended that the entries in the pyramid $\texttt{Zvar}(t)$ be used to construct a corresponding Taylor series with variable x. In view of (3.15), there are the initial conditions

$$c_0(t_0) = z^d(t_0), \tag{S.5.16}$$

$$c_1(t_0) = 1, \tag{S.5.17}$$

$$c_j(t_0) = 0 \text{ for } j > 1. \tag{S.5.18}$$

We next seek the differential equations that determine the time evolution of the $c_j(t)$. Under the inverse replacement rule, there is also the correspondence

$$(d/dt)\texttt{Zvar}(t) \rightsquigarrow \dot{c}_0(t) + \sum_{j\geq 1} \dot{c}_j(t)x^j. \tag{S.5.19}$$

We have found a representation for the left side of (5.14). We need to do the same for the right side. That is, we need the Taylor series associated with the pyramid $\texttt{F}(\texttt{Zvar}, t)$. By the inverse replacement rule, it will be given by the relation

$$\texttt{F}(\texttt{Zvar}, t) \rightsquigarrow f(\sum_{j\geq 0} c_j(t)x^j, t). \tag{S.5.20}$$

Here it is understood that the right side of (5.20) is to be expanded in a Taylor series about $x = 0$. From (5.4), (5.5), and (5.10) we have the relations

$$\begin{aligned} f(\sum_{j\geq 0} c_j(t)x^j, t) &= f(c_0(t)) + g(c_0(t), t, \sum_{j\geq 1} c_j(t)x^j) \\ &= f(c_0(t)) + \sum_{k\geq 1} g^k(t)(\sum_{j\geq 1} c_j(t)x^j))^k \\ &= f(c_0(t)) + \sum_{k\geq 1} g^k(t) \sum_{j\geq 1} U_k^j(c_\ell)x^j. \end{aligned} \tag{S.5.21}$$

Therefore, there is the inverse replacement rule

$$\mathtt{F(Zvar},t) \rightsquigarrow f(c_0(t)) + \sum_{k\geq 1} g^k(t) \sum_{j\geq 1} U_k^j(c_\ell)x^j. \tag{S.5.22}$$

Upon comparing like powers of x in (5.19) and (5.22), we see that the pyramid differential equation (5.14) is equivalent to the set of differential equations

$$\dot{c}_0(t) = f(c_0(t)), \tag{S.5.23}$$

$$\dot{c}_j(t) = \sum_{k\geq 1} g^k(t)U_k^j(c_\ell). \tag{S.5.24}$$

Finally, compare the initial conditions (5.11) through (5.13) with the initial conditions (5.16) through (5.18), and compare the differential equations (5.6) and (5.9) with the differential equations (5.23) and (5.24). We conclude that that there must be the relations

$$c_0(t) = z^d(t), \tag{S.5.25}$$

$$c_j(t) = h^j(t) \text{ for } j \geq 1. \tag{S.5.26}$$

We have verified, in the single variable case, that the use of pyramids in integration routines is equivalent to the solution of the complete variational equations using forward integration. As stated earlier, verification of the analogous m-variable result is left to the reader.

We also observe the wonderful convenience that, when pyramid operations are implemented and employed, it is not necessary to explicitly work out the forcing terms $g_a^r(t)$ of Subsection 10.12.1 and the universal functions $U_r^{r''}(h_n^s)$ of Subsection 10.12.3, nor is it necessary to explicitly set up the complete variational equations (10.12.36). All these complications are handled implicitly and automatically by the pyramid routines.

S.6 Acknowledgment

Dobrin Kaltchev made major contributions to the work of this appendix.

Exercises

S.6.1. Verify, in the general m variable case, that the use of pyramids in integration routines is equivalent to the solution of the complete variational equations using forward integration.

Bibliography

[1] R. Neidinger, "Computing Multivariable Taylor Series to Arbitrary Order", *Proc. of Intern. Conf. on Applied programming languages*, San Antonio, pp. 134-144 (1995).

[2] R. Neidinger, "Introduction to Automatic Differentiation and MATLAB Object-Oriented Programming", *SIAM Review* **52**, 545-563 (2010).

[3] Wolfram Research, Inc., *Mathematica*, Version 7.0, Champaign, IL (2008).

[4] Dobrin Kaltchev (*TRIUMF, 4004 Wesbrook Mall, Vancouver, B.C., Canada V6T 2A3*) designed and wrote all the *Mathematica* code for this appendix, and he and Alex Dragt coauthored the text. D. Kaltchev wishes to thank his colleagues from TRIUMF and CERN, especially Richard Abram Baartman, for their interest and support.

[5] D. Kalman and R. Lindell, "A recursive approach to multivariate automatic differentiation", *Optimization Methods and Software*, Volume 6, Issue 3, pp. 161-192 (1995).

[6] M. Berz, "Differential algebraic description of beam dynamics to very high orders", *Particle Accelerators* 24, p. 109 (1989).

[7] U. Naumann, *The Art of Differentiating Computer Programs: An Introduction to Algorithmic Differentiation*, SIAM (2012).

[8] Alex Haro, "Automatic Differentiation Tools in Computational Dynamical Systems". See the Web site http://www.maia.ub.es/~alex/ad/adhds.pdf

[9] A. Haro, M. Canadell, J-L. Figueras, A. Luque, and J-M. Mondelo, *The Parameterization Method for Invariant Manifolds: From Rigorous Results to Effective Computations*, Applied Mathematical Sciences Volume 195, Springer (2016).

Appendix T

Quadrature and Cubature Formulas

T.1 Quadrature Formulas

T.1.1 Introduction

Suppose we wish to integrate some function $f(x)$ over some (finite) interval. Without loss of generality, by suitable translation and scaling, we may take this interval to be $[0, 1]$. A *quadrature* formula is a set of k *sampling points* x_i in the interval $[0, 1]$ and *weights* w_i such that

$$\int_0^1 dx\ f(x) \simeq \sum_{i=1}^{k} w_i f(x_i). \tag{T.1.1}$$

The challenge is to select the sampling points and weights in such a way that the approximation (1.1) is optimal and to define what is meant by *optimal.* From the *Weierstrass* approximation theorem we know that the monomials are dense on any bounded domain. Also, according to Taylor's theorem, monomials are the building blocks for analytic functions. Therefore, for our purposes, we will define optimal to mean that the relation (1.1) is to hold exactly for polynomials in x of as high a degree as possible. That is, for a given set of sampling points, we select the w_i in such a way that

$$\sum_{i=1}^{k} w_i (x_i)^\ell = \int_0^1 dx\ x^\ell = 1/(\ell + 1) \tag{T.1.2}$$

for $\ell = 0, 1, 2, \cdots$ up to as large an ℓ value (for a given k) as possible.

At this point some discussion is required. Suppose we reason as follows: Let m be an integer with $m \geq 0$. Assume $f(x)$ is a polynomial of maximum degree m. It will then have

$$k = m + 1 \tag{T.1.3}$$

coefficients in its Taylor series representation, and these coefficients can be found by sampling the value of f at k different points x_i on the interval $[0, 1]$. Put another way, suppose we are given k values $v_1, v_2, \cdots, v_k$. Then $f(x)$ is the unique polynomial of degree m whose graph passes through the points $\{x_i, v_i\}$. Now, with the coefficients known, the Taylor series can be integrated to determine the left side of (1.1). Thus, with a knowledge of f at k sampling

sampling points, it is in principle possible to integrate exactly polynomials of degree $\ell \leq m$ with m given by $m = k - 1$.

We next observe that once the k sampling points have been determined, the k weights w_i are uniquely determined. Let $L_i(x)$, called the *Lagrange* polynomial, be the degree m polynomial that takes on the value 1 at the sampling point x_i and has the value 0 at all the other sampling points,

$$L_i(x_j) = \delta_{ij}. \tag{T.1.4}$$

It is given by the construction

$$L_i(x) = \left[\prod_{j \neq i}(x - x_j)\right] \Big/ \left[\prod_{j \neq i}(x_i - x_j)\right]. \tag{T.1.5}$$

Evidently there are k such polynomials. Also, as a result of (1.4), it follows that $f(x)$ given by

$$f(x) = \sum_{i=1}^{k} v_i L_i(x) \tag{T.1.6}$$

has the property

$$f(x_j) = v_j. \tag{T.1.7}$$

Moreover, we see that

$$\int_0^1 dx\ f(x) = \sum_i v_i \int_0^1 dx\ L_i(x). \tag{T.1.8}$$

Therefore, in view of (1.7) and the desire (1.1), we make the definition

$$w_i = \int_0^1 dx\ L_i(x) \tag{T.1.9}$$

to achieve the result

$$\int_0^1 dx\ f(x) = \sum_i w_i f(x_i). \tag{T.1.10}$$

We conclude that given k sampling points, there are k weights w_i, uniquely determined by (1.9), such that (1.2) holds for $\ell \leq m$ with $m = k - 1$.

Given the k sampling points x_i, and the associated weights w_i, can it happen that (1.2) also holds for some ℓ values with $\ell > m$, i.e. $\ell > k - 1$? Let $\ell_{\max}$ be the largest integer for which (1.2) holds. More precisely, we require that (1.2) holds for $\ell \leq \ell_{\max}$, but not for $\ell = \ell_{\max} + 1$. We will see that exactly how large $\ell_{\max}$ can be depends on how the sampling points are chosen. In particular, we will learn that there is a unique optimum sampling procedure (called Legendre Gauss) for which $\ell_{\max}$ has the optimum value

$$\ell_{\max} = 2k - 1. \tag{T.1.11}$$

Upon combining (1.3) and (1.11) we see that for any sampling procedure there is the range relation

$$k - 1 \leq \ell_{\max} \leq 2k - 1. \tag{T.1.12}$$

Put another way, if $\ell_{\max}$ is specified, we will see that there can be sampling procedures such that (1.2) holds for all $\ell \leq \ell_{\max}$ with $k < \ell_{\max} + 1$. Indeed, if $\ell_{\max}$ is odd, k can be as small as $k_{\min}$ with

$$k_{\min} = (\ell_{\max} + 1)/2. \tag{T.1.13}$$

Since we know that $\ell_{\max}+1$ sampling points are necessary to determine the Taylor coefficients of a polynomial of degree $\ell_{\max}$, how can it be that we can exactly integrate over the interval $[0, 1]$ such polynomials using just k sampling points with $k < \ell_{\max} + 1$? The answer is that the *only* thing we want to know is the value of the integral over the *fixed* interval $[0, 1]$, and *not* the values of the individual Taylor coefficients. What happens is that the value of the integral depends only on the value of a certain combination of the Taylor coefficients, and the value of this combination can be found by sampling the function at fewer than $\ell_{\max} + 1$ points providing these points are judiciously chosen. See Exercise 1.2.

T.1.2 Newton Cotes

One sampling option is to space the x_i evenly with $x_1 = 0$ and $x_k = 1$,

$$x_i = (i - 1)/(k - 1). \tag{T.1.14}$$

Doing so gives the family of *Newton-Cotes* quadrature formulas.[1] For example, for the case $k = 3$, the sampling points are

$$(x_1, x_2, x_3) = (0, 1/2, 1), \tag{T.1.15}$$

the associated weights are found, using (1.9), to be

$$(w_1, w_2, w_3) = (1/6, 4/6, 1/6), \tag{T.1.16}$$

thereby yielding the celebrated *Simpson's rule* 1-4-1 formula

$$\int_0^1 dx\ f(x) \simeq (1/6)f(0) + (4/6)f(1/2) + (1/6)f(1). \tag{T.1.17}$$

In this case (1.2) holds for $\ell = 0, 1, 2,$ and 3; and errors first begin to appear for $\ell \geq 4$. For $\ell = 4$ the sum on the left side of (1.2) has the value

$$\sum_{i=1}^{3} w_i(x_i)^4 = 1/(4 + 1) + 4!/[90(2^5)] = 1/5 + 4!/2880. \tag{T.1.18}$$

Correspondingly, assuming that f is sufficiently differentiable, use of Newton Cotes gives (for the case $k = 3$) the approximation

$$\int_0^1 dx\ f(x) = (1/6)f(0) + (4/6)f(1/2) + (1/6)f(1) - (1/2880)f^{(4)}(\xi) \tag{T.1.19}$$

[1] Newton's student Roger Cotes urged and inspired Newton to write a second and enlarged edition of his *Principia*, and wrote the preface to this edition. He died of a violent fever at the early early age of 33. At Cotes' death Newton remarked, "If he had lived we would have known something." It is interesting to note that several years before Euler wrote his famous formula $\exp(i\theta) = \cos\theta + i \sin\theta$, Cotes wrote the inverse relation $\log(\cos\theta + i \sin\theta) = i\theta$.

where $\xi \in [0,1]$.

As a second example, for the case $k = 4$ the sampling points are

$$(x_1, x_2, x_3, x_4) = (0, 1/3, 2/3, 1), \tag{T.1.20}$$

the associated weights are found to be

$$(w_1, w_2, w_3, w_4) = (1/8, 3/8, 3/8, 1/8), \tag{T.1.21}$$

and we have the equally celebrated Simpson's 3/8 rule

$$\begin{aligned}\int_0^1 dx\ f(x) &= (1/8)f(0) + (3/8)f(1/3) + (3/8)f(2/3) + (1/8)f(1) \\ &\quad -(1/6480)f^{(4)}(\xi)\end{aligned} \tag{T.1.22}$$

where again $\xi \in [0,1]$.

Table 1.1 lists $\ell_{\max}$ as a function of k for Newton Cotes.[2] Evidently, when k is odd, there is no increase in order when using instead the next even value of k. Doing so does result in a decrease in the coefficient in the error term, but this decrease is fairly modest. For this reason, odd values of k are frequently preferred. Note that for even k the entries in the Table take the floor value $\ell_{\max} = k - 1$ guaranteed by (1.12), and beat this value for odd k.

Table T.1.1: Maximum Order $\ell_{\max}$ for k Newton-Cotes Sampling Points.

k	1	2	3	4	5	6	7	8	9	10
$\ell_{\max}$	1	1	3	3	5	5	7	7	9	9

T.1.3 Legendre Gauss

Another appealing sampling option is not to space the x_i evenly, but rather to select them in such a way that (for a fixed k) the number of successive ℓ values for which (1.2) holds is maximized. This choice produces the family of *Legendre-Gauss* quadrature formulas. Legendre-Gauss quadrature is most naturally described in terms of the interval $[-1,1]$. It can be shown that the *Legendre* polynomial $P_k(x)$ has k distinct zeroes on the interval $[-1,1]$, and it is these zeroes that are used in k-sampling point Legendre-Gauss quadrature.

Three examples of Legendre Gauss are given below. For ease of comparison with Newton-Cotes quadrature, we have transformed Legendre-Gauss results to the interval $[0,1]$.

For $k = 3$, the (transformed) Legendre-Gauss sampling points are given by

$$(x_1, x_2, x_3) = (1/2 - \sqrt{15}/10,\ 1/2,\ 1/2 + \sqrt{15}/10). \tag{T.1.23}$$

[2] Strictly speaking, for the Newton-Cotes we have been describing, the $k = 1$ table entry is meaningless. However, it does apply in the case of *open* Newton Cotes. See Exercise 1.1.

The corresponding weights, calculated using (1.9), are given by

$$(w_1, w_2, w_3) = (5/18,\ 8/18,\ 5/18). \tag{T.1.24}$$

Correspondingly, there is the formula

$$\int_0^1 dx f(x) \simeq (5/18)f(1/2 - \sqrt{15}/10) + (8/18)f(1/2) + (5/18)f(1/2 + \sqrt{15}/10). \tag{T.1.25}$$

In this case (1.2) holds for $\ell = 0$ through 5, but not $\ell = 6$. Therefore $\ell_{\max} = 5$. And for $\ell = 6$ the sum on the left side of (1.2) has the value

$$\sum_{i=1}^{3} w_i (x_i)^6 = 1/7 - 6!/2016000. \tag{T.1.26}$$

Correspondingly, again assuming that f is sufficiently differentiable, use of Legendre Gauss gives (for the case $k = 3$) the approximation

$$\begin{aligned} \int_0^1 dx\ f(x) \quad &= \quad (5/18)f(1/2 - \sqrt{15}/10) + (8/18)f(1/2) + (5/18)f(1/2 + \sqrt{15}/10) \\ &\quad +(1/2016000)f^{(6)}(\xi). \end{aligned} \tag{T.1.27}$$

As another example, consider the case $k = 2$. Then there is the formula

$$\int_0^1 dx\ f(x) \simeq (1/2)f(1/2 - \sqrt{3}/6) + (1/2)f(1/2 + \sqrt{3}/6). \tag{T.1.28}$$

It corresponds to sampling points and weights given by the rules

$$(x_1, x_2) = (1/2 - \sqrt{3}/6, 1/2 + \sqrt{3}/6), \tag{T.1.29}$$

$$(w_1, w_2) = (1/2, 1/2). \tag{T.1.30}$$

In this case (1.2) holds for $\ell = 0$ through 3, but not for $\ell = 4$. That is, $\ell_{\max} = 3$. For $\ell = 4$ the sum on the left side of (1.2) has the value

$$\sum_{i=1}^{2} w_i (x_i)^4 = 1/5 - 4!/4320. \tag{T.1.31}$$

Correspondingly, use of Legendre Gauss gives (for the case $k = 2$) the approximation

$$\int_0^1 dx\ f(x) = (1/2)f(1/2 - \sqrt{3}/6) + (1/2)f(1/2 + \sqrt{3}/6) + (1/4320)f^{(4)}(\xi). \tag{T.1.32}$$

Compare (1.32), Legendre Gauss for $k = 2$, with (1.19), Newton Cotes for $k = 3$. They are of the same order, but Legendre Gauss has a somewhat smaller error coefficient. Thus, Legendre Gauss for $k = 2$ not only requires one less evaluation of f than Newton Cotes for $k = 3$, it is also slightly more accurate.

Finally, consider the case $k = 1$. It has sampling point and weight given by

$$x_1 = 1/2, \tag{T.1.33}$$

$$w_1 = 1, \tag{T.1.34}$$

and yields the midpoint rule

$$\int_0^1 dx\ f(x) \simeq f(1/2). \tag{T.1.35}$$

In this case (1.2) holds for $\ell = 0$ and 1, but not for $\ell = 2$. Therefore $\ell_{\max} = 1$. And for $\ell = 2$ the sum on the left side of (1.2) has the value

$$w_1(x_1)^2 = 1/3 - 2!/24. \tag{T.1.36}$$

Correspondingly, use of Legendre Gauss gives (for the case $k = 1$) the approximation

$$\int_0^1 dx\ f(x) = f(1/2) + (1/24)f^{(2)}(\xi). \tag{T.1.37}$$

It can be shown that for Legendre Gauss the relation (1.11) holds. Moreover, for a given k value, there is no sampling procedure with larger $\ell_{\max}$ than that of Legendre Gauss. All other sampling-point choices yield a smaller value of $\ell_{\max}$. Comparison of Table 1.1 and (1.11) shows that, with increasing k, Legendre Gauss rapidly becomes far more efficient than Newton Cotes.

T.1.4 Clenshaw Curtis

Like Legendre Gauss, *Clenshaw-Curtis* quadrature is most naturally described on the interval $[-1, 1]$. It uses the *Chebyshev* points as sampling points. For $k = 1$ the Chebyshev point is 0.[3] For $k = 2$ the Chebyshev points are ∓ 1, and for $k = 3$ the Chebyshev points are $\{-1, 0, 1\}$. For $k \geq 2$ select k equally spaced angles θ_i with $\theta_1 = -\pi$ and $\theta_k = \pi$,

$$\theta_i = -\pi + 2\pi(i-1)/(k-1). \tag{T.1.38}$$

For $k \geq 2$ the Chebyshev points are given by the rule

$$x_i = \cos(\theta_i). \tag{T.1.39}$$

Let $T_{k-1}(x)$ be the Chebyshev polynomial of degree $k-1$. It can be shown, for $k \geq 2$, that T_{k-1} has k extrema on the interval $[-1, 1]$ (all of which are ∓ 1). Moreover, for $k \geq 2$, these extrema are the k Chebyshev sampling points defined by (1.38) and (1.39).

Evidently for $k \leq 3$ the Chebyshev points, when transformed to the interval $[0, 1]$, are the same as the sampling points for Newton Cotes, and therefore Table 1 provides the value of $\ell_{\max}$ in these cases. It can be shown that in fact Table 1 provides the correct value of

[3] Some authors omit this point since it is really a Chebyshev point of the second kind.

$\ell_{\max}$ for Clenshaw Curtis in all cases. That is, Table 1 holds for both Newton Cotes and Clenshaw Curtis.[4]

Although the order of Clenshaw-Curtis quadrature is the same as that for Newton Cotes, and therefore much less that that of Legendre Gauss, its use has two advantages. First, if the value of $k-1$ is doubled, thereby doubling the order, essentially half of the sampling points are the same as before. Therefore, in the same spirit as embedded Runge Kutta, it is relatively easy to devise adaptive integration schemes. A second advantage, as described in the next subsection, has to do with convergence.

T.1.5 Convergence

What happens in the limit $k \to \infty$? For the interval $[-1, 1]$ it can be shown that Newton-Cotes quadrature converges to the correct result providing $f(x)$ is *analytic* in a disk centered about $x = 0$ and having radius slightly larger than 1; and is divergent if f fails to be analytic in the open unit disk centered about $x = 0$. (The case where f has singularities on the boundary of this open unit disk requires a more refined analysis.)[5] When the interval is $[0, 1]$, this result for convergence translates to the requirement that f be analytic in a disk centered about $x = 1/2$ and having radius slightly larger than $1/2$.

By contrast, and working in the interval $[-1, 1]$, Legendre-Gauss quadrature is guaranteed to converge under the much less restrictive condition that $f(x)$ simply be *sufficiently smooth* for $x \in [-1, 1]$. An adequate condition for sufficiently smooth, which is generally realized in practice, is that $f(x)$ be *Lipschitz* continuous for $x \in [-1, 1]$. Correspondingly, when transformed to the interval $[0, 1]$, Legendre-Gauss quadrature is guaranteed to converge if $f(x)$ is sufficiently smooth for $x \in [0, 1]$.

The reason for this difference is that Taylor series on the interval $[-1, 1]$ converge in disks about $x = 0$ whereas Legendre polynomial expansions in the interval $[-1, 1]$ converge in ellipses (sometimes called *Bernstein* ellipses) with foci at $x = \mp 1$.

Although for a given k Clenshaw-Curtis quadrature has relatively low order (the same as Newton Cotes) compared to Legendre Gauss, its convergence properties are similar to those for Legendre Gauss. It too, when transformed to the interval $[0, 1]$, is guaranteed to converge under the much less restrictive condition that $f(x)$ simply be sufficiently smooth for $x \in [0, 1]$.

What can be said about Legendre-Gauss and Clenshaw-Curtis convergence in the case that f is analytic? Again it is most convenient to employ the interval $[-1, 1]$ with the understanding that the results obtained for this interval can be easily be transformed to the interval $[0, 1]$. Let ρ be a real number with $\rho > 1$. Consider the points z in the complex plane given by the rule

$$z = \rho \exp(i\phi) \tag{T.1.40}$$

[4] From (1.38) and (1.39) it follows that the Chebyshev points are symmetrically distributed about $x = 0$. Correspondingly the weights for the points $\mp x_i$ are the same. Therefore, by symmetry, Clenshaw Curtis is exact (yields the value 0) for all odd-degree monomials.

[5] Runge considered the function $f(x) = 1/(1 + 25x^2)$, now called the Runge function, on the interval $x \in [-1, 1]$ and showed that Newton-Cotes quadrature diverges for this example. Note that Runge's f is analytic on this interval, but has poles at the points $x = \pm i/5$, and these poles lie inside the unit disk centered about the origin.

with $\phi \in [0, 2\pi)$. Evidently they lie on a circle about the origin with radius ρ. Next consider points w related to points z by the rule

$$w = (1/2)(z + z^{-1}). \tag{T.1.41}$$

It can be verified that the image of the circle (1.40) under the transformation (1.41) is an ellipse in the complex plane (a Bernstein ellipse E_ρ) with foci ∓ 1,

$$\text{semi-major axis } = (1/2)(\rho + 1/\rho), \tag{T.1.42}$$

and

$$\text{semi-minor axis } = (1/2)(\rho - 1/\rho). \tag{T.1.43}$$

Suppose f is analytic on $[-1, 1]$ and that it can be analytically continued (see Figure 32.4.9) into the interior of some E_ρ without encountering a singularity, and also suppose that there are no singularities on the boundary (E_ρ itself). Then it can be shown that for large k the error in Legendre-Gauss quadrature must go to zero at least as fast as

$$\text{Legendre-Gauss quadrature error} \sim \exp[-2k \log(\rho)].$$

And for Clenshaw-Curtis quadrature the error must go to zero at least as fast as

$$\text{Clenshaw-Curtis quadrature error} \sim \exp[-k \log(\rho)].$$

Thus, in both cases, the error goes to zero *exponentially* with increasing k. And the larger the value of ρ can be without there being singularities inside or on E_ρ, the more rapid the exponential decrease.

Exercises

T.1.1. The sampling option (1.14) involves use of the endpoints 0 and 1, and for this reason is more precisely called *closed* Newton Cotes. It is also possible to employ a sampling procedure, called *open* Newton Cotes, for which the x_i are still equally spaced but the endpoints are not used. Consider, for example, the case $k = 3$ and the two equally spaced sampling procedures

$$(x_1, x_2, x_3) = (0, 1/2, 1) \text{ closed Newton Cotes}, \tag{T.1.44}$$

and

$$(x_1, x_2, x_3) = (1/4, 1/2, 3/4) \text{ open Newton Cotes}. \tag{T.1.45}$$

For $k = 3$ open Newton Cotes the weights are

$$(w_1, w_2, w_3) = (2/3, -1/3, 2/3). \tag{T.1.46}$$

Show that, just as for the case of closed Newton Cotes, (1.2) holds for $k = 3$ open Newton Cotes when $\ell = 0, 1, 2, 3$. Show that, for $k = 3$ open Newton Cotes,

$$\sum_{i=1}^{3} w_i (x_i)^4 = 1/(4 + 1) + (14)(4!)/[45(4^5)] = 1/5 + (4!)(7/23040). \tag{T.1.47}$$

Correspondingly, assuming that f is sufficiently differentiable, use of $k = 3$ open Newton Cotes gives the approximation

$$\int_0^1 dx\ f(x) = (2/3)f(1/4) - (1/3)f(1/2) + (2/3)f(3/4) + (7/23040)f^{(4)}(\xi) \qquad \text{(T.1.48)}$$

where $\xi \in [0, 1]$.

For $k = 4$ open Newton Cotes the sampling points are

$$(x_1, x_2, x_3, x_4) = (1/5, 2/5, 3/5, 4/5), \qquad \text{(T.1.49)}$$

and the weights are

$$(w_1, w_2, w_3, w_4) = (11/24, 1/24, 1/24, 11/24). \qquad \text{(T.1.50)}$$

Correspondingly, assuming that f is sufficiently differentiable, use of $k = 4$ open Newton Cotes gives the approximation

$$\begin{aligned} \int_0^1 dx\ f(x) \quad &= \quad (11/24)f(1/5) + (1/24)f(2/5) + (1/24)f(3/5) + (11/24)f(4/5) \\ &\quad +(19/90000)f^{(4)}(\xi) \qquad \text{(T.1.51)} \end{aligned}$$

where again $\xi \in [0, 1]$.

It can be shown that, for a given value of k, both closed and open Newton Cotes have the same $\ell_{\max}$. Thus Table 1.1 holds for both open and closed Newton Cotes. Comparison of (1.19) and (1.48) shows that for $k = 3$ the error coefficient for open Newton Cotes is slightly smaller than that for closed Newton Cotes. This case is an anomaly. For $k \geq 4$ closed Newton Cotes has a smaller error coefficient than open Newton Cotes. Finally, we remark that $k = 1$ Legendre Gauss my also be viewed as being $k = 1$ open Newton Cotes.

T.1.2. Consider quadrature on the interval $[-1, 1]$. Let $f(x)$ be the third-order polynomial

$$f(x) = a + bx + cx^2 + dx^3. \qquad \text{(T.1.52)}$$

Verify that

$$\int_{-1}^1 dx\ f(x) = 2a + (2/3)c. \qquad \text{(T.1.53)}$$

Observe that the right side of (1.53) is a particular linear combination of the Taylor coefficients for f, and the value of the integral on the left side of (1.53) depends only on the value of this particular combination.

The $k = 2$ Legendre-Gauss sampling points and weights on the interval $[-1, 1]$ are

$$(x_1, x_2) = (-1/\sqrt{3}, 1/\sqrt{3}), \qquad \text{(T.1.54)}$$

$$(w_1, w_2) = (1, 1). \qquad \text{(T.1.55)}$$

Verify that

$$\sum_{i=1}^{2} w_i f(x_i) = 2a + (2/3)c, \qquad \text{(T.1.56)}$$

and therefore $k = 2$ Legendre-Gauss quadrature is exact for all polynomials of degree 3 or less. Verify that Legendre Gauss is the unique $k = 2$ quadrature rule with this property. Also verify that $k = 2$ Legendre-Gauss quadrature fails for x^4.

T.1.3. There is another way of viewing quadrature formulas that is more akin to the numerical integration of differential equations. Consider the single-variable differential equation

$$dy/dt = g(t) \tag{T.1.57}$$

with the initial condition

$$y(0) = 0. \tag{T.1.58}$$

Note that g does not depend on y, and therefore (1.57) has the immediate solution

$$y(t) = \int_0^t dt' g(t'). \tag{T.1.59}$$

Suppose, in the spirit of a local stepping formula, we wish to compute $y(h)$ through some order in h. From (1.59) we find

$$y(h) = \int_0^h dt g(t). \tag{T.1.60}$$

Bring the right side of (1.60) to the $\int_0^1$ standard form by making the change of variable and definition

$$t = xh, \tag{T.1.61}$$

$$f(x) = g(xh). \tag{T.1.62}$$

Show that (1.60) then becomes

$$y(h) = h \int_0^1 dx f(x). \tag{T.1.63}$$

Suppose the right side of (1.63) is evaluated using Newton Cotes with k odd. Show that there is the result

$$y(h) = h \sum_{i=1}^{k} w_i f(x_i) + c_{k+1} h f^{k+1}(\xi) \tag{T.1.64}$$

where c_{k+1} is a coefficient that can be read off from formulas such as (1.19) and (1.48). Finally, show that

$$f^{k+1}(\xi) = h^{k+1} g^{k+1}(\tau), \tag{T.1.65}$$

where $\tau \in [0, h]$, so that (1.64) becomes

$$y(h) = h \sum_{i=1}^{k} w_i g(x_i h) + c_{k+1} h^{k+2} g^{k+1}(\tau), \; k \text{ odd.} \tag{T.1.66}$$

We see that, for k function evaluations with k odd, Newton Cotes provides a stepping formula with local error of order h^{k+2}, and therefore local accuracy through terms of order h^{k+1}.

Suppose the right side of (1.63) is evaluated using Newton Cotes with k even. Show that then there is the result

$$y(h) = h \sum_{i=1}^{k} w_i f(x_i) + c_k h f^k(\xi) \tag{T.1.67}$$

where c_k is a coefficient that can be read off from formulas such as (1.22) and (1.51). Finally, show that

$$f^k(\xi) = h^k g^k(\tau), \tag{T.1.68}$$

where $\tau \in [0, h]$, so that (1.67) becomes

$$y(h) = h \sum_{i=1}^{k} w_i g(x_i h) + c_k h^{k+1} g^k(\tau),\ k \text{ even.} \tag{T.1.69}$$

We see that, for k function evaluations with k even, Newton Cotes provides a stepping formula with local error of order h^{k+1}, and therefore local accuracy through terms of order h^k.

Suppose, instead, that the right side of (1.63) is evaluated using Legendre Gauss. Then there is the result

$$y(h) = h \sum_{i=1}^{k} w_i f(x_i) + c_{2k} h f^{2k}(\xi) \tag{T.1.70}$$

where c_{2k} is a coefficient that can be read off from formulas such as (1.27), (1.32), and (1.37). Now show that

$$f^{2k}(\xi) = h^{2k} g^{2k}(\tau), \tag{T.1.71}$$

where $\tau \in [0, h]$, so that (1.70) becomes

$$y(h) = h \sum_{i=1}^{k} w_i g(x_i h) + c_{2k} h^{2k+1} g^{2k}(\tau). \tag{T.1.72}$$

We see that, for k function evaluations, Legendre Gauss provides a stepping formula with local error of order h^{2k+1}, and therefore local accuracy through terms of order h^{2k}.

T.1.4. Verify that E_ρ given by (1.40) and (1.41) is indeed an ellipse with the advertised properties.

T.2 Cubature Formulas

T.2.1 Introduction

Cubature formulas extend the concept of qudrature formulas to the case of domains D having dimension greater than one. All such formulas are called cubature formulas, no matter what the dimension of D, as long as this dimension is greater than one.

Let D be some domain having dimension m. Label points within D by m-dimensional vectors

$$x = (x_1, x_2, \cdots, x_m), \tag{T.2.1}$$

and let $f(x)$ be a function defined on D. Then a cubature formula is a set of k sampling points in D, now call them x^i, and weights w^i such that

$$\int_D d^m x\ f(x) \simeq \sum_{i=1}^{k} w^i f(x^i). \tag{T.2.2}$$

Again the challenge is to select the sampling points and weights in such a way that the approximation (2.2) is optimal and to define what is meant by optimal. Results are known for some standard domains. For our purposes, we are interested in the cases where D is either a square, rectangle, or S^2 (the surface of a 2-sphere).

T.2.2 Cubature on a Square

The unit m-cube, denoted by C^m, is defined to be the domain

$$-1 \leq x_c \leq 1 \quad c = 1, \cdots, m. \tag{T.2.3}$$

A variety of cubature formulas are known for the C^m for all m. We will be particularly interested in the case C^2, which we call the unit square.

Let P_{j_1,j_2} denote the monomial

$$P_{j_1,j_2}(x) = x_1^{j_1} x_2^{j_2}. \tag{T.2.4}$$

It has degree

$$\hat{d} = j_1 + j_2. \tag{T.2.5}$$

From the work of Exercise 7.10.2 we know that the number of such monomials of degree 0 through d is given by

$$S_0(2, d) = (2 + d)!/(2!d!). \tag{T.2.6}$$

See (7.10.17). For convenience, the values of $S_0(2, d)$ are tabulated below for the first few values of d.

Table T.2.1: $S_0(2, d)$ as a Function of d.

d	0	1	2	3	4	5	6	7
$S_0(2, d)$	1	3	6	10	15	21	28	36

Let D be the domain C^2. Again, because of the results of Weierstrass and Taylor, as our definition of optimal we will seek k sampling points x^i and weights w^i such that

$$\int_D d^2x \; P_{j_1,j_2}(x) = \sum_{i=1}^{k} w^i P_{j_1,j_2}(x^i) \tag{T.2.7}$$

for all monomials of degree less than or equal to d. We note that the left side of (2.7) is easily evaluated to give the result

$$\int_D d^2x \; P_{j_1,j_2}(x) = 4/[(j_1 + 1)(j_2 + 1)] \quad j_1, j_2 \text{ both even,} \tag{T.2.8}$$

$$= 0 \quad \text{otherwise.} \tag{T.2.9}$$

Observe that, for the 2-dimensional case, the specification of each sampling point x^i requires 2 values. Consequently, the specification of k sampling points and k weights involves the specification of $2k + k = 3k$ values. On the other hand, there are $S_0(2, d)$ conditions of the form (2.7) that need to be met. Therefore naive counting suggests that (2.7) can be satisfied providing

$$3k \geq S_0(2, d). \tag{T.2.10}$$

For example, for the case $d = 5$, we see from Table 2.1 that (2.10) yields the requirement

$$3k \geq 21, \text{ or } k \geq 7. \tag{T.2.11}$$

We note that a formula that works for $d = 5$ can be constructed by converting the integral over the unit square into two iterated integrals, and these integrals (since they must be exact for single-variable monomials of degrees 0 through 5) can each be approximated using $[-1, 1]$ variants of the 3-point Legendre-Gauss formula (1.25). Doing so produces what is called a *product* cubature fromula, and evidently has $k = 3 \times 3 = 9$.

Remarkably, for the unit square and $d = 5$, there are two cubature formulas for which $k = 7$. *Stroud*, in his book on the approximate calculation of multiple integrals, refers to them as $^*C^2$:5-1 and $^*C^2$:5-2. In these formulas several sampling points have the same weight. The formulas are described below. Each description consists of weights and the sampling points having these weights. Also shown, in Figures 2.1 and 2.2, are the locations of the sampling points in the x_1, x_2 plane. Note that all sampling points lie *within* the unit square, a feature that is essential for our purposes; and all have positive weights, a feature that is numerically desirable.

The Unit Square Cubature Formula $^*C^2$:5-1

weights	points
8/7	$(0, 0)$
20/36	$(\pm r, \pm s)$
20/63	$(0, \pm t)$

$$\begin{aligned} r &= \sqrt{3/5} \simeq .775 \\ s &= \sqrt{1/3} \simeq .577 \\ t &= \sqrt{14/15} \simeq .966 \end{aligned} \tag{T.2.12}$$

The Unit Square Cubature Formula $^*C^2$:5-2

weights	points
$8/7$	$(0,0)$
$100/168$	$\pm\,(r,r)$
$20/48$	$\pm\,(s,-t)$
$20/48$	$\pm\,(t,-s)$

$$\begin{aligned} r &= \sqrt{7/15} \simeq .683 \\ s &= [(7+\sqrt{24})/15]^{1/2} \simeq .891 \\ t &= [(7-\sqrt{24})/15]^{1/2} \simeq .374 \end{aligned} \tag{T.2.13}$$

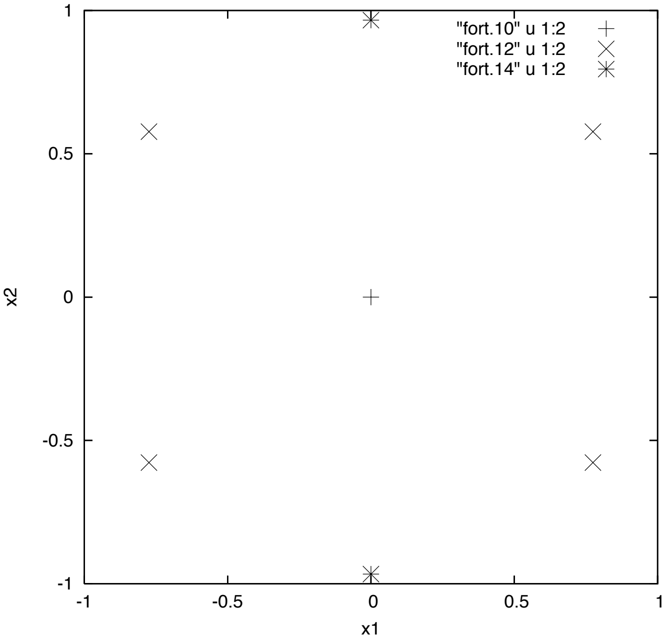


Figure T.2.1: Sampling points for $^*C^2$:5-1

Rough Error Analysis

Analysis of the errors involved with the use of cubature formulas is a complicated subject. However, one simple thing that can be done is to see how well (2.7) works for monomials having degree $d+1$ when the method has been constructed to work for monomials having degrees less than or equal to d. For example, for the two unit-square $d=5$ and $k=7$ cubature formulas we have been discussing, we can examine how well they work for monomials of degree $5+1=6$.

Observe, for these two unit-square cubature formulas, that the $(0,0)$ sampling point does not contribute to the result for any first or higher degree monomial. With regard to the other sampling points, the sampling points for each weight are symmetrically arranged in

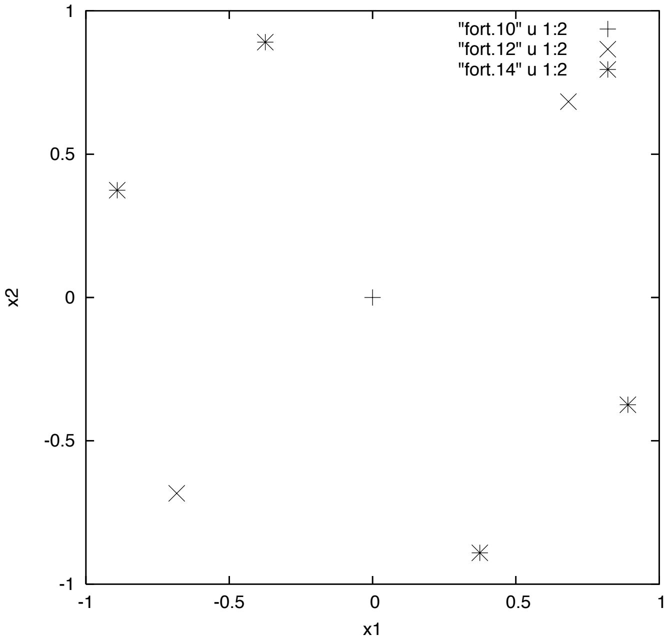


Figure T.2.2: Sampling points for ${}^*C^2$:5-2

such a way that their net contribution for any monomial $x_1^{j_1}x_2^{j_2}$ is identically zero if either j_1 or j_2 is odd, but not both are odd. Therefore these cubature formulas automatically satisfy (2.7) and (2.9) exactly for this subset of monomials,

$$\sum_{i=1}^{5} w^i P_{j_1,j_2}(x^i) = \int_D d^2x\ P_{j_1,j_2}(x) = 0 \quad \text{if either } j_1 \text{ or } j_2 \text{ is odd, but not both are odd.} \tag{T.2.14}$$

Note that, by this symmetry, (2.7) is automatically satisfied for *all* odd degree polynomials.

With regard to monomials of degree 6, we need to examine the performance of the methods (2.12) and (2.13) separately. Comparison of Figures 2.1 and 2.2 shows that the sampling points for the method (2.12) are more symmetrically arranged. In particular, for this method we have the stronger result that for all monomials

$$\sum_{i=1}^{5} w^i P_{j_1,j_2}(x^i) = 0 \quad \text{if either } j_1 \text{ or } j_2 \text{ is odd, or both are odd.} \tag{T.2.15}$$

Therefore this cubature formula automatically satisfies (2.9) exactly for all monomials of any degree if either j_1 or j_2 is odd, or both are odd. What remains to be examined for monomials of degree 6 are the cases x_1^6, $x_1^4x_2^2$, $x_1^2x_2^4$, and x_2^6, for which, according to (2.8), the exact results are 4/7, 4/15, 4/15, and 4/7, respectively. Numerical calculations show that, for ${}^*C^2$:5-1, there are the results

$$\sum_{i=1}^{5} w^i P_{6,0}(x^i) = 4/7 - .0914\cdots = (4/7)(1 - .16\cdots), \tag{T.2.16}$$

$$\sum_{i=1}^{5} w^i P_{4,2}(x^i) = 4/15 + 0, \tag{T.2.17}$$

$$\sum_{i=1}^{5} w^i P_{2,4}(x^i) = 4/15 - .1185\cdots = (4/15)(1 - .44\cdots), \tag{T.2.18}$$

$$\sum_{i=1}^{5} w^i P_{0,6}(x^i) = 4/7 + .0271\cdots = (4/7)(1 + .05\cdots). \tag{T.2.19}$$

Surprisingly, (2.17) shows that ${}^*C^2$:5-1 is exact for $x_1^4 x_2^2$! We conclude, with regard to monomials of degrees 0 through 7, that ${}^*C^2$:5-1 is exact for all such monomials save for the monomials x_1^6, $x_1^2 x_2^4$, and x_2^6 where it makes errors of approximately 16%, 44%, and 5%, respectively.

For ${}^*C^2$:5-2, the sampling points are less symmetrically arranged so that a larger number of degree 6 monomials need to be considered separately. Numerical calculations show that there are the results

$$\sum_{i=1}^{5} w^i P_{6,0}(x^i) = 4/7 - .032\cdots = (4/7)(1 - .06\cdots), \tag{T.2.20}$$

$$\sum_{i=1}^{5} w^i P_{5,1}(x^i) = 0 - .059\cdots, \tag{T.2.21}$$

$$\sum_{i=1}^{5} w^i P_{4,2}(x^i) = 4/15 - .059\cdots = (4/15)(1 - .22\cdots), \tag{T.2.22}$$

$$\sum_{i=1}^{5} w^i P_{3,3}(x^i) = 0 + .059\cdots, \tag{T.2.23}$$

$$\sum_{i=1}^{5} w^i P_{2,4}(x^i) = 4/15 - .059\cdots = (4/15)(1 - .22\cdots), \tag{T.2.24}$$

$$\sum_{i=1}^{5} w^i P_{1,5}(x^i) = 0 - .059\cdots, \tag{T.2.25}$$

$$\sum_{i=1}^{5} w^i P_{0,6}(x^i) = 4/7 - .032\cdots = (4/7)(1 - .06\cdots). \tag{T.2.26}$$

We conclude, with regard to monomials of degrees 0 through 7, that ${}^*C^2$:5-2 is exact for all such monomials save for all the monomials of degree 6 where it makes the errors indicated above.

Finally, comparison of (2.16) through (2.19) for ${}^*C^2$:5-1 with (2.20) through (2.26) for ${}^*C^2$:5-2 shows that, save for the case $x_1^2 x_2^4$, ${}^*C^2$:5-1 generally makes smaller errors than ${}^*C^2$:5-2, and therefore might be slightly preferred.

T.2.3 Cubature on a Rectangle

Let $R(a, b)$ be a rectangle with sides $2a$ and $2b$ parameterized by coordinates ξ_1, ξ_2 such that

$$\xi_1 \in [-a, a], \quad \xi_2 \in [-b, b]. \tag{T.2.27}$$

Let $g(\xi)$ be some function defined on $R(a, b)$ and suppose we wish to evaluate the integral

$$\int_{R(a,b)} d^2\xi \; g(\xi). \tag{T.2.28}$$

Our goal is to find k sampling points ξ^i and k weights $\bar{w}^i$ such that

$$\int_{R(a,b)} d^2\xi \; g(\xi) \simeq \sum_{i=1}^{k} \bar{w}^i g(\xi^i). \tag{T.2.29}$$

Introduce new variables x_1, x_2 by the rules

$$\xi_1 = ax_1, \quad \xi_2 = bx_1 \tag{T.2.30}$$

so that (2.30) maps C^2 into $R(a, b)$. See (2.3). Then there is the relation

$$\int_{R(a,b)} d^2\xi \; g(\xi) = ab \int_{C^2} d^2x \; f(x) \tag{T.2.31}$$

where

$$f(x) = g(ax_1, bx_2). \tag{T.2.32}$$

From (2.2) we have the approximation

$$\int_{C^2} d^2x \; f(x) \simeq \sum_{i=1}^{k} w^i f(x^i). \tag{T.2.33}$$

It follows that

$$\int_{R(a,b)} d^2\xi \; g(\xi) \simeq ab \sum_{i=1}^{k} w^i f(x^i) = ab \sum_{i=1}^{k} w^i g(ax_1^i, bx_2^i). \tag{T.2.34}$$

We conclude that (2.29) holds providing we make the definitions

$$\bar{w}^i = abw^i, \tag{T.2.35}$$

$$\xi_1^i = ax_1^i, \quad \xi_2^i = bx_2^i. \tag{T.2.36}$$

Exercises

T.2.1. The purpose of this exercise is to study how the accuracy of a cubature formula depends on the size of the integration domain and the analytic properties of the function being integrated. For simplicity, we will consider the case (2.29), cubature on a rectangle.

Suppose that g has a Taylor expansion of the form

$$g(\xi) = \sum_{j_1,j_2} g_{j_1,j_2}\xi_1^{j_1}\xi_2^{j_2}. \tag{T.2.37}$$

Rearrange this expansion into one in homogenous polynomials by writing

$$g(\xi) = \sum_{\ell} P_\ell(\xi) \tag{T.2.38}$$

where

$$P_\ell(\xi) = \sum_{j_1+j_2=\ell} g_{j_1,j_2}\xi_1^{j_1}\xi_2^{j_2}. \tag{T.2.39}$$

Since we will be employing cubature formulas that are exact for polynomials through degree d, let use rewrite (2.38) in the form

$$g(\xi) = \sum_{\ell=0}^{d} P_\ell(\xi) + \Delta(\xi) \tag{T.2.40}$$

where

$$\Delta(\xi) = \sum_{\ell=d+1}^{\infty} P_\ell(\xi) = \sum_{\ell=d+1}^{\infty}\sum_{j_1+j_2=\ell} g_{j_1,j_2}\xi_1^{j_1}\xi_2^{j_2}. \tag{T.2.41}$$

In a moment, we will treat Δ as an error term. First, assuming suitable analyticity for g, use the Cauchy bound (33.2.18) on Taylor coefficients to show that, for $\xi \in R(a,b)$, Δ has the bound

$$|\Delta(\xi)| \le \sum_{\ell=d+1}^{\infty}\sum_{j_1+j_2=\ell} |g_{j_1,j_2}||\xi_1^{j_1}||\xi_2^{j_2}| \le L \tag{T.2.42}$$

where

$$L = K\sum_{\ell=d+1}^{\infty}\sum_{j_1+j_2=\ell} (a/R_1')^{j_1}(b/R_2')^{j_2}. \tag{T.2.43}$$

Verify that the series for L converges, when $a < R_1'$ and $b < R_2'$, by showing that there is the relation

$$\begin{aligned}\sum_{\ell=d+1}^{\infty}\sum_{j_1+j_2=\ell} (a/R_1')^{j_1}(b/R_2')^{j_2} \;&\le\; \sum_{\ell=0}^{\infty}\sum_{j_1+j_2=\ell} (a/R_1')^{j_1}(b/R_2')^{j_2} = \sum_{j_1,j_2}(a/R_1')^{j_1}(b/R_2')^{j_2} \\ &=\; [1/(1-a/R_1')][1/(1-b/R_2')].\end{aligned} \tag{T.2.44}$$

Therefore, L is well defined.

Next, from (2.40), verify the relations

$$\int_{R(a,b)} d^2\xi\ g(\xi) = \int_{R(a,b)} d^2\xi\ \sum_{\ell=0}^{d} P_\ell(\xi) + \int_{R(a,b)} d^2\xi\ \Delta(\xi), \tag{T.2.45}$$

$$\sum_{i=1}^{k} \bar{w}^i g(\xi^i) = \sum_{i=1}^{k} \bar{w}^i \sum_{\ell=0}^{d} P_\ell(\xi^i) + \sum_{i=1}^{k} \bar{w}^i \Delta(\xi^i). \tag{T.2.46}$$

Show that, by the construction of the sampling points ξ^i and the weights $\bar{w}^i$, there must be the relation

$$\int_{R(a,b)} d^2\xi\ \sum_{\ell=0}^{d} P_\ell(\xi) = \sum_{i=1}^{k} \bar{w}^i \sum_{\ell=0}^{d} P_\ell(\xi^i). \tag{T.2.47}$$

Then subtract (2.46) from (2.45) to show that

$$\int_{R(a,b)} d^2\xi\ g(\xi) - \sum_{i=1}^{k} \bar{w}^i g(\xi^i) = \int_{R(a,b)} d^2\xi\ \Delta(\xi) - \sum_{i=1}^{k} \bar{w}^i \Delta(\xi^i). \tag{T.2.48}$$

Consequently, verify that there is the inequality

$$|\int_{R(a,b)} d^2\xi\ g(\xi) - \sum_{i=1}^{k} \bar{w}^i g(\xi^i)| \le |\int_{R(a,b)} d^2\xi\ \Delta(\xi)| + |\sum_{i=1}^{k} \bar{w}^i \Delta(\xi^i)|. \tag{T.2.49}$$

To continue, verify the bounds

$$|\int_{R(a,b)} d^2\xi\ \Delta(\xi)| \le 4abL, \tag{T.2.50}$$

$$|\sum_{i=1}^{k} \bar{w}^i \Delta(\xi^i)| \le 4abL. \tag{T.2.51}$$

Thus, show that there is the final inequality

$$|\int_{R(a,b)} d^2\xi\ g(\xi) - \sum_{i=1}^{k} \bar{w}^i g(\xi^i)| \le 8abL. \tag{T.2.52}$$

Your challenge is to determine how the error term L depends on the dimensions of the rectangle $R(a,b)$. To this end suppose a and b are scaled by a common factor of σ,

$$a(\sigma) = \sigma a_1, \quad b(\sigma) = \sigma b_1 \tag{T.2.53}$$

so that $R(a,b)$ is what we may call the original rectangle $R(a_1,b_1)$ when $\sigma = 1$, and $R(a,b)$ becomes ever smaller as $\sigma \to 0$.

Based on (2.43), define $L(\sigma)$ by writing

$$L(\sigma) = K \sum_{\ell=d+1}^{\infty} \sum_{j_1+j_2=\ell} (\sigma a_1/R_1')^{j_1} (\sigma b_1/R_2')^{j_2}. \tag{T.2.54}$$

Show that

$$\begin{aligned} L(\sigma) &= \sigma^{d+1}K\sum_{\ell=d+1}^{\infty}\sigma^{\ell-d-1}\sum_{j_1+j_2=\ell}(a_1/R_1')^{j_1}(b_1/R_2')^{j_2} \\ &= \sigma^{d+1}K' + O(\sigma^{d+2}) \end{aligned} \tag{T.2.55}$$

where

$$K' = K\sum_{j_1+j_2=d+1}(a_1/R_1')^{j_1}(b_1/R_2')^{j_2}. \tag{T.2.56}$$

Insert (2.55) into (2.52) to obtain the inequality

$$|\int_{R(a,b)} d^2\xi\ g(\xi) - \sum_{i=1}^{k}\bar{w}^i g(\xi^i)| \le 8\sigma^2 a_1 b_1 L(\sigma) = \sigma^{d+3}8a_1b_1K' + O(\sigma^{d+4}). \tag{T.2.57}$$

We expect that each of the two terms on the left side of (2.57) scales as σ^2. You have shown that their difference, the error, scales as σ^{d+3}. Therefore, the *relative* error scales as σ^{d+1}.

T.2.2. Review Exercise 2.1. The purpose of this exercise is to find a more refined error bound for the specific case of the cubature formula ${}^*C^2$:5-1 by exploiting the relations (2.16) through (2.19). Define two linear functionals I and Q by the rules

$$I[g] = \int_{R(a,b)} d^2\xi\ g(\xi), \tag{T.2.58}$$

$$Q[g] = \sum_{i=1}^{5}\bar{w}^i g(\xi^i). \tag{T.2.59}$$

Show, when ${}^*C^2$:5-1 is used in (2.35) and (2.36), that

$$Q[P_{j_1,j_2}] = I[P_{j_1,j_2}] \tag{T.2.60}$$

for all P_{j_1,j_2} having degree less than 6, and also for all P_{j_1,j_2} having degree 7.

Monomials of degree 6 need to be treated separately. With reference to (2.16) through (2.19), define constants $\lambda_{6,0}$ etc. by the rules

$$\lambda_{6,0} = .0914\cdots, \tag{T.2.61}$$

$$\lambda_{2,4} = .1185\cdots, \tag{T.2.62}$$

$$\lambda_{0,6} = .0271\cdots. \tag{T.2.63}$$

Show that, when ${}^*C^2$:5-1 is used in (2.35) and (2.36), there are the results

$$Q[P_{6,0}] = I[P_{6,0}] - a^7b\lambda_{6,0}, \tag{T.2.64}$$

$$Q[P_{2,4}] = I[P_{2,4}] - a^3b^5\lambda_{2,4}, \tag{T.2.65}$$

$$Q[P_{0,6}] = I[P_{0,6}] + ab^7\lambda_{0,6}, \tag{T.2.66}$$

and that (2.60) holds for all the remaining monomials of degree 6.

Using the machinery just developed show that, when ${}^*C^2$:5-1 is used in (2.35) and (2.36), there is the result

$$\int_{R(a,b)} d^2\xi \; g(\xi) \;=\; \sum_{i=1}^{5} \bar{w}^i g(\xi^i) + ab[\lambda_{6,0}g_{6,0}a^6 + \lambda_{2,4}g_{2,4}a^2b^4 - \lambda_{0,6}g_{0,6}b^6] + O(\sigma^{10}). \tag{T.2.67}$$

Note that

$$g_{6,0} = (1/6!)g^{[6,0]}(0,0), \;\; g_{2,4} = (1/2!)(1/4!)g^{[2,4]}(0,0), \;\; g_{0,6} = (1/6!)g^{[0,6]}(0,0). \tag{T.2.68}$$

Show that (2.67) can also be written in the form

$$\int_{R(a,b)} d^2\xi \; g(\xi) \;=\; \sum_{i=1}^{5} \bar{w}^i g(\xi^i) + \sigma^8 a_1 b_1[\lambda_{6,0}g_{6,0}a_1^6 + \lambda_{2,4}g_{2,4}a_1^2b_1^4 - \lambda_{0,6}g_{0,6}b_1^6] + O(\sigma^{10}). \tag{T.2.69}$$

Therefore, for the case where ${}^*C^2$:5-1 is used in (2.35) and (2.36), you have found an exact result for the $O(\sigma^{d+3})$ term in (2.57) and have shown that the $O(\sigma^{d+4})$ term, in fact, vanishes.

Examination of (2.67) and (2.68) indicates that there is still some freedom left in the choice of orientation of the sampling point scheme and the aspect ratio of the rectangle $R(a,b)$. In some cases it may be possible to minimize the leading error term in (2.67) by exploiting this freedom. Suppose that something is known about the relative sizes of the terms $g_{6,0}$, $g_{4,2}$, $g_{2,4}$, and $g_{0,6}$. If, for example, $g_{4,2}$ is much smaller than $g_{2,4}$, then it might be advantageous to rotate the sampling points shown in Figure 2.1 by 90° so that the cubature formula error no longer involves $g_{2,4}$ but instead involves $g_{4,2}$. With regard to the aspect ratio of $R(a,b)$, in the case that (2.67) is employed and assuming that $g_{6,0}$ and $g_{0,6}$ are comparable, it might be advantageous to use rectangles for which $a < b$ since $\lambda_{6,0} > \lambda_{0,6}$. Obviously, when contemplating this strategy, attention should also be paid to the size of the $\lambda_{2,4}g_{2,4}a^2b^4$ error term.

T.2.4 Cubature on the Two-Sphere

Bibliography

Quadrature Formulas

[1] P. Davis and P. Rabinowitz, *Methods of Numerical Integration*, Second Edition, Dover (2007).

[2] F.B. Hildebrand, *Introduction to Numerical Analysis*, Second Edition, Dover (1987).

[3] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover (1972). Also available on the Web by Googling "abramowitz and stegun 1972". We note that the original June 1964 edition of this venerable reference has an error with regard to Gaussian quadrature. There formula 25.4.30 for R_n has an erroneous factor of 2^{2n+1}. This error has been corrected in the 1972 edition.

[4] F. Olver, D. Lozier, R. Boisvert, and C. Clark, Editors, *NIST Handbook of Mathematical Functions*, Cambridge (2010). See also the Web site http://dlmf.nist.gov/.

[5] L.N. Trefethen, *Approximation Theory and Approximation Practice*, SIAM (2013).

[6] L.N. Trefethen, *Six Myths of Polynomial Interpolation and Quadrature*, http://people.maths.ox.ac.uk/trefethen/publication/PDF/2011_139.pdf.

[7] B. Beers and A.J. Dragt, "New Theorems about Spherical Harmonic Expansions and $SU(2)$", *J. Math. Phys.* **11**, 2313 (1970).

Cubature Formulas

[8] A.H. Stroud, *Approximate calculation of multiple integrals*, Prentice-Hall (1971).

[9] H. Engels, *Numerical Quadrature and Cubature*, Academic Press (1980).

[10] R. Cools, see, for example, the Web sites

http://nines.cs.kuleuven.be/research/ecf/Teksten/accepted.pdf
http://nines.cs.kuleuven.be/ecf/

Appendix U

Rotational Classification and Properties of Polynomials and Analytic/Polynomial Vector Fields

U.1 Introduction

Suppose $\boldsymbol{A}(\boldsymbol{r})$ is an analytic vector field in three dimensions. That is, suppose there are three component functions $A_x(\boldsymbol{r})$, $A_y(\boldsymbol{r})$, and $A_z(\boldsymbol{r})$, all of which are analytic in some common domain in the variables x, y, and z. Without loss of generality, we may take this domain to be centered on the origin. Doing so brings us to the notationally easier problem of studying polynomial vector fields, vector fields whose components are polynomials in the variables x, y, and z. In this appendix we will study how to use $SO(3)$, the rotation group in 3 dimensions, as a tool for labeling/classifying both such polynomials and such polynomial vector fields. We will also study some of their properties.

U.2 Polynomials and Spherical Polynomials

U.2.1 Polynomials

The first step in studying multi-variable polynomials is to decompose them into homogeneous polynomials. According to (7.3.40) the number of monomials of degree n in $d = 3$ variables is given by $N(n, 3)$. And, according to (7.10.17), the total number of monomials in $d = 3$ variables having degrees 0 through n is given by $S_0(n, 3)$. Table 2.1 below shows values of $N(n, 3)$ and $S_0(n, 3)$ for various values of n. Also shown are the quantities $3N(n, 3)$ and $3S_0(n, 3)$. The quantity $3N(n, 3)$ is the number of parameters required to specify 3 homogeneous polynomials of degree n in $d = 3$ variables. And the quantity $3S_0(n, 3)$ is the number of parameters required to specify a $d = 3$ dimensional vector field in $d = 3$ variables through terms of degree n. Thus, for example, to specify a 3-dimensional vector field through terms of degree $n = 4$ requires $3S_0(4, 3) = 105$ parameters. Finally, the table displays $S_B(n)$, the number of parameters required to specify a source-free magnetic field through terms of degree n. See (2.7). Thus, for example, to specify a source-free magnetic field through terms of degree $n = 4$ requires requires $S_B(4) = 35$ parameters. Note that the requirement that the field be source free, namely divergence and curl free, reduces the parameter count substantially even for modest values of n.

Table U.2.1: $N(n,3)$, $S_0(n,3)$, $3N(n,3)$, $3S_0(n,3)$, and $S_B(n)$ as functions of n.

n	$N(n,3)$	$S_0(n,3)$	$3N(n,3)$	$3S_0(n,3)$	$S_B(n)$
0	1	1	3	3	3
1	3	4	9	12	8
2	6	10	18	30	15
3	10	20	30	60	24
4	15	35	45	105	35
5	21	56	63	168	48
6	28	84	84	252	63
7	36	120	108	360	80
8	45	165	135	495	99

U.2.2 Spherical Polar Coordinates and Harmonic Polynomials

Introduce spherical polar coordinates in the usual way as in Section 15.2.2:

$$r^2 = x^2 + y^2 + z^2, \tag{U.2.1}$$

$$x = r\sin(\theta)\cos(\phi), \tag{U.2.2}$$

$$y = r\sin(\theta)\sin(\phi), \tag{U.2.3}$$

$$z = r\cos\theta. \tag{U.2.4}$$

Let $Y_\ell^m(\theta,\phi)$ denote the usual *spherical* harmonics,

$$\begin{aligned} Y_\ell^m(\theta,\phi) &= \{[(2\ell+1)(\ell-m)!]/[4\pi(\ell+m)!]\}^{1/2} P_\ell^m(\cos\theta)\exp(im\phi) \\ &\quad \text{with } -\ell \le m \le \ell. \end{aligned} \tag{U.2.5}$$

Here the P_ℓ^m are the usual associated Legendre functions. Consider the functions

$$H_\ell^m(\boldsymbol{r}) = r^\ell Y_\ell^m(\theta,\phi). \tag{U.2.6}$$

They are homogeneous polynomials of degree ℓ in the variables x, y, and z. They are also harmonic functions, and are variously called *harmonic polynomials* or *solid* harmonics. For a given ℓ there are $2\ell+1$ such polynomials as m ranges over $-\ell \le m \le \ell$.

At this point we are prepared to compute $S_B(n)$. Evidently a source-free magnetic field homogeneous of degree ℓ results from the gradient of a harmonic polynomial of degree $\ell+1$, and there are $2(\ell+1)+1 = 2\ell+3$ such harmonic polynomials. Therefore we have the result

$$S_B(n) = \sum_{\ell=0}^{n}(2\ell+3) = 3\sum_{\ell=0}^{n}1 + 2\sum_{\ell=0}^{n}\ell = 3(n+1)+2(n/2)(n+1) = (n+1)(n+3). \tag{U.2.7}$$

U.2.3 Examples of Harmonic Polynomials and Missing Homogeneous Polynomials

Continuing with the discussion of harmonic polynomials we find, for example, the results

$$H_0^0(\boldsymbol{r}) = 1/\sqrt{4\pi}; \tag{U.2.8}$$

$$\begin{aligned}
H_1^1(\boldsymbol{r}) &= \sqrt{3/(4\pi)}(-1/\sqrt{2})(x+iy) = -\sqrt{3/(8\pi)}(x+iy),\\
H_1^0(\boldsymbol{r}) &= \sqrt{3/(4\pi)}z,\\
H_1^{-1}(\boldsymbol{r}) &= \sqrt{3/(4\pi)}(1/\sqrt{2})(x-iy) = \sqrt{3/(8\pi)}(x-iy);
\end{aligned} \tag{U.2.9}$$

$$\begin{aligned}
H_2^2(\boldsymbol{r}) &= \sqrt{15/(32\pi)}(x+iy)^2,\\
H_2^1(\boldsymbol{r}) &= -\sqrt{15/(8\pi)}(x+iy)z,\\
H_2^0(\boldsymbol{r}) &= \sqrt{5/(16\pi)}(2z^2-x^2-y^2),\\
H_2^{-1}(\boldsymbol{r}) &= \sqrt{15/(8\pi)}(x-iy)z,\\
H_2^{-2}(\boldsymbol{r}) &= \sqrt{15/(32\pi)}(x-iy)^2.
\end{aligned} \tag{U.2.10}$$

Note that there is one function for the case $\ell = 0$, three functions for the case $\ell = 1$, and five functions in the case $\ell = 2$. Comparison of this function count with the first thee lines of Table 2.1 shows that all the homogeneous monomials of degrees 0 and 1 have been accounted for, but one homogenous polynomial of degree 2 is missing. This missing polynomial is evidently proportional to $r^2 = x^2 + y^2 + z^2$, and may be taken to be $r^2 H_0^0(\boldsymbol{r})$.[1]

What about the case $n = 3$? There are the polynomials $H_3^m(\boldsymbol{r})$, and there are $2\ell + 1 = 7$ such polynomials when $\ell = 3$. But from Table 2.1 we see that there should be, in total, ten polynomials when $n = 3$. What are the remaining three? The remaining three third-degree polynomials can be taken to be the polynomials $r^2 H_1^m(\boldsymbol{r})$.

U.2.4 Spherical Polynomials

The general picture should now be clear. Form the functions $S_{n\ell}^m(\boldsymbol{r})$, which we will call *spherical polynomials*, by the rule

$$S_{n\ell}^m(\boldsymbol{r}) = r^n Y_\ell^m(\theta,\phi) \text{ with } \ell = n, n-2, \cdots, 0 \text{ for } n \text{ even}, \tag{U.2.11}$$

$$S_{n\ell}^m(\boldsymbol{r}) = r^n Y_\ell^m(\theta,\phi) \text{ with } \ell = n, n-2, \cdots, 1 \text{ for } n \text{ odd}. \tag{U.2.12}$$

So doing produces polynomials of degree n, and these polynomials form a basis for the set of all polynomials of degree n. But still more can be said. The functions Y_ℓ^m have well-defined transformation properties under rotations, and r is unchanged by rotations. It follows that the $S_{n\ell}^m(\boldsymbol{r})$ have the same transformation properties as the Y_ℓ^m.

Listed below, for possible future use, are the spherical polynomials for the cases $n = 0$, $n = 1$, and $n = 2$:

$$S_{00}^0(\boldsymbol{r}) = 1/\sqrt{4\pi}; \tag{U.2.13}$$

[1] Note that quantities of the form $r^{2k} = (x^2 + y^2 + x^2)^k$ are *polynomials* in the variables x, y, and z.

$$\begin{aligned}
&S_{11}^{1}(\boldsymbol{r}) = -\sqrt{3/(8\pi)}(x+iy),\\
&S_{11}^{0}(\boldsymbol{r}) = \sqrt{3/(4\pi)}z,\\
&S_{11}^{-1}(\boldsymbol{r}) = \sqrt{3/(8\pi)}(x-iy);
\end{aligned} \tag{U.2.14}$$

$$S_{20}^{0}(\boldsymbol{r}) = 1/\sqrt{4\pi}(x^2+y^2+z^2); \tag{U.2.15}$$

$$\begin{aligned}
&S_{22}^{2}(\boldsymbol{r}) = \sqrt{15/(32\pi)}(x+iy)^2,\\
&S_{22}^{1}(\boldsymbol{r}) = -\sqrt{15/(8\pi)}(x+iy)z,\\
&S_{22}^{0}(\boldsymbol{r}) = \sqrt{5/(16\pi)}(2z^2-x^2-y^2),\\
&S_{22}^{-1}(\boldsymbol{r}) = \sqrt{15/(8\pi)}(x-iy)z,\\
&S_{22}^{-2}(\boldsymbol{r}) = \sqrt{15/(32\pi)}(x-iy)^2.
\end{aligned} \tag{U.2.16}$$

U.3 Analytic/Polynomial Vector Fields and Spherical Polynomial Vector Fields

With the $S_{n\ell}^{m}(\boldsymbol{r})$ in hand, we next turn to the problem of classifying/labeling all vector fields whose components are polynomials in x, y, and z. Our construction will be analogous to that for vector spherical harmonics.

U.3.1 Vector Spherical Harmonics

Define a spherical basis $\boldsymbol{e}_{\pm 1}$, $\boldsymbol{e}_0$ by the rule

$$\begin{aligned}
&\boldsymbol{e}_{+1} = -(1/\sqrt{2})(\boldsymbol{e}_x + i\boldsymbol{e}_y),\\
&\boldsymbol{e}_0 = \boldsymbol{e}_z,\\
&\boldsymbol{e}_{-1} = (1/\sqrt{2})(\boldsymbol{e}_x - i\boldsymbol{e}_y).
\end{aligned} \tag{U.3.1}$$

Note the resemblance between (3.1) and (2.9) and (2.14).[2] Indeed, suppose we define three functions $r_m(\boldsymbol{r})$ by the rules

$$\begin{aligned}
&r_{+1}(\boldsymbol{r}) = \boldsymbol{r}\cdot\boldsymbol{e}_{+1} = -(1/\sqrt{2})(x+iy),\\
&r_0(\boldsymbol{r}) = \boldsymbol{r}\cdot\boldsymbol{e}_0 = z,\\
&r_{-1}(\boldsymbol{r}) = \boldsymbol{r}\cdot\boldsymbol{e}_{-1} = (1/\sqrt{2})(x-iy).
\end{aligned} \tag{U.3.2}$$

Then (2.13) can be rewritten in the form

$$S_{11}^{m}(\boldsymbol{r}) = \sqrt{3/(4\pi)}\; r_m(\boldsymbol{r}) = \sqrt{3/(4\pi)}\; \boldsymbol{r}\cdot\boldsymbol{e}_m. \tag{U.3.3}$$

[2] Note also that the basis (3.1) differs slightly from that of Exercise 3.7.22, which was selected to make the $so(3)$ structure constants real.

Moreover, we have the relation

$$\begin{aligned} r_{-1}(\boldsymbol{r})\boldsymbol{e}_{+1} + r_{+1}(\boldsymbol{r})\boldsymbol{e}_{-1} &= -(1/2)[(x-iy)(\boldsymbol{e}_x + i\boldsymbol{e}_y) + (x+iy)(\boldsymbol{e}_x - i\boldsymbol{e}_y)] \\ &= -x\boldsymbol{e}_x - y\boldsymbol{e}_y. \end{aligned} \tag{U.3.4}$$

It follows that there is the relation

$$-r_{-1}(\boldsymbol{r})\boldsymbol{e}_{+1} + r_0(\boldsymbol{r})\boldsymbol{e}_0 - r_{+1}(\boldsymbol{r})\boldsymbol{e}_{-1} = x\boldsymbol{e}_x + y\boldsymbol{e}_y + z\boldsymbol{e}_z = \boldsymbol{r}. \tag{U.3.5}$$

But, we have digressed. The *vector spherical harmonics* $\boldsymbol{Y}_{\ell J}^M(\theta,\phi)$ are defined by the rules

$$\boldsymbol{Y}_{\ell J}^M(\theta,\phi) = \sum_{m_1,m_2} C_{m_1 m_2 M}^{\ell 1 J} Y_\ell^{m_1}(\theta,\phi)\boldsymbol{e}_{m_2}. \tag{U.3.6}$$

Here the $C_{m_1 m_2 M}^{\ell 1 J}$ denote the *Clebsch-Gordan* coefficients that couple the angular momenta ℓ and 1 to produce angular momentum J.[3] In particular, there are *range* rules:

$$\text{when } \ell = 0, \text{ then } J = 1; \tag{U.3.7}$$

$$\text{when } \ell > 0, \text{ then } J \text{ can have the values } J = \ell-1, \ell, \ell+1. \tag{U.3.8}$$

The particular Clebsch-Gordan coefficients needed for our purposes are given by the relations

$$C_{M-1,1,M}^{\ell,1,\ell+1} = \sqrt{(\ell+M)(\ell+M+1)/[(2\ell+1)(2\ell+2)]}, \tag{U.3.9}$$

$$C_{M,0,M}^{\ell,1,\ell+1} = \sqrt{(\ell-M+1)(\ell+M+1)/[(2\ell+1)(\ell+1)]}, \tag{U.3.10}$$

$$C_{M+1,-1,M}^{\ell,1,\ell+1} = \sqrt{(\ell-M)(\ell-M+1)/[(2\ell+1)(2\ell+2)]}; \tag{U.3.11}$$

$$C_{M-1,1,M}^{\ell,1,\ell} = -\sqrt{(\ell+M)(\ell-M+1)/[2\ell(\ell+1)]}, \tag{U.3.12}$$

$$C_{M,0,M}^{\ell,1,\ell} = M/\sqrt{\ell(\ell+1)}, \tag{U.3.13}$$

$$C_{M+1,-1,M}^{\ell,1,\ell} = \sqrt{(\ell-M)(\ell+M+1)/[2\ell(\ell+1)]}; \tag{U.3.14}$$

$$C_{M-1,1,M}^{\ell,1,\ell-1} = \sqrt{(\ell-M)(\ell-M+1)/[2\ell(2\ell+1)]}, \tag{U.3.15}$$

$$C_{M,0,M}^{\ell,1,\ell-1} = -\sqrt{(\ell-M)(\ell+M)/[\ell(2\ell+1)]}, \tag{U.3.16}$$

$$C_{M+1,-1,M}^{\ell,1,\ell-1} = \sqrt{(\ell+M+1)(\ell+M)/[2\ell(2\ell+1)]}. \tag{U.3.17}$$

[3] Note that we write the subscripts on the vector spherical harmonics and elsewhere in the order ℓJ, the same order in which they appear in the Clebsch-Gordan coefficients. Thus, the last lower index and the upper index are *paired* and obey the rule $-J \le M \le J$. Many authors employ the opposite order, namely $J\ell$. We also remark that the Clebsch-Gordan coefficients are also sometimes called *Wigner* or vector-addition coefficients. Finally, we use the more compact notation $C_{m_1 m_2 M}^{\ell 1 J}$ for what some authors write as $C(\ell 1 J; m_1 m_2 M)$.

U.3.2 Spherical Polynomial Vector Fields

In a corresponding manner, we define *spherical polynomial vector fields* $\boldsymbol{S}_{n\ell J}^M(\boldsymbol{r})$ by the rule

$$\begin{aligned}\boldsymbol{S}_{n\ell J}^M(\boldsymbol{r}) &= \sum_{m_1,m_2} C_{m_1 m_2 M}^{\ell 1 J} S_{n\ell}^{m_1}(\boldsymbol{r})\boldsymbol{e}_{m_2} \\ &= r^n \sum_{m_1,m_2} C_{m_1 m_2 M}^{\ell 1 J} Y_\ell^{m_1}(\theta,\phi)\boldsymbol{e}_{m2} = r^n \mathbf{Y}_{\ell J}^M(\theta,\phi). \qquad \text{(U.3.18)}\end{aligned}$$

Note that by construction the components of $\boldsymbol{S}_{n\ell J}^M(\boldsymbol{r})$ are polynomial (analytic) functions of the variables x, y, and z.

For convenience, Table 3.1 lists the allowed values of the triplets $n\ell J$ in accord with the relations (2.10), (2.11), (3.7), and (3.8). And, of course, M lies in the range $-J \leq M \leq J$.

Table U.3.1: Allowed values of $n\ell J$

n	ℓ	J
0	0	1
1	1	0
1	1	1
1	1	2
2	0	1
2	2	1
2	2	2
2	2	3
3	1	0
3	1	1
3	1	2
3	3	2
3	3	3
3	3	4
·	·	·
·	·	·

U.3.3 Examples of and Counting Spherical Polynomial Vector Fields

Let us work out a first few examples. The simplest are those for $n = 0$. In this case we must have $\ell = 0$, see (2.10), and $J = 1$, see (3.7). For the $\boldsymbol{S}_{001}^M$ we find from (3.18) the results

$$\boldsymbol{S}_{001}^M(\boldsymbol{r}) = C_{0MM}^{011} S_{00}^0(\boldsymbol{r})\boldsymbol{e}_M. \qquad \text{(U.3.19)}$$

From (3.9) through (3.11) we see that all the C_{0MM}^{011} have value 1, and $S_{00}^0(\boldsymbol{r})$ is given by (2.12). Therefore, there is the final result

$$\boldsymbol{S}_{001}^M(\boldsymbol{r}) = (1/\sqrt{4\pi})\boldsymbol{e}_M. \qquad \text{(U.3.20)}$$

Note that M can take the three values $-1, 0, +1$ corresponding to the fact that a constant vector field has 3 constant components.

The next simplest case is $n = 1$. Then we must have $\ell = 1$, see (2.11), and there are the possibilities $J = 0, 1, 2$. See (3.8). For the case $J = 0$ we find from (3.18) the result

$$\boldsymbol{S}^0_{110}(\boldsymbol{r}) = C^{110}_{-1,1,0} S^{-1}_{11}(\boldsymbol{r})\boldsymbol{e}_{+1} + \ C^{110}_{000} S^0_{11}(\boldsymbol{r})\boldsymbol{e}_0 + \ C^{110}_{1,-1,0} S^1_{11}(\boldsymbol{r})\boldsymbol{e}_{-1}. \tag{U.3.21}$$

From (3.15) through (3.17) we see that the required Clebsch-Gordan coefficients have the values

$$C^{110}_{-1,1,0} = \sqrt{1/3}, \tag{U.3.22}$$

$$C^{110}_{000} = -\sqrt{1/3}, \tag{U.3.23}$$

$$C^{110}_{1,-1,0} = \sqrt{1/3}. \tag{U.3.24}$$

Also, the required $S^m_{11}(\boldsymbol{r})$ can be written in the form (3.3). Therefore, (3.21) can be rewritten in the final form

$$\boldsymbol{S}^0_{110}(\boldsymbol{r}) = [\sqrt{1/3}][\sqrt{3/(4\pi)}][r_{-1}(\boldsymbol{r})\boldsymbol{e}_{+1} - r_0(\boldsymbol{r})\boldsymbol{e}_0 + r_{+1}(\boldsymbol{r})\boldsymbol{e}_{-1}] = -\sqrt{1/(4\pi)}\ \boldsymbol{r}. \tag{U.3.25}$$

Here we have used (3.5).

For the case $J = 1$ we find from (3.18) and (3.3) the results

$$\boldsymbol{S}^M_{111}(\boldsymbol{r}) = \sum_{m_1,m_2} C^{111}_{m_1 m_2 M} S^{m_1}_{11}(\boldsymbol{r})\boldsymbol{e}_{m_2} = \sqrt{3/(4\pi)} \sum_{m_1,m_2} C^{111}_{m_1 m_2 M} r_{m_1}(\boldsymbol{r})\boldsymbol{e}_{m_2}. \tag{U.3.26}$$

It follows that

$$\boldsymbol{S}^1_{111}(\boldsymbol{r}) = [\sqrt{3/(4\pi)}][C^{111}_{011} r_0(\boldsymbol{r})\boldsymbol{e}_{+1} + C^{111}_{101} r_{+1}(\boldsymbol{r})\boldsymbol{e}_0], \tag{U.3.27}$$

$$\boldsymbol{S}^0_{111}(\boldsymbol{r}) = [\sqrt{3/(4\pi)}][C^{111}_{-1,1,0} r_{-1}(\boldsymbol{r})\boldsymbol{e}_{+1} + C^{111}_{000} r_0(\boldsymbol{r})\boldsymbol{e}_0 + C^{111}_{1,-1,0} r_{+1}(\boldsymbol{r})\boldsymbol{e}_{-1}], \tag{U.3.28}$$

$$\boldsymbol{S}^{-1}_{111}(\boldsymbol{r}) = [\sqrt{3/(4\pi)}][C^{111}_{-1,0,-1} r_{-1}(\boldsymbol{r})\boldsymbol{e}_0 + C^{111}_{0,-1,-1} r_0(\boldsymbol{r})\boldsymbol{e}_{-1}]. \tag{U.3.29}$$

In this case the relevant Clebsch-Gordan coefficients are given by the following relations:

$$C^{111}_{011} = -1/\sqrt{2}, \tag{U.3.30}$$

$$C^{111}_{101} = 1/\sqrt{2}, \tag{U.3.31}$$

see (3.12) and (3.13);

$$C^{111}_{-1,1,0} = -1/\sqrt{2}, \tag{U.3.32}$$

$$C^{111}_{000} = 0, \tag{U.3.33}$$

$$C^{111}_{1,-1,0} = 1/\sqrt{2}, \tag{U.3.34}$$

see (3.12) through (3.14);

$$C^{111}_{-1,0,-1} = -1/\sqrt{2}, \tag{U.3.35}$$

$$C^{111}_{0,-1,-1} = 1/\sqrt{2}, \tag{U.3.36}$$

see (3.13) and (3.14). Consequently, the relations (3.27) through (3.29) can be rewritten in the form

$$\boldsymbol{S}^{1}_{111}(\boldsymbol{r}) = [\sqrt{3/(8\pi)}][-r_0(\boldsymbol{r})\boldsymbol{e}_{+1} + r_{+1}(\boldsymbol{r})\boldsymbol{e}_0], \tag{U.3.37}$$

$$\boldsymbol{S}^{0}_{111}(\boldsymbol{r}) = [\sqrt{3/(8\pi)}][-r_{-1}(\boldsymbol{r})\boldsymbol{e}_{+1} + r_{+1}(\boldsymbol{r})\boldsymbol{e}_{-1}], \tag{U.3.38}$$

$$\boldsymbol{S}^{-1}_{111}(\boldsymbol{r}) = [\sqrt{3/(8\pi)}][-r_{-1}(\boldsymbol{r})\boldsymbol{e}_0 + r_0(\boldsymbol{r})\boldsymbol{e}_{-1}]. \tag{U.3.39}$$

After a little struggle, we recognize that the expressions appearing on the right sides of (3.37) through (3.39) can be written more compactly in terms of the vector cross product. Doing so, we find the neat result

$$\boldsymbol{S}^{M}_{111}(\boldsymbol{r}) = -i[\sqrt{3/(8\pi)}][\boldsymbol{r} \times \boldsymbol{e}_M]. \tag{U.3.40}$$

In retrospect, the cross-product result we have found should not be too surprising since we have, in effect, been combining two spin 1 objects to produce another spin 1 object, and that is just what the vector cross product operation does.

We also note, in view of (3.20), that there is the relation

$$\boldsymbol{S}^{M}_{111}(\boldsymbol{r}) = -i[\sqrt{3/2}][\boldsymbol{r} \times \boldsymbol{S}^{M}_{001}(\boldsymbol{r})]. \tag{U.3.41}$$

Finally, consider the case $J = 2$. Now, from (3.18) and (3.3), we find that

$$\boldsymbol{S}^{M}_{112}(\boldsymbol{r}) = \sum_{m_1,m_2} C^{112}_{m_1 m_2 M} S^{m_1}_{11}(\boldsymbol{r})\boldsymbol{e}_{m_2} = \sqrt{3/(4\pi)} \sum_{m_1,m_2} C^{112}_{m_1 m_2 M} r_{m_1}(\boldsymbol{r})\boldsymbol{e}_{m_2}. \tag{U.3.42}$$

It follows that

$$\boldsymbol{S}^{2}_{112}(\boldsymbol{r}) = [\sqrt{3/(4\pi)}][C^{112}_{112} r_1(\boldsymbol{r})\boldsymbol{e}_1], \tag{U.3.43}$$

$$\boldsymbol{S}^{1}_{112}(\boldsymbol{r}) = [\sqrt{3/(4\pi)}][C^{112}_{011} r_0(\boldsymbol{r})\boldsymbol{e}_1 + C^{112}_{101} r_1(\boldsymbol{r})\boldsymbol{e}_0], \tag{U.3.44}$$

$$\boldsymbol{S}^{0}_{112}(\boldsymbol{r}) = [\sqrt{3/(4\pi)}][C^{112}_{-1,1,0} r_{-1}(\boldsymbol{r})\boldsymbol{e}_1 + C^{112}_{000} r_0(\boldsymbol{r})\boldsymbol{e}_0 + C^{112}_{1,-1,0} r_1(\boldsymbol{r})\boldsymbol{e}_{-1}], \tag{U.3.45}$$

$$\boldsymbol{S}^{-1}_{112}(\boldsymbol{r}) = [\sqrt{3/(4\pi)}][C^{112}_{-1,0,-1} r_{-1}(\boldsymbol{r})\boldsymbol{e}_0 + C^{112}_{0,-1,-1} r_0(\boldsymbol{r})\boldsymbol{e}_{-1}], \tag{U.3.46}$$

$$\boldsymbol{S}^{-2}_{112}(\boldsymbol{r}) = [\sqrt{3/(4\pi)}][C^{112}_{-1,-1,-2} r_{-1}(\boldsymbol{r})\boldsymbol{e}_{-1}]. \tag{U.3.47}$$

To complete this calculation we need the Clebsch-Gordan coefficient values listed below:

$$C^{112}_{112} = 1, \tag{U.3.48}$$

see (3.9);

$$C^{112}_{011} = 1/\sqrt{2}, \tag{U.3.49}$$

$$C^{112}_{101} = 1/\sqrt{2}, \tag{U.3.50}$$

see (3.9) and (3.10);

$$C^{112}_{-1,1,0} = 1/\sqrt{6}, \tag{U.3.51}$$

$$C^{112}_{000} = \sqrt{2/3} = 2/\sqrt{6}, \tag{U.3.52}$$

$$C^{112}_{1,-1,0} = 1/\sqrt{6}, \tag{U.3.53}$$

see (3.9) through (3.11);

$$C^{112}_{-1,0,-1} = 1/\sqrt{2}, \tag{U.3.54}$$

$$C^{112}_{0,-1,-1} = 1/\sqrt{2}, \tag{U.3.55}$$

see (3.10) and (3.11);

$$C^{112}_{-1,-,1-2} = 1, \tag{U.3.56}$$

see (3.11)).

Putting everything together gives the final results

$$\begin{aligned} \boldsymbol{S}^2_{112}(\boldsymbol{r}) &= [\sqrt{3/(4\pi)}][C^{112}_{112} r_1(\boldsymbol{r})\boldsymbol{e}_1] = [\sqrt{3/(4\pi)}] r_1(\boldsymbol{r})\boldsymbol{e}_1 \\ &= [\sqrt{3/(16\pi)}](x+iy)(\boldsymbol{e}_x + i\boldsymbol{e}_y), \end{aligned} \tag{U.3.57}$$

$$\begin{aligned} \boldsymbol{S}^1_{112}(\boldsymbol{r}) &= [\sqrt{3/(4\pi)}][C^{112}_{011} r_0(\boldsymbol{r})\boldsymbol{e}_1 + C^{112}_{101} r_1(\boldsymbol{r})\boldsymbol{e}_0] \\ &= [\sqrt{3/(8\pi)}][r_0(\boldsymbol{r})\boldsymbol{e}_1 + r_1(\boldsymbol{r})\boldsymbol{e}_0] \\ &= -[\sqrt{3/(16)\pi)}][z(\boldsymbol{e}_x + i\boldsymbol{e}_y) + (x+iy)\boldsymbol{e}_z], \end{aligned} \tag{U.3.58}$$

$$\begin{aligned} \boldsymbol{S}^0_{112}(\boldsymbol{r}) &= [\sqrt{3/(4\pi)}][C^{112}_{-1,1,0} r_{-1}(\boldsymbol{r})\boldsymbol{e}_1 + C^{112}_{000} r_0(\boldsymbol{r})\boldsymbol{e}_0 + C^{112}_{1,-1,0} r_1(\boldsymbol{r})\boldsymbol{e}_{-1}] \\ &= [\sqrt{1/(8\pi)}][r_{-1}(\boldsymbol{r})\boldsymbol{e}_1 + 2r_0(\boldsymbol{r})\boldsymbol{e}_0 + r_1(\boldsymbol{r})\boldsymbol{e}_{-1}] \\ &= [\sqrt{1/(8\pi)}](-x\boldsymbol{e}_x - y\boldsymbol{e}_y + 2z\boldsymbol{e}_z), \end{aligned} \tag{U.3.59}$$

$$\begin{aligned} \boldsymbol{S}^{-1}_{112}(\boldsymbol{r}) &= [\sqrt{3/(4\pi)}][C^{112}_{-1,0,-1} r_{-1}(\boldsymbol{r})\boldsymbol{e}_0 + C^{112}_{0,-1,-1} r_0(\boldsymbol{r})\boldsymbol{e}_{-1}] \\ &= [\sqrt{3/(8\pi)}][r_{-1}(\boldsymbol{r})\boldsymbol{e}_0 + r_0(\boldsymbol{r})\boldsymbol{e}_{-1}] \\ &= [\sqrt{3/(16\pi)}][(x-iy)\boldsymbol{e}_z + z(\boldsymbol{e}_x - i\boldsymbol{e}_y)], \end{aligned} \tag{U.3.60}$$

$$\begin{aligned} \boldsymbol{S}^{-2}_{112}(\boldsymbol{r}) &= [\sqrt{3/(4\pi)}][C^{112}_{-1,-1,-2} r_{-1}(\boldsymbol{r})\boldsymbol{e}_{-1}] = [\sqrt{3/(4\pi)}] r_{-1}(\boldsymbol{r})\boldsymbol{e}_{-1} \\ &= [\sqrt{3/(16\pi)}](x-iy)(\boldsymbol{e}_x - i\boldsymbol{e}_y). \end{aligned} \tag{U.3.61}$$

Let us do a count of the $n = 1$ spherical polynomial vector fields we have found. Observe that, when considering all $n = 1$ cases, there are $1 + 3 + 5 = 9$ possibilities, which is to be expected: When $n = 1$, $N(1, 3) = 3$. See table 2.1. Moreover there are three components to be specified, and therefore there are $3N(1, 3) = 9$ parameters to be specified.

At this point the dubious reader may wonder at our counting calculations because complex numbers require two real numbers for their specification, and we appear to be working over the complex field. Should, therefore, all our counts be doubled? The answer is *no* because in our case there are implicit built-in constraints. Let * denote complex conjugation. Then, for the vectors $\boldsymbol{e}_m$, there are the conjugation relations

$$(\boldsymbol{e}_m)^* = (-1)^m \boldsymbol{e}_{-m}. \tag{U.3.62}$$

And for the functions Y_ℓ^m there are the conjugation relations

$$[Y_\ell^m(\theta,\phi)]^* = (-1)^m Y_l^{-m}(\theta,\phi). \tag{U.3.63}$$

Finally, since the Clebsch-Gordan coefficients are real and satisfy the relation

$$C^{j_1 j_2 j_3}_{m_1 m_2 m_3} = (-1)^{j_1+j_2-j_3} C^{j_1 j_2 j_3}_{-m_1,-m_2,-m_3}, \tag{U.3.64}$$

it follows from (3.18) and (3.62) through (3.64) that the $\boldsymbol{S}^M_{n\ell J}(\boldsymbol{r})$ satisfy the conjugation relations

$$[\boldsymbol{S}^M_{n\ell J}(\boldsymbol{r})]^* = (-1)^{\ell+J-M+1}\boldsymbol{S}^{-M}_{n\ell J}(\boldsymbol{r}). \tag{U.3.65}$$

U.4 Independence/Orthogonality/Integral Properties of Spherical Polynomials and Spherical Polynomial Vector Fields

The spherical polynomials $S^m_{n\ell}(\boldsymbol{r})$ and $S^{m'}_{n'\ell'}(\boldsymbol{r})$ are evidently linearly independent if their degrees n and n' differ. What can be said if $n = n'$? The spherical harmonics have the orthogonality properties

$$\int d\Omega\ [Y_\ell^m(\theta,\phi)]^* Y_{\ell'}^{m'}(\theta,\phi) = \delta_{\ell\ell'}\delta_{mm'}. \tag{U.4.1}$$

Here

$$\int d\Omega = \int_0^\pi \sin(\theta)d\theta \int_0^{2\pi} d\phi. \tag{U.4.2}$$

It follows from (2.9), (2.10), and (4.1) that the various $S^m_{n\ell}(\boldsymbol{r})$ are all linearly independent.

The spherical polynomial vector fields $\boldsymbol{S}^M_{n\ell J}(\boldsymbol{r})$ and $\boldsymbol{S}^{M'}_{n'\ell' J'}(\boldsymbol{r})$ are also evidently linearly independent if their degrees n and n' differ. What can be said if $n = n'$? The vector spherical harmonics have the orthogonality properties

$$\int d\Omega\ [\boldsymbol{Y}^M_{\ell J}(\theta,\phi)]^* \cdot \boldsymbol{Y}^{M'}_{\ell' J'}(\theta,\phi) = \delta_{\ell\ell'}\delta_{JJ'}\delta_{MM'}. \tag{U.4.3}$$

It follows from (3.18) and (4.3) that the various $\boldsymbol{S}^M_{n\ell J}(\boldsymbol{r})$ are all linearly independent.

U.5 Differential Properties of Spherical Polynomials and Spherical Polynomial Vector Fields

The purpose of the section is to list various effects of the differential operator ∇ when acting on spherical polynomials and spherical polynomial vector fields.

U.5.1 Gradient Action on Spherical Polynomials

We begin with the action of ∇ on spherical polynomials. Suppose $f(r)$ is any function of r, and suppose $\ell \geq 1$. Then it can be shown that

$$\begin{aligned}\nabla[f(r)Y_\ell^m(\theta,\phi)] &= \sqrt{\ell/(2\ell+1)}\{f'(r)+[(\ell+1)/r]f(r)\}\boldsymbol{Y}^m_{\ell-1,\ell}(\theta,\phi)\\ &\quad -\sqrt{(\ell+1)/(2\ell+1)}\{f'(r)-(\ell/r)f(r)\}\boldsymbol{Y}^m_{\ell+1,\ell}(\theta,\phi).\end{aligned} \tag{U.5.1}$$

For the case of the spherical polynomial functions $S^m_{n\ell}(\boldsymbol{r})$ we have

$$f(r)=r^n. \tag{U.5.2}$$

See (2.10) and (2.11). It follows (again supposing $\ell \geq 1$) that

$$\begin{aligned}\nabla S^m_{n\ell}(\boldsymbol{r}) &= \sqrt{\ell/(2\ell+1)}(n+\ell+1)\boldsymbol{S}^m_{n-1,\ell-1,\ell}(\boldsymbol{r})\\ &\quad -\sqrt{(\ell+1)/(2\ell+1)}(n-\ell)\boldsymbol{S}^m_{n-1,\ell+1,\ell}(\boldsymbol{r}).\end{aligned} \tag{U.5.3}$$

What about the special case $\ell = 0$? Then n must be even. So we write

$$n=2k. \tag{U.5.4}$$

Also we must have $m=0$. If $\ell=0$, we might imagine evaluating (5.3) with the first term omitted since it contains $\sqrt{\ell}$. Doing so gives the result

$$\nabla S^0_{2k,0}(\boldsymbol{r}) = -2k\boldsymbol{S}^0_{2k-1,1,0}(\boldsymbol{r}). \tag{U.5.5}$$

This result is, in fact, correct, and can be verified directly. See Exercise 6.11.

We close this subsection by observing that a special case of (5.3) is the relation

$$\nabla S^m_{nn}(\boldsymbol{r}) = \sqrt{n(2n+1)}\boldsymbol{S}^m_{n-1,n-1,n}(\boldsymbol{r}). \tag{U.5.6}$$

U.5.2 Divergence Action on Spherical Polynomial Vector Fields

We continue with the case of spherical polynomial vector fields. Suppose again that $f(r)$ is any function of r. Then (assuming $\ell \geq 1$) it can be shown that

$$\nabla\cdot[f(r)\boldsymbol{Y}^M_{\ell,J}(\theta,\phi)] = \sqrt{J/(2J+1)}\{f'(r)-[(J-1)/r]f(r)\}Y^M_J(\theta,\phi) \text{ when } J=\ell+1, \tag{U.5.7}$$

$$\nabla\cdot[f(r)\boldsymbol{Y}^M_{\ell,J}(\theta,\phi)] = 0 \text{ when } J=\ell, \tag{U.5.8}$$

$$\nabla\cdot[f(r)\boldsymbol{Y}^M_{\ell,J}(\theta,\phi)] = -\sqrt{(J+1)/(2J+1)}\{f'(r)+[(J+2)/r]f(r)\}Y^M_J(\theta,\phi) \text{ when } J=\ell-1. \tag{U.5.9}$$

For the case of the spherical polynomial vector fields $\boldsymbol{S}^M_{n,\ell,J}(\boldsymbol{r})$ the relation (5.2) again holds. It follows (again assuming $\ell \geq 1$) that

$$\nabla\cdot\boldsymbol{S}^M_{n,\ell,J}(\boldsymbol{r}) = \sqrt{J/(2J+1)}(n-J+1)S^M_{n-1,J}(\boldsymbol{r}) \text{ when } J=\ell+1, \tag{U.5.10}$$

$$\nabla \cdot \boldsymbol{S}^M_{n,\ell,J}(\boldsymbol{r}) = 0 \text{ when } J = \ell, \tag{U.5.11}$$

$$\nabla \cdot \boldsymbol{S}^M_{n,\ell,J}(\boldsymbol{r}) = -\sqrt{(J+1)/(2J+1)}(n+J+2)S^M_{n-1,J}(\boldsymbol{r}) \text{ when } J = \ell - 1. \tag{U.5.12}$$

What about the special case $\ell = 0$? Then we must have $J = 1$. Moreover n must be even so that (5.4) again holds. Evidently when $\ell = 0$ and $J = 1$ the conditions relating J and ℓ in (5.11) and (5.12) do not hold. However, the condition in (5.10) does hold and so we might speculate that (5.10) should be evaluated with $J = 1$ to give the result

$$\nabla \cdot \boldsymbol{S}^M_{2k,0,1}(\boldsymbol{r}) = (\sqrt{1/3})2kS^M_{2k-1,1}(\boldsymbol{r}). \tag{U.5.13}$$

This speculation is correct, and can be proved directly. See Exercise 6.13.

U.5.3 Curl Action on Spherical Polynomial Vector Fields

It can also be shown (assuming $\ell \geq 1$) that

$$\begin{aligned}\nabla \times [f(r)\boldsymbol{Y}^M_{\ell,J}(\theta,\phi)] &= i\sqrt{(J+1)/(2J+1)}\{f'(r) - [(J-1)/r]f(r)\}\boldsymbol{Y}^M_{J,J}(\theta,\phi)\\ &\quad \text{when } J = \ell + 1,\end{aligned} \tag{U.5.14}$$

$$\begin{aligned}\nabla \times [f(r)\boldsymbol{Y}^M_{\ell,J}(\theta,\phi)] &= i\sqrt{(J+1)/(2J+1)}\{f'(r) + [(J+1)/r]f(r)\}\boldsymbol{Y}^M_{J-1,J}(\theta,\phi)\\ &\quad +i\sqrt{J/(2J+1)}\{f'(r) - (J/r)f(r)\}\boldsymbol{Y}^M_{J+1,J}(\theta,\phi)\\ &\quad \text{when } J = \ell,\end{aligned} \tag{U.5.15}$$

$$\begin{aligned}\nabla \times [f(r)\boldsymbol{Y}^M_{\ell,J}(\theta,\phi)] &= i\sqrt{J/(2J+1)}\{f'(r) + [(J+2)/r]f(r)\}\boldsymbol{Y}^M_{J,J}(\theta,\phi)\\ &\quad \text{when } J = \ell - 1.\end{aligned} \tag{U.5.16}$$

For the case of the spherical polynomial vector fields $\boldsymbol{S}^M_{n,\ell J}(\boldsymbol{r})$ the relation (5.2) remains true. It follows (again assuming $\ell \geq 1$) that there are the relations:

$$\begin{aligned}&\nabla \times \boldsymbol{S}^M_{n,\ell,J}(\boldsymbol{r}) = i\sqrt{(J+1)/(2J+1)}(n-J+1)\boldsymbol{S}^M_{n-1,J,J}(\boldsymbol{r})\\ &\text{when } J = \ell + 1. \text{ Equivalently, we have}\\ &\nabla \times \boldsymbol{S}^M_{n,\ell,\ell+1}(\boldsymbol{r}) = i\sqrt{(\ell+2)/(2\ell+3)}(n-\ell)\boldsymbol{S}^M_{n-1,\ell+1,\ell+1}(\boldsymbol{r}),\end{aligned} \tag{U.5.17}$$

$$\begin{aligned}&\nabla \times \boldsymbol{S}^M_{n,\ell,J}(\boldsymbol{r}) = i\sqrt{(J+1)/(2J+1)}(n+J+1)\boldsymbol{S}^M_{n-1,J-1,J}(\boldsymbol{r})\\ &\qquad\qquad +i\sqrt{J/(2J+1)}(n-J)\boldsymbol{S}^M_{n-1,J+1,J}(\boldsymbol{r})\\ &\text{when } J = \ell. \text{ Equivalently, we have}\\ &\nabla \times \boldsymbol{S}^M_{n,\ell,\ell}(\boldsymbol{r}) = i\sqrt{(\ell+1)/(2\ell+1)}(n+\ell+1)\boldsymbol{S}^M_{n-1,\ell-1,\ell}(\boldsymbol{r})\\ &\qquad\qquad +i\sqrt{\ell/(2\ell+1)}(n-\ell)\boldsymbol{S}^M_{n-1,\ell+1,\ell}(\boldsymbol{r}),\end{aligned} \tag{U.5.18}$$

$$\begin{aligned}&\nabla \times \boldsymbol{S}^M_{n,\ell,J}(\boldsymbol{r}) = i\sqrt{J/(2J+1)}(n+J+2)\boldsymbol{S}^M_{n-1,J,J}(\boldsymbol{r})\\ &\text{when } J = \ell - 1. \text{ Equivalently, we have}\\ &\nabla \times \boldsymbol{S}^M_{n,\ell,\ell-1}(\boldsymbol{r}) = i\sqrt{(\ell-1)/(2\ell-1)}(n+\ell+1)\boldsymbol{S}^M_{n-1,\ell-1,\ell-1}(\boldsymbol{r}).\end{aligned} \tag{U.5.19}$$

Again we must consider the special case $\ell = 0$, in which case $J = 1$ and (5.4) holds. The conditions relating J and ℓ associated with (5.18) and (5.19) do not hold in this case, but the one associated with (5.17) does hold. We therefore speculate that (5.17) should be employed in the case $\ell = 0$ and $J = 1$ to give the result

$$\nabla \times \boldsymbol{S}^{M}_{2k,0,1}(\boldsymbol{r}) = i(\sqrt{2/3})(2k)\boldsymbol{S}^{M}_{2k-1,1,1}(\boldsymbol{r}). \tag{U.5.20}$$

This speculation is also correct, and can be proved directly. See Exercise 6.17.

Note that in all cases there is the pleasant fact that the $\nabla\times$ operator *preserves* the top index and the last bottom index, the M and J indices, on $\boldsymbol{S}^{M}_{n,\ell,J}$. It can be shown that there are *total* angular momentum operators $\mathscr{J}_1$, $\mathscr{J}_2$, and $\mathscr{J}_3$, and this preservation is a consequence of the fact that the operator $\nabla\times$ commutes with the total angular momentum operators.

We close this subsection by observing that a special case of (5.18) is the relation

$$\nabla \times \boldsymbol{S}^{M}_{n,n,n}(\boldsymbol{r}) = i\sqrt{(n+1)(2n+1)}\boldsymbol{S}^{M}_{n-1,n-1,n}(\boldsymbol{r}). \tag{U.5.21}$$

U.6 Multiplicative Properties of Spherical Polynomials and Spherical Polynomial Vector Fields

This section deals with the effects of multiplication by $\boldsymbol{r}$.

U.6.1 Ordinary Multiplication

We begin with the case of spherical polynomials and consider ordinary multiplication. Suppose $\ell \geq 1$. Then it can be shown that

$$\boldsymbol{r}Y^{m}_{\ell}(\theta,\phi) = \sqrt{\ell/(2\ell+1)}\; r\boldsymbol{Y}^{m}_{\ell-1,\ell}(\theta,\phi) - \sqrt{(\ell+1)/(2\ell+1)}\; r\boldsymbol{Y}^{m}_{\ell+1,\ell}(\theta,\phi). \tag{U.6.1}$$

In view of (3.18), it follows (again supposing $\ell \geq 1$) that

$$\boldsymbol{r}S^{m}_{n\ell}(\boldsymbol{r}) = \sqrt{\ell/(2\ell+1)}\; \boldsymbol{S}^{m}_{n+1,\ell-1,\ell}(\boldsymbol{r}) - \sqrt{(\ell+1)/(2\ell+1)}\; \boldsymbol{S}^{m}_{n+1,\ell+1,\ell}(\boldsymbol{r}). \tag{U.6.2}$$

What about the special case $\ell = 0$? Then n must be even. So we write

$$n = 2k. \tag{U.6.3}$$

Also we must have $m = 0$. If $\ell = 0$, we might imagine evaluating (6.1) with the first term omitted since it contains $\sqrt{\ell}$. Doing so gives the result

$$\boldsymbol{r}S^{0}_{2k,0}(\boldsymbol{r}) = -\boldsymbol{S}^{0}_{2k+1,1,0}(\boldsymbol{r}). \tag{U.6.4}$$

This result is, in fact, correct, and can be verified directly. See Exercise 6.23.

U.6.2 Dot Product Multiplication

We continue with the case of spherical polynomial vector fields, and consider the case of dot product multiplication. Assume that $\ell \geq 1$. Then it can be shown that

$$\boldsymbol{r} \cdot \boldsymbol{Y}_{\ell,J}^M(\theta,\phi) = \sqrt{J/(2J+1)}\, rY_J^M(\theta,\phi) \text{ when } J = \ell + 1, \tag{U.6.5}$$

$$\boldsymbol{r} \cdot \boldsymbol{Y}_{\ell,J}^M(\theta,\phi) = 0 \text{ when } J = \ell, \tag{U.6.6}$$

$$\boldsymbol{r} \cdot \boldsymbol{Y}_{\ell,J}^M(\theta,\phi) = -\sqrt{(J+1)/(2J+1)}\, rY_J^M(\theta,\phi) \text{ when } J = \ell - 1. \tag{U.6.7}$$

It follows from (3.18), again assuming $\ell \geq 1$, that

$$\boldsymbol{r} \cdot \boldsymbol{S}_{n,\ell,J}^M(\boldsymbol{r}) = \sqrt{J/(2J+1)}\, S_{n+1,J}^M(\boldsymbol{r}) \text{ when } J = \ell + 1, \tag{U.6.8}$$

$$\boldsymbol{r} \cdot \boldsymbol{S}_{n,\ell,J}^M(\boldsymbol{r}) = 0 \text{ when } J = \ell, \tag{U.6.9}$$

$$\boldsymbol{r} \cdot \boldsymbol{S}_{n,\ell,J}^M(\boldsymbol{r}) = -\sqrt{(J+1)/(2J+1)}\, S_{n+1,J}^M(\boldsymbol{r}) \text{ when } J = \ell - 1. \tag{U.6.10}$$

What about the special case $\ell = 0$? Then we must have $J = 1$. Moreover n must be even so that (5.4) again holds. Evidently when $\ell = 0$ and $J = 1$ the conditions relating J and ℓ in (6.9) and (6.10) do not hold. However, the condition in (6.8) does hold and so we might speculate that (6.8) should be evaluated with $J = 1$ to give the result

$$\boldsymbol{r} \cdot \boldsymbol{S}_{2k,0,1}^M(\boldsymbol{r}) = \sqrt{1/3}\, S_{2k+1,1}^M(\boldsymbol{r}). \tag{U.6.11}$$

This speculation is correct, and can be proved directly. See Exercise 6.25.

U.6.3 Cross Product Multiplication

Lastly, we consider the case of cross product multiplication of spherical polynomial vector fields. It can also be shown (assuming $\ell \geq 1$) that

$$\begin{aligned}\boldsymbol{r} \times \boldsymbol{Y}_{\ell,J}^M(\theta,\phi) &= i\sqrt{(J+1)/(2J+1)}\, r\boldsymbol{Y}_{J,J}^M(\theta,\phi) \\ &\quad \text{when } J = \ell + 1,\end{aligned} \tag{U.6.12}$$

$$\begin{aligned}\boldsymbol{r} \times \boldsymbol{Y}_{\ell,J}^M(\theta,\phi) &= i\sqrt{(J+1)/(2J+1)}\, r\boldsymbol{Y}_{J-1,J}^M(\theta,\phi) \\ &\quad + i\sqrt{J/(2J+1)}\, r\boldsymbol{Y}_{J+1,J}^M(\theta,\phi) \\ &\quad \text{when } J = \ell,\end{aligned} \tag{U.6.13}$$

$$\begin{aligned}\boldsymbol{r} \times \boldsymbol{Y}_{\ell,J}^M(\theta,\phi)] &= i\sqrt{J/(2J+1)}\, r\boldsymbol{Y}_{J,J}^M(\theta,\phi) \\ &\quad \text{when } J = \ell - 1.\end{aligned} \tag{U.6.14}$$

It follows from (3.18), again assuming $\ell \geq 1$, that there are the following results:

$\boldsymbol{r} \times \boldsymbol{S}_{n,\ell,J}^M(\boldsymbol{r}) = i\sqrt{(J+1)/(2J+1)}\boldsymbol{S}_{n+1,J,J}^M(\boldsymbol{r})$
when $J = \ell + 1$. Equivalently, we have

$$\boldsymbol{r} \times \boldsymbol{S}_{n,\ell,\ell+1}^M(\boldsymbol{r}) = i\sqrt{(\ell+2)/(2\ell+3)}\boldsymbol{S}_{n+1,\ell+1,\ell+1}^M(\boldsymbol{r}), \tag{U.6.15}$$

$$\boldsymbol{r} \times \boldsymbol{S}^M_{n,\ell,J}(\boldsymbol{r}) = i\sqrt{(J+1)/(2J+1)}\boldsymbol{S}^M_{n+1,J-1,J}(\boldsymbol{r})$$
$$+i\sqrt{J/(2J+1)}\boldsymbol{S}^M_{n+1,J+1,J}(\boldsymbol{r})$$

when $J = \ell$. Equivalently, we have

$$\boldsymbol{r} \times \boldsymbol{S}^M_{n,\ell,\ell}(\boldsymbol{r}) = i\sqrt{(\ell+1)/(2\ell+1)}\boldsymbol{S}^M_{n+1,\ell-1,\ell}(\boldsymbol{r})$$
$$+i\sqrt{\ell/(2\ell+1)}\boldsymbol{S}^M_{n+1,\ell+1,\ell}(\boldsymbol{r}), \qquad \text{(U.6.16)}$$

$$\boldsymbol{r} \times \boldsymbol{S}^M_{n,\ell,J}(\boldsymbol{r}) = i\sqrt{J/(2J+1)}\boldsymbol{S}^M_{n+1,J,J}(\boldsymbol{r})$$

when $J = \ell - 1$. Equivalently, we have

$$\boldsymbol{r} \times \boldsymbol{S}^M_{n,\ell,\ell-1}(\boldsymbol{r}) = i\sqrt{(\ell-1)/(2\ell-1)}\boldsymbol{S}^M_{n+1,\ell-1,\ell-1}(\boldsymbol{r}). \qquad \text{(U.6.17)}$$

Again we must consider the special case $\ell = 0$, in which case $J = 1$ and (5.4) holds. The conditions relating J and ℓ associated with (6.16) and (6.17) do not hold in this case, but the one associated with (6.15) does hold. We therefore speculate that (6.15) should be employed in the case $\ell = 0$ and $J = 1$ to give the result

$$\boldsymbol{r} \times \boldsymbol{S}^M_{2k,0,1}(\boldsymbol{r}) = i(\sqrt{2/3})\boldsymbol{S}^M_{2k+1,1,1}(\boldsymbol{r}). \qquad \text{(U.6.18)}$$

This speculation is also correct, and can be proved directly. See Exercise 6.27.

Note that in all cases there is also the pleasant fact that the $\boldsymbol{r}\times$ operator also preserves the top index and the last bottom index, the M and J indices, on $\boldsymbol{S}^M_{n,\ell,J}$.

We close this subsection by making two useful observations. The first observation is that a special case of (6.15) yields the relation

$$\boldsymbol{S}^M_{nnn}(\boldsymbol{r}) = [-i\sqrt{(2n+1)/(n+1)}][\boldsymbol{r} \times \boldsymbol{S}^M_{n-1,n-1,n}(\boldsymbol{r})]. \qquad \text{(U.6.19)}$$

To verify this claim, evaluate (6.15) for the case

$$n = n' - 1 \qquad \text{(U.6.20)}$$

and

$$\ell = n' - 1; \qquad \text{(U.6.21)}$$

from which it follows that

$$\ell + 1 = n' \qquad \text{(U.6.22)}$$

and

$$(\ell+2)/(2\ell+3) = (n'+1)/(2n'+1). \qquad \text{(U.6.23)}$$

So doing gives the result

$$\boldsymbol{r} \times \boldsymbol{S}^M_{n'-1,n'-1,n'}(\boldsymbol{r}) = i\sqrt{(n'+1)/(2n'+1)}\boldsymbol{S}^M_{n'n'n'}(\boldsymbol{r}), \qquad \text{(U.6.24)}$$

from which (6.19) follows. Note that (3.41) is special case of (6.19).

The second observation is that combining (5.6) and (6.19) gives the relation

$$\begin{aligned}\boldsymbol{S}^M_{nnn}(\boldsymbol{r}) &= [-i\sqrt{(2n+1)/(n+1)}][\boldsymbol{r} \times \boldsymbol{S}^M_{n-1,n-1,n}(\boldsymbol{r})] \\ &= [-i/\sqrt{n(n+1)}][\boldsymbol{r} \times \nabla S^M_{nn}(\boldsymbol{r})]. \end{aligned} \qquad \text{(U.6.25)}$$

Note that if we define an *orbital* angular momentum operator $\boldsymbol{L}$ by the rule

$$\boldsymbol{L} = \boldsymbol{r} \times \nabla, \tag{U.6.26}$$

then (6.25) can be written in the form

$$\boldsymbol{S}^M_{nnn}(\boldsymbol{r}) = [-i/\sqrt{n(n+1)}]\boldsymbol{L}S^M_{nn}(\boldsymbol{r}). \tag{U.6.27}$$

Exercises

U.6.1. Cognizant of the relation between ℓ and n given by (2.10) and (2.11) and the rule $-\ell \leq m \leq \ell$, count how many $S^m_{n\ell}$ there are for a given n. Show the result agrees with $N(n, 3)$.

U.6.2. Verify (3.3) through (3.5).

U.6.3. Verify (3.19) and (3.20).

U.6.4. Verify (3.21) through (3.25).

U.6.5. Verify (3.26) through (3.41).

U.6.6. Verify (3.42) through (3.61).

U.6.7. Verify (3.65) given (3.18), (3.62), (3.63), and (3.65). Verify (3.65) directly for the cases $\boldsymbol{S}^M_{001}$, $\boldsymbol{S}^0_{110}$, $\boldsymbol{S}^M_{111}$, and $\boldsymbol{S}^M_{112}$ worked out explicitly in Subsection 3.3.

U.6.8. Recall the relation between n, ℓ, and J given by (2.10), (2.11), (3.3), and (3.4). See Table 3.1. Recall also the rule $-J \leq M \leq J$. Count how many $\boldsymbol{S}^M_{n\ell J}$ there are for a given n. Show the result agrees with $3N(n, 3)$.

U.6.9. Show that

$$\int d\Omega\ S^m_{n\ell} = \sqrt{4\pi}\ r^n \delta_{\ell 0}\delta_{m0}. \tag{U.6.28}$$

Recall Exercise 16.1.1.

U.6.10. Given (5.1) and (5.2), derive (5.3). Verify (5.6).

U.6.11. Show from the definition (2.10) that

$$S^0_{2k,0}(\boldsymbol{r}) = (1/\sqrt{4\pi})(x^2 + y^2 + z^2)^k. \tag{U.6.29}$$

Show from the definition (3.18) and the result (3.25) that

$$\boldsymbol{S}^0_{2k-1,1,0}(\boldsymbol{r}) = (-1/\sqrt{4\pi})(x^2 + y^2 + z^2)^{k-1}\boldsymbol{r} = (-1/\sqrt{4\pi})\ r^{2k-2}\boldsymbol{r}. \tag{U.6.30}$$

Verify (5.5) by direct computation.

U.6.12. Given (5.2) and (5.8) through (5.9), derive (5.10) through (5.12).

U.6.13. Show from the definition (2.11) that

$$S^M_{2k-1,1}(\boldsymbol{r}) = r^{2k-2} S^M_{11}(\boldsymbol{r}). \tag{U.6.31}$$

Show from the definition (3.18) and the relation (3.20) that

$$\boldsymbol{S}^M_{2k,0,1}(\boldsymbol{r}) = r^{2k}\boldsymbol{S}^M_{001}(\boldsymbol{r}) = (1/\sqrt{4\pi}) r^{2k}\boldsymbol{e}_M. \tag{U.6.32}$$

Verify (5.13) by direct computation. Hint: Use (3.3).

U.6.14. Verify that $S^m_{nn}(\boldsymbol{r})$ is a harmonic polynomial. Verify, using the rules (5.3) and (5.5) and (5.9) through (5.12), that

$$\nabla \cdot [\nabla S^m_{nn}(\boldsymbol{r})] = 0, \tag{U.6.33}$$

as expected.

U.6.15. Show that

$$\nabla^2 S^m_{n\ell} = [n(n+1) - \ell(\ell+1)] S^m_{n-2,\ell}. \tag{U.6.34}$$

U.6.16. Given (5.2) and (5.14) through (5.16), derive (5.17) through (5.19). Verify (5.21).

U.6.17. Show from the definition (3.18) and the relation (3.40) that

$$\boldsymbol{S}^M_{2k-1,1,1}(\boldsymbol{r}) = -i\sqrt{3/(4\pi)}\, r^{2k-2}\boldsymbol{r} \times \boldsymbol{e}_M. \tag{U.6.35}$$

Review Exercise 6.13. Using the results (6.32) and (6.35) for $\boldsymbol{S}^M_{2k,0,1}$ and $\boldsymbol{S}^M_{2k-1,1,1}$, verify (5.20) by direct computation.

U.6.18. Using the rules (5.3) and (5.5) and (5.17) through (5.20), verify that

$$\nabla \times [\nabla S^m_{n\ell}(\boldsymbol{r})] = 0, \tag{U.6.36}$$

as expected.

U.6.19. Using the rules (5.10) through (5.13) and (5.17) through (5.20), verify that

$$\nabla \cdot [\nabla \times \boldsymbol{S}^M_{n\ell J}(\boldsymbol{r})] = 0, \tag{U.6.37}$$

as expected.

U.6.20. Verify the relations

$$\nabla \times \boldsymbol{S}^0_{110}(\boldsymbol{r}) = 0, \tag{U.6.38}$$

$$\nabla \times \boldsymbol{S}^M_{111}(\boldsymbol{r}) = i\sqrt{6}\boldsymbol{S}^M_{001}(\boldsymbol{r}), \tag{U.6.39}$$

$$\nabla \times \boldsymbol{S}^M_{112}(\boldsymbol{r}) = 0. \tag{U.6.40}$$

U.6.21. Show that

$$\nabla \times \boldsymbol{S}_{201}^{M}(\boldsymbol{r}) = i\sqrt{8/3}\boldsymbol{S}_{111}^{M}(\boldsymbol{r}), \tag{U.6.41}$$

$$\nabla \times \boldsymbol{S}_{223}^{M}(\boldsymbol{r}) = 0, \tag{U.6.42}$$

$$\nabla \times \boldsymbol{S}_{222}^{M}(\boldsymbol{r}) = i\sqrt{15}\boldsymbol{S}_{112}^{M}(\boldsymbol{r}), \tag{U.6.43}$$

$$\nabla \times \boldsymbol{S}_{221}^{M}(\boldsymbol{r}) = i\sqrt{25/3}\boldsymbol{S}_{111}^{M}(\boldsymbol{r}). \tag{U.6.44}$$

Show that

$$\nabla \times \nabla \times \boldsymbol{S}_{201}^{M}(\boldsymbol{r}) = i\sqrt{8/3}\nabla \times \boldsymbol{S}_{111}^{M}(\boldsymbol{r}) = -4\boldsymbol{S}_{001}^{M}(\boldsymbol{r}), \tag{U.6.45}$$

$$\nabla \times \nabla \times \boldsymbol{S}_{223}^{M}(\boldsymbol{r}) = 0, \tag{U.6.46}$$

$$\nabla \times \nabla \times \boldsymbol{S}_{222}^{M}(\boldsymbol{r}) = i\sqrt{15}\nabla \times \boldsymbol{S}_{112}^{M}(\boldsymbol{r}) = 0 \tag{U.6.47}$$

$$\nabla \times \nabla \times \boldsymbol{S}_{221}^{M}(\boldsymbol{r}) = i\sqrt{25/3}\nabla \times \boldsymbol{S}_{111}^{M}(\boldsymbol{r}) = -\sqrt{50}\, \boldsymbol{S}_{001}^{M}(\boldsymbol{r}). \tag{U.6.48}$$

U.6.22. Given (6.1), derive (6.2).

U.6.23. Verify (6.4) using (2.10) and (3.25).

U.6.24. Given (6.5) through (6.7), derive (6.8) through (6.10).

U.6.25. Review Exercise 6.13. Verify (6.11) using (6.32), (3.3), and (2.11).

U.6.26. Given (6.12) through (6.14), derive (6.15) through (6.17).

U.6.27. Review Exercise 6.13. Verify (6.18) using (6.32), (3.40), (3.18), and (2.11).

U.6.28. Verify the steps that connect (6.14) to (6.19).

U.6.29. Verify that combining (5.6) and (6.19) yields (6.25).

Bibliography

Group Theory of Angular Momentum

[1] M. Rose, *Elementary Theory of Angular Momentum*, John Wiley & Sons (1957).

[2] A. Edmonds, *Angular Momentum in Quantum Mechanics*, Princeton University Press (1957).

[3] E. Condon and G. Shortley, *The Theory of Atomic Spectra*, Cambridge University Press (1935). A corrected 1999 version is available in paperback.

[4] E.P. Wigner, *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra*, Academic Press (1959).

[5] H. Weyl, *The Theory of Groups and Quantum Mechanics*, Dover (1950).

Harmonic Functions and Vector Spherical Harmonics

[6] E. Stein and G. Weiss, *Introduction to Fourier analysis on Euclidean Spaces*, Princeton University Press (1971).

[7] M. Rose, *Multipole Fields*, John Wiley & Sons (1955).

[8] J. Mathews, *Tensor Spherical Harmonics*, California Institute of Technology (1981).

[9] J. Blatt and V. Weisskopf, *Theoretical Nuclear Physics*, Appendix B, John Wiley (1958) and Dover (2010).

[10] E. Hill, “Theory of Vector Spherical Harmonics”, *Am. J. Phys.* **22**, 211 (1954).

[11] Google Vector Spherical Harmonics. See, for example, the University of Texas (at Austin) Professor Austin Gleeson Web site

http://www.ph.utexas.edu/~gleeson/ElectricityMagnetismAppendixE.pdf

Appendix V

PROT without and in the Presence of a Magnetic Field

V.1 The Case of No Magnetic Field

The material to be covered in this section is standard fare with results known through at least third order, but not yet completely documented.

V.2 The Constant Magnetic Field Case

V.2.1 Preliminaries

Recall from Exercise 1.6.2 that the Hamiltonian for charged-particle motion in an electromagnetic field, when employing cylindrical coordinates with the angle ϕ as the independent variable, is given by the relation

$$K = -\rho[(p_t + q\psi)^2/c^2 - m^2c^2 - (p_\rho - qA_\rho)^2 - (p_y - qA_y)^2]^{1/2} - q\rho A_\phi. \tag{V.2.1}$$

Assume that $\psi = 0$ in accord with the desire that there be no electric field. Also stipulate that $\boldsymbol{A}$ have the components

$$A_\rho = 0, \tag{V.2.2}$$

$$A_y = 0, \tag{V.2.3}$$

$$A_\phi = -(\rho/2)B. \tag{V.2.4}$$

According to Exercise 1.5.8 this choice for $\boldsymbol{A}$ results in a constant magnetic field

$$\boldsymbol{B} = B\boldsymbol{e}_y. \tag{V.2.5}$$

With these provisos it follows that K takes the form

$$K = -\rho[(p_t/c)^2 - m^2c^2 - p_\rho^2 - p_y^2]^{1/2} + q(\rho^2/2)B. \tag{V.2.6}$$

Note that (1.5.49) can be written in the vector/matrix form

$$\begin{pmatrix} A_\phi \\ A_\rho \end{pmatrix} = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} A_z \\ A_x \end{pmatrix}, \tag{V.2.7}$$

from which it follows immediately that there is the inverse relation

$$\begin{pmatrix} A_z \\ A_x \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} A_\phi \\ A_\rho \end{pmatrix}. \tag{V.2.8}$$

Upon combining (1.2) through (1.4), (1.8), and (1.5.33), we see that there is the relation

$$\boldsymbol{A} = (B/2)(-x\boldsymbol{e}_z + z\boldsymbol{e}_x). \tag{V.2.9}$$

Evidently, $\boldsymbol{A}$ is in the Poincaré-Coulomb gauge.

V.2.2 Dimensionless Variables and Limiting Hamiltonian

Change to new scaled variables by writing

$$\rho = \rho_0\ell + \xi\ell, \tag{V.2.10}$$

$$p_\rho = P_\xi p_0, \tag{V.2.11}$$

$$y = Y\ell, \tag{V.2.12}$$

$$p_y = P_y p_0, \tag{V.2.13}$$

$$t = \tau\ell/c, \tag{V.2.14}$$

$$p_t = p_t^0 + P_\tau p_0 c. \tag{V.2.15}$$

Here ℓ is a scale length and p_0 is the design momentum. [Note that the variable Y in (2.12) is not to be confused with the Y associated with the $\boldsymbol{R}$ of Section 15.9.] Correspondingly, there are the Poisson bracket relations

$$[\xi, P_\xi] = (p_0\ell)^{-1}[\rho, p_\rho], \tag{V.2.16}$$

$$[Y, P_y] = (p_0\ell)^{-1}[y, p_y], \tag{V.2.17}$$

$$[\tau, P_\tau] = (p_0\ell)^{-1}[t, p_t]. \tag{V.2.18}$$

Let $\tilde{K}$ be the new Hamiltonian for these new variables. It is given by the relation

$$\tilde{K} = \lambda\ell\{-(\rho_0+\xi)[(p_t^0 + P_\tau p_0 c)^2/c^2 - m^2c^2 - p_0^2P_\xi^2 - p_0^2P_y^2]^{1/2} + (qB\ell/2)(\rho_0+\xi)^2\}, \tag{V.2.19}$$

or, equivalently,

$$\tilde{K} = \lambda\ell p_0\{-(\rho_0+\xi)[(p_t^0 p_0^{-1}c^{-1} + P_\tau)^2 - m^2c^2/p_0^2 - P_\xi^2 - P_y^2]^{1/2} + [qB\ell/(2p_0)](\rho_0+\xi)^2\}. \tag{V.2.20}$$

Here

$$\lambda = (\ell p_0)^{-1}. \tag{V.2.21}$$

Observe that there are the relations

$$p_0 = \gamma m v_0 = \gamma\beta mc, \tag{V.2.22}$$

$$p_t^0 = -\gamma mc^2. \tag{V.2.23}$$

It follows that there are the relations

$$m^2c^2/p_0^2 = m^2c^2/(\beta\gamma mc)^2 = 1/(\beta\gamma)^2, \tag{V.2.24}$$

$$p_t^0/(p_0c) = -(\gamma mc^2)/(\gamma\beta mc^2) = -1/\beta. \tag{V.2.25}$$

Consequently, $\tilde{K}$ can also be written on the form

$$\tilde{K} = -(\rho_0 + \xi)[(-1/\beta + P_\tau)^2 - (\beta\gamma)^{-2} - P_\xi^2 - P_y^2]^{1/2} + (b/2)(\rho_0 + \xi)^2 \tag{V.2.26}$$

where

$$b = qB\ell/p_0. \tag{V.2.27}$$

We also observe that

$$1/\beta^2 - 1/(\beta\gamma)^2 = 1. \tag{V.2.28}$$

It follows that $\tilde{K}$ can also be written as

$$\tilde{K} = -(\rho_0 + \xi)[1 - 2P_\tau/\beta + P_\tau^2 - P_\xi^2 - P_y^2]^{1/2} + (b/2)(\rho_0 + \xi)^2. \tag{V.2.29}$$

Finally, we take the limit $\rho_0 \to 0$ to obtain the limiting Hamiltonian

$$\tilde{K}^{\text{lim}} = -\xi[1 - 2P_\tau/\beta + P_\tau^2 - P_\xi^2 - P_y^2]^{1/2} + (b/2)\xi^2. \tag{V.2.30}$$

V.2.3 Design Trajectory

Evidently P_y and P_τ are integrals of motion and vanish on the design trajectory. Therefore the variables ξ, P_ξ *on* the *design trajectory* are governed by the Hamiltonian

$$\tilde{K}^{\text{dt}} = -\xi[1 - P_\xi^2]^{1/2} + (b/2)\xi^2. \tag{V.2.31}$$

Correspondingly, the associated equations of motion for these variables on the design trajectory are given by the relations

$$\xi' = \partial\tilde{K}^{\text{dt}}/\partial P_\xi = \xi P_\xi[1 - P_\xi^2]^{-1/2}, \tag{V.2.32}$$

$$P_\xi' = -\partial\tilde{K}^{\text{dt}}/\partial\xi = [1 - P_\xi^2]^{1/2} - b\xi. \tag{V.2.33}$$

They have the particular solution

$$\xi = 0, \tag{V.2.34}$$

$$P_\xi(\phi) = \sin\Delta, \tag{V.2.35}$$

where

$$\Delta = \phi - \phi^{\text{in}}. \tag{V.2.36}$$

We will take (2.34) through (2.36) to be the ξ, P_ξ results for the design trajectory. Note that the design trajectory does not depend on b. That is, it does not depend on the magnetic field.

As assumed earlier, and consistent with the full equations of motion associated with the full Hamiltonian (2.30), the remaining variables on the design trajectory vanish,

$$Y = P_y = \tau = P_\tau = 0. \tag{V.2.37}$$

It follows, from (2.34) through (2.37), that all the variables save P_ξ are deviation variables.

V.2.4 Deviation Variables

To proceed further we wish to replace ξ and P_ξ by deviation variables, which we will call $\hat{\xi}$ and $\hat{P}_\xi$, so that all variables are deviation variables. This is simply done by making the definitions

$$\xi = \hat{\xi}, \tag{V.2.38}$$

$$P_\xi = \sin\Delta + \hat{P}_\xi, \tag{V.2.39}$$

and leaving all other variables in peace. The relations (2.38) and (2.39) are a canonical transformation and can be obtained from the F_2 generating function given by

$$F_2(\xi, \hat{P}_\xi) = \xi(\sin\Delta + \hat{P}_\xi). \tag{V.2.40}$$

Indeed, employing the standard machinery (6.5.5) yields the results

$$P_\xi = \partial F_2/\partial\xi = \sin\Delta + \hat{P}_\xi, \tag{V.2.41}$$

$$\hat{\xi} = \partial F_2/\partial\hat{P}_\xi = \xi, \tag{V.2.42}$$

as desired.

V.2.5 Deviation Variable Hamiltonian

We may regard the deviation variables as new variables. Associated with the use of these new variables will be a new Hamiltonian H given by the relation

$$H = \tilde{K}^{\text{lim}} + \partial F_2/\partial\phi = \tilde{K}^{\text{lim}} + \xi\cos\Delta = \tilde{K}^{\text{lim}} + \hat{\xi}\cos\Delta. \tag{V.2.43}$$

Use of (2.43) yields the result

$$H = -\hat{\xi}[1 - 2P_\tau/\beta + P_\tau^2 - (\hat{P}_\xi + \sin\Delta)^2 - P_y^2]^{1/2} + (b/2)\hat{\xi}^2 + \hat{\xi}\cos\Delta, \tag{V.2.44}$$

or

$$H = -\hat{\xi}[1 - \sin^2\Delta - 2P_\tau/\beta + P_\tau^2 - \hat{P}_\xi^2 - 2\hat{P}_\xi\sin\Delta - P_y^2]^{1/2} + (b/2)\hat{\xi}^2 + \hat{\xi}\cos\Delta, \tag{V.2.45}$$

or

$$H = -\hat{\xi}[\cos^2\Delta - 2P_\tau/\beta + P_\tau^2 - \hat{P}_\xi^2 - 2\hat{P}_\xi\sin\Delta - P_y^2]^{1/2} + (b/2)\hat{\xi}^2 + \hat{\xi}\cos\Delta. \tag{V.2.46}$$

V.2.6 Computation of Transfer Map

Our aim is to find the transfer map associated with H. According to Section 10.4, this entails expanding H in terms of homogeneous polynomials,

$$H = H_0 + H_1 + H_2 + H_3 + H_4 + \cdots . \tag{V.2.47}$$

So doing gives for H_0 through H_2 the results

$$H_0 = 0, \tag{V.2.48}$$

$$H_1 = 0, \tag{V.2.49}$$

$$H_2 = \hat{\xi}P_\tau/(\beta\cos\Delta) + \hat{\xi}\hat{P}_\xi\tan\Delta + (b/2)\hat{\xi}^2. \tag{V.2.50}$$

Note that H_1 vanishes as expected. That is, the design trajectory is given by the relations (2.37) supplemented by the relations

$$\hat{\xi} = \hat{P}_\xi = 0. \tag{V.2.51}$$

Linear Part of Transfer Map

The first step is to find $\mathcal{R}$, the linear part of the transfer map. This requires solving the equations of motion associated with H_2. They read

$$\hat{\xi}' = \partial H_2/\partial\hat{P}_\xi = \hat{\xi}\tan\Delta, \tag{V.2.52}$$

$$\hat{P}_\xi' = -\partial H_2/\partial\hat{\xi} = -\hat{P}_\xi\tan\Delta - P_\tau/(\beta\cos\Delta) - b\hat{\xi}, \tag{V.2.53}$$

$$Y' = \partial H_2/\partial P_y = 0, \tag{V.2.54}$$

$$P_y' = -\partial H_2/\partial Y = 0, \tag{V.2.55}$$

$$\tau' = \partial H_2/\partial P_\tau = \hat{\xi}/(\beta\cos\Delta), \tag{V.2.56}$$

$$P_\tau' = -\partial H_2/\partial\tau = 0. \tag{V.2.57}$$

The solutions to (2.54), (2.55), and (2.57) can be written immediately,

$$Y(\phi) = Y^{\text{in}}, \tag{V.2.58}$$

$$P_Y(\phi) = P_Y^{\text{in}}, \tag{V.2.59}$$

$$P_\tau(\phi) = P_\tau^{\text{in}}. \tag{V.2.60}$$

The solution to (2.52), which is less trivial, is

$$\hat{\xi}(\phi) = \hat{\xi}^{\text{in}}/\cos\Delta. \tag{V.2.61}$$

The results (2.60) and (2.61) can now be inserted into (2.53) to yield the differential equation

$$\hat{P}_\xi' = -\hat{P}_\xi\tan\Delta - P_\tau^{\text{in}}/(\beta\cos\Delta) - b\hat{\xi}^{\text{in}}/\cos\Delta = -\hat{P}_\xi\tan\Delta - (b\hat{\xi}^{\text{in}} + P_\tau^{\text{in}}/\beta)/\cos\Delta. \tag{V.2.62}$$

It has the solution

$$\hat{P}_\xi(\phi) = \hat{P}_\xi^{\text{in}}\cos\Delta - (b\hat{\xi}^{\text{in}} + P_\tau^{\text{in}}/\beta)\sin\Delta. \tag{V.2.63}$$

Finally, insertion of (2.61) into (2.56) yields the differential equation

$$\tau' = \hat{\xi}^{\text{in}}/(\beta\cos^2\Delta). \tag{V.2.64}$$

It has the solution

$$\tau(\phi) = \tau^{\text{in}} + \hat{\xi}^{\text{in}}(1/\beta)\tan\Delta. \tag{V.2.65}$$

From these solutions we can read off the matrix R associated with $\mathcal{R}$. From (2.58) through (2.61), (2.63), and (2.65) we see that there is the vector/matrix relation

$$\begin{pmatrix} \hat{\xi}(\phi) \\ \hat{P}_\xi(\phi) \\ Y(\phi) \\ P_y(\phi) \\ \tau(\phi) \\ P_\tau(\phi) \end{pmatrix} = \begin{pmatrix} 1/\cos\Delta & 0 & 0 & 0 & 0 & 0 \\ -b\sin\Delta & \cos\Delta & 0 & 0 & 0 & -(1/\beta)\sin\Delta \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ (1/\beta)\tan\Delta & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\xi}^{\rm in} \\ \hat{P}_\xi^{\rm in} \\ Y^{\rm in} \\ P_y^{\rm in} \\ \tau^{\rm in} \\ P_\tau^{\rm in} \end{pmatrix}. \tag{V.2.66}$$

The matrix R is the matrix appearing in (2.66).

Nonlinear Parts of Transfer Map

To compute the nonlinear parts of the transfer map it is necessary to continue the expansion (2.47) begun in (2.48) through (2.50) to find H_3, $H_4 \cdots$ and to then apply the machinery of Section 10.5 to find the associated f_3, $f_4 \cdots$. For example, one finds from (2.46) and (2.47) the result

$$H_3 = . \tag{V.2.67}$$

Exercises

V.2.1. Verify that the solutions given by (2.34) through (2.36) do indeed satisfy the differential equations (2.32) and (2.33).

V.2.2. Verify the expansion (2.48) through (2.50).

V.2.3. Verify that the solutions (2.58) through (2.61), (2.63), and (2.65) do indeed satisfy the differential equations (2.52) through (2.57).

V.2.4. Verify that R, the matrix appearing in (1.65), is symplectic.

V.2.5. Verify (2.67).

V.3 The Inhomogeneous Field Case

The work so far has dealt with the case of a constant magnetic field. We now consider the general case.

V.3.1 Vector Potential for the General Inhomogeneous Field Case

We begin by expanding the vector potential in the Poincaré-Coulomb gauge and in homogeneous polynomials employing Cartesian coordinates and Cartesian unit vectors. That is we write

$$\boldsymbol{A}(\boldsymbol{r}) = \boldsymbol{A}^{\min 1}(\boldsymbol{r}) + \boldsymbol{A}^{\min 2}(\boldsymbol{r}) + \boldsymbol{A}^{\min 3}(\boldsymbol{r}) + \cdots . \tag{V.3.1}$$

For $\boldsymbol{A}^{\min 1}(\boldsymbol{r})$ we use (2.9) to account for the constant part of the magnetic field and write

$$\boldsymbol{A}^{\min 1}(\boldsymbol{r}) = (B/2)(-x\boldsymbol{e}_z + z\boldsymbol{e}_x). \tag{V.3.2}$$

For example, in the case of a magnetic monopole doublet, there is the result

$$\boldsymbol{A}^{\min 1}(\boldsymbol{r}) = [ga/(X_0^2 + Z_0^2 + a^2)^{3/2}](-z\boldsymbol{e}_x + x\boldsymbol{e}_z), \tag{V.3.3}$$

and there is the relation

$$B = -2[ga/(X_0^2 + Z_0^2 + a^2)^{3/2}]. \tag{V.3.4}$$

And, for the same example and again employing Cartesian coordinates and Cartesian unit vectors, there is the result

$$\begin{gathered}\boldsymbol{A}^{\min 2}(\boldsymbol{r}) = [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ [(Z_0y^2 - Z_0z^2 - X_0xz)\boldsymbol{e}_x + (X_0yz - Z_0xy)\boldsymbol{e}_y + (X_0x^2 + Z_0xz - X_0y^2)\boldsymbol{e}_z].\end{gathered} \tag{V.3.5}$$

See (15.9.7) and (15.9.8) in Section 15.9. For other examples the $\boldsymbol{A}^{\min n}(\boldsymbol{r})$ with $n \geq 2$ will be different, but still homogeneous of degree n. We will continue to assume that the constant part of the magnetic field is of the form (2.5), and therefore (3.2) will always be assumed to hold.

V.3.2 Transition to Cylindrical Coordinates

Next, as in Exercise 1.5.4, introduce polar coordinates in the x, z plane by the relations

$$x = \rho \ \cos \ \phi, \tag{V.3.6}$$

$$z = \rho \ \sin \ \phi.$$

That is, we will again employ the cylindrical coordinates ρ, y, ϕ and also the unit vectors $\boldsymbol{e}_\rho, \boldsymbol{e}_y, \boldsymbol{e}_\phi$ of Exercise 1.5.4. See (3.6) and (1.5.52) through (1.5.54). Let us express $\boldsymbol{A}$ in terms of these cylindrical coordinates and unit vectors.

Begin, for example, with (3.2). With the aid of (3.6) and (1.5.53) the relation (3.2) can be rewritten in the form

$$\begin{aligned}\boldsymbol{A}^{\min 1}(\boldsymbol{r}) &= -(B/2)\rho(-\sin\phi\ \boldsymbol{e}_x + \cos\phi\ \boldsymbol{e}_z) \\ &= -(B/2)\rho\boldsymbol{e}_\phi.\end{aligned} \tag{V.3.7}$$

Since $\boldsymbol{e}_\rho, \boldsymbol{e}_y, \boldsymbol{e}_\phi$ form an orthonormal triad, it follows from (1.5.44) and (3.2) that there are the results

$$A_\rho^{\min 1}(\boldsymbol{r}) = \boldsymbol{e}_\rho \cdot \boldsymbol{A}^{\min 1}(\boldsymbol{r}) = 0, \tag{V.3.8}$$

$$A_y^{\min 1}(\boldsymbol{r}) = \boldsymbol{e}_y \cdot \boldsymbol{A}^{\min 1}(\boldsymbol{r}) = 0, \tag{V.3.9}$$

$$A_\phi^{\min 1}(\boldsymbol{r}) = \boldsymbol{e}_\phi \cdot \boldsymbol{A}^{\min 1}(\boldsymbol{r}) = -(B/2)\rho, \tag{V.3.10}$$

which are to be expected in accord with (2.2) through (2.4) and (2.9).

As a second illustrative example, let us work on (3.5), which would be the $\boldsymbol{A}^{\min 2}(\boldsymbol{r})$ in the case of a magnetic monopole doublet. With the aid of (3.6) it takes the form

$$\begin{gathered}\boldsymbol{A}^{\min 2}(\boldsymbol{r}) = [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ \{[Z_0 y^2 - \rho^2(Z_0 \sin^2\phi - X_0 \cos\phi \sin\phi)]\boldsymbol{e}_x + [\rho y(X_0 \sin\phi - Z_0 \cos\phi)]\boldsymbol{e}_y \\ +[\rho^2(X_0 \cos^2\phi + Z_0 \cos\phi \sin\phi) - X_0 y^2 \boldsymbol{e}_z\}.\end{gathered} \tag{V.3.11}$$

From (1.5.35), the definitions (1.5.44), and (3.11) there are the results

$$\begin{gathered}A_\rho^{\min 2}(\boldsymbol{r}) = \boldsymbol{e}_\rho \cdot \boldsymbol{A}^{\min 2}(\boldsymbol{r}) = [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ \{[\cos\phi\ \boldsymbol{e}_x + \sin\phi\ \boldsymbol{e}_z] \cdot [Z_0 y^2 \boldsymbol{e}_x - \rho^2(Z_0 \sin^2\phi - X_0 \cos\phi \sin\phi)\boldsymbol{e}_x] \\ +[\cos\phi\ \boldsymbol{e}_x + \sin\phi\ \boldsymbol{e}_z] \cdot [\rho^2(X_0 \cos^2\phi + Z_0 \cos\phi \sin\phi)\boldsymbol{e}_z - X_0 y^2 \boldsymbol{e}_z]\} \\ = [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ \{(\cos\phi)[Z_0 y^2 - \rho^2(Z_0 \sin^2\phi - X_0 \cos\phi \sin\phi)] \\ +(\sin\phi)[\rho^2(X_0 \cos^2\phi + Z_0 \cos\phi \sin\phi) - X_0 y^2]\},\end{gathered} \tag{V.3.12}$$

$$A_y^{\min 2}(\boldsymbol{r}) = \boldsymbol{e}_y \cdot \boldsymbol{A}^{\min 2}(\boldsymbol{r}) = [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}][\rho y(X_0 \sin\phi - Z_0 \cos\phi)], \tag{V.3.13}$$

$$\begin{gathered}A_\phi^{\min 2}(\boldsymbol{r}) = \boldsymbol{e}_\phi \cdot \boldsymbol{A}^{\min 2}(\boldsymbol{r}) = [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ \{[-\sin\phi\ \boldsymbol{e}_x + \cos\phi\ \boldsymbol{e}_z] \cdot [Z_0 y^2 \boldsymbol{e}_x - \rho^2(Z_0 \sin^2\phi - X_0 \cos\phi \sin\phi)\boldsymbol{e}_x] \\ +[-\sin\phi\ \boldsymbol{e}_x + \cos\phi\ \boldsymbol{e}_z] \cdot [\rho^2(X_0 \cos^2\phi + Z_0 \cos\phi \sin\phi)\boldsymbol{e}_z - X_0 y^2 \boldsymbol{e}_z]\} \\ = [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ \{-(\sin\phi)[Z_0 y^2 - \rho^2(Z_0 \sin^2\phi - X_0 \cos\phi \sin\phi)] \\ +(\cos\phi)[\rho^2(X_0 \cos^2\phi + Z_0 \cos\phi \sin\phi) - X_0 y^2]\}.\end{gathered} \tag{V.3.14}$$

V.3.3 Dimensionless Variables and Limiting Vector Potential

We now make the substitutions (2.10) through (2.15) and take the limit $\rho_0 \to 0$ to obtain the limiting vector potential whose components we will denote by letters with breve marks $\breve{}$ above. So doing yields for the constant part of the magnetic field the limiting result

$$\breve{A}_\phi^{\min 1} = -\ell(B/2)\xi. \tag{V.3.15}$$

And for the leading term of the nonconstant part of the magnetic field, again taking for illustrative purposes the magnetic monopole doublet example, there are the results

$$\begin{gathered}\breve{A}_\rho^{\min 2} = \ell^2[-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ \{(\cos\phi)[Z_0 Y^2 - \xi^2(Z_0 \sin^2\phi - X_0 \cos\phi \sin\phi)] \\ +(\sin\phi)[\xi^2(X_0 \cos^2\phi + Z_0 \cos\phi \sin\phi) - X_0 Y^2]\},\end{gathered} \tag{V.3.16}$$

$$\breve{A}_y^{\min 2} = \ell^2[-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}][\xi Y(X_0 \sin\phi - Z_0 \cos\phi)], \tag{V.3.17}$$

$$\begin{gathered}\breve{A}_\phi^{\min 2} = \ell^2[-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ \{-(\sin\phi)[Z_0 Y^2 - \xi^2(Z_0 \sin^2\phi - X_0 \cos\phi \sin\phi)] \\ +(\cos\phi)[\xi^2(X_0 \cos^2\phi + Z_0 \cos\phi \sin\phi) - X_0 Y^2]\}.\end{gathered} \tag{V.3.18}$$

V.3.4 Computation of Limiting Hamiltonian in Dimensionless Variables

At this point we are again ready to compute $\tilde{K}^{\text{lim}}$, but this time with possible inhomogeneity in the magnetic field included. Let us introduce the notation

$$\breve{\boldsymbol{A}}^{\text{min non}} = \breve{\boldsymbol{A}}^{\text{min 2}} + \breve{\boldsymbol{A}}^{\text{min 3}} + \breve{\boldsymbol{A}}^{\text{min 4}} + \cdots \tag{V.3.19}$$

to denote the *nonlinear* part of the vector potential. We now find for the limiting Hamiltonian the result

$$\begin{aligned}\tilde{K}^{\text{lim}} &= -\xi[1 - 2P_\tau/\beta + P_\tau^2 - (P_\xi - \mathcal{A}_\rho^{\text{min non}})^2 - (P_y - \mathcal{A}_y^{\text{min non}})^2]^{1/2} \\ &\quad +(b/2)\xi^2 - \xi\mathcal{A}_\phi^{\text{min non}}\end{aligned} \tag{V.3.20}$$

where

$$\mathcal{A}_\rho^{\text{min non}} = (q/p_0)\breve{A}_\rho^{\text{min non}}, \tag{V.3.21}$$

$$\mathcal{A}_y^{\text{min non}} = (q/p_0)\breve{A}_y^{\text{min non}}, \tag{V.3.22}$$

$$\mathcal{A}_\phi^{\text{min non}} = (q/p_0)\breve{A}_\phi^{\text{min non}}. \tag{V.3.23}$$

V.3.5 Deviation Variable Hamiltonian

Introduce, as before, deviation variables $(\hat{\xi}, \tau, Y; \hat{P}_\xi, P_\tau, P_y)$ with $\hat{\xi}$ and $\hat{P}_\xi$ defined by (2.38) and (2.39). Doing so, and employing the rule (2.43), yields the new Hamiltonian

$$\begin{aligned}H &= -\hat{\xi}[1 - 2P_\tau/\beta + P_\tau^2 - (\sin\Delta + \hat{P}_\xi - \hat{\mathcal{A}}_\rho^{\text{min non}})^2 - (P_y - \hat{\mathcal{A}}_y^{\text{min non}})^2]^{1/2} \\ &\quad + (b/2)\hat{\xi}^2 - \hat{\xi}\hat{\mathcal{A}}_\phi^{\text{min non}} + \hat{\xi}\cos\Delta.\end{aligned} \tag{V.3.24}$$

Here we have used the notation $\hat{\mathcal{A}}_\rho^{\text{min non}}$ to indicate that the variable ξ in $\mathcal{A}_\rho^{\text{min non}}$ has been replaced by $\hat{\xi}$, etc. For example, in the case of a magnetic monopole doublet, there are the results

$$\begin{aligned}\hat{\mathcal{A}}_\rho^{\text{min 2}} &= (q/p_0)\ell^2[-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ &\{(\cos\phi)[Z_0Y^2 - \hat{\xi}^2(Z_0\sin^2\phi - X_0\cos\phi\sin\phi)] \\ &+(\sin\phi)[\hat{\xi}^2(X_0\cos^2\phi + Z_0\cos\phi\sin\phi) - X_0Y^2]\},\end{aligned} \tag{V.3.25}$$

$$\hat{\mathcal{A}}_y^{\text{min 2}} = (q/p_0)\ell^2[-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}][\hat{\xi}Y(X_0\sin\phi - Z_0\cos\phi)], \tag{V.3.26}$$

$$\begin{aligned}\hat{\mathcal{A}}_\phi^{\text{min 2}} &= (q/p_0)\ell^2[-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ &\{-(\sin\phi)[Z_0Y^2 - \hat{\xi}^2(Z_0\sin^2\phi - X_0\cos\phi\sin\phi)] \\ &+(\cos\phi)[\hat{\xi}^2(X_0\cos^2\phi + Z_0\cos\phi\sin\phi) - X_0Y^2]\}.\end{aligned} \tag{V.3.27}$$

V.3.6 Expansion of Deviation Variable Hamiltonian and Computation of Transfer Map

As done before in Subsection 2.6, we expand H in terms of homogeneous polynomials. Begin by observing that there is the relation

$$1-(\sin\Delta+\hat{P}_\xi-\hat{\mathcal{A}}_\rho^{\text{min non}})^2 =$$
$$\cos^2\Delta-\hat{P}_\xi^2-(\hat{\mathcal{A}}_\rho^{\text{min non}})^2-2(\sin\Delta)\hat{P}_\xi+2(\sin\Delta)\hat{\mathcal{A}}_\rho^{\text{min non}}+2\hat{P}_\xi\hat{\mathcal{A}}_\rho^{\text{min non}}. \tag{V.3.28}$$

Consequently H, as given by (3.24), can be rewritten in the form

$$\begin{aligned} H &= -\hat{\xi}[\cos^2\Delta-2P_\tau/\beta+P_\tau^2-\hat{P}_\xi^2-(\hat{\mathcal{A}}_\rho^{\text{min non}})^2 \\ &\quad -2(\sin\Delta)\hat{P}_\xi+2(\sin\Delta)\hat{\mathcal{A}}_\rho^{\text{min non}}+2\hat{P}_\xi\hat{\mathcal{A}}_\rho^{\text{min non}} \\ &\quad -P_y^2+2P_y\hat{\mathcal{A}}_y^{\text{min non}}-(\hat{\mathcal{A}}_y^{\text{min non}})^2]^{1/2} \\ &\quad +(b/2)\hat{\xi}^2-\hat{\xi}\hat{\mathcal{A}}_\phi^{\text{min non}}+\ \hat{\xi}\cos\Delta. \end{aligned} \tag{V.3.29}$$

We are now prepared to expand H in the form (2.47). Compare (2.46) and (3.29). Since $\hat{\mathcal{A}}_\rho^{\text{min non}}$ and $\hat{\mathcal{A}}_y^{\text{min non}}$ consist entirely of terms of degree 2 and higher, and $\hat{\xi}\hat{\mathcal{A}}_\phi^{\text{min non}}$ consists entirely of terms of degree 3 and higher, it follows that they make *no* contribution to H_0 through H_2. {Note that the term of the form $[***]^{1/2}$ in (3.29) is multiplied by $\hat{\xi}$.} Therefore the H_0, H_1, and H_2 terms in the expansion are the *same* as those given by (2.48) through (2.50). Consequently the design orbit and $\mathcal{R}$, the linear part of the transfer map about the design orbit, are the same as those found earlier. That is, the design orbit does *not* depend on the magnetic field, and the linear part of the transfer map depends *only* on the uniform part of the magnetic field, described by $A_\phi^{\text{min }1}$ or b. The design orbit and the linear part of the transfer map do *not* depend on field inhomogeneities described by the $\boldsymbol{A}^{\text{min }n}$ with $n \geq 2$. Field inhomogeneities play a role only in the calculation of the H_m with $m \geq 3$. Correspondingly, field inhomogeneities play a role in the transfer map only for the generators f_m with $m \geq 3$.

We close this subsection by computing, for example, the H_3 that occurs in the expansion of (3.29). We find the result

$$H_3 = . \tag{V.3.30}$$

See Exercise 3.1. Note that (2.67) and (3.30) agree when there are no field inhomogeneities.

Exercises

V.3.1. Verify that, in the computation of the H_3 term that occurs in the expansion of (3.29), the terms * are of too high an order to play a role, and therefore may be neglected.

Bibliography

[1] É. Forest, *Beam Dynamics: A New Attitude and Framework*, Harwood Academic Publishers (1998).

Appendix W

Smoothing for Harmonic Functions

W.1 Introduction

W.2 The Line in Two Space

Consider in x, y space the line $y = 0$ and suppose a potential $\psi_0(x)$ is specified on this line. Define its Fourier transform $\tilde{\psi}_0(k_x)$ by the rule

$$\tilde{\psi}_0(k_x) = [1/\sqrt{(2\pi)}] \int dx \ \exp(-ik_x x)\psi_0(x). \tag{W.2.1}$$

Make the Ansatz

$$\psi(x, y) = [1/\sqrt{(2\pi)}] \int dk_x \exp(ik_x x) \exp(-ky)\tilde{\psi}_0(k_x) \tag{W.2.2}$$

where

$$k = \sqrt{k_x^2} = |k_x|. \tag{W.2.3}$$

Evidently this $\psi(x, y)$ is harmonic and vanishes as $y \rightarrow +\infty$. We also have the result

$$\psi(x, 0) = [1/\sqrt{(2\pi)}] \int dk_x \exp(ik_x x)\tilde{\psi}_0(k_x) = \psi_0(x). \tag{W.2.4}$$

It follows that we have found the solution to Laplace's equation in the upper half plane $y \geq 0$ associated with the $y = 0$ boundary value $\psi_0(x)$.

Note that the operation defined by (2.2) is smoothing for $y > 0$. High spatial frequencies are suppressed by the factor $\exp(-ky)$, and this *exponential* suppression/damping is ever more effective the larger the value of y. The higher the y observation line is above the $y = 0$ line, the smoother $\psi(x, y)$ on this observation line becomes as a function of x.

We also observe, in passing, two facts. First, suppose ψ_0, now to be called ψ_0^c, is a *constant* function,

$$\psi_0^c(x) = c. \tag{W.2.5}$$

Then, by (2.1),

$$\tilde{\psi}_0^c(k_x) = c\sqrt{2\pi}\delta(k_x). \tag{W.2.6}$$

It follows from (2.2) that there is the relation

$$\psi^c(x,y) = c. \tag{W.2.7}$$

As expected, if ψ is constant on the boundary $y = 0$, it will have the same constant value in the upper half plane $y \geq 0$. Second, for any solution, there is the relation

$$\begin{aligned} \int dx\ \psi(x,y) &= [1/\sqrt{(2\pi)}] \int dk_x \exp(-ky)\tilde{\psi}_0(k_x) \int dx\ \exp(ik_x x) \\ &= [1/\sqrt{(2\pi)}] \int dk_x \exp(-ky)\tilde{\psi}_0(k_x)(2\pi)\delta(k_x) \\ &= \sqrt{(2\pi)}\tilde{\psi}_0(0) = \int dx\ \psi_0(x). \end{aligned} \tag{W.2.8}$$

That is, the dx integral of $\psi(x,y)$ over any line of constant y is independent of y.

To further study smoothing in the case of a line, suppose ψ_0, now to be called ψ_0^δ, is a delta function centered on the origin,

$$\psi_0^\delta(x) = \delta(x). \tag{W.2.9}$$

Then, by (2.1),

$$\tilde{\psi}_0^\delta(k_x) = 1/\sqrt{(2\pi)}, \tag{W.2.10}$$

and (2.2) takes the form

$$\psi^\delta(x,y) = [1/(2\pi)] \int dk_x \exp(ik_x x)\exp(-ky). \tag{W.2.11}$$

This integral can be evaluated to give the result

$$\psi^\delta(x,y) = (1/\pi)[y/(x^2+y^2)]. \tag{W.2.12}$$

We next observe directly that, as expected, the function $\psi^\delta(x,y)$ given by (2.12) is harmonic. Define ρ by the rule

$$\rho = \sqrt{x^2+y^2}. \tag{W.2.13}$$

From 2-D potential theory we know that the function $\log(\rho)$ is harmonic. By the properties of the logarithm function there is the relation

$$\log(\rho^2) = 2\log(\rho), \tag{W.2.14}$$

and therefore the function $\log(\rho^2)$ is also harmonic. We next observe that the operators ∂_y and ∇^2 commute. It follows that the function $\partial_y \log(\rho^2)$ is also harmonic. Finally, there is the result

$$\partial_y \log(\rho^2) = \partial_y \log(x^2+y^2) = 2y/(x^2+y^2). \tag{W.2.15}$$

Upon comparing (2.12) and (2.15) we see that $\psi^\delta(x,y)$ is indeed harmonic.

Let us now, with the aid of (2.12), illustrate the general behavior of $\psi^\delta(x,y)$. Figure 2.1 displays $\psi^\delta(x,y)$ as a function of x for various values of y. Figure 2.2 displays $\psi^\delta(x,y)$ as a function of y for various values of x.

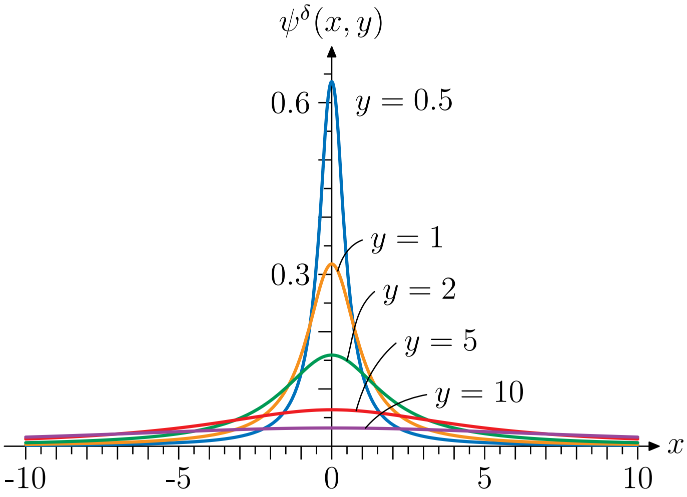


Figure W.2.1: The function $\psi^\delta(x,y)$ as a function of x for various values of y.

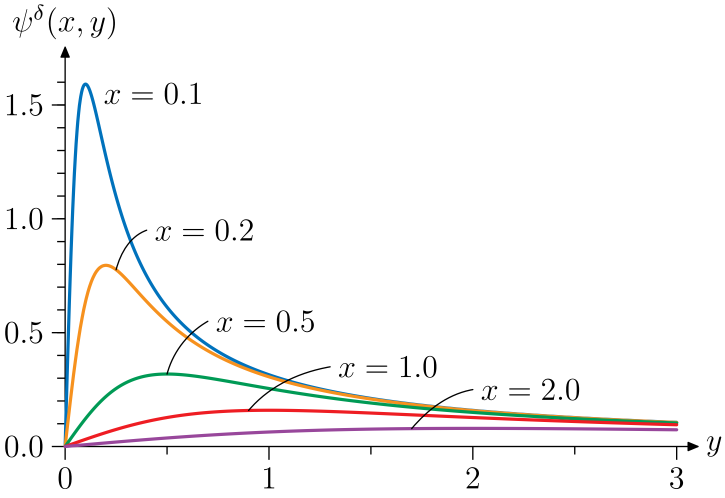


Figure W.2.2: The function $\psi^\delta(x,y)$ as a function of y for various values of x.

From Figure 2.1 we see that the delta function potential spike on the $y = 0$ line becomes an ever lower and broader bump on lines of increasing y. And we know from (2.8) that the weighted area under the bump remains the same for all y,

$$\int dx\ \psi^\delta(x,y) = \int dx\ \delta(x) = 1. \tag{W.2.16}$$

Thus the effect of a disturbance in the potential on the $y = 0$ line "decays" away as one moves to lines with successively larger values of y. Figure 2.2 illustrates this decay as a function of y for various values of x. Indeed, there are the expansions

$$\psi^\delta(x,y) = (1/\pi)(1/y)[1 - (x^2/y^2) + (x^2/y^2)^2 - \cdots] \text{ for } x < y, \tag{W.2.17}$$

$$\psi^\delta(x,y) = [1/(2\pi)](1/y) \text{ for } x = y, \tag{W.2.18}$$

$$\psi^\delta(x,y) = (1/\pi)(y/x^2)[1 - (y^2/x^2) + (y^2/x^2)^2 - \cdots] \text{ for } y < x. \tag{W.2.19}$$

Evidently, as expected, the sequence of functions $\psi^\delta(x,y)$ for varying y converges to the delta function,

$$\lim_{y\to 0^+} \psi^\delta(x,y) = \delta(x). \tag{W.2.20}$$

Finally, we see from (2.17) that $\psi^\delta(x,y)$ falls of as y^{-1} for fixed x and large y, and observe that the dimension of a line is 1. And, from (2.19), we see that $\psi^\delta(x,y)$ falls of as x^{-2} for fixed y and large x. For yet more insight, see Exercise 2.3.

What is the mechanism for this decay? In agreement with (2.8) and (2.16), this decay occurs entirely due to spreading. Indeed, the relations (2.5) and (2.7) illustrate that if the initial/boundary potential distribution is completely "spread out", i.e. constant, then no decay occurs.

We have seen an example of how the effects of a local disturbance in the potential diminish with distance from the disturbance.

From the response (2.12) to a delta function disturbance (2.9) we can derive the response to a general disturbance. With (2.12) in mind, define a kernel $G(x;x';y)$ by the rule

$$G(x;x';y) = (1/\pi)\{y/[y^2 + (x-x')^2]. \tag{W.2.21}$$

Observe that a general disturbance $\psi_0(x)$ has the integral representation

$$\psi_0(x) = \int dx'\psi_0(x')\delta(x-x'). \tag{W.2.22}$$

It follows that the response to $\psi_0(x)$ is given by the integral

$$\psi(x,y) = \int dx'\psi_0(x')G(x;x';y). \tag{W.2.23}$$

For what it's worth we remark that, according to the connection between Laplace and Monte Carlo, the quantity $G(x;x';y)$ is the probability that a random walk initiated at the point x, y will reach the point $x', 0$.

Exercises

W.2.1. Evaluate the integral (2.11) to verify the claim (2.12).

W.2.2. Evaluate directly the integral on the left side of (2.16) using (2.12). Verify (2.23) for the case (2.5).

W.2.3. The purpose of this exercise is to verify and employ an observation made by *Dan Abell.* Using (2.12), consider the *level curves* of ψ^δ having heights h by writing

$$\psi^\delta(x, y) = h. \tag{W.2.24}$$

Show that so doing yields the relation

$$x^2 + [y - 1/(2\pi h)]^2 = [1/(2\pi h)]^2. \tag{W.2.25}$$

Observe that the level curves (*equipotential lines*) are all circles, and that all the circles pass through the origin. Sketch them for yourself, and label them according to the values of h! Employ this result to explain the features of Figures 2.1 and 2.2.

W.3 The Plane in Three Space

Consider in x, y, z space the plane $z = 0$ and suppose a potential $\psi_0(x, y)$ is specified on this plane. Define its Fourier transform $\tilde{\psi}_0(k_x, k_y)$ by the rule

$$\tilde{\psi}_0(k_x, k_y) = [1/(2\pi)] \int dxdy\ \exp(-ik_x x)\exp(-ik_y y)\psi_0(x, y). \tag{W.3.1}$$

Make the Ansatz

$$\psi(x, y, z) = [1/(2\pi)] \int dk_x dk_y \exp(ik_x x)\exp(ik_y y)\exp(-kz)\tilde{\psi}_0(k_x, k_y) \tag{W.3.2}$$

where

$$k = \sqrt{k_x^2 + k_y^2}. \tag{W.3.3}$$

Evidently this $\psi(x, y, z)$ is harmonic and vanishes as $z \to +\infty$. We also have the result

$$\psi(x, y, 0) = [1/(2\pi)] \int dk_x dk_y \exp(ik_x x)\exp(ik_y y)\tilde{\psi}_0(k_x, k_y) = \psi_0(x, y). \tag{W.3.4}$$

It follows that we have found the solution to Laplace's equation in the upper half space $z \geq 0$ associated with the $z = 0$ boundary value $\psi_0(x, y)$.

Note that the operation defined by (3.2) is smoothing for $z > 0$. High spatial frequencies are suppressed by the factor $\exp(-kz)$, and this *exponential* suppression/damping is ever more effective the larger the value of z. The higher the z observation plane is above the $z = 0$ plane, the smoother $\psi(x, y, z)$ on this observation plane becomes as a function of x and y.

We also observe, in passing, two facts. First, suppose ψ_0, now to be called ψ_0^c, is a *constant* function,

$$\psi_0^c(x, y) = c. \tag{W.3.5}$$

Then, by (3.1),

$$\tilde{\psi}_0^c(k_x) = c(2\pi)\delta(k_x)\delta(k_y). \tag{W.3.6}$$

It follows from (3.2) that there is the relation

$$\psi^c(x, y, z) = c. \tag{W.3.7}$$

As expected, if ψ is constant on the boundary $z = 0$, it will have the same constant value in the upper half space $z \geq 0$. Second, for any solution, there is the relation

$$\begin{aligned}\int dxdy\ \psi(x, y, z) &= [1/(2\pi)] \int dk_x dk_y \exp(-kz)\tilde{\psi}_0(k_x, k_y) \int dxdy\ \exp(ik_x x)\exp(ik_y y)\\ &= [1/(2\pi)] \int dk_x dk_y \exp(-kz)\tilde{\psi}_0(k_x, k_y)(2\pi)^2\delta(k_x)\delta(k_y)\\ &= (2\pi)\tilde{\psi}_0(0, 0) = \int dxdy\ \psi_0(x, y). \end{aligned} \tag{W.3.8}$$

That is, the $dxdy$ integral of $\psi(x, y, z)$ over any plane of constant z is independent of z.

To further study smoothing in the case of a plane, suppose ψ_0, now to be called ψ_0^δ, is a delta function centered on the origin,

$$\psi_0^\delta(x, y) = \delta(x)\delta(y). \tag{W.3.9}$$

Then, by (3.1),

$$\tilde{\psi}_0^\delta(k_x, k_y) = 1/(2\pi), \tag{W.3.10}$$

and (3.2) takes the form

$$\psi^\delta(x, y, z) = [1/(2\pi)^2] \int dk_x dk_y \exp(ik_x x)\exp(ik_y y)\exp(-kz). \tag{W.3.11}$$

Let work to evaluate this double integral. Introduce polar variables by writing

$$\begin{aligned} x &= \rho\cos(\theta),\\ y &= \rho\sin(\theta); \end{aligned} \tag{W.3.12}$$

$$\begin{aligned} k_x &= k\cos(\phi),\\ k_y &= k\sin(\phi). \end{aligned} \tag{W.3.13}$$

Then we have the relations

$$k_x x + k_y y = k\rho[\cos(\phi)\cos(\theta) + \sin(\phi)\sin(\theta)] = k\rho\cos(\phi - \theta), \tag{W.3.14}$$

$$dk_x dk_y = kdkd\phi. \tag{W.3.15}$$

Correspondingly, (3.11) takes the form

$$\psi^\delta(x,y,z) = [1/(2\pi)^2]\int_0^\infty kdk\exp(-kz)\int_0^{2\pi} d\phi\ \exp[ik\rho\cos(\phi-\theta)]. \tag{W.3.16}$$

Next perform further manipulations. By periodicity we have the result

$$\int_0^{2\pi} d\phi\ \exp[ik\rho\cos(\phi-\theta)] = \int_0^{2\pi} d\phi\ \exp[ik\rho\cos(\phi)]. \tag{W.3.17}$$

Also there is the result

$$\exp[ik\rho\cos(\phi)] = \cos[k\rho\cos(\phi)] + i\sin[k\rho\cos(\phi)]. \tag{W.3.18}$$

Moreover, we recall the relations

$$\cos[k\rho\cos(\phi)] = J_0(k\rho) + 2\sum_{k=1}^\infty(-1)^k J_{2k}(k\rho)\cos(2k\phi), \tag{W.3.19}$$

$$\sin[k\rho\cos(\phi)] = 2\sum_{k=0}^\infty(-1)^k J_{2k+1}(k\rho)\cos[(2k+1)\phi]. \tag{W.3.20}$$

It follows that

$$\int_0^{2\pi} d\phi\ \cos[k\rho\cos(\phi)] = (2\pi)J_0(k\rho), \tag{W.3.21}$$

$$\int_0^{2\pi} d\phi\ \sin[k\rho\cos(\phi)] = 0, \tag{W.3.22}$$

and therefore

$$\int_0^{2\pi} d\phi\ \exp[ik\rho\cos(\phi)] = (2\pi)J_0(k\rho). \tag{W.3.23}$$

Upon combining the fruits of our labor we find the pleasant result

$$\psi^\delta(x,y,z) = [1/(2\pi)]\int_0^\infty kdkJ_0(k\rho)\exp(-kz). \tag{W.3.24}$$

Note that $\psi^\delta(x,y,z)$ depends on x and y only through the rotationally invariant quantity ρ, as expected by axial symmetry about the z axis.

Yet more can be accomplished. There is the general Bessel function relation

$$\int_0^\infty tdt\exp(-at)J_0(bt) = a/(a^2+b^2)^{3/2}. \tag{W.3.25}$$

Consequently, we have the final result

$$\psi^\delta(x,y,z) = [1/(2\pi)][z/(z^2+\rho^2)^{3/2}]. \tag{W.3.26}$$

We next observe directly that, as expected, the function $\psi^\delta(x,y,z)$ given by (3.26) is harmonic. Define r by the rule

$$r = \sqrt{x^2+y^2+z^2} = \sqrt{z^2+\rho^2}. \tag{W.3.27}$$

From 3-D potential theory we know that the function $1/r$ is harmonic. We next observe that the operators ∂_z and ∇^2 commute. It follows that the function $\partial_z(1/r)$ is also harmonic. Finally, there is the result

$$\partial_z(1/r) = \partial_z(1/\sqrt{z^2+\rho^2}) = z/(z^2+\rho^2)^{3/2}. \tag{W.3.28}$$

Upon comparing (3.26) and (3.28) we see that $\psi^\delta(x,y,z)$ is indeed harmonic.

Let us now, with the aid of (3.26), illustrate the general behavior of $\psi^\delta(x,y,z)$. Figure 3.1 displays $\psi^\delta(x,y,z)$ as a function of ρ for various values of z. Figure 3.2 displays $\psi^\delta(x,y,z)$ as a function of z for various values of ρ.

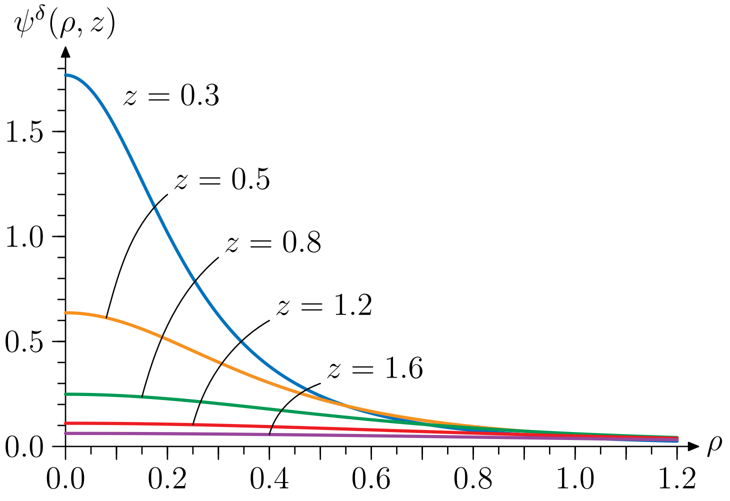


Figure W.3.1: The function $\psi^\delta(x,y,z) = \psi^\delta(\rho,z)$ as a function of ρ for various values of z.

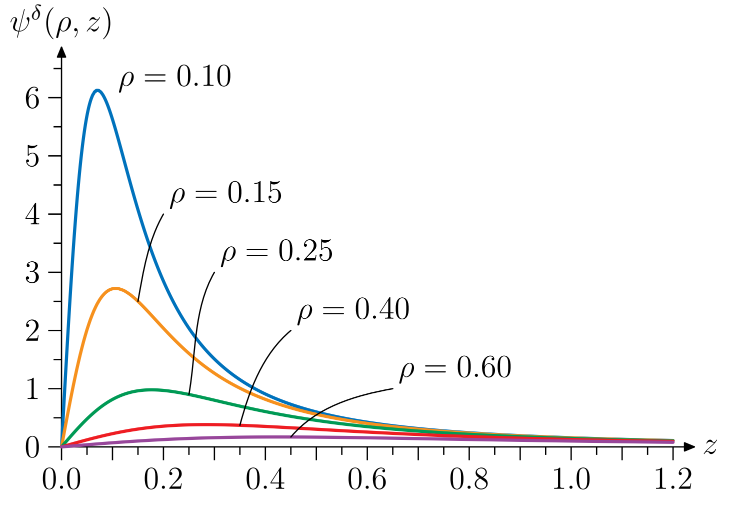


Figure W.3.2: The function $\psi^\delta(x,y,z) = \psi^\delta(\rho,z)$ as a function of z for various values of ρ.

From Figure 3.1 we see that the delta function potential spike in the $z = 0$ plane becomes an ever lower and broader bump in planes with increasing z. And we know from (3.8) that the weighted area under the bump remains the same for all z,

$$\int dxdy\ \psi^{\delta}(x,y,z) = \int dxdy\ \delta(x)\delta(y) = 1. \tag{W.3.29}$$

Thus the effect of a disturbance in the potential in the $z = 0$ plane "decays" away as one moves to planes with successively larger values of z. Figure 3.2 illustrates this decay as a function of z for various values of ρ. Indeed, there are the expansions

$$\psi^{\delta}(x,y,z) = [1/(2\pi)](1/z^2)[1-(3/2)(\rho^2/z^2)+(15/8)(\rho^2/z^2)^2-\cdots] \text{ for } \rho < z, \tag{W.3.30}$$

$$\psi^{\delta}(x,y,z) = [1/(2\pi)](1/2^{3/2})(1/z^2) \text{ for } \rho = z, \tag{W.3.31}$$

$$\psi^{\delta}(x,y,z) = [1/(2\pi)](z/\rho^3)[1-(3/2)(z^2/\rho^2)+(15/8)(z^2/\rho^2)^2-\cdots] \text{ for } z < \rho. \tag{W.3.32}$$

Evidently, as expected, the sequence of functions $\psi^{\delta}(x,y,z)$ for varying z converges to the delta function,

$$\lim_{z\to 0^+} \psi^{\delta}(x,y,z) = \delta(x)\delta(y). \tag{W.3.33}$$

Finally, we see from (3.30) that $\psi^{\delta}(x,y,z)$ falls of as z^{-2} for fixed ρ and large z, and observe that the dimension of a plane is 2. And, from (3.32), we see that $\psi^{\delta}(x,y,z)$ falls of as ρ^{-3} for fixed z and large ρ. What is the mechanism for this decay? In agreement with (3.8) and (3.29), this decay occurs entirely due to spreading. Indeed, the relations (3.5) and (3.7) illustrate that if the initial/boundary potential distribution is completely "spread out", i.e. constant, then no decay occurs.

We have again seen an example of how the effects of a local disturbance in the potential diminish with distance from the disturbance.

From the response (3.26) to a delta function disturbance (3.9) we can derive the response to a general disturbance. With (3.26) in mind, define a kernel $G(x,y;x',y';z)$ by the rule

$$G(x,y;x',y';z) = [1/(2\pi)]\{z/[z^2+(x-x')^2+(y-y')^2]^{3/2}\}. \tag{W.3.34}$$

Observe that a general disturbance $\psi_0(x,y)$ has the integral representation

$$\psi_0(x,y) = \int dx'dy'\ \psi_0(x',y')\delta(x-x')\delta(y-y'). \tag{W.3.35}$$

It follows that the response to $\psi_0(x,y)$ is given by the integral

$$\psi(x,y,z) = \int dx'dy'\ \psi_0(x',y')G(x,y;x',y';z). \tag{W.3.36}$$

For what it's worth we remark that, according to the connection between Laplace and Monte Carlo, the quantity $G(x,y;x',y';z)$ is the probability that a random walk initiated at the point x,y,z will reach the point $x',y',0$.

Exercises

W.3.1. Evaluate directly the integral on the left side of (3.29) using (3.26). Verify (3.36) for the case (3.5).

W.4 The Circle in Two Space

Consider in x, y space a circle of radius R centered on the origin, and suppose a potential ψ_R is specified on this circle. More specifically, employ the polar variables (3.9) so that

$$\psi(x, y) = \psi(\rho, \theta) \tag{W.4.1}$$

and

$$\psi(R, \theta) = \psi_R(\theta). \tag{W.4.2}$$

Define the angular Fourier transform of $\psi_R(\theta)$ by the rule

$$\tilde{\psi}_R(m) = ([1/(2\pi)] \int_0^{2\pi} d\theta \ \exp(-im\theta)\psi_R(\theta). \tag{W.4.3}$$

Make the Ansatz

$$\psi(\rho, \theta) = \sum_{m=-\infty}^{m=\infty} (\rho/R)^{|m|} \exp(im\theta)\tilde{\psi}_R(m). \tag{W.4.4}$$

Evidently this $\psi(\rho, \theta)$ is harmonic. We also have the result

$$\psi(R, \theta) = \sum_{m=-\infty}^{m=\infty} \exp(im\theta)\tilde{\psi}_R(m) = \psi_R(\theta). \tag{W.4.5}$$

It follows that we have found the solution to Laplace's equation in the disk of radius R with the boundary value $\psi_R(\theta)$.

Note that the operation defined by (4.4) is smoothing for $(\rho/R) < 1$. We see that high angular frequencies are suppressed by the factor

$$(\rho/R)^{|m|} = \exp[|m| \log(\rho/R)], \tag{W.4.6}$$

and observe that $\log(\rho/R) < 0$. This *exponential* suppression/damping is ever more effective the larger the value of R and/or the smaller the value of ρ. The larger the radius R of the boundary circle and/or the smaller the radius ρ of the observation circle, the smoother $\psi(\rho, \theta)$ becomes as a function of θ.

We also observe, in passing, three facts. First, suppose ψ_R, now to be called ψ_R^c, is a *constant* function,

$$\psi_R^c(\theta) = c. \tag{W.4.7}$$

Then, by (4.3),

$$\tilde{\psi}_R^c(m) = c\delta_{m,0}. \tag{W.4.8}$$

It follows from (4.4) that there is the relation

$$\psi^c(\rho, \theta) = c. \tag{W.4.9}$$

As expected, if ψ is constant on the boundary $\rho = R$, it will have the same constant value inside the circle. Second, for any solution, there is the relation

$$\int_0^{2\pi} d\theta \ \psi(\rho, \theta) = (2\pi) \sum_{m=-\infty}^{m=\infty} (\rho/R)^{|m|}\tilde{\psi}_R(m)\delta_{m,0} = (2\pi)\tilde{\psi}_R(0) = \int_0^{2\pi} d\theta \ \psi_R(\theta). \tag{W.4.10}$$

That is, the $d\theta$ integral of $\psi(\rho, \theta)$ over any circle of constant ρ is independent of ρ. Third, we also see from (4.4) that there is the relation

$$\psi(0, \theta) = \tilde{\psi}_R(0) = [1/(2\pi)] \int_0^{2\pi} d\theta \psi_R(\theta), \tag{W.4.11}$$

which shows that the average value of an harmonic function over a circle equals its value at the center of the circle, a result which in turn is a special case of the connection between Laplace and Monte Carlo.

To further study smoothing in the case of a circle in two space, suppose ψ_R, now to be called ψ_R^δ, is a delta function centered on $\theta = 0$,

$$\psi_R^\delta(\theta) = \delta(\theta). \tag{W.4.12}$$

Then, by (4.3),

$$\tilde{\psi}_R^\delta(m) = 1/(2\pi), \tag{W.4.13}$$

and (4.4) takes the form

$$\psi^\delta(\rho, \theta) = [1/(2\pi)] \sum_{m=-\infty}^{m=\infty} (\rho/R)^{|m|} \exp(im\theta). \tag{W.4.14}$$

Let us work to evaluate this sum. Introduce the simplifying notation

$$\lambda = \rho/R \tag{W.4.15}$$

with the understanding that, for our purposes,

$$\lambda \in [0, 1]. \tag{W.4.16}$$

Correspondingly, make the definition

$$\hat{\psi}^\delta(\lambda, \theta) = \psi^\delta(\rho, \theta) = [1/(2\pi)] \sum_{m=-\infty}^{m=\infty} \lambda^{|m|} \exp(im\theta). \tag{W.4.17}$$

Observe that there is the result

$$\sum_{m=-\infty}^{m=\infty} \lambda^{|m|} \exp(im\theta) = -1 + \sum_{m=0}^{m=\infty} \lambda^m \exp(im\theta) + \sum_{m=0}^{m=\infty} \lambda^m \exp(-im\theta). \tag{W.4.18}$$

Each of the series appearing on the right side of (4.18) is a geometric series, and can therefore be evaluated. We find the results

$$\sum_{m=0}^{m=\infty} \lambda^m \exp(im\theta) = 1/[1 - \lambda \exp(i\theta)], \tag{W.4.19}$$

$$\sum_{m=0}^{m=\infty} \lambda^m \exp(-im\theta) = 1/[1 - \lambda \exp(-i\theta)]. \tag{W.4.20}$$

Consequently,

$$\hat{\psi}^{\delta}(\lambda,\theta)=[1/(2\pi)]\{-1+1/[1-\lambda\exp(i\theta)]+1/[1-\lambda\exp(-i\theta)]\}. \tag{W.4.21}$$

The three terms on the right side of (4.21) can be put over a common denominator. Doing so, and recalling that

$$\cos(\theta)=(1/2)[\exp(i\theta)+\exp(-i\theta)], \tag{W.4.22}$$

give the final result

$$\hat{\psi}^{\delta}(\lambda,\theta)=[1/(2\pi)]\{[\lambda^{-1}-\lambda]/[\lambda^{-1}+\lambda-2\cos(\theta)]\}. \tag{W.4.23}$$

We know that, by construction, the function $\hat{\psi}^{\delta}(\lambda,\theta)$ is harmonic. See (4.14). We will next observe *directly* that the function $\hat{\psi}^{\delta}(\lambda,\theta)$, as given by (4.23), is harmonic. To do so we will exploit a fact about analytic functions. Suppose f is an *analytic* function of the complex variable $z=x+iy$. (Here z is *not* a Cartesian coordinate.) Define a function $u(x,y)$ by writing

$$u(x,y)=f(z)=f(x+iy). \tag{W.4.24}$$

Then, by the chain rule, it follows that

$$(\partial_x)^2u(x,y)=f''(z) \tag{W.4.25}$$

and

$$(\partial_y)^2u(x,y)=(i^2)f''(z)=-f''(z), \tag{W.4.26}$$

from which it follows that

$$[(\partial_x)^2+(\partial_y)^2]u(x,y)=0; \tag{W.4.27}$$

the function $u(x,y)$ defined by (4.24) is harmonic. Similarly the function $v(x,y)$ defined by

$$v(x,y)=f(\bar{z})=f(x-iy) \tag{W.4.28}$$

is also harmonic. Now look at the right sides of (4.19) and (4.20). They can be rewritten in the forms

$$1/[1-\lambda\exp(i\theta)]=1/[1-(1/R)(x+iy)]=1/[1-(1/R)z], \tag{W.4.29}$$

$$1/[1-\lambda\exp(-i\theta)]=1/[1-(1/R)(x-iy)]=1/[1-(1/R)\bar{z}]. \tag{W.4.30}$$

It follows that both these functions are harmonic. Finally, we see from (4.23) that $\hat{\psi}^{\delta}(\lambda,\theta)$ is the sum of a constant (which is a harmonic function) and multiples of the harmonic functions in (4.29) and (4.30). Therefore $\hat{\psi}^{\delta}(\lambda,\theta)$ is harmonic.

Let us now, with the aid of (4.23), illustrate the general behavior of $\hat{\psi}^{\delta}$. Figure 4.1 displays the function $\hat{\psi}^{\delta}(\lambda,\theta)$ as a function of $\theta\in(-\pi,\pi)$ for various values of $\lambda\in[0,1]$. Note that there are the relations

$$(\lambda^{-1}+\lambda)>2 \text{ for } \lambda\in(0,1) \tag{W.4.31}$$

and

$$(\lambda^{-1}+\lambda)=2 \text{ for } \lambda=1. \tag{W.4.32}$$

From the figure two facts are evident:

- For $\lambda \simeq 1$, $\hat{\psi}^{\delta}(\lambda,\theta)$ is highly peaked about $\theta = 0$ and is small for $\theta \neq 0$.
- For $\lambda \simeq 0$, $\hat{\psi}^{\delta}(\lambda,\theta)$ is nearly 1. Indeed, by (4.14), there is the small λ expansion

$$\hat{\psi}^{\delta}(\lambda,\theta) = [1/(2\pi)][1 + 2\lambda\cos(\theta) + 2\lambda^2\cos(2\theta) + \cdots]. \tag{W.4.33}$$

Also, again by (4.14), there is the integral relation

$$\int_{-\pi}^{\pi} d\theta\ \hat{\psi}^{\delta}(\lambda,\theta) = 1. \tag{W.4.34}$$

[Note that (4.34) is a special case of (4.10).] Putting all these facts together, we conclude that the sequence of functions $\hat{\psi}^{\delta}(\lambda,\theta)$ for varying λ converges to the delta function,

$$\lim_{\lambda\to 1^-} \hat{\psi}^{\delta}(\lambda,\theta) = \delta(\theta). \tag{W.4.35}$$

We see that the delta function spike about $\theta = 0$ on the circle $\rho = R$ (which corresponds to $\lambda = 1$) becomes an ever lower and broader bump with decreasing ρ. Thus the effect of a disturbance in the potential on the $\rho = R$ circle "decays" away as one moves to circles with successively smaller values of ρ. Finally we note that, in agreement with (4.10), the "decay" we have been observing is entirely due to spreading. Indeed, the relations (4.7) and (4.9) illustrate that if the initial/boundary potential distribution is completely "spread out", i.e. constant, then no decay occurs.

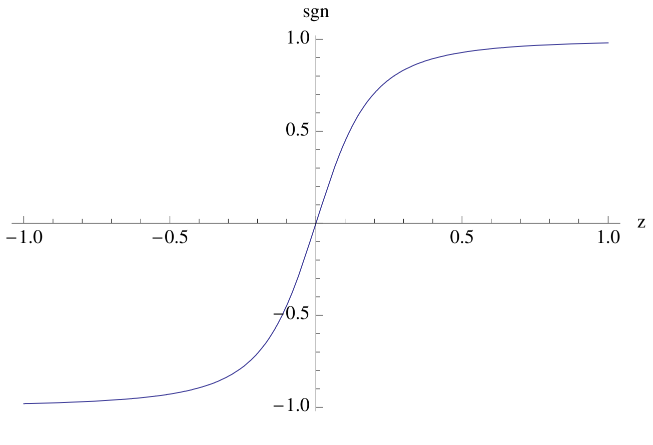


Figure W.4.1: (Place Holder) The function $\hat{\psi}^{\delta}(\lambda,\theta)$ as a function of θ for various values of λ.

We have again seen an example of how the effects of a local disturbance in the potential diminish with distance from the disturbance.

From the response (4.23) to a delta function disturbance (4.12) we can derive the response to a general disturbance. With (4.23) in mind, define a kernel $G(\theta;\theta';\lambda)$ by the rule

$$G(\theta;\theta';\lambda) = [1/(2\pi)]\{[\lambda^{-1} - \lambda]/[\lambda^{-1} + \lambda - 2\cos(\theta - \theta')]\}. \tag{W.4.36}$$

Observe that a general disturbance $\psi_R(\theta)$ has the integral representation

$$\psi_R(\theta) = \int_{-\pi}^{\pi} d\theta' \psi_R(\theta')\delta(\theta - \theta'). \tag{W.4.37}$$

It follows that the response to $\psi_R(\theta)$ is given by the integral

$$\psi(\rho,\theta) = \int_{-\pi}^{\pi} d\theta' \psi_R(\theta') G(\theta;\theta';\lambda). \tag{W.4.38}$$

For what it's worth we remark that, according to the connection between Laplace and Monte Carlo, the quantity $G(\theta;\theta';\lambda)$ is the probability that a random walk initiated at the point ρ, θ will reach the point R, θ'.

We close this section by making two calculations that will be of future use. First we will present a previous result in terms of the Cartesian coordinates x, y. From (4.33) we see that there is the result

$$\psi^\delta(x,y) = \psi^\delta(\rho,\theta) = [1/(2\pi)][1 + 2(\rho/R)\cos(\theta) + 2(\rho/R)^2\cos(2\theta) + \cdots]. \tag{W.4.39}$$

There are also the relations

$$\rho\cos(\theta) = x, \tag{W.4.40}$$

$$\rho^2\cos(2\theta) = \rho^2[\cos^2(\theta) - \sin^2(\theta)] = x^2 - y^2. \tag{W.4.41}$$

It follows that there is the result

$$\psi^\delta(x,y) = [1/(2\pi)][1 + 2(1/R)x + 2(1/R^2)(x^2 - y^2) + \cdots]. \tag{W.4.42}$$

As a second complementary case, suppose ψ_R, now to be called ψ_R^Δ, is a delta function centered on $\theta = \pi/2$,

$$\psi_R^\Delta(\theta) = \delta(\theta - \pi/2). \tag{W.4.43}$$

Then, by (4.3),

$$\tilde{\psi}_R^\Delta(m) = 1/(2\pi)\exp(-im\pi/2), \tag{W.4.44}$$

and (4.4) takes the form

$$\begin{aligned}\psi^\Delta(x,y) &= \psi^\Delta(\rho,\theta) = [1/(2\pi)] \sum_{m=-\infty}^{m=\infty} (\rho/R)^{|m|} \exp[im(\theta - \pi/2)] \\ &= [1/(2\pi)\{1 + 2(\rho/R)\cos(\theta - \pi/2) + 2(\rho/R)^2\cos[2(\theta - \pi/2)] + \cdots\} \\ &= [1/(2\pi)\{1 + 2(\rho/R)\sin(\theta) - 2(\rho/R)^2\cos(2\theta) + \cdots\} \\ &= [1/(2\pi)\{1 + 2(1/R)y - 2(1/R^2)(x^2 - y^2) + \cdots\}.\end{aligned} \tag{W.4.45}$$

Exercises

W.4.1. Verify the steps that led from (4.14) to (4.23).

W.4.2. Verify (4.29) and (4.30).

W.4.3. Verify (4.31).

W.4.4. Show that

$$\lim_{\lambda \to 1^-} \hat{\psi}^\delta(\lambda, \theta) = 0 \text{ for } \theta \neq 0. \tag{W.4.46}$$

Show that

$$\lim_{\lambda \to 1^-} \hat{\psi}^\delta(\lambda, 0) = +\infty. \tag{W.4.47}$$

W.4.5. Verify directly the relation (4.34) using (4.23). Verify (4.38) for the case (4.7).

W.5 The Circular Cylinder in Three Space

Consider in x, y, z space a circular cylinder of radius R centered on the z axis, and suppose a potential $\psi_R(\phi, z)$ is specified on this cylinder. [Here we have used the cylindrical coordinates ρ, ϕ, and z specified by the rules (15.2.12) through (15.2.16).] More specifically, write

$$\psi(x, y, z) = \psi(\rho, \phi, z) \tag{W.5.1}$$

and

$$\psi(R, \phi, z) = \psi_R(\phi, z). \tag{W.5.2}$$

Given the boundary potential $\psi_R(\phi, z)$, we wish to find the interior harmonic function (solution to Laplace's equation) $\psi(\rho, \phi, z)$ associated with this boundary potential.

From (15.3.7) we know that the most general ψ that is harmonic and finite within the cylinder is of the form

$$\psi(\rho, \phi, z) = \sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} dk \, G_m(k) \exp(ikz) \exp(im\phi) I_m(k\rho). \tag{W.5.3}$$

Next define the *double* Fourier transform $\tilde{\tilde{\psi}}(R, m', k')$ of the boundary potential by the rule

$$\tilde{\tilde{\psi}}(R, m', k') = [1/(2\pi)]^2 \int_{-\infty}^{\infty} dz \; \exp(-ik'z) \int_0^{2\pi} d\phi \; \exp(-im'\phi) \psi_R(\phi, z). \tag{W.5.4}$$

See (17.2.2). Then we know from (17.2.5) that

$$G_m(k) = \tilde{\tilde{\psi}}(R, m, k)/I_m(kR). \tag{W.5.5}$$

Consequently, the desired ψ is given by the relation

$$\psi(\rho, \phi, z) = \sum_{m=-\infty}^{\infty} \exp(im\phi) \int_{-\infty}^{\infty} dk \; \exp(ikz) \tilde{\tilde{\psi}}(R, m, k)[I_m(k\rho)/I_m(kR)]. \tag{W.5.6}$$

Let us verify that our criteria have been met. By construction the ψ given by (5.6) is a superposition of the functions $\exp(ikz)\exp(im\phi)I_m(k\rho)$ and therefore is harmonic. Also it has the property

$$\psi(R, \phi, z) = \sum_{m=-\infty}^{\infty} \exp(im\phi) \int_{-\infty}^{\infty} dk \; \exp(ikz) \tilde{\tilde{\psi}}(R, m, k) = \psi_R(\phi, z). \tag{W.5.7}$$

We have found the solution to Laplace's equation in the circular cylinder in three space having boundary $\rho = R$ and the boundary value $\psi_R(\phi, z)$.

We also observe, in passing, two facts. First, suppose ψ_R, now to be called ψ_R^c, is a *constant* function,

$$\psi_R^c(\phi, z) = c. \tag{W.5.8}$$

Then, by (5.4),

$$\tilde{\tilde{\psi}}^c(R, m', k') = c\delta(k')\delta_{m',0}. \tag{W.5.9}$$

It follows from (5.6) that there is the relation

$$\psi^c(\rho, \phi, z) = c. \tag{W.5.10}$$

As expected, if ψ is constant on the cylinder boundary $\rho = R$, it will have the same constant value everywhere within the cylinder. Second, for any solution, there is the relation

$$\begin{aligned}\int_0^{2\pi} d\phi \int_{-\infty}^{\infty} dz\ \psi(\rho, \phi, z) &= (2\pi)^2 \sum_{m=-\infty}^{\infty} \delta_{m,0} \int_{-\infty}^{\infty} dk\ \delta(k) \tilde{\tilde{\psi}}(R, m, k)[I_m(k\rho)/I_m(kR)] \\ &= (2\pi)^2 \tilde{\tilde{\psi}}(R, 0, 0) = \int_0^{2\pi} d\phi \int_{-\infty}^{\infty} dz\ \psi_R(\phi, z). \end{aligned} \tag{W.5.11}$$

That is, the $d\phi dz$ integral of $\psi(\rho, \phi, z)$ over any cylinder of constant ρ is independent of ρ.

We are now prepared to address the general subject of smoothing. We will examine the asymptotic behavior of the kernel $K(m, k; \rho, R)$, defined by the rule

$$K(m, k; \rho, R) = [I_m(k\rho)/I_m(kR)], \tag{W.5.12}$$

that appears in (5.6). Note that, because

$$I_{-m}(w) = I_m(w) \tag{W.5.13}$$

and

$$I_m(-w) = (-1)^m I_m(w), \tag{W.5.14}$$

the kernel $K(m, k; \rho, R)$ is evidently an *even* function of both m and k. Therefore we only need consider the cases $m \geq 0$ and $k \geq 0$. Finally, by (15.3.11), we see that

$$K(m, 0; \rho, R) = (\rho/R)^{|m|}. \tag{W.5.15}$$

For fixed m and large w the Bessel functions $I_m(w)$ have the asymptotic property

$$|I_m(w)| \simeq (1/\sqrt{2\pi w}) \exp(w) \text{ as } w \to \infty. \tag{W.5.16}$$

Consequently, for fixed m, there is the asymptotic relation

$$K(m, k; \rho, R) \simeq (\sqrt{R/\rho}) \exp[k(\rho - R)] \text{ as } k \to \infty. \tag{W.5.17}$$

We see that, for each fixed m, there is smoothing/damping in the longitudinal variable z when $\rho < R$. Note that this smoothing is analogous to that for the line in two space and the plane in three space.

For fixed w and large m the Bessel functions $I_m(w)$ have the asymptotic property

$$\begin{aligned} |I_m(w)| &\simeq (1/\sqrt{2\pi m})[(e|w|)/(2m)]^m \\ &\simeq (1/2)^m[\sqrt{2\pi m}(m/e)^m]^{-1}|w|^m \\ &\simeq (1/2)^m(1/m!)|w|^m \text{ as } m \to \infty. \end{aligned} \tag{W.5.18}$$

Here we have used the Stirling large m approximation

$$m! \simeq \sqrt{2\pi m}(m/e)^m, \tag{W.5.19}$$

which is actually already quite accurate for $m \geq 2$.[1] Consequently, for fixed k, there is the asymptotic relation

$$K(m,k;\rho,R) \simeq (\rho/R)^m \text{ as } m \to \infty. \tag{W.5.20}$$

We see that, for each fixed k, there is smoothing/damping in the angular variable ϕ when $\rho < R$. Note that this smoothing is analogous to that for the circle in two space.

With regard to angular smoothing there is also the consideration that the angular Fourier transform filters out all angular Fourier modes save for the one of interest. Moreover, the disturbance in the angular Fourier mode of interest produced by an error in any given grid-point value is suppressed by $1/N$ where N is the number of sampling points used in the discrete angular Fourier transform.

What happens if k and m increase simultaneously? This is a more difficult question. We will explore the case were k and m are proportional,

$$k = \lambda m \tag{W.5.21}$$

where λ is some proportionality constant having the dimensions of inverse length.

If $w = \tau m$, where τ is some proportionality constant, there is the uniform doubly asymptotic relation

$$I_m(\tau m) \simeq (1/\sqrt{2\pi m})[1/(1+\tau^2)^{1/4}]\exp(m\eta) \text{ as } m \to \infty. \tag{W.5.22}$$

Here η is a function of τ given by the relation

$$\eta(\tau) = \sqrt{1+\tau^2} + \log[\tau/(1+\sqrt{1+\tau^2})]. \tag{W.5.23}$$

We will use the assumption (5.21) and the result (5.22) to estimate $I_m(\lambda m\rho)$ and $I_m(\lambda mR)$, the numerator and denominator appearing in (5.9).

For the numerator define a quantity $\hat{\tau}$ by the rule

$$\hat{\tau} = \lambda\rho, \tag{W.5.24}$$

and for the denominator define a quantity $\check{\tau}$ by the rule

$$\check{\tau} = \lambda R. \tag{W.5.25}$$

[1] Note that the final result in (5.18) also follows from retaining only the $\ell = 0$ term in (15.3.11).

(Note that both $\hat{\tau}$ and $\check{\tau}$ are dimensionless.) In terms of these definitions we have, as a consequence of (5.22), the large m results

$$I_m(\lambda m\rho) \simeq (1/\sqrt{2\pi m})[1/(1+\hat{\tau}^2)^{1/4}]\exp(m\hat{\eta}), \tag{W.5.26}$$

$$I_m(\lambda mR) \simeq (1/\sqrt{2\pi m})[1/(1+\check{\tau}^2)^{1/4}]\exp(m\check{\eta}). \tag{W.5.27}$$

Here we have used the notation

$$\hat{\eta} = \eta(\hat{\tau}), \tag{W.5.28}$$

$$\check{\eta} = \eta(\check{\tau}). \tag{W.5.29}$$

It follows that there is the large m result

$$K(m, \lambda m; \rho, R) \simeq [(1+\check{\tau}^2)^{1/4}/(1+\hat{\tau}^2)^{1/4}]\exp[m(\hat{\eta}-\check{\eta})] \text{ as } m \to \infty. \tag{W.5.30}$$

We see that there is *exponential* smoothing/damping if

$$\hat{\eta} < \check{\eta}. \tag{W.5.31}$$

When does the smoothing condition (5.31) hold? Let us examine the function $\eta(\tau)$ given by (5.23). Its behavior is displayed in Figure 5.1. Evidently, for $\tau \geq 0$, it appears to be *monotonically increasing*. This surmise is proved in Exercise 5.1. Consequently, (5.31) holds if $\hat{\tau} < \check{\tau}$ and hence $\rho < R$. We conclude, assuming $\rho < R$, that there is exponential smoothing/damping as one goes out in any direction from the origin in the m k plane; and the damping rate depends on the direction. For example, Figure 5.2 displays $K(m, k; \rho, R)$ as function of m and k for the case $\rho = 2$ cm and $R = 2.5$ cm.

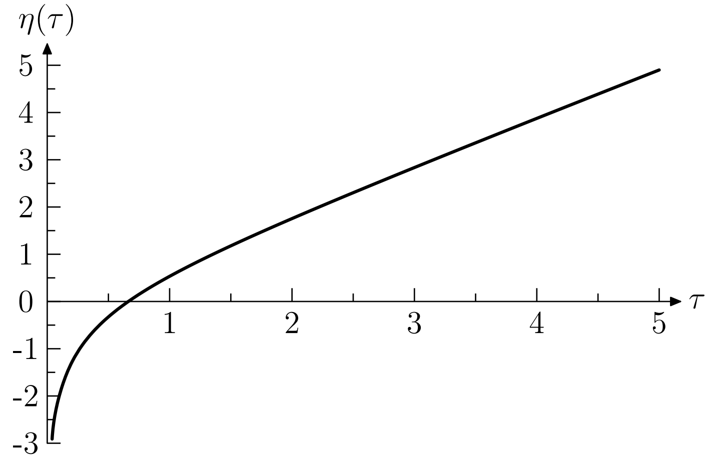


Figure W.5.1: The function $\eta(\tau)$. It appears to be monotonically increasing.

To further study smoothing in the case of a circular cylinder in three space, suppose ψ_R, now to be called ψ_R^δ, is a delta function centered on $(\phi, z) = (0, 0)$,

$$\psi_R^\delta(\phi, z) = \delta(\phi)\delta(z). \tag{W.5.32}$$

Then, by (5.4),

$$\tilde{\tilde{\psi}}^\delta(R, m', k') = [1/(2\pi)]^2, \tag{W.5.33}$$

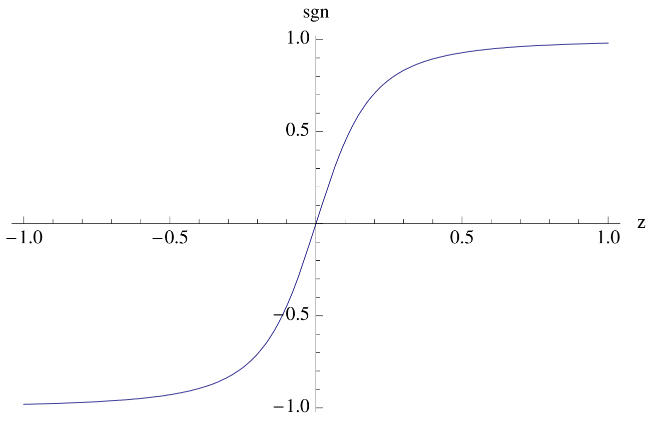


Figure W.5.2: (Place Holder) The kernel $K(m, k; \rho, R)$ as function of m and k for the case $\rho = 2$ cm and $R = 2.5$ cm. The quantity k has units of inverse centimeters.

and (5.6) takes the form

$$\psi^\delta(\rho, \phi, z) = [1/(2\pi)]^2 \sum_{m=-\infty}^{\infty} \exp(im\phi) \int_{-\infty}^{\infty} dk\ \exp(ikz)[I_m(k\rho)/I_m(kR)]. \tag{W.5.34}$$

Our task now is to study the properties of $\psi^\delta(\rho, \phi, z)$. We begin by observing that, as a consequence of (5.11), there is the integral relation

$$\int_0^{2\pi} d\phi \int_{-\infty}^{\infty} dz\ \psi^\delta(\rho, \phi, z) = (2\pi)^2 \tilde{\tilde{\psi}}^\delta(R, 0, 0) = 1. \tag{W.5.35}$$

To proceed further, and in analogy to what was done in previous sections for ψ^δ, it would be ideal if the representation (5.34) could be evaluated analytically in terms of known functions. However, this seems to be a difficult. What we can do is to evaluate (5.34) numerically for various values of ρ and R. Figure 5.3 shows $\psi^\delta(\rho, \phi, z)$ as a function of ϕ and z when $\rho = 2$ cm and $R = 2.5$ cm. And Figure 5.4 shows $\psi^\delta(\rho, \phi, z)$ as a function of ϕ and z when $\rho = 1$ cm and $R = 2.5$ cm. Evidently $\psi^\delta(\rho, \phi, z)$ falls off for large z. Moreover, it is smaller and less peaked about $(\phi, z) = (0, 0)$ for the smaller value of ρ. The effect of a disturbance in the potential at the point $(\phi, z) = (0, 0)$ "decays" away as one moves to cylinders with successively smaller values of ρ. Conversely, as ρ increases, the sequence of functions $\psi^\delta(\rho, \phi, z)$ for varying ρ converges to the delta function,

$$\lim_{\rho \to R^-} \psi^\delta(\rho, \phi, z) = \delta(\phi)\delta(z). \tag{W.5.36}$$

Finally we note that, in agreement with (5.11), the "decay" we have been observing is entirely due to spreading. Indeed, the relations (5.8) and (5.10) illustrate that if the initial/boundary potential distribution is completely "spread out", i.e. constant, then no decay occurs.

At this point we might wonder how fast the effect of a disturbance falls off as a function of z. Figure 5.5 shows $\psi^\delta(1, 0, z)$ as a function of z when $R = 2.5$ cm. We can also study the on-axis case $\rho = 0$ for which (5.34) takes the simpler form

$$\psi^\delta(0, \phi, z) = [1/(2\pi)]^2 \int_{-\infty}^{\infty} dk\ \exp(ikz)[1/I_0(kR)]. \tag{W.5.37}$$

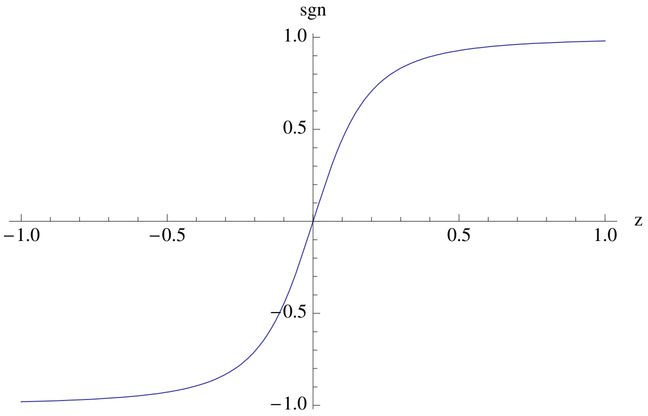


Figure W.5.3: (Place Holder) The function $\psi^{\delta}(\rho,\phi,z)$ as a function of ϕ and z when $\rho = 2$ cm and $R = 2.5$ cm.

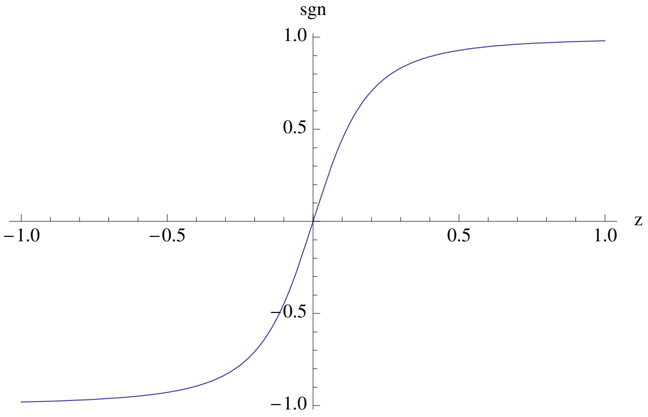


Figure W.5.4: (Place Holder) The function $\psi^{\delta}(\rho,\phi,z)$ as a function of ϕ and z when $\rho = 1$ cm and $R = 2.5$ cm.

Make the change of variables $\lambda = kR$. So doing brings (5.37) to the form

$$\psi^{\delta}(0, \phi, z) = [1/(2\pi)]^2 (1/R) \int_{-\infty}^{\infty} d\lambda \; \exp(i\lambda z/R)/I_0(\lambda). \tag{W.5.38}$$

We have already studied the integral appearing on the right side of (5.38). Reference to (21.1.37) shows that there is the result

$$\psi^{\delta}(0, \phi, z) = [1/(2\pi)](1/R) F(z/R, 0). \tag{W.5.39}$$

It follows from the work of Section 21.1.3, see Figure 21.1.2, that there is the asymptotic behavior

$$\psi^{\delta}(0, \phi, z) \propto \exp[-\pi |z|/(2R)] \text{ as } |z| \to \infty. \tag{W.5.40}$$

When viewed from on axis the effect of a disturbance falls off exponentially. Reference to Figure 5.5 shows that there is a similar rapid fall off with z in the off-axis case.

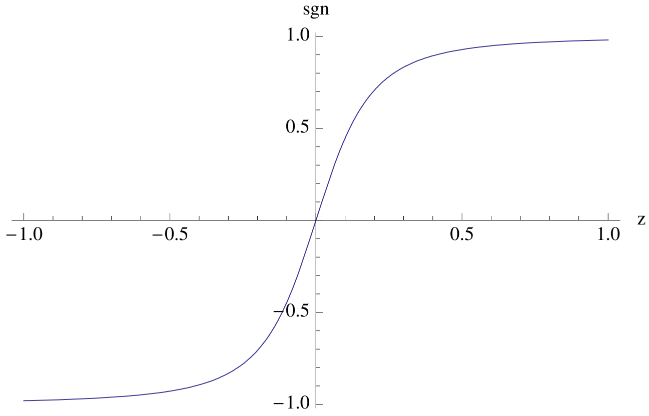


Figure W.5.5: (Place Holder) The function $\psi^{\delta}(1, 0, z)$ as a function of z when $R = 2.5$ cm.

We have again seen an example of how the effects of a local disturbance in the potential diminish with distance from the disturbance.

From the response (5.34) to a delta function disturbance (5.32) we can derive the response to a general disturbance. With (5.34) in mind, define a kernel $G(\phi, z; \phi', z'; \rho)$ by the rule

$$G(\phi, z; \phi', z'; \rho) = \psi^{\delta}(\rho, \phi - \phi', z - z'). \tag{W.5.41}$$

Observe that a general disturbance $\psi_R(\phi, z)$ has the integral representation

$$\psi_R(\phi, z) = \int_{-\pi}^{\pi} d\phi' \int_{-\infty}^{\infty} dz' \psi_R(\phi', z') \delta(\phi - \phi') \delta(z - z'). \tag{W.5.42}$$

It follows that the response to $\psi_R(\phi, z)$ is given by the integral

$$\psi(\rho, \phi, z) = \int_{-\pi}^{\pi} d\phi' \int_{-\infty}^{\infty} dz' \psi_R(\phi', z') G(\phi, z; \phi', z'; \rho). \tag{W.5.43}$$

For what it's worth we remark that, according to the connection between Laplace and Monte Carlo, the quantity $G(\phi, z; \phi', z'; \rho)$ is the probability that a random walk initiated at the point ρ, ϕ, z will reach the point R, ϕ', z'.

Exercises

W.5.1. The purpose of this exercise is to prove that $\eta(\tau)$ is a monotonically increasing function of τ. We begin by observing that the $\sqrt{1+\tau^2}$ term just to the right of the equal sign in (5.23) is a monotonically increasing function of τ, and we know that the log function is a monotonically increasing function of its argument. It remains to be shown that the function $f(\tau)$ defined by

$$f(\tau) = \tau/(1+\sqrt{1+\tau^2}), \tag{W.5.44}$$

the argument of the log function, is monotonically increasing. If this can be verified, then $\eta(\tau)$ is a monotonically increasing function of τ.

To complete the proof, show that

$$\begin{aligned} f'(\tau) &= 1/(1+\sqrt{1+\tau^2}) - (\tau^2/\sqrt{1+\tau^2})/(1+\sqrt{1+\tau^2})^2 \\ &= [1/(1+\sqrt{1+\tau^2})^2][(1+\sqrt{1+\tau^2}) - (\tau^2/\sqrt{1+\tau^2})] \\ &= [1/(1+\sqrt{1+\tau^2})^2](1/\sqrt{1+\tau^2})(\sqrt{1+\tau^2}+1+\tau^2-\tau^2) \\ &= [1/(1+\sqrt{1+\tau^2})^2](1/\sqrt{1+\tau^2})(\sqrt{1+\tau^2}+1) \\ &= [1/(1+\sqrt{1+\tau^2})](1/\sqrt{1+\tau^2}). \end{aligned} \tag{W.5.45}$$

Evidently all factors on the far right side of (5.45) are positive, and therefore $f(\tau)$ is monotonically increasing.

W.5.2. Verify (5.15) given (15.3.11).

W.5.3. Verify (5.17) given (5.16).

W.5.4. Verify (5.20) given (5.18).

W.5.5. Verify (5.30) given (5.22).

W.5.6. Verify (5.38) given (5.37).

W.5.7. Compare the fall off in the case of a plane in three space given by (3.32) with the fall off in the case of a circular cylinder in three space given (5.40). Explain why the fall off is so much more rapid in the case of a cylinder.

W.6 The Ellipse in Two Space

For the ellipse in two space let us employ the coordinates given by (17.4.1) and (17.4.2) and illustrated in Figure 17.4.2. Then, upon writing the relation

$$\psi(x, y) = \psi(u, v), \tag{W.6.1}$$

we find from (17.4.12) that

$$\nabla^2\psi = (1/f^2)[\cosh^2(u) - \cos^2(v)]^{-1}[(\partial_u)^2 + (\partial_v)^2]\psi. \tag{W.6.2}$$

Consequently, ψ will be harmonic provided it satisfies the relation

$$[(\partial_u)^2 + (\partial_v)^2]\psi = 0. \tag{W.6.3}$$

In analogy with (17.4.35) through (17.4.37) let us define functions $\mathrm{c}_n(v)$ and $\mathrm{s}_n(v)$ by the rules

$$\mathrm{c}_0(v) = 1/\sqrt{2}, \tag{W.6.4}$$

$$\mathrm{c}_n(v) = \cos(nv) \text{ for } n \geq 1, \tag{W.6.5}$$

$$\mathrm{s}_0(v) = 0, \tag{W.6.6}$$

$$\mathrm{s}_n(v) = \sin(nv) \text{ for } n \geq 1. \tag{W.6.7}$$

Evidently they form a complete orthogonal set and are normalized so that

$$\int_0^{2\pi} dv \; \mathrm{c}_m(v) \; \mathrm{c}_n(v) = \pi\delta_{mn}, \tag{W.6.8}$$

$$\int_0^{2\pi} dv \; \mathrm{s}_m(v) \; \mathrm{s}_n(v) = \pi\delta_{mn}, \tag{W.6.9}$$

$$\int_0^{2\pi} dv \; \mathrm{c}_m(v) \; \mathrm{s}_n(v) = 0. \tag{W.6.10}$$

Also, in analogy with (17.4.70) and (7.4.71), let us define functions $\mathrm{C}_n(u)$ and $\mathrm{S}_n(u)$ by the rules

$$\mathrm{C}_n(u) = \mathrm{c}_n(iu), \tag{W.6.11}$$

$$\mathrm{S}_n(u) = -i\mathrm{s}_n(iu). \tag{W.6.12}$$

In view of (6.4) through (6.7) we have the results

$$\mathrm{C}_0(u) = 1/\sqrt{2}, \tag{W.6.13}$$

$$\mathrm{C}_n(u) = \cosh(nu) \text{ for } n \geq 1, \tag{W.6.14}$$

$$\mathrm{S}_0(u) = 0, \tag{W.6.15}$$

$$\mathrm{S}_n(u) = \sinh(nu) \text{ for } n \geq 1. \tag{W.6.16}$$

Evidently the functions $\mathrm{C}_n(u)$ and $\mathrm{S}_n(u)$ are entire functions of u, and the functions $\mathrm{c}_n(v)$ and $\mathrm{s}_n(v)$ are entire functions of v. However, they are not entire functions of x and y because of the singularities described in Exercise 17.4.2.

It is easily verified that functions $\psi_n^c(u,v)$ and $\psi_n^s(u,v)$ of the form

$$\psi_n^c(u,v) \propto \mathrm{C}_n(u)\mathrm{c}_n(v), \tag{W.6.17}$$

$$\psi_n^s(u,v) \propto \mathrm{S}_n(u)\mathrm{s}_n(v) \tag{W.6.18}$$

satisfy (6.3) and are therefore harmonic functions. We claim that they are also polynomial, and therefore entire analytic, functions of x and y. For example, there are the relations

$$\mathrm{C}_0(u)\mathrm{c}_0(v) = 1/2, \tag{W.6.19}$$

$$\mathrm{S}_0(u)\mathrm{s}_0(v) = 0, \tag{W.6.20}$$

$$\mathrm{C}_1(u)\mathrm{c}_1(v) = \cosh(u)\cos(v) = x/f, \tag{W.6.21}$$

$$\mathrm{S}_1(u)\mathrm{s}_1(v) = \sinh(u)\sin(v) = y/f, \tag{W.6.22}$$

$$\begin{aligned}
\mathrm{C}_2(u)\mathrm{c}_2(v) &= \cosh(2u)\cos(2v) = (1/2)\cosh(2u)\cos(2v) + (1/2)\cosh(2u)\cos(2v) \\
&= (1/2)[2\cosh^2(u) - 1][2\cos^2(v) - 1] \\
&\quad -(1/2)[2\sinh^2(u) + 1][2\sin^2(v) - 1] \\
&= (1/2)\{4\cosh^2(u)\cos^2(v) - 2[\cosh^2(u) + \cos^2(v)] + 1\} \\
&\quad -(1/2)\{4\sinh^2(u)\sin^2(v) - 2[\sinh^2(u) - \sin^2(v)] - 1\} \\
&= 2\cosh^2(u)\cos^2(v) - 2\sinh^2(u)\sin^2(v) \\
&\quad - \cosh^2(u) + \sinh^2(u) - \cos^2(v) - \sin^2(v) + 1/2 + 1/2 \\
&= 2\cosh^2(u)\cos^2(v) - 2\sinh^2(u)\sin^2(v) - 1 \\
&= 2(x^2 - y^2)/f^2 - 1.
\end{aligned} \tag{W.6.23}$$

$$\begin{aligned}
\mathrm{S}_2(u)\mathrm{s}_2(v) &= \sinh(2u)\sin(2v) = 4\sinh(u)\cosh(u)\sin(v)\cos(v) \\
&= 4xy/f^2.
\end{aligned} \tag{W.6.24}$$

For a general proof for all n, see Exercise 6.3.

With the above background in mind, consider the ellipse $u = U$ and suppose a potential $\psi_U(v)$ is specified on this ellipse. Since the functions $\mathrm{c}_n(v)$ and $\mathrm{s}_n(v)$ form a complete set, we may make the expansion

$$\psi_U(v) = \sum_{n=0}^{\infty} \tilde{\psi}_U^c(n)\mathrm{c}_n(v) + \sum_{n=1}^{\infty} \tilde{\psi}_U^s(n)\mathrm{s}_n(v) \tag{W.6.25}$$

with

$$\tilde{\psi}_U^c(n) = (1/\pi)\int_0^{2\pi} dv\ \mathrm{c}_n(v)\psi_U(v), \tag{W.6.26}$$

$$\tilde{\psi}_U^s(n) = (1/\pi)\int_0^{2\pi} dv\ \mathrm{s}_n(v)\psi_U(v). \tag{W.6.27}$$

Now make the Ansatz

$$\psi(u,v) = \sum_{n=0}^{\infty} \tilde{\psi}_U^c(n)[\mathrm{C}_n(u)/\mathrm{C}_n(U)]\mathrm{c}_n(v) + \sum_{n=1}^{\infty} \tilde{\psi}_U^s(n)[\mathrm{S}_n(u)/\mathrm{S}_n(U)]\mathrm{s}_n(v). \tag{W.6.28}$$

By construction $\psi(u,v)$ is a harmonic function, and also has the property

$$\psi(U,v) = \psi_U(v). \tag{W.6.29}$$

We have found the solution to Laplace's equation in the ellipse having boundary $u = U$ and the boundary value $\psi_U(v)$.

Note that the operation (6.28) is smoothing for $u < U$. Indeed, according (6.14) and (6.16), there are the asymptotic results

$$\mathrm{C}_n(u) \propto \exp(nu) \text{ as } n \to \infty, \tag{W.6.30}$$

$$\mathrm{S}_n(u) \propto \exp(nu) \text{ as } n \to \infty. \tag{W.6.31}$$

It follows that there are the asymptotic results

$$[\mathrm{C}_n(u)/\mathrm{C}_n(U)] \propto \exp[-n(U-u)] \text{ as } n \to \infty, \tag{W.6.32}$$

$$[\mathrm{S}_n(u)/\mathrm{S}_n(U)] \propto \exp[-n(U-u)] \text{ as } n \to \infty. \tag{W.6.33}$$

Consequently, there is exponential smoothing provided $u < U$.

We also observe, in passing, two facts. First, suppose ψ_U, now to be called ψ_U^d, is a *constant* function with value d,

$$\psi_U^d(v) = d. \tag{W.6.34}$$

Then, by (6.26) and (6.27),

$$\tilde{\psi}_U^{dc}(n) = d\sqrt{2}\delta_{n,0}, \tag{W.6.35}$$

$$\tilde{\psi}_U^{ds}(n) = 0. \tag{W.6.36}$$

It follows from (6.28) that there is the relation

$$\psi^d(u,v) = d. \tag{W.6.37}$$

As expected, if ψ is constant on the boundary $u = U$, it will have the same constant value inside the ellipse. Second, for any solution, we find from (6.28) that there is the relation

$$\int_0^{2\pi} dv\ \psi(u,v) = (\pi\sqrt{2})\sum_{n=0}^{\infty}\tilde{\psi}_U^c(n)[\mathrm{C}_n(u)/\mathrm{C}_n(U)]\delta_{n,0} = (\pi\sqrt{2})\tilde{\psi}_U^c(0) = \int_0^{2\pi} dv\ \psi_U(v). \tag{W.6.38}$$

That is, the dv integral of $\psi(u,v)$ over any ellipse of constant u is independent of u.

To further study smoothing in the case of an ellipse in two space, suppose ψ_U, now to be called ψ_U^δ, is a delta function centered on $v = 0$,

$$\psi_U^\delta(v) = \delta(v). \tag{W.6.39}$$

Then, by (6.26) and (6.27),

$$\tilde{\psi}_U^{\delta c}(0) = 1/(\pi\sqrt{2}), \tag{W.6.40}$$

$$\tilde{\psi}_U^{\delta c}(n) = 1/\pi \text{ for } n \geq 1, \tag{W.6.41}$$

$$\tilde{\psi}_U^{\delta s}(n) = 0, \tag{W.6.42}$$

and (6.28) takes the form

$$\begin{aligned}
\psi^\delta(u,v) &= \sum_{n=0}^{\infty}\tilde{\psi}_U^{\delta c}(n)[1/\mathrm{C}_n(U)]\mathrm{C}_n(u)\mathrm{c}_n(v) \\
&= [1/(\pi\sqrt{2})][1/\mathrm{C}_0(U)]\mathrm{C}_0(u)\mathrm{c}_0(v) + (1/\pi)\sum_{n=1}^{\infty}[1/\mathrm{C}_n(U)]\mathrm{C}_n(u)\mathrm{c}_n(v). \\
&= [1/(2\pi)] + (1/\pi)\sum_{n=1}^{\infty}[1/\cosh(nU)]\cosh(nu)\cos(nv).
\end{aligned} \tag{W.6.43}$$

Figure 6.1 shows $\psi^{\delta}(u,v)$ as a function of v for various values of u. For this example $U = 0.5$. We see that the delta function spike about $v = 0$ on the ellipse $u = U$ becomes an ever lower and broader bump with decreasing u. Thus the effect of a disturbance in the potential on the $u = U$ ellipse "decays" away as one moves to ellipses with successively smaller values of u. Finally we note that, in agreement with (6.38), the "decay" we have been observing is entirely due to spreading. Indeed, the relations (6.34) and (6.37) illustrate that if the initial/boundary potential distribution is completely "spread out", i.e. constant, then no decay occurs.

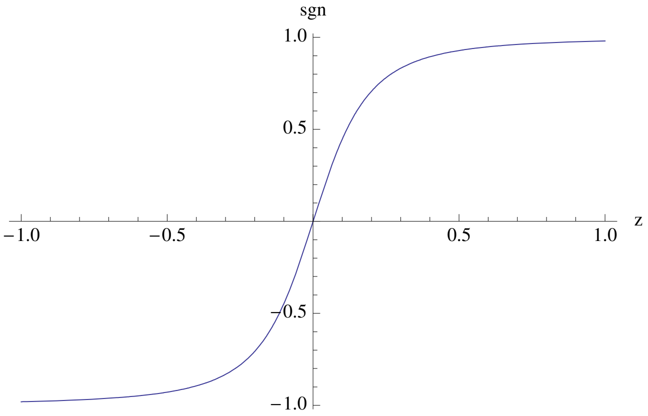


Figure W.6.1: (Place Holder) The function $\psi^{\delta}(u,v)$ as a function of v for various values of u when $U = 0.5$ and therefore $\tanh(U) = 0.46\cdots$.

For future use let us examine the behavior of $\psi^{\delta}(u,v)$ about the origin. Evidently there is the expansion

$$\begin{aligned}
\psi^{\delta}(u,v) &= \sum_{n=0}^{2} \tilde{\psi}_U^{\delta c}(n)[1/\mathrm{C}_n(U)]\mathrm{C}_n(u)\mathrm{c}_n(v) + \cdots \\
&= [1/(\pi\sqrt{2})][1/\mathrm{C}_0(U)]\mathrm{C}_0(u)\mathrm{c}_0(v) + (1/\pi)[1/\mathrm{C}_1(U)]\mathrm{C}_1(u)\mathrm{c}_1(v) \\
&\quad +(1/\pi)[1/\mathrm{C}_2(U)]\mathrm{C}_2(u)\mathrm{c}_2(v) + \cdots \\
&= [1/(\pi\sqrt{2})][1/\sqrt{2}] + (1/\pi)[1/\mathrm{C}_1(U)](x/f) \\
&\quad +(1/\pi)[1/\mathrm{C}_2(U)][2(x^2-y^2)/f^2 - 1] + \cdots . \\
&= 1/(2\pi) + (1/\pi)[1/\cosh(U)](x/f) \\
&\quad +(1/\pi)[1/\cosh(2U)][2(x^2-y^2)/f^2 - 1] + \cdots . \qquad \text{(W.6.44)}
\end{aligned}$$

Next, suppose ψ_U, now to be called ψ_U^{Δ}, is a delta function centered on $v = \pi/2$,

$$\psi_U^{\Delta}(v) = \delta(v - \pi/2). \qquad \text{(W.6.45)}$$

Then, by (6.26) and (6.27), the first few Fourier coefficients results are

$$\tilde{\psi}_U^{\Delta c}(0) = 1/(\pi\sqrt{2}), \qquad \text{(W.6.46)}$$

$$\tilde{\psi}_U^{\Delta c}(1) = 0, \qquad \text{(W.6.47)}$$

$$\tilde{\psi}_U^{\Delta c}(2) = -1/\pi; \tag{W.6.48}$$

$$\tilde{\psi}_U^{\Delta s}(0) = 0, \tag{W.6.49}$$

$$\tilde{\psi}_U^{\Delta s}(1) = 1/\pi, \tag{W.6.50}$$

$$\tilde{\psi}_U^{\Delta s}(2) = 0. \tag{W.6.51}$$

Correspondingly, (6.28) now takes the form

$$\begin{aligned}
\psi^{\Delta}(u,v) &= \sum_{n=0}^{2} \tilde{\psi}_U^{\Delta c}(n)[\mathrm{C}_n(u)/\mathrm{C}_n(U)]\mathrm{c}_n(v) + \cdots \\
&\quad + \sum_{n=1}^{2} \tilde{\psi}_U^{\Delta s}(n)[\mathrm{S}_n(u)/\mathrm{S}_n(U)]\mathrm{s}_n(v) + \cdots \\
&= [1/(\pi\sqrt{2})][1/\mathrm{C}_0(U)]\mathrm{C}_0(u)\mathrm{c}_0(v) - (1/\pi)[1/\mathrm{C}_2(U)]\mathrm{C}_2(u)\mathrm{c}_2(v) + \cdots \\
&\quad + (1/\pi)[1/\mathrm{S}_1(U)]\mathrm{S}_1(u)\mathrm{s}_1(v) + \cdots \\
&= 1/(2\pi) + (1/\pi)[1/\sinh(U)](y/f) \\
&\quad - (1/\pi)[1/\cosh(2U)][2(x^2-y^2)/f^2 - 1] + \cdots .
\end{aligned} \tag{W.6.52}$$

Let us compare the terms in ψ^{δ} given by (6.44) with the terms in ψ^{Δ} given by (6.52). In particular, let us begin by making the comparison

$$[1/\cosh(U)](x/f) \text{ versus } [1/\sinh(U)](y/f). \tag{W.6.53}$$

For ψ^{δ} the delta function disturbance in the boundary potential is made at the point $(x,y) = (x^{\delta},0)$ with

$$x^{\delta} = f\cosh(U). \tag{W.6.54}$$

See (17.4.1) and Figure 17.4.2. And for ψ^{Δ} the delta function disturbance in the boundary potential is made at the point $(x,y) = (0,y^{\Delta})$ with

$$y^{\Delta} = f\sinh(U). \tag{W.6.55}$$

See (17.4.2). With this observation in mind, we see that the comparison (6.53) can be rewritten in the form

$$x/x^{\delta} \text{ versus } y/y^{\Delta}. \tag{W.6.56}$$

Now suppose the bounding ellipse $u = U$ has been chosen so that

$$y^{\Delta} < x^{\delta}. \tag{W.6.57}$$

See Figure 17.4.3. Then, according to (6.56), near the origin the effect of a disturbance at $(x^{\delta},0)$ is diminished from the effect of a disturbance at $(0,y^{\Delta})$ by a factor of

$$y^{\Delta}/x^{\delta} = \tanh(U). \tag{W.6.58}$$

By contrast, comparison of (4.42) and (4.45) shows, as expected, there is no such effect in the case of a circular boundary. Our findings are in accord with the expectation described in Section 17.4.1 to the effect that, for wigglers or dipoles with small gaps and wide pole faces, use of a cylinder with elliptical cross section should give improved error insensitivity.

Let us also examine the next higher-order (and non constant) terms in ψ^{δ} and ψ^{Δ} which, according to (6.44) and (6.52), are

$$\pm (1/\pi)[1/\cosh(2U)][2(x^2-y^2)/f^2]. \tag{W.6.59}$$

Note that there is the relation

$$f^2\cosh(2U) = f^2[\cosh^2(U)+\sinh^2(U)] = (x^{\delta})^2 + (y^{\Delta})^2. \tag{W.6.60}$$

Thus, (6.59) can also be written in the form

$$\pm (1/\pi)(x^2-y^2)/\{(1/2)[(x^{\delta})^2+(y^{\Delta})^2]\}. \tag{W.6.61}$$

The comparable term for the circle in two-space case, that given in (4.42) or (4.49), is

$$\pm (1/\pi)(x^2-y^2)/R^2. \tag{W.6.62}$$

Thus, to contrast the use of a circle with the use of an ellipse, we should make the comparison

$$R^2 \text{ versus } \{(1/2)[(x^{\delta})^2+(y^{\Delta})^2]\}. \tag{W.6.63}$$

Moreover, when contrasting the use of a circle to the use of an ellipse, it is reasonable to presume that the ellipse just contains the circle so that

$$y^{\Delta} = R. \tag{W.6.64}$$

In this case (6.63) becomes

$$(1/2)(y^{\Delta})^2 \text{ versus } (1/2)(x^{\delta})^2. \tag{W.6.65}$$

In view of (6.57) the left term in the comparison (6.65) is smaller than the term on the right. Correspondingly, the denominator in (6.61) is larger than that in (6.62) thereby again illustrating that the use of a cylinder with elliptical cross section should give improved error insensitivity.

Exercises

W.6.1. Verify that the functions (6.17) and (6.18) are harmonic.

W.6.2. Since deriving the result (6.23) involved considerable algebra, it is useful to check a few specific cases. Consider the points $(x,y) = (0,0)$ and $(x,y) = (\pm f, 0)$ for which $(u,v) = (0, \pi/2 \text{ or } 3\pi/2)$ and $(u,v) = (0, 0 \text{ or } \pi)$. See Figure 17.4.3. Verify that (6.23) holds at these points.

W.6.3. The purpose of this exercise is to prove the claim that, for all n, functions ψ_n^c and ψ_n^s of the form (6.17) and (6.18) are polynomial functions of x and y. Recall (17.4.7). Show that from this relation it follows that

$$(x+iy)^n/f^n = [\cosh(w)]^n. \tag{W.6.66}$$

Next verify the expansion

$$\begin{aligned}[\cosh(w)]^n &= (1/2^n)\{\exp(w)+\exp(-w)\}^n \\ &= (1/2^n)\{\exp(nw)+n\exp[(n-2)w]+[n(n-1)/2!]\exp[(n-4)w] \\ &\quad +\cdots+[n(n-1)/2!]\exp[-(n-4)w]+n\exp[-(n-2)w]+\exp(-nw)\}.\end{aligned} \tag{W.6.67}$$

Show that the terms in this expansion can be combined to yield the result

$$[\cosh(w)]^n = (1/2^{n-1})\{\cosh(nw)+n\cosh[(n-2)w]+[n(n-1)/2!]\cosh[(n-4)w]+\cdots\}. \tag{W.6.68}$$

Verify that the last term on the right side of (6.68) is

$$(1/2)^n \begin{pmatrix} n \\ n/2 \end{pmatrix} \text{ if } n \text{ is even,} \tag{W.6.69}$$

and is

$$(1/2)^{n-1} \begin{pmatrix} n \\ (n-1)/2 \end{pmatrix} \cosh(w) \text{ if } n \text{ is odd.} \tag{W.6.70}$$

Next verify that

$$\begin{aligned}\cosh(mw) &= \cosh(mu+imv) = \cosh(mu)\cosh(imv)+\sinh(mu)\sinh(imv) \\ &= \cosh(mu)\cos(mv)+i\sinh(mu)\sin(mv),\end{aligned} \tag{W.6.71}$$

from which it follows that

$$\cosh(mu)\cos(mv) = \Re[\cosh(mw)], \tag{W.6.72}$$

$$\sinh(mu) sin(mv) = \Im[\cosh(mw)]. \tag{W.6.73}$$

Using the results found so far, take real and imaginary parts to rewrite (6.66) in the form

$$(1/2^{n-1})\cosh(nu)\cos(nv) = \Re[(x+iy)^n/f^n]-(1/2^{n-1})\Re\{n\cosh[(n-2)w]+\cdots\}, \tag{W.6.74}$$

$$(1/2^{n-1})\sinh(nu)\sin(nv) = \Im[(x+iy)^n/f^n]-(1/2^{n-1})\Im\{n\cosh[(n-2)w]+\cdots\}. \tag{W.6.75}$$

Conclude that $\cosh(nu)\cos(nv)$ is a polynomial in x and y provided the same is true of $\cosh(mu)\cos(mv)$ for $m = n-2, n-4, \cdots$. Make a similar conclusion for $\sinh(nu)\sin(nv)$. Finally prove by induction, starting with (6.19) through (6.22), the claim stated at the beginning of this exercise.

W.6.4. Use some of the machinery of Exercise 6.3 above to produce an easy derivation of the relations (6.23) and (6.24).

W.7 The Elliptical Cylinder in Three Space

W.8 The Rectangle in Two Space

W.9 The Rectangular Cylinder in Three Space

W.10 The Sphere in Three Space

Higher angular modes are suppressed by the exponential factor $(r/R)^\ell = \exp[\ell \log(r/R)]$. Note that $\log(r/R) < 0$.

W.11 The Ellipsoid in Three Space

Bibliography

[1] F. Olver, D. Lozier, R. Boisvert, and C. Clark, Editors, *NIST Handbook of Mathematical Functions*, Cambridge (2010). For properties of Bessel functions, see Chapter 10 at the Web site http://dlmf.nist.gov/.

Appendix X

Lie Algebraic Theory of Light Optics

Overview

Need text here.

X.1 Hamiltonian Formulation

Consider the optical system illustrated schematically in Figure 1.1. A ray originates at the general *initial* point P^i with spatial coordinate $\boldsymbol{r}^i$ and moves in an initial direction specified by the unit vector $\hat{\boldsymbol{s}}^i$. After passing through an optical device it arrives at the *final* point P^f with with spatial coordinate $\boldsymbol{r}^f$ and moves in a final direction specified by the unit vector $\hat{\boldsymbol{s}}^f$. Given the initial quantities $(\boldsymbol{r}^i, \hat{\boldsymbol{s}}^i)$, the fundamental problem of geometrical optics is to determine the final quantities $(\boldsymbol{r}^f, \hat{\boldsymbol{s}}^f)$ and to design an optical device in such a way that the relation between the initial and final ray quantities has various desired properties.

Suppose the z coordinates of the initial and final points P^i and $P^f$are held fixed. In some instances the planes $z = z^i$ and $z = z^f$ can be viewed as object and image planes, respectively. But in other cases they simply serve as convenient reference planes. Further, suppose the general light ray from P^i to P^f is parameterized using z as an *independent/time-like* variable. That is, the path of a general ray is described by specifying the two functions $x(z)$ and $y(z)$. Then the element of path length ds along a ray is given by the expression

$$ds = [(dz)^2 + (dx)^2 + (dy)^2]^{1/2} = [1 + (x')^2 + (y')^2]^{1/2} dz. \tag{X.1.1}$$

Here a prime denotes the differentiation d/dz. Consequently the optical path length along a ray from P^i to P^f is given by the integral

$$A = \int_{z^i}^{z^f} n(x, y, z)[1 + (x')^2 + (y')^2]^{1/2} dz. \tag{X.1.2}$$

Here the function $n(x, y, z) = n(\boldsymbol{r})$ specifies the index of refraction at each point before and after the optical device and in the device itself.

Fermat's principle requires that A be an extremum for the path of an actual ray. Therefore the ray path satisfies the Euler-Lagrange equations

$$d/dz(\partial L/\partial x') - \partial L/\partial x = 0, \tag{X.1.3}$$

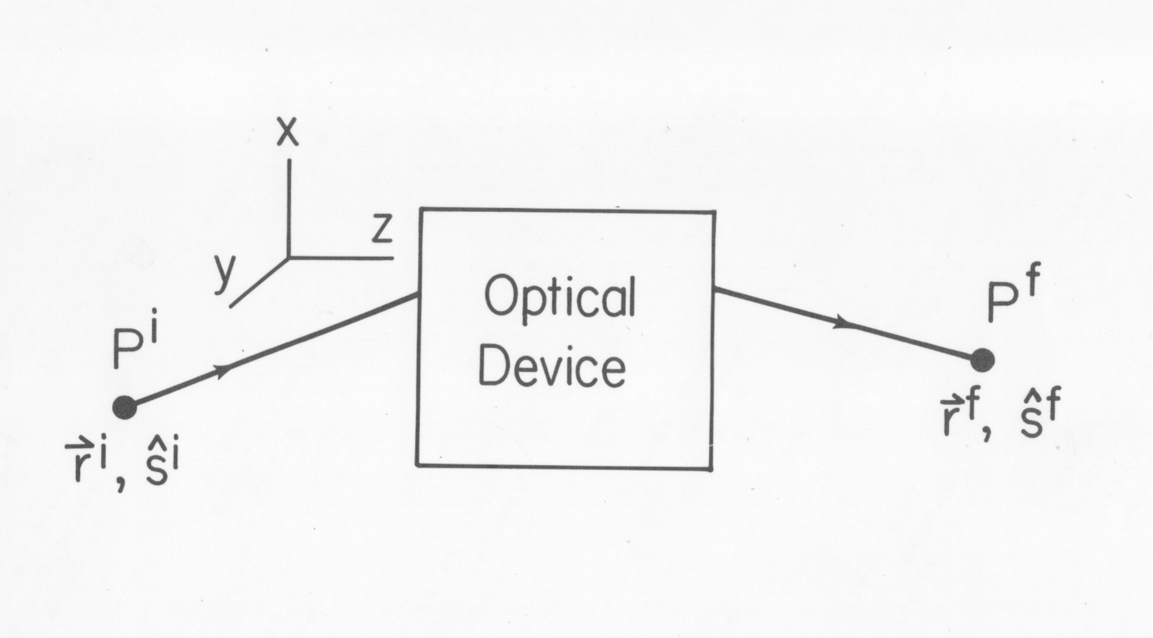


Figure X.1.1: Optical system consisting of an optical device preceded and followed by simple transit. A ray originates at P^i with *initial* location $\boldsymbol{r}^i$ and *initial* direction $\hat{\boldsymbol{s}}^i$, and terminates at P^f with *final* location $\boldsymbol{r}^f$ and *final* direction $\hat{\boldsymbol{s}}^f$

.

$$d/dz(\partial L/\partial y') - \partial L/\partial y = 0, \tag{X.1.4}$$

with a Lagrangian L given by the expression

$$L = n(x, y, z)[1 + (x')^2 + (y')^2]^{1/2}. \tag{X.1.5}$$

To proceed further, it is useful to pass from a Lagrangian formulation to a Hamiltonian formulation. Introduce two momenta p_x and p_y conjugate to the coordinates x and y by the rule

$$p_x = \partial L/\partial x', \tag{X.1.6}$$

$$p_y = \partial L/\partial y', \tag{X.1.7}$$

with the explicit results that

$$p_x = n(\boldsymbol{r})x'/[1 + (x')^2 + (y')^2]^{1/2}, \tag{X.1.8}$$

$$p_y = n(\boldsymbol{r})y'/[1 + (x')^2 + (y')^2]^{1/2}. \tag{X.1.9}$$

The Hamiltonian H is defined in terms of the Lagrangian L by the Legendre transformation

$$H(x, y, p_x, p_y; z) = p_x x' + p_y y' - L. \tag{X.1.10}$$

It follows from (1.5) through (1.10) that in our case H is given by the relation

$$H = -[n^2(\boldsymbol{r}) - p_x^2 - p_y^2]^{1/2}. \tag{X.1.11}$$

Let $\boldsymbol{q}$ be a two-component vector with entries $q_x = x$ and $q_y = y$, and let $\boldsymbol{p}$ be a two-component vector with entries p_x and p_y. Evidently, a ray leaving the initial point P^i is characterized by the quantities z^i, $\boldsymbol{q}^i$, and $\boldsymbol{p}^i$. The quantity $\boldsymbol{q}^i$ specifies the initial point of origin on the plane $z = z^i$ and, according to (1.8) and (1.9), $\boldsymbol{p}^i$ describes the initial direction of the ray. Similarly, $\boldsymbol{q}^f$ and $\boldsymbol{p}^f$ characterize the ray as it arrives at the final point P^f in the plane $z = z^f$. Finally, the relation between the initial conditions $\boldsymbol{q}^i$ and $\boldsymbol{p}^i$ and the final conditions $\boldsymbol{q}^f$ and $\boldsymbol{p}^f$ is given by following from $z = z^i$ to $z = z^f$ a trajectory $\boldsymbol{q}(z)$, $\boldsymbol{p}(z)$ governed by the Hamiltonian H.

At this point it is convenient to introduce a four-component vector $\boldsymbol{w}$ with entries $\boldsymbol{q}$, $\boldsymbol{p}$:

$$\boldsymbol{w} = (w_1, w_2, w_3, w_4) = (q_x, p_x, q_y, p_y). \tag{X.1.12}$$

Also, let $\boldsymbol{w}^i$ and $\boldsymbol{w}^f$ denote initial and final values of $\boldsymbol{w}$. The fact that initial conditions *determine* the final conditions can be expressed in terms of a functional relationship or *mapping* $\mathcal{M}$. This relationship can be defined formally by writing the expression

$$\boldsymbol{w}^f = \mathcal{M}\boldsymbol{w}^i. \tag{X.1.13}$$

Hamilton's equations of motion for the canonical variables $\boldsymbol{q}$ and $\boldsymbol{p}$ read

$$q'_\alpha = \partial H/\partial p_\alpha =: -H : q_\alpha, \tag{X.1.14}$$

$$p'_\alpha = -\partial H/\partial q_\alpha =: -H : p_\alpha. \tag{X.1.15}$$

Correspondingly, there is an equation of motion for $\mathcal{M}$ given by the relation

$$\mathcal{M}' = \mathcal{M} : -H : \tag{X.1.16}$$

with the initial condition

$$\mathcal{M}|_{z=z^i} = \mathcal{I} \tag{X.1.17}$$

where $\mathcal{I}$ is the identity map. Recall Subsection 10.1.1.

As has been seen, Fermat's principle is equivalent to the statement that the initial conditions $\boldsymbol{w}^i$ and the final conditions $\boldsymbol{w}^f$ are related by following a trajectory governed by a Hamiltonian, namely the Hamiltonian (1.11). From the work of Subsection 6.4.1 this statement is equivalent in turn to the statement that $\mathcal{M}$ is a *symplectic* map.

Let us recapitulate briefly for the light optics context some of what we have learned about symplectic maps. Let M be the Jacobian matrix associated with the mapping $\mathcal{M}$. It is defined by the relation

$$M_{\alpha\beta} = \partial w^f_\alpha/\partial w^i_\beta, \tag{X.1.18}$$

and describes how small changes in $\boldsymbol{w}^i$ produce small changes in $\boldsymbol{w}^f$. Also, let J be the four-by-four matrix defined by the equation

$$J = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}. \tag{X.1.19}$$

Then from earlier work we know that M satisfies the matrix equation

$$M^T J M = J. \tag{X.1.20}$$

Equation (1.20) is the condition that M be a *symplectic* matrix (and in the context of optics is sometimes called the *lens equation*). Correspondingly, as described earlier, a map $\mathcal{M}$ whose Jacobian matrix M is symplectic is said to be a *symplectic map.* Note that, as indicated by the notation $M(\boldsymbol{w}^i; z^i, z^f)$, the matrix M depends in general on the variables $\boldsymbol{w}^i$, z^i, and z^f. Observe, however, that the right side of (1.20), namely the matrix J given by (1.19), does not depend on these variables and is in fact a constant matrix. The requirement that (1.20) holds for all values of $\boldsymbol{w}^i$, z^i, and z^f places strong restrictions on the nature of symplectic maps. These restrictions were first studied by Hamilton (in the context of light optics!) and led to the introduction/invention of *characteristic/generating* functions to describe and manage symplectic maps. In this appendix we will see, in the context of light optics, how Lie methods can also be used for this purpose.

Exercises

X.1.1. Verify that H as given by (1.11) is indeed the Hamiltonian associated with the Lagrangian L given by (1.5).

X.1.2. Recall Liouville's theorem. See Subsection 6.8.1. Google the word *etendue.* Work out the consequences of Liouville's theorem when applied to the case of light optics.

X.2 Assumption of Axial Symmetry and Lie-algebraic Consequences

Although framed in the context of light optics, the discussion so far is applicable to general Hamiltonian systems having a four-dimensional phase space. We are dealing with symplectic maps $\mathcal{M}$ whose linear parts M about any trajectory/ray are elements of $Sp(4, \mathbb{R})$. We now turn to the specific case of the optical Hamiltonian (1.11). Moreover, at this point we also assume that the optical device has *axial/rotational symmetry* about the z axis. Introduce the definitions

$$q^2 = (q_x)^2 + (q_y)^2 = \boldsymbol{q} \cdot \boldsymbol{q}, \tag{X.2.1}$$

$$p^2 = (p_x)^2 + (p_y)^2 = \boldsymbol{p} \cdot \boldsymbol{p}, \tag{X.2.2}$$

$$\boldsymbol{p} \cdot \boldsymbol{q} = p_x q_x + p_y q_y. \tag{X.2.3}$$

(Note that this notation can be misleading since, for example, q^2 is not the square of any Cartesian coordinate.) From (1.11) we see that the optical Hamiltonian depends on $\boldsymbol{p}$ only through the quantity p^2. To enforce axial symmetry, we assume that $n(\boldsymbol{r})$ is of the functional form

$$n(\boldsymbol{r}) = \hat{n}(q^2, z) \tag{X.2.4}$$

so that the index of refraction also has axial symmetry. Now imagine that H as given by (1.11) and the assumption (2.4) is expanded in a power series in the components of $\boldsymbol{q}$ and $\boldsymbol{p}$. By the assumption of axial symmetry, such an expansion must be of the form

$$H = H_0 + H_2 + H_4 + H_6 + \cdots \tag{X.2.5}$$

where the H_n are homogeneous polynomials of degree n in the components of $\boldsymbol{q}$ and $\boldsymbol{p}$. That is, only even powers can occur. Indeed, the H_n depend only on powers of q^2 and p^2 and products of these powers. Since only even powers of the components of $\boldsymbol{w}$ can occur in H, it follows that the z axis, $\boldsymbol{w}(z) = 0$, is a trajectory/ray for the equations of motion generated by H, and the phase-space origin is a fixed point of $\mathcal{M}$, $\mathcal{M}0 = 0$.

Let L_z denote the second degree homogeneous polynomial

$$L_z = (\boldsymbol{q} \times \boldsymbol{p}) \cdot \boldsymbol{e}_z = q_x p_y - q_y p_x \tag{X.2.6}$$

and let $\mathcal{L}_z$ be the associated Lie operator

$$\mathcal{L}_z =: L_z : . \tag{X.2.7}$$

Then, as a Lie-algebraic expression of the condition of axial symmetry for the quantities q^2, p^2, and $(\boldsymbol{p} \cdot \boldsymbol{q})$, we have the relations

$$\mathcal{L}_z q^2 = \mathcal{L}_z p^2 = \mathcal{L}_z(\boldsymbol{p} \cdot \boldsymbol{q}) = 0. \tag{X.2.8}$$

And, as a Lie-algebraic expression of the condition of axial symmetry for H, we have the relations

$$\mathcal{L}_z H_n = 0 \tag{X.2.9}$$

and

$$\mathcal{L}_z H = 0. \tag{X.2.10}$$

We conclude from (2.10) that L_z is an integral of motion,

$$\mathcal{L}_z H =: L_z : H = [L_z, H] = 0 \tag{X.2.11}$$

from which it follows that there is the relation

$$L_z^f = L_z^i \tag{X.2.12}$$

which takes the explicit form

$$q_x^f p_y^f - q_y^f p_x^f = q_x^i p_y^i - q_y^i p_x^i. \tag{X.2.13}$$

We note, for future use, that there is the relation

$$(L_z)^2 = (q_x p_y - q_y p_x)^2 = q_x^2 p_y^2 - 2 q_x p_x q_y p_y + p_x^2 q_y^2 = p^2 q^2 - (\boldsymbol{p} \cdot \boldsymbol{q})^2, \tag{X.2.14}$$

and the relation

$$\begin{aligned} (\mathcal{L}_z)^\dagger &= : L_z :^\dagger = (: q_x p_y - q_y p_x :)^\dagger =: q_x p_y :^\dagger - : q_y p_x :^\dagger \\ &= : q_y p_x : - : q_x p_y := -\mathcal{L}_z. \end{aligned} \tag{X.2.15}$$

For the steps made in obtaining the latter relation, recall (7.3.16) through (7.3.18).

We also pause to make the definitions

$$\mathcal{L}_+ =: -p^2/2 :, \tag{X.2.16}$$

$$\mathcal{L}_0 =: (\boldsymbol{p}\cdot\boldsymbol{q})/2 :, \tag{X.2.17}$$

$$\mathcal{L}_- =: q^2/2 : . \tag{X.2.18}$$

From (7.3.16) through (7.3.18) we see that

$$(\mathcal{L}_+)^\dagger =: -p_x^2/2 - p_y^2/2 :^\dagger =: q_x^2/2 + q_y^2/2 := \mathcal{L}_-, \tag{X.2.19}$$

$$(\mathcal{L}_0)^\dagger =: (\boldsymbol{p}\cdot\boldsymbol{q})/2 :^\dagger =: p_x q_x + p_y q_y :^\dagger =: p_x q_x + p_y q_y := \mathcal{L}_0, \tag{X.2.20}$$

$$(\mathcal{L}_-)^\dagger =: q_x^2/2 + q_y^2/2 :^\dagger =: -p_x^2/2 - p_y^2/2 := \mathcal{L}_+. \tag{X.2.21}$$

The Lie operators (2.16) through (2.18) obey the commutation rules

$$\{\mathcal{L}_+, \mathcal{L}_-\} = 2\mathcal{L}_0, \tag{X.2.22}$$

$$\{\mathcal{L}_0, \mathcal{L}_+\} = \mathcal{L}_+, \tag{X.2.23}$$

$$\{\mathcal{L}_0, \mathcal{L}_-\} = -\mathcal{L}_-. \tag{X.2.24}$$

According to Exercise 27.5.5 these commutation rules are a variant of the commutation rules for $sp(2,\mathbb{R})$. (Only some labeling and normalizations have been changed.) Therefore we will be able to work with a particular $Sp(2,\mathbb{R})$ subgroup of $Sp(4,\mathbb{R})$, thereby simplifying/organizing many computations and results.

The rules (2.22) through (2.24) are also the commutation rules for $su(2)$ in its raising and lowering operator form. [Recall that $su(2)$ and $sp(2,\mathbb{R})$ are equivalent over the complex field, and therefore a relation of this form between them should not be a surprise. See Subsection 3.7.6.] Finally note, as a consequence of (5.3.14) and (2.8), that $\mathcal{L}_z$ commutes with $\mathcal{L}_\pm$ and $\mathcal{L}_0$,

$$\{\mathcal{L}_z, \mathcal{L}_\pm\} = \{\mathcal{L}_z, \mathcal{L}_0\} = 0. \tag{X.2.25}$$

Next we invoke a Lie algebraic fact: It can be shown that the solution of (1.16) involves only the ingredients of H and quantities that can be formed by taking Poisson brackets of the ingredients of H. See Chapter 10. It follows that, under the assumption of axial symmetry, the solution to (1.16) must be of the form

$$\begin{aligned} \mathcal{M} &= \exp(: f_2^c :)\exp(: f_2^a :)\exp(: f_4 :)\exp(: f_6 :)\exp(: f_8 :)\cdots \\ &= \mathcal{R}\exp(: f_4 :)\exp(: f_6 :)\exp(: f_8 :)\cdots . \end{aligned} \tag{X.2.26}$$

That is, only the f_n with *even* n can occur in the factored product representation of $\mathcal{M}$. Moreover, all the f_n must satisfy the axial symmetry (rotational invariance) relation

$$\mathcal{L}_z f_n = 0. \tag{X.2.27}$$

Also, there is the Poisson bracket relation

$$[q^2, p^2] = 4\boldsymbol{q}\cdot\boldsymbol{p} = 4\boldsymbol{p}\cdot\boldsymbol{q}. \tag{X.2.28}$$

Therefore f_2^c and f_2^a, and hence $\mathcal{R}$, can depend only on the quantities q^2, $\boldsymbol{p}\cdot\boldsymbol{q}$, and p^2. (We note, as can be easily verified, that p^2-q^2 and $\boldsymbol{p}\cdot\boldsymbol{q}$ are f_2^a polynomials, and p^2+q^2 is an f_2^c polynomial.) Similarly, f_4 can depend only on the six quantities $(p^2)^2$, $p^2(\boldsymbol{p}\cdot\boldsymbol{q})$, $(\boldsymbol{p}\cdot\boldsymbol{q})^2$, p^2q^2, $(\boldsymbol{p}\cdot\boldsymbol{q})q^2$, and $(q^2)^2$. That is, under the assumption of axial symmetry, we may write

$$f_4 = A(p^2)^2 + Bp^2(\boldsymbol{p}\cdot\boldsymbol{q}) + C(\boldsymbol{p}\cdot\boldsymbol{q})^2 + Dp^2q^2 + E(\boldsymbol{p}\cdot\boldsymbol{q})q^2 + F(q^2)^2 \tag{X.2.29}$$

where the coefficients A through F are to be determined. Note that all these ingredients of f_4 are rotationally invariant [as required by (2.27)] and are powers and products of powers of the ingredients for f_2. Similarly, the ingredients of all the f_n for $n \geq 4$ are axially symmetric and are powers and products of powers of the ingredients for f_2. Finally we remark it can be demonstrated that (for an imaging system) the polynomials in (2.29), shown multiplied by the coefficients A through E, are related to the Seidel aberrations of spherical aberration, coma, astigmatism, curvature of field, and distortion, respectively:[1]

$$A(p^2)^2 \leftrightarrow \text{spherical aberration,} \tag{X.2.30}$$

$$Bp^2(\boldsymbol{p}\cdot\boldsymbol{q}) \leftrightarrow \text{coma,} \tag{X.2.31}$$

$$C(\boldsymbol{p}\cdot\boldsymbol{q})^2 \leftrightarrow \text{astigmatism,} \tag{X.2.32}$$

$$Dp^2q^2 \leftrightarrow \text{curvature of field,} \tag{X.2.33}$$

$$E(\boldsymbol{p}\cdot\boldsymbol{q})q^2 \leftrightarrow \text{distortion,} \tag{X.2.34}$$

$$F(q^2)^2 \leftrightarrow \text{nameless.} \tag{X.2.35}$$

The nameless aberration does not affect the ability of an imaging system to form images, but may be important for other systems.

Suppose we present a general/arbitrary f_4 in terms of monomials by writing

$$f_4 = \sum_{|k|=4} f(k_1,k_2,k_3,k_4)q_x^{k_1}p_x^{k_2}q_y^{k_3}p_y^{k_4} \tag{X.2.36}$$

where the coefficients $f(k_1,k_2,k_3,k_4)$ are arbitrary and we have used the notation

$$|k| = k_1+k_2+k_3+k_4. \tag{X.2.37}$$

Under the assumption of axial symmetry, as exemplified by (2.27), there will be relations among the various $f(k_1,k_2,k_3,k_4)$. For example, we see from (2.2) that there is the monomial expansion

$$(p^2)^2 = [(p_x)^2+(p_y)^2]^2 = (p_x)^4 + 2(p_x)^2(p_y)^2 + (p_y)^4. \tag{X.2.38}$$

Also, we see from (2.29) that only the $A(p^2)^2$ term in (2.29) contributes to monomials that are of degree four in the components of $\boldsymbol{p}$. Therefore comparison of (2.29) and (2.36) gives the relations

$$f(0,4,0,0)=A,\ \ f(0,2,0,2)=2A,\ \ f(0,0,0,4)=A. \tag{X.2.39}$$

[1] Philipp Ludwig von Seidel (1821-1896) classified and described the possible third-order geometric aberrations for axially symmetric optical systems.

In Exercise 2.2 you will have the pleasure of working out the further relations of this kind that are implied by (2.29). From these relations one can extract formulas for the coefficients A through F in terms of the $f(k_1, k_2, k_3, k_4)$ with $|k| = 4$, and vice versa. In particular, you will verify the relations

$$A = f(0, 4, 0, 0), \tag{X.2.40}$$

$$B = f(1, 3, 0, 0), \tag{X.2.41}$$

$$C = (1/2)f(1, 1, 1, 1), \tag{X.2.42}$$

$$D = f(0, 2, 2, 0), \tag{X.2.43}$$

$$E = f(3, 1, 0, 0), \tag{X.2.44}$$

$$F = f(4, 0, 0, 0). \tag{X.2.45}$$

Exercises

X.2.1. Verify (2.14).

X.2.2. Verify that (2.27) follows from (2.9) and the work of Chapter 10.

X.2.3. Consider the ingredients of (2.29). We have already found the monomial decomposition (2.38). Verify that all the ingredients of (2.29) have the monomial decompositions

$$(p^2)^2 = [(p_x)^2 + (p_y)^2]^2 = (p_x)^4 + 2(p_x)^2(p_y)^2 + (p_y)^4, \tag{X.2.46}$$

$$p^2(\boldsymbol{p} \cdot \boldsymbol{q}) = [(p_x)^2 + (p_y)^2](p_x q_x + p_y q_y) = q_x(p_x)^3 + (p_x)^2 q_y p_y + q_x p_x (p_y)^2 + q_y (p_y)^3, \tag{X.2.47}$$

$$(\boldsymbol{p} \cdot \boldsymbol{q})^2 = (p_x q_x + p_y q_y)^2 = (q_x)^2(p_x)^2 + 2q_x p_x q_y p_y + (q_y)^2(p_y)^2, \tag{X.2.48}$$

$$p^2 q^2 = [(p_x)^2+(p_y)^2][(q_x)^2+(q_y)^2] = (q_x)^2(p_x)^2+(p_x)^2(q_y)^2+(q_x)^2(p_y)^2+(q_y)^2(p_y)^2, \tag{X.2.49}$$

$$(\boldsymbol{p} \cdot \boldsymbol{q})q^2 = (p_x q_x + p_y q_y)[(q_x)^2 + (q_y)^2] = (q_x)^3 p_x + q_x p_x (q_y)^2 + q_x (q_y)^2 p_y + (q_y)^3 p_y, \tag{X.2.50}$$

$$(q^2)^2 = [(q_x)^2 + (q_y)^2]^2 = (q_x)^4 + 2(q_x)^2(q_y)^2 + (q_y)^4. \tag{X.2.51}$$

From (2.29), (2.36), and (2.40) we have already found the results

$$f(0, 4, 0, 0) = A,\ f(0, 2, 0, 2) = 2A,\ f(0, 0, 0, 4) = A. \tag{X.2.52}$$

Verify from (2.41) that there are the results

$$f(1, 3, 0, 0) = B,\ f(2, 0, 1, 1) = B,\ f(1, 1, 0, 2) = B,\ f(0, 0, 1, 3) = B. \tag{X.2.53}$$

Verify from (2.42) and (2.43) that there are the results

$$f(2, 2, 0, 0) = C + D,\ f(0, 0, 2, 2) = C + D, \tag{X.2.54}$$

$$f(1, 1, 1, 1) = 2C, \tag{X.2.55}$$

$$f(0, 2, 2, 0) = D,\ f(2, 0, 0, 2) = D. \tag{X.2.56}$$

Verify from (2.44) that there are the results

$$f(3, 1, 0, 0) = E,\ f(1, 1, 2, 0) = E,\ f(1, 0, 2, 1) = E,\ f(0, 0, 3, 1) = E. \tag{X.2.57}$$

Verify from (2.45) that there are the results

$$f(4, 0, 0, 0) = F,\ f(2, 0, 2, 0) = 2F,\ f(0, 0, 4, 0) = F. \tag{X.2.58}$$

Verify (2.40) through (2.45).

X.3 Lie-Algebraic Decomposition of Polynomials

X.3.1 Fourth Degree Homogeneous Polynomials

Define the *fourth* degree homogeneous polynomials ${}^4\chi_0^0(\boldsymbol{w})$ and ${}^4\chi_m^2(\boldsymbol{w})$ by the rules

$$ {}^4\chi_0^0 = (L_z)^2 = q_x^2 p_y^2 - 2q_x p_x q_y p_y + p_x^2 q_y^2 = p^2 q^2 - (\boldsymbol{p}\cdot\boldsymbol{q})^2; \tag{X.3.1}$$

$$ {}^4\chi_2^2 = (p^2)^2, \tag{X.3.2}$$

$$ {}^4\chi_1^2 = 2p^2(\boldsymbol{p}\cdot\boldsymbol{q}), \tag{X.3.3}$$

$$ {}^4\chi_0^2 = (2/3)^{1/2}[p^2q^2 + 2(\boldsymbol{p}\cdot\boldsymbol{q})^2], \tag{X.3.4}$$

$$ {}^4\chi_{-1}^2 = 2q^2(\boldsymbol{p}\cdot\boldsymbol{q}), \tag{X.3.5}$$

$$ {}^4\chi_{-2}^2 = (q^2)^2. \tag{X.3.6}$$

(Note that all these polynomials are axially symmetric.) Then it can be verified that there are the operator results

$$ \mathcal{L}_\pm {}^4\chi_0^0 = 0, \tag{X.3.7}$$

$$ \mathcal{L}_0 {}^4\chi_0^0 = 0; \tag{X.3.8}$$

$$ \mathcal{L}_+ {}^4\chi_m^2 = [(2-m)(3+m)]^{1/2}\, {}^4\chi_{m+1}^2, \tag{X.3.9}$$

$$ \mathcal{L}_- {}^4\chi_m^2 = [(2+m)(3-m)]^{1/2}\, {}^4\chi_{m-1}^2, \tag{X.3.10}$$

$$ \mathcal{L}_0\, {}^4\chi_m^2 = m\, {}^4\chi_m^2. \tag{X.3.11}$$

Note that, in a manner identical to that found in the subject of quantum-mechanical angular momentum [which amounts to a study of the representations of $su(2)$], the operator $\mathcal{L}_0$ extracts the m value, and the operators $\mathcal{L}_+$ and $\mathcal{L}_-$ raise and lower m values, respectively. In particular, the results (3.7) through (3.11) can be written in the form

$$ \mathcal{L}_+ {}^n\chi_m^j = [(j-m)(j+m+1)]^{1/2}\, {}^n\chi_{m+1}^j, \tag{X.3.12}$$

$$ \mathcal{L}_- {}^n\chi_m^j = [(j+m)(j-m+1)]^{1/2}\, {}^n\chi_{m-1}^j, \tag{X.3.13}$$

$$ \mathcal{L}_0\, {}^n\chi_m^j = m\, {}^n\chi_m^j. \tag{X.3.14}$$

with j, which (as will become evident subsequently) plays the role of “spin”, having the values $j = 0$ or $j = 2$.[2] Thus, under the action of $\mathcal{L}_\pm$ and $\mathcal{L}_0$, ${}^4\chi_0^0$ behaves as a singlet and the $5 = 2j + 1$ with $j = 2$ quantities ${}^4\chi_{m}^2.$ behave as a quintuplet. Lastly, n denotes the degree of the homogeneous polynomial, with $n = 4$ in this case. This similarity arises because the underlying Lie algebra/group theory is the same here and in the treatment of quantum-mechanical angular momentum.

[2] We have placed the word *spin* in quotation marks because here j does not arise in the context of physical rotations, but rather in this instance is an aspect of the symplectic Lie algebra $sp(2,\mathbb{R})$. Note also that, because of the $(j - m)$ term on the right side of (3.12), the raising operation terminates when $m = j$. Similarly, the lowering operation terminates when $m = -j$.

Finally we observe that, in the present context, f_4 decomposes into a singlet spanned by ${}^4\chi_0^0$ and a quintuplet spanned by the ${}^4\chi_m^2$. That is, we may write

$$f_4 = {}^4c_0^0\ {}^4\chi_0^0 + \sum_{m=-2}^{2} {}^4c_m^2\ {}^4\chi_m^2 \tag{X.3.15}$$

where ${}^4c_0^0$ and the ${}^4c_m^2$ are uniquely defined coefficients.[3]

Let us compare the presentations (2.29) and (3.15). Upon equating like powers we obtain the the relations

$$A(p^2)^2 = {}^4c_2^2\,{}^4\chi_2^2 \Leftrightarrow A = {}^4c_2^2 \Leftrightarrow {}^4c_2^2 = A, \tag{X.3.16}$$

$$Bp^2(\boldsymbol{p}\cdot\boldsymbol{q}) = {}^4c_1^2\,{}^4\chi_1^2 \Leftrightarrow B = 2\,{}^4c_1^2 \Leftrightarrow {}^4c_1^2 = B/2, \tag{X.3.17}$$

$$C(\boldsymbol{p}\cdot\boldsymbol{q})^2 + Dp^2q^2 = {}^4c_0^0\,{}^4\chi_0^0 + {}^4c_0^2\,{}^4\chi_0^2, \tag{X.3.18}$$

$$E(\boldsymbol{p}\cdot\boldsymbol{q})q^2 = {}^4c_{-1}^2\,{}^4\chi_{-1}^2 \Leftrightarrow E = 2\,{}^4c_{-1}^2 \Leftrightarrow {}^4c_{-1}^2 = E/2, \tag{X.3.19}$$

$$F(q^2)^2 = {}^4c_{-2}^2\,{}^4\chi_{-2}^2 \Leftrightarrow F = {}^4c_{-2}^2 \Leftrightarrow {}^4c_{-2}^2 = F. \tag{X.3.20}$$

Further equating of like terms in (3.18 yields the results

$$C = -{}^4c_0^0 + 2(2/3)^{1/2}\ {}^4c_0^2, \tag{X.3.21}$$

$$D = {}^4c_0^0 + (2/3)^{1/2}\ {}^4c_0^2. \tag{X.3.22}$$

See Exercise 3.4. Finally, the relations (3.21) and (3.22) can be inverted to yield the relations

$${}^4c_0^0 = (1/3)(-C + 2D), \tag{X.3.23}$$

$${}^4c_0^2 = (1/6)^{1/2}(C + D). \tag{X.3.24}$$

If we make use of (2.34) through (2.39), and the results of the previous paragraph, we can also express the c_m^j in terms of various $f(k_1, k_2, k_3, k_4)$. So doing gives the results

$$\begin{aligned} {}^4c_0^0 &= (1/3)(-C + 2D) = (1/3)[-(1/2)f(1,1,1,1) + 2f(0,2,2,0)] \\ &= (1/6)[-f(1,1,1,1) + 4f(0,2,2,0)]; \end{aligned} \tag{X.3.25}$$

$${}^4c_2^2 = A = f(0,4,0,0), \tag{X.3.26}$$

$${}^4c_1^2 = B/2 = (1/2)f(1,3,0,0), \tag{X.3.27}$$

$$\begin{aligned} {}^4c_0^2 &= (1/6)^{1/2}(C + D) = (1/6)^{1/2}[(1/2)f(1,1,1,1) + f(0,2,2,0)] \\ &= (1/2)(1/6)^{1/2}[f(1,1,1,1) + 2f(0,2,2,0)], \end{aligned} \tag{X.3.28}$$

$${}^4c_{-1}^2 = E/2 = (1/2)f(3,1,0,0), \tag{X.3.29}$$

$${}^4c_{-2}^2 = F = f(4,0,0,0). \tag{X.3.30}$$

[3] It is interesting to note that, while in the case of $su(2)$ the construction of representations involves the mathematical use of two-variable polynomials, these variables otherwise play no direct physical role. By contrast, in the construction/representation of symplectic maps, polynomials in the phase-space variables play a direct physical role and have specific physical interpretations.

There is another way of obtaining explicit formulas for the c_m^j in terms of f_4 that is both instructive and convenient. Let $\langle *, * \rangle$ denote the scalar product defined in Section 7.3. See (7.3.1) through (7.3.9). Upon employing this scalar product we find, as will be seen subsequently, the results

$$\langle {}^4\chi_0^0, {}^4\chi_0^0 \rangle = 12, \tag{X.3.31}$$

$$\langle {}^4\chi_m^2, {}^4\chi_{m'}^2 \rangle = 64\delta_{mm'}, \tag{X.3.32}$$

$$\langle {}^4\chi_m^2, {}^4\chi_0^0 \rangle = 0. \tag{X.3.33}$$

Consequently, there are the formulas

$${}^4c_0^0 = (1/12)\langle {}^4\chi_0^0, f_4 \rangle, \tag{X.3.34}$$

$${}^4c_m^2 = (1/64)\langle {}^4\chi_m^2, f_4 \rangle. \tag{X.3.35}$$

Finally, we remark that the quantity

$${}^4c_0^0\ {}^4\chi_0^0 = {}^4c_0^0\,(L_z)^2 = {}^4c_0^0\,[p^2q^2 - (\boldsymbol{p}\cdot\boldsymbol{q})^2] = {}^4c_0^0\,p^2q^2 - {}^4c_0^0\,(\boldsymbol{p}\cdot\boldsymbol{q})^2, \tag{X.3.36}$$

which evidently is a combination of astigmatism and curvature of field, is called the (Joseph Maximilian) Petzval (1807-1891) in honor of his early analytic work on geometric aberrations. From (3.23) we see that the Petzval vanishes when

$$-C + 2D = 0. \tag{X.3.37}$$

X.3.2 Second Degree Homogeneous Polynomials

Let us continue with the Lie-algebraic decomposition of polynomials. The polynomials q^2, p^2, and $\boldsymbol{p}\cdot\boldsymbol{q}$ that go into making up f_2 may be labelled according to a scheme that is analogous to that used in (3.2) through (3.6). Define the axially-symmetric *second* degree homogeneous polynomials ${}^2\chi_m^1(\boldsymbol{w})$ by the rules

$${}^2\chi_1^1 = p^2, \tag{X.3.38}$$

$${}^2\chi_0^1 = (2)^{1/2}(\boldsymbol{p}\cdot\boldsymbol{q}), \tag{X.3.39}$$

$${}^2\chi_{-1}^1 = q^2. \tag{X.3.40}$$

Then we find the operator results

$$\mathcal{L}_+ {}^2\chi_m^1 = [(1-m)(2+m)]^{1/2}\ {}^2\chi_{m+1}^1, \tag{X.3.41}$$

$$\mathcal{L}_- {}^2\chi_m^1 = [(1+m)(2-m)]^{1/2}\ {}^2\chi_{m-1}^1, \tag{X.3.42}$$

$$\mathcal{L}_0\ {}^2\chi_m^1 = m\ {}^2\chi_m^1. \tag{X.3.43}$$

Consequently the rules (3,12) through (3.14) continue to hold with, in this case, $j = 1$ and $n = 2$.

We may also define an axially-symmetric second degree homogeneous polynomial ${}^2\psi_0^0(\boldsymbol{w})$ by the rule

$${}^2\psi_0^0 = L_z. \tag{X.3.44}$$

It meets our requirement for axial symmetry because it satisfies

$$\mathcal{L}_z\,{}^2\psi^0_0 =: L_z : L_z = 0. \tag{X.3.45}$$

It also satisfies the relation

$$\mathcal{L}_0\,{}^2\psi^0_0 = (1/2) : (\boldsymbol{p}\cdot\boldsymbol{q}) : L_z = (1/2)[(\boldsymbol{p}\cdot\boldsymbol{q}), L_z] = -(1/2)\mathcal{L}_z(\boldsymbol{p}\cdot\boldsymbol{q}) = 0. \tag{X.3.46}$$

Similarly, there are the relations

$$\mathcal{L}_\pm\,{}^2\psi^0_0 = 0. \tag{X.3.47}$$

(The polynomial ${}^2\psi^0_0$ is therefore entitled to carry the index values $j = 0$ and $m = 0$.) According to our previous discussion, $L_z = {}^2\psi^0_0$ does not appear as an odd power in f_2 or any other f_n. Only the ${}^2\chi^1_m$ can appear in the f_n. That is why we have used the symbol ψ for it rather than χ.

However that is not the end of the matter. According to (2.14) L_z^2 depends on p^2, q^2, and $(\boldsymbol{p}\cdot\boldsymbol{q})$, which are allowed ingredients for the f_n. Therefore, although *odd* powers of L_z are not allowed to appear in the f_n, *even* powers are allowed as, for example, in ${}^4\chi^0_0$. See (3.1).[4]

At this point it may be observed that there are the relations

$${}^4\chi^0_0 = (4/3)^{1/2}[({}^2\chi^1_1)({}^2\chi^1_{-1}) - ({}^2\chi^1_0)^2]; \tag{X.3.48}$$

$${}^4\chi^2_2 = ({}^2\chi^1_1)^2, \tag{X.3.49}$$

$${}^4\chi^2_1 = (2)^{1/2}({}^2\chi^1_1)({}^2\chi^1_0), \tag{X.3.50}$$

$${}^4\chi^2_0 = (2/3)^{1/2}[({}^2\chi^1_1)({}^2\chi^1_{-1}) + ({}^2\chi^1_0)^2], \tag{X.3.51}$$

$${}^4\chi^2_{-1} = (2)^{1/2}({}^2\chi^1_{-1})({}^2\chi^1_0), \tag{X.3.52}$$

$${}^4\chi^2_{-2} = ({}^2\chi^1_{-1})^2. \tag{X.3.53}$$

These relations are examples of the Clebsch-Gordan series for $sp(2,\mathbb{R})$, and are identical to the relations in quantum mechanics for coupling spin 1 and spin 1 to achieve net spin 0 or net spin 2. Agreement between the Clebsch-Gordan series for $sp(2,\mathbb{R})$ and $su(2)$ is to be expected because the commutation rules (2.22) through (2.24) for $sp(2,\mathbb{R})$ are the same as those for $su(2)$ in raising and lowering operator form, and the state relations (3.12) through (3.14) are the same in both cases.

X.3.3 Sixth and Eighth Degree Homogeneous Polynomials

The homogeneous polynomials that go into making f_6, f_8, etc., may be classified in similar fashion. One finds, for example, that f_6 decomposes into a triplet and a septuplet given by the relations

$${}^6\chi^1_m = ({}^4\chi^0_0)({}^2\chi^1_m); \tag{X.3.54}$$

$${}^6\chi^3_3 = (p^2)^3, \tag{X.3.55}$$

[4] *Odd* (as well as even) powers of L_z are allowed in the f_n for the magnetic optics case of a solenoid, which also has axial symmetry. See Section 16.2.

$${}^6\chi^3_2 = (6)^{1/2}(p^2)^2(\boldsymbol{p}\cdot\boldsymbol{q}), \tag{X.3.56}$$

$${}^6\chi^3_1 = (3/5)^{1/2}[4p^2(\boldsymbol{p}\cdot\boldsymbol{q})^2 + (p^2)^2q^2], \tag{X.3.57}$$

$${}^6\chi^3_0 = (4/5)^{1/2}[2(\boldsymbol{p}\cdot\boldsymbol{q})^3 + 3(\boldsymbol{p}\cdot\boldsymbol{q})p^2q^2], \tag{X.3.58}$$

$${}^6\chi^3_{-1} = (3/5)^{1/2}[4q^2(\boldsymbol{p}\cdot\boldsymbol{q})^2 + (q^2)^2p^2], \tag{X.3.59}$$

$${}^6\chi^3_{-2} = (6)^{1/2}(q^2)^2(\boldsymbol{p}\cdot\boldsymbol{q}), \tag{X.3.60}$$

$${}^6\chi^3_{-3} = (q^2)^3. \tag{X.3.61}$$

Consequently we may present a general (axially symmetric) f_6 in the form

$$f_6 = \sum_{m=-1}^{1} {}^6c^1_m\ {}^6\chi^1_m + \sum_{m=-3}^{3} {}^6c^3_m\ {}^6\chi^3_m \tag{X.3.62}$$

where the ${}^6c^1_m$ and ${}^6c^3_m$ are uniquely defined coefficients.

Similarly, f_8 decomposes into a singlet, a quintuplet, and a 9-tuplet given by the relations

$${}^8\chi^0_0 = ({}^4\chi^0_0)^2; \tag{X.3.63}$$

$${}^8\chi^2_m = ({}^4\chi^0_0)({}^4\chi^2_m); \tag{X.3.64}$$

$${}^8\chi^4_4 = ({}^2\chi^1_1)^4 = (p^2)^4, \tag{X.3.65}$$

$${}^8\chi^4_3 = (8)^{1/2}(p^2)^3(\boldsymbol{p}\cdot\boldsymbol{q}), \tag{X.3.66}$$

$${}^8\chi^4_2 = (4/7)^{1/2}[(p^2)^3q^2 + 6(p^2)^2(\boldsymbol{p}\cdot\boldsymbol{q})^2], \tag{X.3.67}$$

$${}^8\chi^4_1 = (8/7)^{1/2}[3(p^2)^2(\boldsymbol{p}\cdot\boldsymbol{q})q^2 + 4(p^2)(\boldsymbol{p}\cdot\boldsymbol{q})^3], \tag{X.3.68}$$

$${}^8\chi^4_0 = (2/35)^{1/2}[24p^2q^2(\boldsymbol{p}\cdot\boldsymbol{q})^2 + 3(q^2)^2(p^2)^2 + 8(\boldsymbol{p}\cdot\boldsymbol{q})^4], \tag{X.3.69}$$

$${}^8\chi^4_{-1} = (8/7)^{1/2}[3(q^2)^2(\boldsymbol{p}\cdot\boldsymbol{q})p^2 + 4(q^2)(\boldsymbol{p}\cdot\boldsymbol{q})^3], \tag{X.3.70}$$

$${}^8\chi^4_{-2} = (4/7)^{1/2}[(q^2)^3p^2 + 6(q^2)^2(\boldsymbol{p}\cdot\boldsymbol{q})^2, \tag{X.3.71}$$

$${}^8\chi^4_{-3} = (8)^{1/2}(q^2)^3(\boldsymbol{p}\cdot\boldsymbol{q}), \tag{X.3.72}$$

$${}^8\chi^4_{-4} = ({}^2\chi^1_{-1})^4 = (q^2)^4. \tag{X.3.73}$$

Consequently we may present a general (axially symmetric) f_8 in the form

$$f_8 = {}^8c^0_0\ {}^8\chi^0_0 + \sum_{m=-2}^{2} {}^8c^2_m\ {}^8\chi^2_m + \sum_{m=-4}^{4} {}^8c^4_m\ {}^8\chi^4_m \tag{X.3.74}$$

where ${}^8c^0_0$ and the ${}^8c^2_m$ and the ${}^8c^4_m$ are uniquely defined coefficients.

It can be checked, as implied by the presentations (3.15), (3.62), and (3.74), that the decompositions given above are exhaustive. That is, the various ${}^n\chi^j_m$ span each (axially symmetric) f_n. Moreover, the ${}^n\chi^j_m$ have been defined in such a way that one has the general relations (3.12) through (3.14).

X.3.4 Proof of Orthogonality and Definition/Use of the Quadratic Casimir Operator

The aim of this subsection is to prove the scalar product relations

$$\langle {}^{n'}\chi^{j'}_{m'}\,,\,{}^{n}\chi^{j}_{m}\rangle = N(n,j)\delta_{n'n}\delta_{j'j}\delta_{m'm} \tag{X.3.75}$$

where the $N(n,j)$ are normalization constants *independent* of m. One of the tools for doing so will be the quadratic Casimir operator for our realization of $sp(2,\mathbb{R})$.

That (3.75) should contain the delta function $\delta_{n'n}$ is obvious because, by definition, $\langle *, *\rangle$ is zero for unlike monomial pairs, and hence vanishes for homogeneous polynomials of different degrees. See Subsection 7.3.1 and Exercise 7.3.25.

Let us next verify the $\delta_{m'm}$ factor. Consider the quantity $\langle {}^{n'}\chi^{j'}_{m'}\,,\,\mathcal{L}_0\,{}^{n}\chi^{j}_{m}\rangle$. With the aid of (3.14) we may write

$$\langle {}^{n'}\chi^{j'}_{m'}\,,\,\mathcal{L}_0\,{}^{n}\chi^{j}_{m}\rangle = m\langle {}^{n'}\chi^{j'}_{m'}\,,\,{}^{n}\chi^{j}_{m}\rangle. \tag{X.3.76}$$

But, with the aid of (2.20) and (3.14), we may also write

$$\langle {}^{n'}\chi^{j'}_{m'}\,,\,\mathcal{L}_0\,{}^{n}\chi^{j}_{m}\rangle = \langle (\mathcal{L}_0)^{\dagger}\,{}^{n'}\chi^{j'}_{m'}\,,\,{}^{n}\chi^{j}_{m}\rangle = \langle \mathcal{L}_0\,{}^{n'}\chi^{j'}_{m'}\,,\,{}^{n}\chi^{j}_{m}\rangle = m'\langle {}^{n'}\chi^{j'}_{m'}\,,\,{}^{n}\chi^{j}_{m}\rangle. \tag{X.3.77}$$

It follows that

$$(m'-m)\langle {}^{n'}\chi^{j'}_{m'}\,,\,{}^{n}\chi^{j}_{m}\rangle = 0 \tag{X.3.78}$$

from which we conclude that

$$\langle {}^{n'}\chi^{j'}_{m'}\,,\,{}^{n}\chi^{j}_{m}\rangle = 0 \text{ when } m' \neq m. \tag{X.3.79}$$

To see that $N(n,j)$ is (as the notation asserts) independent of m, consider the quantity $\langle \mathcal{L}_+\,{}^{n}\chi^{j}_{m}\,,\,\mathcal{L}_+\,{}^{n}\chi^{j}_{m}\rangle$. Making use of (3.12) gives the result

$$\langle \mathcal{L}_+\,{}^{n}\chi^{j}_{m}\,,\,\mathcal{L}_+\,{}^{n}\chi^{j}_{m}\rangle = [(j-m)(j+m+1)]^{1/2}[(j-m)(j+m+1)]^{1/2}\langle {}^{n}\chi^{j}_{m+1}\,,\,{}^{n}\chi^{j}_{m+1}\rangle. \tag{X.3.80}$$

From (2.19) we conclude that

$$\langle \mathcal{L}_+\,{}^{n}\chi^{j}_{m}\,,\,\mathcal{L}_+\,{}^{n}\chi^{j}_{m}\rangle = \langle {}^{n}\chi^{j}_{m}\,,\,(\mathcal{L}_+)^{\dagger}\mathcal{L}_+\,{}^{n}\chi^{j}_{m}\rangle = \langle {}^{n}\chi^{j}_{m}\,,\,\mathcal{L}_-\mathcal{L}_+\,{}^{n}\chi^{j}_{m}\rangle. \tag{X.3.81}$$

But from (3.12) and (3.13) we see that

$$\begin{aligned} \mathcal{L}_-\mathcal{L}_+\,{}^{n}\chi^{j}_{m} &= [(j-m)(j+m+1)]^{1/2}\mathcal{L}^{n}_{-}\chi^{j}_{m+1} \\ &= [(j-m)(j+m+1)]^{1/2}[(j+m+1)(j-m)]^{1/2}\,{}^{n}\chi^{j}_{m} \end{aligned} \tag{X.3.82}$$

so that

$$\langle \mathcal{L}_+\,{}^{n}\chi^{j}_{m}\,,\,\mathcal{L}_+\,{}^{n}\chi^{j}_{m}\rangle = [(j-m)(j+m+1)]^{1/2}[(j+m+1)(j-m)]^{1/2}\,\langle {}^{n}\chi^{j}_{m}\,,\,{}^{n}\chi^{j}_{m}\rangle. \tag{X.3.83}$$

Upon comparing (3.80) with (3.83) we see that there is the relation

$$\begin{aligned} &[(j-m)(j+m+1)]^{1/2}[(j-m)(j+m+1)]^{1/2}\langle {}^{n}\chi^{j}_{m+1}\,,\,{}^{n}\chi^{j}_{m+1}\rangle = \\ &[(j-m)(j+m+1)]^{1/2}[(j+m+1)(j-m)]^{1/2}\,\langle {}^{n}\chi^{j}_{m}\,,\,{}^{n}\chi^{j}_{m}\rangle. \end{aligned} \tag{X.3.84}$$

Therefore, as long as $[(j-m)(j+m+1)] \neq 0$ (which will be true if raising of m is possible), we have found the relation

$$\langle {}^n\chi^j_{m+1}\,,\,{}^n\chi^j_{m+1}\rangle = \langle {}^n\chi^j_m\,,\,{}^n\chi^j_m\rangle. \tag{X.3.85}$$

With this result we can repeatedly increase m starting from $m=-j$ to obtain the result

$$\langle {}^n\chi^j_m\,,\,{}^n\chi^j_m\rangle = \langle {}^n\chi^j_{-j}\,,\,{}^n\chi^j_{-j}\rangle \text{ for } m \in [-j,j], \tag{X.3.86}$$

which verifies that N does not depend on m.

The last step is to verify the $\delta_{j'j}$ factor in (3.75). This can be done with the aid of the quadratic Casimir operator $\mathcal{C}_2$ for our realization of $sp(2,\mathbb{R})$. For the purposes of this appendix, it is defined by the rule

$$\mathcal{C}_2 = (\mathcal{L}_+\mathcal{L}_- + \mathcal{L}_-\mathcal{L}_+ + 2\mathcal{L}_0^2)/2. \tag{X.3.87}$$

(For a discussion of Casimir operators, see Section 27.11.) It follows from (2.19) through (2.21) that $\mathcal{C}_2$ is Hermitian,

$$\mathcal{C}_2^\dagger = \mathcal{C}_2. \tag{X.3.88}$$

And it follows from (2.22) through (2.24) that $\mathcal{C}_2$ commutes with all the $sp(2,\mathbb{R})$ generators, $\mathcal{L}_\pm$ and $\mathcal{L}_0$,

$$\{\mathcal{C}_2, \mathcal{L}_\pm\} = \{\mathcal{C}_2, \mathcal{L}_0\} = 0, \tag{X.3.89}$$

as expected for a Casimir operator.

Let us compute $\mathcal{C}_2\,{}^n\chi^j_j$. Evidently,

$$\mathcal{C}_2\,{}^n\chi^j_j = (1/2)(\mathcal{L}_+\mathcal{L}_- + \mathcal{L}_-\mathcal{L}_+ + 2\mathcal{L}_0^2)\,{}^n\chi^j_j. \tag{X.3.90}$$

Evaluate each of the three terms appearing in (3.90). For $\mathcal{L}_0^2\,{}^n\chi^j_j$ there is the result

$$\mathcal{L}_0^2\,{}^n\chi^j_j = j^2\,{}^n\chi^j_j. \tag{X.3.91}$$

Recall (3.14). For $(1/2)\mathcal{L}_-\mathcal{L}_+\,{}^n\chi^j_j$ there is the result

$$(1/2)\mathcal{L}_-\mathcal{L}_+\,{}^n\chi^j_j = 0. \tag{X.3.92}$$

Recall (3.82) evaluated for $m=j$. Finally, for $(1/2)\mathcal{L}_+\mathcal{L}_-\,{}^n\chi^j_j$, there is the result

$$(1/2)\mathcal{L}_+\mathcal{L}_-\,{}^n\chi^j_j = (1/2)\mathcal{L}_+(2j)^{1/2}\,{}^n\chi^j_{j-1} = (1/2)(2j)^{1/2}(2j)^{1/2}\,{}^n\chi^j_j = j\,{}^n\chi^j_j. \tag{X.3.93}$$

Recall (3.12) and (3.13). Consequently, as in the analogous quantum-mechanical calculation, there is the net result

$$\mathcal{C}_2\,{}^n\chi^j_j = (j+j^2)\,{}^n\chi^j_j = [j(j+1)]\,{}^n\chi^j_j. \tag{X.3.94}$$

To continue our exploration, multiply both sides of (3.94) by $\mathcal{L}_-^\ell$ where $\ell = j-m$. So doing gives the result

$$[j(j+1)]\mathcal{L}_-^\ell\,{}^n\chi^j_j = \mathcal{L}_-^\ell\mathcal{C}_2\,{}^n\chi^j_j = \mathcal{C}_2\mathcal{L}_-^\ell\,{}^n\chi^j_j \tag{X.3.95}$$

where (3.89) has been used. But from (repeated, if necessary) use of (3.13) it follows that there is a relation of the form

$$\mathcal{L}_-^\ell \, {}^n\chi_j^j = \lambda(j,\ell) \, {}^n\chi_m^j \tag{X.3.96}$$

where $\lambda(j,\ell)$ is a non-vanishing proportionality constant. Combining (3.95) and (3.96) gives the final result

$$\mathcal{C}_2 \, {}^n\chi_m^j = [j(j+1)] \, {}^n\chi_m^j. \tag{X.3.97}$$

We are now prepared to verify the $\delta_{j'j}$ factor in (3.75). Using (3.97) gives the relation

$$\langle {}^n\chi_m^{j'}, \mathcal{C}_2 \, {}^n\chi_m^j \rangle = [j(j+1)] \langle {}^n\chi_m^{j'}, {}^n\chi_m^j \rangle = [(j+1/2)^2 - 1/4] \langle {}^n\chi_m^{j'}, {}^n\chi_m^j \rangle. \tag{X.3.98}$$

But, from (3.88), we also have the relation

$$\langle {}^n\chi_m^{j'}, \mathcal{C}_2 \, {}^n\chi_m^j \rangle = \langle \mathcal{C}_2 \, {}^n\chi_m^{j'}, {}^n\chi_m^j \rangle = [(j'+1/2)^2 - 1/4] \langle {}^n\chi_m^{j'}, {}^n\chi_m^j \rangle. \tag{X.3.99}$$

Upon combining (3.98) and (3.99) we see that

$$[(j'+1/2)^2 - (j+1/2)^2] \langle {}^n\chi_m^{j'}, {}^n\chi_m^j \rangle = 0. \tag{X.3.100}$$

It is easily verified that

$$[(j'+1/2)^2 - (j+1/2)^2] = 0 \Leftrightarrow j' = j \text{ or } j' + j = -1. \tag{X.3.101}$$

The second possibility on the right side of (3.101) cannot occur since we are only working with nonnegative values of j and j'. We conclude that for our purposes the factor

$$[(j'+1/2)^2 - (j+1/2)^2] \neq 0 \text{ for } j' \neq j, \tag{X.3.102}$$

and therefore (3.100) requires that

$$\langle {}^n\chi_m^{j'}, {}^n\chi_m^j \rangle = 0 \text{ for } j' \neq j. \tag{X.3.103}$$

What remains is to evaluate/specify the $N(n,j)$. Presumably there is a relatively simple formula that does so. But for our purposes the Table below suffices for values of n and j of present interest.

Table X.3.1: Some Values of $N(n, j)$.

n	$N(n,0)$	$N(n,1)$	$N(n,2)$	$N(n,3)$	$N(n,4)$
2	*	4	*	*	*
4	12	*	64	*	*
6	*	160	*	2304	*
8	?	*	?	*	?

Exercises

X.3.1. Verify (3.7) through (3.11) and (3.41) through (3.43).

X.3.2. Verify (3.25) through (3.30).

X.3.3. Review Exercise 2.3. Equation (3.1) provides the monomial decomposition for ${}^4\chi^0_0$. Verify that the relations below provide the monomial decompositions for the ${}^4\chi^2_m$:

$$ {}^4\chi^2_2 = (p^2)^2 = (p_x)^4 + 2(p_x)^2(p_y)^2 + (p_y)^4, \tag{X.3.104}$$

$$ {}^4\chi^2_1 = 2p^2(\boldsymbol{p}\cdot\boldsymbol{q}) = 2q_x(p_x)^3 + 2(p_x)^2 q_y p_y + 2q_x p_x (p_y)^2 + 2q_y(p_y)^3, \tag{X.3.105}$$

$$\begin{aligned} {}^4\chi^2_0 &= (2/3)^{1/2}[p^2q^2 + 2(\boldsymbol{p}\cdot\boldsymbol{q})^2] \\ &= (2/3)^{1/2}\{[(p_x)^2+(p_y)^2][(q_x)^2+(q_y)^2] + 2(q_xp_x+q_yp_y)^2\} \\ &= (2/3)^{1/2}[(q_x)^2(p_x)^2 + (p_x)^2(q_y)^2 + (q_x)^2(p_y)^2 + (q_y)^2(p_y)^2 \\ &\quad +2(q_x)^2(p_x)^2 + 4q_xp_xq_yp_y + 2(q_y)^2(p_y)^2] \\ &= (2/3)^{1/2}[3(q_x)^2(p_x)^2 + (p_x)^2(q_y)^2 + (q_x)^2(p_y)^2 + 3(q_y)^2(p_y)^2 + 4q_xp_xq_yp_y], \end{aligned} \tag{X.3.106}$$

$$ {}^4\chi^2_{-1} = 2(\boldsymbol{p}\cdot\boldsymbol{q})q^2 = 2(q_x)^3p_x + 2q_xp_x(q_y)^2 + 2(q_x)^2q_yp_y + 2(q_y)^3p_y, \tag{X.3.107}$$

$$ {}^4\chi^2_{-2} = (q^2)^2 = (q_x)^4 + 2(q_x)^2(q_y)^2 + (q_y)^4. \tag{X.3.108}$$

X.3.4. The aim of this exercise is to verify the relations (3.21) through (3.24). Review the results (2.46) through (2.51), (3.1), and (3.104) through (3.108). Verify that

$$\begin{aligned} C(\boldsymbol{p}\cdot\boldsymbol{q})^2 + Dp^2q^2 &= C[(q_x)^2(p_x)^2 + 2q_xp_xq_yp_y + (q_y)^2(p_y)^2] \\ &+ D[(q_x)^2(p_x)^2 + (p_x)^2(q_y)^2 + (q_x)^2(p_y)^2 + (q_y)^2(p_y)^2] \\ &= (C+D)[(q_x)^2(p_x)^2 + (q_y)^2(p_y)^2] \\ &+ 2Cq_xp_xq_yp_y + D[(p_x)^2(q_y)^2 + (q_x)^2(p_y)^2]. \end{aligned} \tag{X.3.109}$$

Verify that

$$\begin{aligned}
&{}^4c_0^0\,{}^4\chi_0^0 + {}^4c_0^2\,{}^4\chi_0^2 = {}^4c_0^0\,[q_x^2p_y^2 - 2q_xp_xq_yp_y + p_x^2q_y^2]\\
&+{}^4c_0^2(2/3)^{1/2}[3(q_x)^2(p_x)^2 + (p_x)^2(q_y)^2 + (q_x)^2(p_y)^2 + 3(q_y)^2(p_y)^2 + 4q_xp_xq_yp_y]\\
&= (6)^{1/2}\,{}^4c_0^2[(q_x)^2(p_x)^2 + (q_y)^2(p_y)^2] + [-2\,{}^4c_0^0 + 4(2/3)^{1/2}\,{}^4c_0^2]q_xp_xq_yp_y\\
&+[{}^4c_0^0 + (2/3)^{1/2}\,{}^4c_0^2][(p_x)^2(q_y)^2 + (q_x)^2(p_y)^2]. \qquad \text{(X.3.110)}
\end{aligned}$$

Since the monomials appearing on the right sides of (3.109) and (3.110) are linearly independent, we may equate the coefficients of like terms. In particular, we may equate the coefficients of the polynomials

$$[(q_x)^2(p_x)^2 + (q_y)^2(p_y)^2],\ q_xp_xq_yp_y,\ \text{and}\ [(p_x)^2(q_y)^2 + (q_x)^2(p_y)^2]. \qquad \text{(X.3.111)}$$

Conclude that (3.18), when combined with (3.109) and (3.110), implies the relations

$${}^4c_0^2 = 6^{-1/2}(C+D) \Leftrightarrow C + D = (6)^{1/2}\,{}^4c_0^2, \qquad \text{(X.3.112)}$$

$$2C = [-2\,{}^4c_0^0 + 4(2/3)^{1/2}\,{}^4c_0^2] \Leftrightarrow C = [-\,{}^4c_0^0 + 2(2/3)^{1/2}\,{}^4c_0^2], \qquad \text{(X.3.113)}$$

$$D = [{}^4c_0^0 + (2/3)^{1/2}\,{}^4c_0^2], \qquad \text{(X.3.114)}$$

from which it follows that

$${}^4c_0^0 = (1/3)(-C+2D) \Leftrightarrow (-C+2D) = 3\,{}^4c_0^0. \qquad \text{(X.3.115)}$$

Observe that (3.113) and (3.114) agree with the claims (3.21) and (3.22). Also, observe that taking the sum of (3.21) and (3.22) produces (3.24), which agrees with (3.112). Finally, verify that (3.23) follows from (3.21) and (3.22).

X.3.5. Observe that the operators $\mathcal{L}_\pm$ and $\mathcal{L}_0$ are derivations. Work out their effects on ${}^4\chi_0^0$ and the ${}^4\chi_m^2$ using (3.1) and (3.48) through (3.53) and (3.41) through (3.43). Verify that your results agree with (3.7) through (3.11).

X.3.6. Review the derivation of the result (3.85) obtained with the use of the operator $\mathcal{L}_+$. Using the operator $\mathcal{L}_-$ in a similar way, derive the relation

$$\langle {}^n\chi_{m-1}^j\,,\,{}^n\chi_{m-1}^j\rangle = \langle {}^n\chi_m^j\,,\,{}^n\chi_m^j\rangle. \qquad \text{(X.3.116)}$$

X.3.7. Verify that the ${}^n\chi_m^j$ given by (3.1) through (3.6), (3.38) through (3.40), (3.54) through (3.61), and (3.63) through (3.73) all obey the rules (3.12) through (3.14).

X.3.8. Verify the relations

$$\mathcal{L}_+ =: -p^2/2 := -(1/2) : {}^2\chi_1^1 :, \qquad \text{(X.3.117)}$$

$$\mathcal{L}_0 =: (\boldsymbol{p}\cdot\boldsymbol{q})/2 := (1/8)^{1/2} : {}^2\chi_0^1 :, \qquad \text{(X.3.118)}$$

$$\mathcal{L}_- =: q^2/2 := (1/2) : {}^2\chi_{-1}^1 : . \qquad \text{(X.3.119)}$$

X.3.9. Verify the correctness of the entries in Table 3.1.

X.4 Application of Multiplet Decomposition

It has been shown that the various f_n can be decomposed into multiplets with members ${}^n\chi^j_m$. What is this decomposition good for? Suppose an optical system is composed of N elements, and let $\mathcal{M}_i$ be the optical transfer map for the i'th element. Then the net optical transfer map $\mathcal{M}$ for the entire system can be written as the product

$$\mathcal{M}_{\rm net} = \mathcal{M}_1\mathcal{M}_2\cdots\mathcal{M}_N. \tag{X.4.1}$$

Next observe that each of the $\mathcal{M}_i$ has a factorization of the form (2.26). Suppose further that the various f's for the various $\mathcal{M}_i$ are all known. Then the only problem involved in computing $\mathcal{M}_{\rm net}$ is that of combining a collection of known maps and writing the result in factorized form.

The general problem of combining/concatenating maps has been treated in Section 8.4. Let us briefly summarize the consequences of this treatment for the present discussion. Suppose $\mathcal{M}_f$ and $\mathcal{M}_g$ are two optical transfer maps written in factored product form,

$$\mathcal{M}_f = \mathcal{R}_f \exp(:f_4:)\exp(:f_6:)\exp(:f_8:)\cdots\,, \tag{X.4.2}$$

$$\mathcal{M}_g = \mathcal{R}_g \exp(:g_4:)\exp(:g_6:)\exp(:g_8:)\cdots\,. \tag{X.4.3}$$

Let $\mathcal{M}_h$ be their product,

$$\mathcal{M}_h = \mathcal{M}_f\mathcal{M}_g, \tag{X.4.4}$$

and write $\mathcal{M}_h$ in the factorized form

$$\mathcal{M}_h = \mathcal{R}_h \exp(:h_4:)\exp(:h_6:)\exp(:h_8:)\cdots\,. \tag{X.4.5}$$

Employ (4.2) and (4.3) in (4.4), and judiciously insert factors of $\mathcal{I} = \mathcal{R}_g\mathcal{R}_g^{-1}$, to write

$$\begin{aligned}\mathcal{M}_h &= \mathcal{R}_f \exp(:f_4:)\exp(:f_6:)\exp(:f_8:)\cdots\times\mathcal{R}_g \exp(:g_4:)\exp(:g_6:)\exp(:g_8:)\cdots\\ &= \mathcal{R}_f\mathcal{R}_g\mathcal{R}_g^{-1}\exp(:f_4:)\mathcal{R}_g\mathcal{R}_g^{-1}\exp(:f_6:)\mathcal{R}_g\mathcal{R}_g^{-1}\exp(:f_8:)\mathcal{R}_g\cdots\\ &\quad\times\exp(:g_4:)\exp(:g_6:)\exp(:g_8:)\cdots\,.\end{aligned} \tag{X.4.6}$$

Next manipulate and make the definition

$$\mathcal{R}_g^{-1}\exp(:f_n:)\mathcal{R}_g = \exp[\mathcal{R}_g^{-1}:f_n:\mathcal{R}_g] = \exp(:\mathcal{R}_g^{-1}f_n:) = \exp(:f_n^{\rm tr}:). \tag{X.4.7}$$

Here the *transformed* polynomials $f_n^{\rm tr}$ are defined in terms of the original f_n by the relations

$$f_n^{\rm tr} = \mathcal{R}_g^{-1}f_n, \tag{X.4.8}$$

from which it follows that

$$f_n^{\rm tr}(\boldsymbol{w}) = f_n[(R^g)^{-1}\boldsymbol{w}] \tag{X.4.9}$$

where R^g is the matrix associated with $\mathcal{R}_g$. [See (8.2.26).] Upon combining the results of our manipulation and definition we conclude that

$$\mathcal{M}_h = \mathcal{R}_f\mathcal{R}_g \exp(:f_4^{\rm tr}:)\exp(:f_6^{\rm tr}:)\exp(:f_8^{\rm tr}:)\cdots\times\exp(:g_4:)\exp(:g_6:)\exp(:g_8:)\cdots\,. \tag{X.4.10}$$

Now compare (4.5) and (4.10) to conclude that

$$\mathcal{R}_h = \mathcal{R}_f \mathcal{R}_g, \tag{X.4.11}$$

and

$$\begin{aligned}&\exp(:h_4:)\exp(:h_6:)\exp(:h_8:)\cdots = \\ &\exp(:f_4^{\mathrm{tr}}:)\exp(:f_6^{\mathrm{tr}}:)\exp(:f_8^{\mathrm{tr}}:)\cdots \times \exp(:g_4:)\exp(:g_6:)\exp(:g_8:)\cdots .\end{aligned} \tag{X.4.12}$$

It follows from (4.11) that the associated matrices obey the relation

$$R^h = R^g R^f. \tag{X.4.13}$$

And, from the work of Section 8.4, we know that (4.12) yields the aberration generator relations

$$h_4 = f_4^{\mathrm{tr}} + g_4, \tag{X.4.14}$$

$$h_6 = f_6^{\mathrm{tr}} + g_6 + [f_4^{\mathrm{tr}}, g_4]/2, \tag{X.4.15}$$

$$h_8 = f_8^{\mathrm{tr}} + g_8 + [f_6^{\mathrm{tr}}, g_4] - [f_4^{\mathrm{tr}}, [f_4^{\mathrm{tr}}, g_4]]/6 + [g_4, [g_4, f_4^{\mathrm{tr}}]]/3, \text{ etc.} \tag{X.4.16}$$

Inspection of (4.14) through (4.16) shows that the determination of the aberration generators involves carrying out the transformations (4.9) and the evaluation of certain Poisson brackets. It is these two tasks that may, in some cases, be simplified by multiplet decomposition.

Observe that any $\mathcal{R}_g^{-1}$ is generated by $\mathcal{L}_\pm$ and $\mathcal{L}_0$. It follows from (3.12) through (3.14) that any $\mathcal{R}_g^{-1}$ acting on any element of a given multiplet must give a result in the *same* multiplet. Specifically, one must have results of the form

$$\mathcal{R}_g^{-1}\ {}^n\chi_m^j = \sum_{m'=-j}^{j} D_{m'm}^j[(R^g)^{-1}]\ {}^n\chi_{m'}^j. \tag{X.4.17}$$

Here the $D_{m'm}^j[(R^g)^{-1}]$ are the transformation functions associated with the symplectic matrices $(R^g)^{-1}$ and are the analytic continuation from $SU(2)$ to $Sp(2,\mathbb{R})$ of the $SU(2)$ Wigner functions (which are entire so that unique analytic continuation is always well defined). Special cases of the relations (4.17) are the results

$$\mathcal{R}_g^{-1}\ {}^n\chi_0^0 \ = \ {}^n\chi_0^0 \text{ for } n = 4, 8, \tag{X.4.18}$$

which are immediately evident consequences of (3.7) and (3.8), and analogous relations for ${}^8\chi_0^0$.

The relations (4.17) in analytic form may or may not be computationally useful.[5] However, they do show what contributes to what. Suppose, for example, that f_4 is given by (3.15), and similarly g_4 is given by

$$g_4 = {}^4d_0^0\ {}^4\chi_0^0 + \sum_{m=-2}^{2} {}^4d_m^2\ {}^4\chi_m^2. \tag{X.4.19}$$

[5] What is essentially involved here is the operation (4.9). It can be realized numerically utilizing the efficient and fast algorithm described in Section 39.9. The operation (4.9) is of direct/immediate use if one wishes to compute the f_n^{tr}. It is of intermediate use if one wishes to compute the $D_{m'm}^j$ using (4.52).

Then, from the definition (4.8) and using (4.17) and (4.18), we find that

$$\begin{aligned} f_4^{\rm tr} &= \mathcal{R}_g^{-1} f_4 = \mathcal{R}_g^{-1}[{}^4c_0^0\ {}^4\chi_0^0 + \sum_{m=-2}^{2} {}^4c_m^2\ {}^4\chi_m^2] = \\ &= {}^4c_0^0\ {}^4\chi_0^0 + \sum_{m'=-2}^{2}\{\sum_{m=-2}^{2} D^2_{m'm}[(R^g)^{-1}]\ {}^4c_m^2\ \}\ {}^4\chi_{m'}^2 \\ &= {}^4c_0^0\ {}^4\chi_0^0 + \sum_{m=-2}^{2}\{\sum_{m'=-2}^{2} D^2_{mm'}[(R^g)^{-1}]\ {}^4c_{m'}^2\ \}\ {}^4\chi_m^2 \\ &= {}^4c_0^0\ {}^4\chi_0^0 + \sum_{m=-2}^{2} {}^4e_m^2\ {}^4\chi_m^2 \end{aligned} \tag{X.4.20}$$

where the coefficients ${}^4e_m^2$ are given by the relation

$${}^4e_m^2 = \sum_{m'=-2}^{2} D^2_{mm'}[(R^g)^{-1}]\ {}^4c_{m'}^2. \tag{X.4.21}$$

All the ingredients are now available for insertion into (4.14) to yield the result

$$h_4 = f_4^{\rm tr} + g_4 = ({}^4c_0^0 + {}^4d_0^0)\ {}^4\chi_0^0\ + \sum_{m=-2}^{2}({}^4e_m^2 + {}^4d_m^2)\ {}^4\chi_m^2. \tag{X.4.22}$$

We see that the singlet contributions (the Petzval contributions) to h_4 are *purely additive*, and depend *only* on the singlet content of f_4 and g_4, Similarly, the quintuplet content of h_4 depends *only* on the quintuplet content of f_4 and g_4, although in a somewhat more complicated way: The quintuplet content of f_4 first has to be transformed by $\mathcal{R}_g^{-1}$ before its addition to the quintuplet content of g_4 to yield the net quintuplet content for h_4.

The discussion so far has dealt with the combining of third-order aberrations as described by (4.14). Now look at (4.15) which describes how fifth-order aberrations combine/arise. Evidently, by an argument similar to that made for third-order aberrations, the triplet and septuplet components of f_6 and g_6 contribute separately and independently to the triplet and septuplet components, respectively, of h_6. Moreover, according to (4.15), there is a contribution arising from third-order aberrations due to the Poisson bracket term. We will discuss Poisson bracket terms shortly,

We end this part of the discussion with the observation that there are some similarities in the computation of h_8 and the computation of h_4. From (3.63) and (3.74) we see that eight-order polynomials of the form f_8 may also have a singlet content ${}^8\chi_0^0$, which may be viewed as a higher-order Petzval. And these singlet contributions to h_8 will also be purely additive. Moreover, there are quintuplet, and 9-tuple components of f_8 and g_8 that contribute separately and independently to the quintuplet, and 9-tuple components, respectively, of h_8

What remains is to study the Poisson bracket terms in the expressions (4.15), (4.16) etc. for h_6, h_8, etc. These terms describe how lower-order aberrations combine (feed up) to produce higher-order aberrations. We will begin with the Poisson bracket term $[f_4^{\rm tr}, g_4]$ in

(4.15). Before going into specifics, there are two general observations. First, the ordinary product and the Lie product (the Poisson bracket) of any two axially symmetric polynomials must also be axially symmetric. See Exercise 4.3. Second, there is the rule (7.6.14) relating the degree of a Poisson bracket to the degrees of its ingredients. From these observations it follows, for example, that $[f_4^{\rm tr}, g_4]$ must be some linear combination of the ${}^6\chi_m^1$ and the ${}^6\chi_m^3$.

Let us now be more specific. From (4.19) and (4.20) we see that $[f_4^{\rm tr}, g_4]$ is some linear combination of the Poisson brackets

$$[{}^4\chi_0^0, {}^4\chi_0^0], \tag{X.4.23}$$

$$[{}^4\chi_0^0, {}^4\chi_m^2], \tag{X.4.24}$$

$$[{}^4\chi_m^2, {}^4\chi_{m'}^2]. \tag{X.4.25}$$

Evidently the Poisson bracket term (4.23) vanishes due to antisymmetry. We will soon see that the Poisson bracket terms (4.24) vanish due to axial symmetry. All that remains are the Poisson brackets (4.25). It follows that the only feed-up terms contributing to h_6 arise from quintuplet terms in f_4 interacting with quintuplet terms in g_4. There are no feed up terms arising from a singlet term interacting with a singlet term, nor from the interaction of singlet and quintuplet terms.

To see that (4.24) vanishes, observe that

$$[{}^4\chi_0^0, {}^4\chi_m^2] = [(L_z)^2, {}^4\chi_m^2] = 2L_z[L_z, {}^4\chi_m^2] = 0. \tag{X.4.26}$$

Here we have used (3.1), the derivation property (1.7.7), and the axial symmetry of the ${}^4\chi_m^2$.

We now study the remaining quantities (4.25). Define polynomials $\theta_{m,m'}^{22}$ by the rule

$$\theta_{m,m'}^{22} = [{}^4\chi_m^2, {}^4\chi_{m'}^2]. \tag{X.4.27}$$

Then, since $\mathcal{L}_0$ and $\mathcal{L}_\pm$ are derivations with respect to the Poisson bracket Lie product, recall (5.3.9), we find the results

$$\mathcal{L}_0\theta_{m,m'}^{22} = [\mathcal{L}_0\ {}^4\chi_m^2, {}^4\chi_{m'}^2] + [{}^4\chi_m^2, \mathcal{L}_0\ {}^4\chi_{m'}^2] = (m+m')\theta_{m,m'}^{22}, \tag{X.4.28}$$

$$\begin{aligned}
\mathcal{L}_+\theta_{m,m'}^{22} &= [\mathcal{L}_+\ {}^4\chi_m^2, {}^4\chi_{m'}^2] + [{}^4\chi_m^2, \mathcal{L}_+\ {}^4\chi_{m'}^2] \\
&= [(2-m)(2+m+1)]^{1/2}[{}^4\chi_{m+1}^2, {}^4\chi_{m'}^2] + [(2-m')(2+m'+1)]^{1/2}[{}^4\chi_m^2, {}^4\chi_{m'+1}^2] \\
&= [(2-m)(2+m+1)]^{1/2}\theta_{m+1,m'}^{22} + [(2-m')(2+m'+1)]^{1/2}\theta_{m,m'+1}^{22},
\end{aligned} \tag{X.4.29}$$

$$\begin{aligned}
\mathcal{L}_-\theta_{m,m'}^{22} &= [\mathcal{L}_-\ {}^4\chi_m^2, {}^4\chi_{m'}^2] + [{}^4\chi_m^2, \mathcal{L}_-\ {}^4\chi_{m'}^2] \\
&= [(2+m)(2-m+1)]^{1/2}[{}^4\chi_{m-1}^2, {}^4\chi_{m'}^2] + [(2+m')(2-m'+1)]^{1/2}[{}^4\chi_m^2, {}^4\chi_{m'-1}^2] \\
&= [(2+m)(2-m+1)]^{1/2}\theta_{m-1,m'}^{22} + [(2+m')(2-m'+1)]^{1/2}\theta_{m,m'-1}^{22}.
\end{aligned} \tag{X.4.30}$$

Inspection of the relations (4.28) through (4.30) shows that they are analogous to the behavior of the (tensor) product of two $j = 2$ entities. Therefore all the standard Clebsch-Gordan and Wigner-Eckart $su(2)$ machinery is available. Also, following earlier reasoning, all the entries in (4.25) must be some linear combination of the ${}^6\chi^1_m$ and the ${}^6\chi^3_m$. Consequently there must be relations of the form

$$\sum_{m_1 m_2} C(22j; m_1, m_2, m)\ [{}^4\chi^2_{m_1}, {}^4\chi^2_{m_2}] = \sum_{m_1 m_2} C(22j; m_1, m_2, m)\theta^{22}_{m_1 m_2}$$
$$= \delta_{j1}\alpha(221)\ {}^6\chi^1_m + \delta_{j3}\alpha(223)\ {}^6\chi^3_m, \tag{X.4.31}$$

where the coefficients C are $su(2)$ Clebsch-Gordan coefficients and the coefficients $\alpha(221)$ and $\alpha(223)$ are to be determined. See Exercise 4.5. The relations (4.31) can be inverted using the completeness properties of the Clebsch-Gordan coefficients to yield the final result

$$\begin{aligned}[{}^4\chi^2_m, {}^4\chi^2_{m'}] &= \alpha(221)\ C(221; m, m', m+m')\ {}^6\chi^1_{m+m'} \\ &+ \alpha(223)\ C(223; m, m', m+m')\ {}^6\chi^3_{m+m'},\end{aligned} \tag{X.4.32}$$

where the coefficients α are seen to play the role of the reduced matrix elements that occur in applications of the Wigner-Eckart theorem. Again see Exercise 4.5. The needed Clebsch-Gordan coefficients are listed in Tables 4.1 and 4.2 below.

Table X.4.1: Some values of $C(221; m, m', m+m')$ and $C(223; m, m', m+m')$

.

m	m'	$m+m'$	$C(221; ***)$	$C(223; ***)$
2	2	4	0	0
2	1	3	0	$\sqrt{1/2}$
2	0	2	0	$\sqrt{1/2}$
2	-1	1	$\sqrt{1/5}$	$\sqrt{3/10}$
2	-2	0	$\sqrt{2/5}$	$\sqrt{1/10}$
1	2	3	0	$-\sqrt{1/2}$
1	1	2	0	0
1	0	1	$-\sqrt{3/10}$	$\sqrt{1/5}$
1	-1	0	$-\sqrt{1/10}$	$\sqrt{2/5}$
1	-2	-1	$\sqrt{1/5}$	$\sqrt{3/10}$
0	2	2	0	$-\sqrt{1/2}$
0	1	1	$\sqrt{3/10}$	$-\sqrt{1/5}$
0	0	0	0	0

Table X.4.2: Remaining values of $C(221;m,m',m+m')$ and $C(223;m,m',m+m')$

.

m	m'	$m+m'$	$C(221;***)$	$C(223;***)$
0	-1	-1	$-\sqrt{3/10}$	$\sqrt{1/5}$
0	-2	-2	0	$\sqrt{1/2}$
-1	2	1	$-\sqrt{1/5}$	$-\sqrt{3/10}$
-1	1	0	$\sqrt{1/10}$	$-\sqrt{2/5}$
-1	0	-1	$\sqrt{3/10}$	$-\sqrt{1/5}$
-1	-1	-2	0	0
-1	-2	-3	0	$\sqrt{1/2}$
-2	2	0	$-\sqrt{2/5}$	$-\sqrt{1/10}$
-2	1	-1	$-\sqrt{1/5}$	$-\sqrt{3/10}$
-2	0	-2	0	$-\sqrt{1/2}$
-2	-1	-3	0	$-\sqrt{1/2}$
-2	-2	-4	0	0

What remains is to find $\alpha(221)$ and $\alpha(223)$. An easy computation gives the result

$$[{}^4\chi^2_2, {}^4\chi^2_{-2}] = -(384/25)^{1/2}\ {}^6\chi^1_0 - (64/5)^{1/2}\ {}^6\chi^3_0. \tag{X.4.33}$$

For the same j and m values use of (4.32) gives the result

$$\begin{aligned}[{}^4\chi^2_2, {}^4\chi^2_{-2}] &= \alpha(221)\ C(221; 2, -2, 0)\ {}^6\chi^1_0 \\ &+ \alpha(223)\ C(223; 2, -2, 0)\ {}^6\chi^3_0,\end{aligned} \tag{X.4.34}$$

Upon comparing (4.33) and (4.34) and with the knowledge that ${}^6\chi^1_0$ and ${}^6\chi^3_0$ are linearly independent, see (3.75), we conclude that

$$\alpha(221)\ C(221; 2, -2, 0) = -(384/25)^{1/2}, \tag{X.4.35}$$

$$\alpha(221)\ C(223; 2, -2, 0) = -(64/5)^{1/2}. \tag{X.4.36}$$

According to Table 4.1 the Clebsch-Gordan coefficients associated with (4.35) and (4.36) are given by the relations

$$C(221; 2, -2, 0) = \sqrt{2/5}, \tag{X.4.37}$$

$$C(223; 2, -2, 0) = \sqrt{1/10}. \tag{X.4.38}$$

It follows from (4.35) through (4.38) that the constants $\alpha(221)$ and $\alpha(223)$ have the values

$$\alpha(221) = -(384/25)^{1/2}/(2/5)^{1/2} = -(192/5)^{1/2}, \tag{X.4.39}$$

$$\alpha(223) = -(64/5)^{1/2}/(1/10)^{1/2} = -(128)^{1/2}. \tag{X.4.40}$$

Correspondingly, (4.32) takes the final form

$$\begin{aligned}[{}^4\chi^2_m, {}^4\chi^2_{m'}] &= -(192/5)^{1/2}\ C(221; m, m', m+m')\ {}^6\chi^1_{m+m'} \\ &\quad -(128)^{1/2}\ C(223; m, m', m+m')\ {}^6\chi^3_{m+m'}.\end{aligned} \tag{X.4.41}$$

Upon reflection, we see that what has been illustrated is that the evaluation of Poisson brackets can be carried out in general in terms of $su(2)$ Clebsch-Gordan coefficients and a few simply computed numbers analogous to reduced matrix elements.

The discussion of the combining of fifth-order aberrations, and the feed-up effect of lower-order aberrations to contribute to fifth-order aberrations, is now complete. Moreover, it is clear from (4.16) and Poisson bracket relations analogous to (4.41) that the the combining of seventh and still higher-order aberrations, and the feed-up effect of lower-order aberrations to contribute to these higher-order aberrations, follow a similar pattern. All that is needed is the computation of the $f^{\rm tr}_n$ and various single and multiple Poisson brackets. Finally we remark that, although the existence and knowledge of explicit formulas, such as (4.41), for Poisson brackets may be illuminating, they are not required for actual numerical computation. Their rapid numerical evaluation may be performed using the methods described in Section 39.8.

Exercises

X.4.1. Show that (4.14) through (4.16) are special cases of (8.4.32), (8.4.34), and (8.4.36).

X.4.2. Show, using (3.75) and (4.17), that there is the relation

$$D^j_{m'm}[(R^g)^{-1}] = [1/N(n,j)]\langle {}^n\chi^j_{m'} \, , \mathcal{R}_g^{-1} \, {}^n\chi^j_m\rangle. \tag{X.4.42}$$

X.4.3. Recall (1.7.7), review Exercise 5.2.3, and recall the axial symmetry relation/condition (2.27). Show that if two functions f and g have axial symmetry, then so does their ordinary product fg and their Lie product (Poisson bracket) $[f,g]$.

X.4.4. Verify (4.33).

X.4.5. The purpose of this exercise is to establish (4.31) and its inverse (4.32). We have seen from (4.28) through (4.30) that the behavior of the $\theta^{22}_{m_1m_2}$ is analogous to the behavior of the product of two $j = 2$ entities. From the Quantum Theory of angular momentum, we know that two entities of spin 2 can be combined/coupled to produce entities of spins $0, 1, 2, 3,$ and 4. That is what the left side of (4.31) seeks to do for the values $j = 0, 1, 2, 3, 4$. The right side of (4.31) states the expected results for these same j values. The expected results seem sensible for the cases $j = 1$ and $j = 3$. But what about the cases $j = 0, 2, 4$? Do we expect the left side of (4.31) to actually *vanish* in these cases as the right side states? We do. It can be shown that the $su(2)$ Clebsch-Gordan coefficients have the symmetry property

$$C(j_1j_2j;m_1,m_2,m_1+m_2) = (-1)^{j_1+j_2-j}C(j_2j_1j;m_2,m_1,m_1+m_2). \tag{X.4.43}$$

A special case of (4.43) is the relation

$$C(22j;m_1,m_2,m_1+m_2) = (-1)^{-j}C(22j;m_2,m_1,m_1+m_2). \tag{X.4.44}$$

That is, these C values are *even* under the interchange of m_1 and m_2 for even values of j, and *odd* under the interchange for odd values of j. But from the antisymmetry property of the Poisson bracket we know that

$$\theta^{22}_{m_1m_2} = -\theta^{22}_{m_2m_1}. \tag{X.4.45}$$

Verify, therefore, that there must be the result

$$\sum_{m_1m_2} C(22j;m_1,m_2,m)\theta^{22}_{m_1m_2} = 0 \text{ for } j = 0,2,4. \tag{X.4.46}$$

And this desired result follows simply from symmetry considerations alone without any additional information.

It can be shown that the $su(2)$ Clebsch-Gordan coefficients satisfy the *completeness relation*

$$\sum_j C(j_1j_2j;m_1,m_2,m_1+m_2)C(j_1j_2j;m_1',m_2',m_1'+m_2') = \delta_{m_1m_1'}\delta_{m_2m_2'}. \tag{X.4.47}$$

Use this result to derive (4.32) from (4.31).

Finally note that, because of the described symmetry properties of the Clebsch-Gordan coefficients and Poisson brackets, both sides of (4.32) are antisymmetric under the interchange of m and m', as desired, and the Clebsch-Gordan coefficients involved in (4.32) vanish when $m = m'$. Scan the entries of Tables 4.1 and 4.2 to verify that the listed C values do indeed have these advertised symmetry properties.

X.4.6. Using (3.75) and (4.32) show that there is the relation

$$\begin{aligned}\langle {}^6\chi^j_{m+m'}, [{}^4\chi^2_m, {}^4\chi^2_{m'}]\rangle &= -(192/5)^{1/2}\, C(221; m, m', m+m')N(6,1)\delta_{j1} \\ &\quad -(128)^{1/2}\, C(223; m, m', m+m')N(6,3)\delta_{j3}. \qquad \text{(X.4.48)}\end{aligned}$$

X.5 Maps/Lie Generators for Continuous Systems

X.6 Maps/Lie Generators for Discontinuous Systems

X.7 Two Sample Designs

X.7.1 Aberration Corrected Doublet

In this subsection we will illustrate how some of out earlier results can be used to design an imaging doublet system that is free of all third-order aberrations and four fifth-order aberrations. This system is illustrated schematically in Figure 7.1 below.

The system consists of four surfaces separated by drift spaces either in air or two possibly different refractive media. Between the object plane and *Surface* S^1 there is a *left-side* drift space (in air) of on-axis length D_L. Surfaces S^1 and S^2, with an on-axis separation of thickness t_L, constitute a first lens made of a medium with refractive index n_L. Surfaces S^3 and S^4, with an on-axis separation of thickness t_R, constitute a second lens made of a medium with refractive index n_R. Between surfaces S^2 and S^3 there is a drift space (in air) of on-axis length D. Finally, between S^4 and the image plane there is a *right-side* drift space (in air) of on-axis length D_R. The surfaces S^1 and S^2 will be chosen so that (in paraxial approximation) the first lens is converging, and the surfaces S^3 and S^4 will be chosen so that (in paraxial approximation) the second lens is diverging.

Elimination of Petzval

We begin our design with the requirement that the net (third-order) Petzval aberration vanish. From our earlier work we know that only the maps associated with the surfaces S^1 through S^4 contribute to the Petzval, and their contributions are additive. According to *, passage through a surface S from a medium with refraction index n^- to a medium with refraction index n^+ makes a contribution to the Petzval coefficient given by the relation

$$\text{contribution } = \beta_2[(1/n^+) - (1/n^-)]. \qquad \text{(X.7.1)}$$

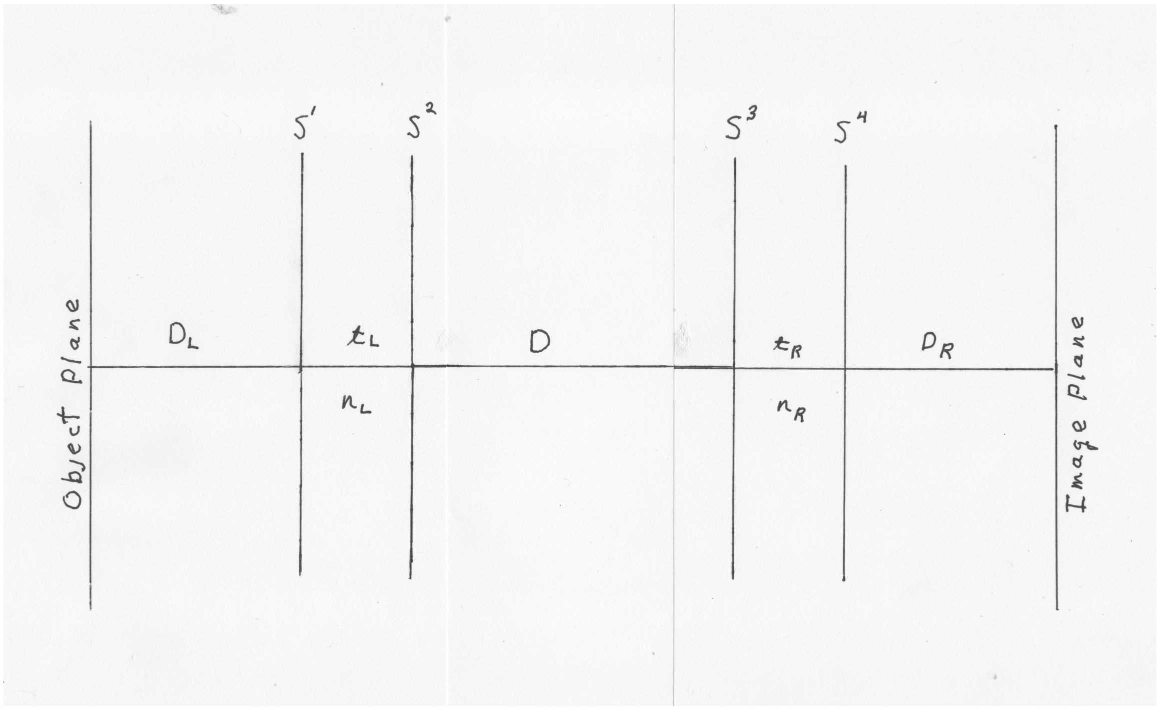


Figure X.7.1: Schematic layout of imaging doublet system that is free of all third-order aberrations and four fifth-order aberrations. The object plane is on the left and the image plane is on the right.

Here β_2 is the quadratic parameter of the surface.[6] Consequently, for the four surfaces, we find the Petzval coefficient contributions to be as follows:

$$\text{For } S^1,\ n^- = 1 \text{ and } n^+ = n_L \Rightarrow \text{contribution } = \beta_2^1[(1/n_L) - 1], \tag{X.7.2}$$

$$\text{For } S^2,\ n^- = n_L \text{ and } n^+ = 1 \Rightarrow \text{contribution } = \beta_2^2[1 - (1/n_L)], \tag{X.7.3}$$

$$\text{For } S^3,\ n^- = 1 \text{ and } n^+ = n_R \Rightarrow \text{contribution } = \beta_2^3[(1/n_R) - 1], \tag{X.7.4}$$

$$\text{For } S^4,\ n^- = n_R \text{ and } n^+ = 1 \Rightarrow \text{contribution } = \beta_2^4[1 - (1/n_R)]. \tag{X.7.5}$$

Here the quantity β_2^j is the quadratic parameter for the j^{th} surface. The net Petzval coefficient is the sum of these terms, and we require that it vanish,

$$\text{Net Petzval coefficient} = (\beta_2^1 - \beta_2^2)[(1/n_L) - 1] + (\beta_2^3 - \beta_2^4)[(1/n_R) - 1] = 0. \tag{X.7.6}$$

There are several ways to satisfy (7.6). For simplicity, we specify that

$$n_L = n_R = n. \tag{X.7.7}$$

[6] Unlike the third-order aberrations associated with the ${}^4\chi_m^2$, the Petzval aberration (associated with ${}^4\chi_0^0$) is independent of the quartic parameter β_4 of surfaces. We may view parameters that govern paraxial behavior as being "paraxial" parameters. Consequently, the β_2 parameters, as well as indices of refraction and lengths, are paraxial parameters. By contrast, the β_4, β_6, etc. have no effect on paraxial behavior and therefore are not paraxial parameters. The Petzval is different from other third-order aberrations in that is governed by paraxial parameters, and is independent of the β_4 parameters.

Also we specify that

$$\beta_2^1 > 0 \text{ and } \beta_2^2 = -\beta_2^1 \tag{X.7.8}$$

so that the first lens is (symmetrically) biconvex and converging. And we specify that

$$\beta_2^3 < 0 \text{ and } \beta_2^4 = -\beta_2^3 \tag{X.7.9}$$

so that the second lens is (symmetrically) biconcave and diverging. (Our intuition, which can be checked, is that minimizing the curvatures of all lens surfaces by making lenses symmetrical, which essentially amounts to sharing power equally between leading and trailing lens surfaces save for finite lens thickness effects, should on average help minimize aberrations.) With these specifications the requirement (7.6) takes the simpler form

$$\text{Net Petzval coefficient} = 2\beta_2^1[(1/n) - 1] + 2\beta_2^3[(1/n) - 1] = 0, \tag{X.7.10}$$

and we see that (7.10) is satisfied providing

$$\beta_2^3 = -\beta_2^1. \tag{X.7.11}$$

See Figure 7.2 where these specifications and the requirements (7.7) through (7.9) and (7.11) are depicted graphically. We conclude that, in order to be Petzval aberration free, a system must have both focusing and defocusing elements.

Paraxial Properties

The next design step is to examine, in the paraxial approximation, the optical transfer map associated with the drifts and lenses depicted in Figure 7.2. These maps can all be written as products of maps of the form exp(: f_2 :). Listed below are the f_2 polynomials for the various items depicted in Figure 7.2.

$$\text{Drift space of length } d \text{ in air: } f_2 = -(d/2)p^2. \tag{X.7.12}$$

Here d takes the values

$$d = D_L, \ d = D, \text{ and } d = D_R. \tag{X.7.13}$$

$$\text{Drift space of length } d \text{ in medium with index } n\text{: } f_2 = -[d/(2n)]p^2. \tag{X.7.14}$$

Here we assume that both lenses in the doublet have on-axis thickness t so that

$$d = t. \tag{X.7.15}$$

According to (*) the f_2 associated with passage through a surface S from a medium with refraction index n^- to a medium with refraction index n^+ is given by the relation

$$f_2 = \beta_2(n^- - n^+)q^2. \tag{X.7.16}$$

Here again β_2 is the quadratic parameter for the surface. Therefore, for the surfaces S^1 through S^4, there are the following general results:

$$\text{For } S^1, \ n^- = 1 \text{ and } n^+ = n_L \Rightarrow f_2 = \beta_2^1(1 - n_L)q^2, \tag{X.7.17}$$

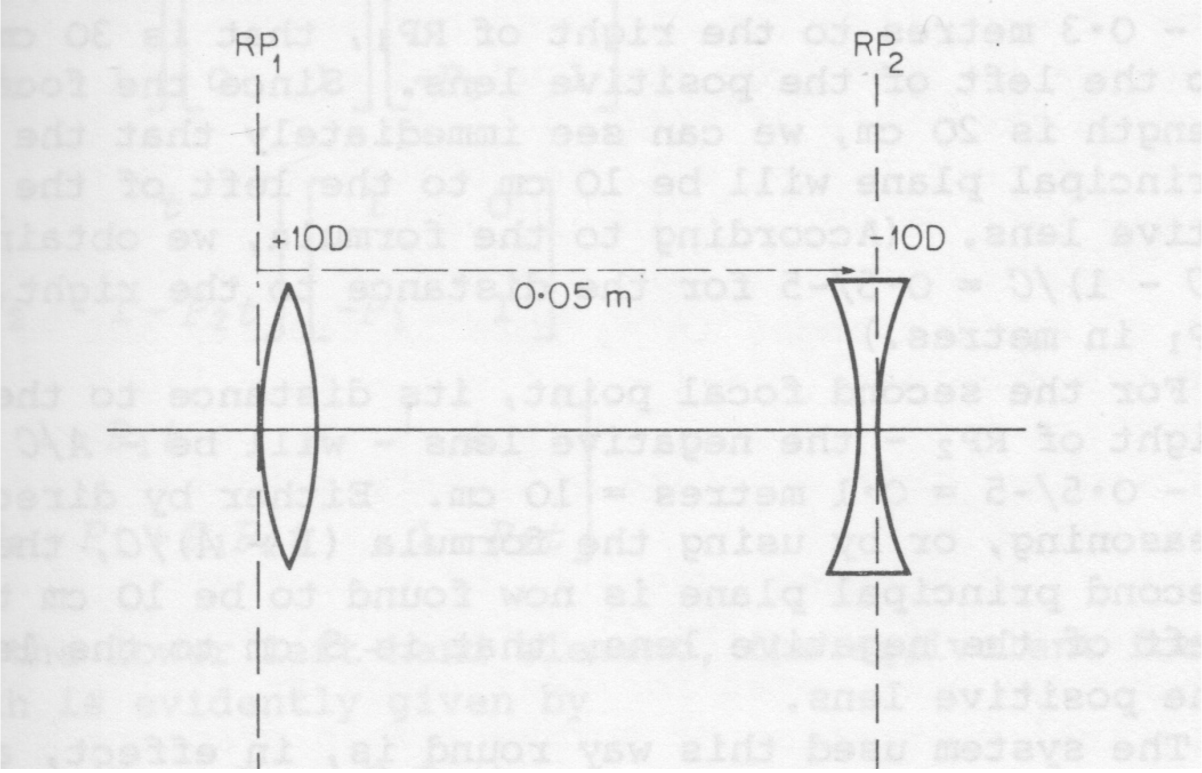


Figure X.7.2: Less schematic layout of imaging doublet system that is free of all third-order aberrations and four fifth-order aberrations. The object plane is on the far left and the image plane is on the far right (so that both are not visible), and only the shapes of the various lens surfaces and the lens thicknesses and spacings are illustrated.

$$\text{For } S^2,\ n^- = n_L \text{ and } n^+ = 1 \Rightarrow f_2 = \beta_2^2(n_L - 1)]q^2, \tag{X.7.18}$$

$$\text{For } S^3,\ n^- = 1 \text{ and } n^+ = n_R \Rightarrow\ f_2 = \beta_2^3(1 - n_R)q^2, \tag{X.7.19}$$

$$\text{For } S^4,\ n^- = n_R \text{ and } n^+ = 1 \Rightarrow f_2 = \beta_2^4(n_R - 1)q^2. \tag{X.7.20}$$

We will use these general results for the specific cases described by (7.7) through (7.9), (7.11), (7.13), and (7.15).

We are now prepared to compute $\mathcal{R}$, the linear part of the transfer map for the optical system illustrated in Figure 7.2. Based on the results summarized in *, it is given by the product

$$\begin{aligned}&\mathcal{R} = \exp[-(D_L/2) : p^2 :] \exp[\beta_2^1(1-n) : q^2 :] \exp\{-[t/(2n)] : p^2 :\} \exp[\beta_2^1(1-n) : q^2 :] \times \\ &\exp[-(D/2) : p^2 :] \exp[-\beta_2^1(1-n) : q^2 :] \exp\{-[t/(2n)] : p^2 :\} \exp[-\beta_2^1(1-n) : q^2 :] \times \\ &\exp[-(D_R/2) : p^2 :].\end{aligned} \tag{X.7.21}$$

Let R be the matrix associated with $\mathcal{R}$. Since only the Lie operators $: p^2 :$ and $: q^2 :$ appear in $\mathcal{R}$, and since these operators map the pairs q_x, p_x and q_y, p_y into themselves and in the same way, it follows that R must be of the block form

$$R = \begin{pmatrix} G & O \\ O & G \end{pmatrix} \tag{X.7.22}$$

where each block is 2×2, the block G is symplectic, and the block O is the zero matrix. Therefore, for the computation of R, there is the simplification of only needing to work with various 2×2 matrices corresponding to the various $\exp(: f_2 :)$. Let us list these matrices, call them K: For f_2 of the form (7.12) there is the correspondence

$$f_2 = -(d/2)p^2 \leftrightarrow K = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix}. \tag{X.7.23}$$

For f_2 of the form (7.14) there is the correspondence

$$f_2 = -[(d/(2n)p^2] \leftrightarrow K = \begin{pmatrix} 1 & d/n \\ 0 & 1 \end{pmatrix}. \tag{X.7.24}$$

For f_2 of the form (7.16) there is the correspondence

$$f_2 = \beta_2(n^- - n^+)q^2 \leftrightarrow K = \begin{pmatrix} 1 & 0 \\ 2\beta_2(n^- - n^+) & 1 \end{pmatrix}. \tag{X.7.25}$$

Since (7.21) has nine factors, it follows that the G associated with $\mathcal{R}$ is the product of nine 2×2 matrices of the form (7.23) through (7.25). We will eventually compute this G numerically. But we will first make some preliminary observations/calculations.

Map $\mathcal{R}$ for the System and Map $\mathcal{R}'$ for the Device

In practical applications, we may imagine that most of the parameter values in (7.21) are fixed save for D_L and D_R, which could be fairly readily adjusted to achieve imaging and the desired magnification. This circumstance suggests that we should understand the nature of the map that these leading and trailing drifts surround. That is, we are interested in the map $\mathcal{R}'$ defined by the product

$$\mathcal{R}' = \exp[\beta_2^1(1-n) : q^2 :] \exp\{-[t/(2n)] : p^2 :\} \exp[\beta_2^1(1-n) : q^2 :] \times \\ \exp[-(D/2) : p^2 :] \exp[-\beta_2^1(1-n) : q^2 :] \exp\{-[t/(2n)] : p^2 :\} \exp[-\beta_2^1(1-n) : q^2 :]. \tag{X.7.26}$$

Compare (7.21) and (7.26). That is, we have the relation

$$\mathcal{R} = \exp[-(D_L/2) : p^2 :]\mathcal{R}' \exp[-(D_R/2) : p^2 :]. \tag{X.7.27}$$

Put another way, we may view $\mathcal{R}'$ as being the linear part of the map for the *optical device* and $\mathcal{R}$ as being the linear part of the map for the complete *optical system.*

Normal Form

What would we like to know about $\mathcal{R}'$ or, equivalently, the matrices R' and G'? We will see that it is possible to associate with $\mathcal{R}'$ a kind of *normal form.* Suppose, as a mathematical trick, we consider the map $\mathcal{R}''$ defined by relation

$$\mathcal{R}'' = \exp[(d_L/2) : p^2 :]\mathcal{R}' \exp[(d_R/2) : p^2 :]. \tag{X.7.28}$$

We have "sandwiched" $\mathcal{R}'$ between two *negative* length drift maps where d_L and d_R are to be determined. Let us compute the matrix G'' associated with $\mathcal{R}''$. With the aid of (7.23) we see that it is given by the relation

$$\begin{aligned} G'' &= \begin{pmatrix} 1 & -d_R \\ 0 & 1 \end{pmatrix} G' \begin{pmatrix} 1 & -d_L \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & -d_R \\ 0 & 1 \end{pmatrix} \begin{pmatrix} G'_{11} & G'_{12} \\ G'_{21} & G'_{22} \end{pmatrix} \begin{pmatrix} 1 & -d_L \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & -d_R \\ 0 & 1 \end{pmatrix} \begin{pmatrix} G'_{11} & -G'_{11}d_L + G'_{12} \\ G'_{21} & -G'_{21}d_L + G'_{22} \end{pmatrix} \\ &= \begin{pmatrix} G'_{11} - d_R G'_{21} & -G'_{11}d_L + G'_{12} - d_R(-G'_{21}d_L + G'_{22}) \\ G'_{21} & -G'_{21}d_L + G'_{22} \end{pmatrix} \\ &= \begin{pmatrix} G''_{11} & G''_{12} \\ G''_{21} & G''_{22} \end{pmatrix}. \end{aligned} \tag{X.7.29}$$

Upon comparing the last two lines in (7.29) we see that

$$G''_{21} = G'_{21}. \tag{X.7.30}$$

Next we seek values of d_L and d_R such that

$$1 = G''_{11} = G'_{11} - d_R G'_{21} \Rightarrow d_R = (G'_{11} - 1)/G'_{21}, \tag{X.7.31}$$

$$1 = G''_{22} = -G'_{21} d_L + G'_{22} \Rightarrow d_L = (G'_{22} - 1)/G'_{21}. \tag{X.7.32}$$

We see that the goals $G''_{11} = 1$ and $G''_{22} = 1$ can be achieved provided

$$G'_{21} \neq 0. \tag{X.7.33}$$

For the values of d_R and d_L given by (7.31) and (7.32) we find that

$$\begin{aligned} G''_{12} &= -G'_{11} d_L + G'_{12} - d_R(-G'_{21} d_L + G'_{22}) \\ &= -G'_{11}(G'_{22} - 1)/G'_{21} + G'_{12} - [(G'_{11} - 1)/G'_{21}]\{-G'_{21}[(G'_{22} - 1)/G'_{21}] + G'_{22}\} \\ &= -G'_{11}(G'_{22} - 1)/G'_{21} + G'_{12} - [(G'_{11} - 1)/G'_{21}] \\ &= [-G'_{11}(G'_{22} - 1) + G'_{12}G'_{21} - (G'_{11} - 1)]/G'_{21} \\ &= [-G'_{11}G'_{22} + G'_{12}G'_{21} + 1]/G'_{21} \\ &= [-\det(G') + 1]/G'_{21} \\ &= 0. \end{aligned} \tag{X.7.34}$$

Here we have used the fact that G' is symplectic.[7] We have verified the remarkable result that there is a (unique) choice for the pair d_L, d_R such that G'' takes the simple/normal form

$$G'' = \begin{pmatrix} 1 & 0 \\ G'_{21} & 1 \end{pmatrix}. \tag{X.7.35}$$

Upon solving (7.29) for G', we find the result

$$G' = \begin{pmatrix} 1 & d_R \\ 0 & 1 \end{pmatrix} G'' \begin{pmatrix} 1 & d_L \\ 0 & 1 \end{pmatrix}. \tag{X.7.36}$$

We conclude that, in paraxial approximation, the device acts like a thin lens preceded by a drift of length d_L and followed by a drift of length d_R. And the thin lens has a focal length f given by

$$1/f = -G'_{21} \Leftrightarrow f = -1/(G'_{21}). \tag{X.7.37}$$

That is, (7.35) can be rewritten in the form

$$G'' = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \tag{X.7.38}$$

with f given by (7.37). For the problem at hand we will eventually find that

$$f > 0. \tag{X.7.39}$$

From (7.35) we see that $\mathcal{R}''$ has the Lie form

$$\mathcal{R}'' = \exp[(G'_{21}/2) : q^2 :]. \tag{X.7.40}$$

Correspondingly, from (7.28) and (7.40), we see that $\mathcal{R}'$ has the factorization

$$\mathcal{R}' = \exp[-(d_L/2) : p^2 :] \exp[(G'_{21}/2) : q^2 :] \exp[-(d_R/2) : p^2 :].$$

Note that this result is consistent with (7.35) and (7.36).

[7] As rewarding as the messy calculation (7.34) ultimately proved to be, it is/was actually not necessary. Once (7.31) through (7.33) are established, the relation $G''_{12} = 0$ must hold in order for G'' to be symplectic, which we already know to be the case.

Imaging Condition and Computation of Magnification

Let us use the normal form for $\mathcal{R}'$ and the associated matrix G' to discuss the possible imaging and magnification properties of $\mathcal{R}$. We will do this by working with the associated matrix G. According to (7.27), it is given by the product

$$\begin{aligned} G &= \begin{pmatrix} 1 & D_R \\ 0 & 1 \end{pmatrix} G' \begin{pmatrix} 1 & D_L \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & D_R \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & d_R \\ 0 & 1 \end{pmatrix} G'' \begin{pmatrix} 1 & d_L \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & D_L \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & D_R + d_R \\ 0 & 1 \end{pmatrix} G'' \begin{pmatrix} 1 & D_L + d_L \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & D_R + d_R \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \begin{pmatrix} 1 & D_L + d_L \\ 0 & 1 \end{pmatrix}. \end{aligned} \tag{X.7.41}$$

At this point, to simplify continuation of this calculation, it is convenient to define *effective* drift lengths D_L^e and D_R^e by the rules

$$D_L^e = D_L + d_L, \tag{X.7.42}$$

$$D_R^e = D_R + d_R, \tag{X.7.43}$$

With these definitions we can move the calculation (7.41) forward to find the result

$$\begin{aligned} G &= \begin{pmatrix} 1 & D_R + d_R \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \begin{pmatrix} 1 & D_L + d_L \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & D_R^e \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \begin{pmatrix} 1 & D_L^e \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & D_R^e \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & D_L^e \\ -1/f & -D_L^e/f + 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 - D_R^e/f & D_L^e - D_R^e D_L^e/f + D_R^e \\ -1/f & -D_L^e/f + 1 \end{pmatrix}. \end{aligned} \tag{X.7.44}$$

Suppose further we attempt to select D_L^e and D_R^e in such a way that

$$G_{12} = 0 \tag{X.7.45}$$

so that R is imaging. Enforcing the relation (7.45) produces the chain of relations

$$\begin{aligned} &0 = D_L^e - D_R^e D_L^e/f + D_R^e \Leftrightarrow \\ &0 = 1/D_R^e - 1/f + 1/D_L^e \Leftrightarrow \\ &1/D_L^e + 1/D_R^e = 1/f. \end{aligned} \tag{X.7.46}$$

Note that the last line of (7.46) is the familiar elementary imaging condition except that it involves effective quantities.

When (7.45) is enforced, (7.44) takes the form

$$G = \begin{pmatrix} 1 - D_R^e/f & 0 \\ -1/f & -D_L^e/f + 1 \end{pmatrix}. \tag{X.7.47}$$

As a sanity check, let us verify that this G is symplectic. We find that for this G

$$\begin{aligned} G_{11}G_{22} &= (1 - D_R^e/f)(1 - D_L^e/f) = 1 - (D_R^e/f + D_L^e/f) + (D_R^e/f)(D_L^e/f) \\ &= 1 - (D_L^e d_R^e/f)(1/D_L^e + 1/D_R^e) + (D_R^e D_L^e/f)(1/f) \\ &= 1 - (D_L^e d_R^e/f)(1/D_L^e + 1/D_R^e - 1/f^e) = 1 \end{aligned} \tag{X.7.48}$$

as expected. [Here we have used the last line of (7.46).] It follows that (7.47) can be rewritten in the form

$$G = \begin{pmatrix} m & 0 \\ -1/f & 1/m \end{pmatrix} \tag{X.7.49}$$

where m is the magnification given by the upper left entry in (7.47),

$$\begin{aligned} m &= -(D_R^e/f - 1) \\ &= -(D_R/f + d_R/f - 1). \end{aligned} \tag{X.7.50}$$

From (7.50) it is evident that, for sufficiently large values of D_R, m is negative (the image is inverted as expected) and can become large in magnitude as $D_R \to \infty$ providing physically possible values of D_L can be found such that (7.46 is satisfied. Expressing the imaging condition (7.46) in terms of D_L and D_R yields the result

$$1/(D_L + d_L) + 1/(D_R + d_R) = 1/f. \tag{X.7.51}$$

In the limit $D_R \to \infty$ we find from (7.51) that

$$D_L \to f - d_L. \tag{X.7.52}$$

The right side of (7.52) is a physically possible value for D_L provided

$$f - d_L \geq 0 \Leftrightarrow f \geq d_L. \tag{X.7.53}$$

We conclude that

$$m \to +\infty \text{ as } D_R \to \infty \tag{X.7.54}$$

provided (7.53) holds. In this case the magnification can be made arbitrarily large in magnitude.

How small can the magnification be? Can it be made vanishingly small for physical values of D_L and D_R? From the second line of (7.50) we see that

$$m = 0 \Leftrightarrow D_R = f - d_R \tag{X.7.55}$$

so that D_R is non-negative (physically possible) provided

$$f - d_R \geq 0 \Leftrightarrow f \geq d_R. \tag{X.7.56}$$

But does the D_R given by (7.55) lead to a physical value of D_L when employed in (7.51)? Inserting the right side of (7.55) into (7.51) yields the relation

$$1/(D_L + d_L) + 1/f = 1/f \Rightarrow 1/(D_L + d_L) = 0 \Rightarrow D_L = +\infty. \tag{X.7.57}$$

We conclude that $D_R \to (f - d_R)$ and $D_L \to +\infty$ is consistent with imaging and results in vanishing magnification. This conclusion is valid provided (7.56) holds.

Thin-Lens Approximation

Eventually we will want to compute to compute G' and hence R'. According to (7.26) this computation involves the product of 7 matrices, and is therefore best done numerically. But before doing so it would be useful to have an approximate result to get some feeling for the expected nature of the exact result. This can be done by treating the two lenses in the thin-lens approximation. That is, we will set $t = 0$ in (7.26). When this is done, what remains is to compute the map $\bar{\mathcal{R}}'$ defined by the product

$$\bar{\mathcal{R}}' = \exp[2\beta_2^1(1-n) : q^2 :] \exp[-(D/2) : p^2 :] \exp[-2\beta_2^1(1-n) : q^2 :]. \tag{X.7.58}$$

Since the Lie transformation $\exp[2\beta_2^1(1-n) : q^2 :]$ is that for a thin lens, let us make the correspondence

$$\exp[2\beta_2^1(1-n) : q^2 :] \leftrightarrow \begin{pmatrix} 1 & 0 \\ -1/F & 1 \end{pmatrix} \tag{X.7.59}$$

where F is the focal length of the first lens in the thin-lens approximation. So doing yields the relation

$$-1/F = 4\beta_2^1(1-n). \tag{X.7.60}$$

Note, according to our prescription that the first lens in the doublet be focusing, recall (7.8), it follows that

$$F > 0. \tag{X.7.61}$$

Now, to compute the matrix $\bar{G}'$ associated with $\bar{\mathcal{R}}'$, we only need to compute the matrix product

$$\begin{aligned} \bar{G}' &= \begin{pmatrix} 1 & 0 \\ 1/F & 1 \end{pmatrix} \begin{pmatrix} 1 & D \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1/F & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 1/F & 1 \end{pmatrix} \begin{pmatrix} 1-D/F & D \\ -1/F & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1-D/F & D \\ -D/F^2 & 1+D/F \end{pmatrix}. \end{aligned} \tag{X.7.62}$$

What can we conclude in the thin-lens approximation? Let us apply the normal-form procedure to $\bar{G}'$. First, in analogy to (7.35) and from (7.62), we see that

$$\bar{G}'' = \begin{pmatrix} 1 & 0 \\ \bar{G}'_{21} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -D/F^2 & 1 \end{pmatrix}. \tag{X.7.63}$$

Also we know that $D > 0$ and therefore

$$\bar{G}'_{21} = -D/F^2 < 0. \tag{X.7.64}$$

That is, in the thin-lens approximation, the net effect of the doublet is *focusing*.[8] In analogy to (7.37) we make the definition

$$1/\bar{f} = -\bar{G}'_{21} \Leftrightarrow \bar{f} = -1/(\bar{G}'_{21}) = F^2/D. \tag{X.7.65}$$

[8] Note that this conclusion holds no matter the sign F. This observation is the essence of *strong focussing* in Accelerator Physics applications. Although only obtained here in the thin-lens approximation, there are analogous results for thick lenses and thick magnetic elements.

That is, (7.63) can be rewritten in the form

$$G'' = \begin{pmatrix} 1 & 0 \\ -1/\bar{f} & 1 \end{pmatrix} \tag{X.7.66}$$

with $\bar{f}$ given by (7.65). Moreover, from (7.31) and (7.32) we find that

$$\bar{d}_L = (\bar{G}'_{22} - 1)/\bar{G}'_{21} = (D/F)/(-D/F^2) = -F \tag{X.7.67}$$

$$\bar{d}_R = (\bar{G}'_{11} - 1)/\bar{G}'_{21} = -(D/F)/(-D/F^2) = F. \tag{X.7.68}$$

Let us use these thin-lens results to examine what magnifications can be achieved in the thin-lens approximation. According to (7.54) the magnification can be made arbitrarily large in magnitude providing (7.53) holds. But, in the thin-lens approximation, we find that

$$\bar{f} - \bar{d}_L = \bar{f} + F > 0 \tag{X.7.69}$$

because both $\bar{f}$ and F are positive. Thus, if the thin-lens approximation is to be believed, the magnification can be made arbitrarily large in magnitude. What about making the magnification arbitrarily small? According to (7.55) the magnification can be made to vanish providing (7.56) holds. But, in the thin-lens approximation we find

$$\bar{f} - \bar{d}_R = F^2/D - F. \tag{X.7.70}$$

Therefore

$$\bar{f} - \bar{d}_R \geq 0 \Leftrightarrow F^2/D - F \geq 0 \Leftrightarrow F/D - 1 \geq 0 \Leftrightarrow F \geq D \Leftrightarrow D \leq F. \tag{X.7.71}$$

The inequality on the far right side of (7.71) provides a design criterion: D must not be too large. And if this criterion is met and the thin-lens approximation is to be believed, then arbitrarily small magnification can also be achieved.

Case for which Both Lenses Have Finite Thickness

We have verified that a doublet can be designed to have satisfactory paraxial performance in the thin-lens approximation. The next step is to verify that a doublet system can be found that has satisfactory paraxial performance when the two lenses have finite thickness. As an example, we suppose the doublet has the parameter values

$$\beta_2^1 = \text{to be determined by fitting}, \tag{X.7.72}$$

$$n = 1.5, \tag{X.7.73}$$

$$t = 0.75, \tag{X.7.74}$$

$$D = 4.75, \tag{X.7.75}$$

and that the remaining β_2^j are given by (7.8), (7.9), and (7.11).

Fitting the Focal Length

We now compute G' while varying β_2^1 to achieve some desired value for the focal length as given by (7.37). For example, suppose we wish/aim to have

$$f = 20 \Leftrightarrow G'_{21} = -.05. \tag{X.7.76}$$

For the aim ($G'_{21} = -.05$) we find for the doublet that

$$G' = \begin{pmatrix} 4.69171E-01 & 5.74703E+00 \\ -5.00000E-02 & 1.51895E+00 \end{pmatrix} \tag{X.7.77}$$

and

$$\beta_2^1 = 5.0003711901875095E-02. \tag{X.7.78}$$

Verification that Any Magnification can be Achieved

From (7.77) we see that

$$-1/f = G'_{21} = -5.00000E-02 \Leftrightarrow f = 20 \text{ as desired,} \tag{X.7.79}$$

$$d_L = (G'_{22} - 1)/G'_{21} = (1.51895 - 1.0)/(-.05) = -10.3790424, \tag{X.7.80}$$

$$d_R = (G'_{11} - 1)/G'_{21} = (0.469171 - 1.0)/(-.05) = 10.6165777. \tag{X.7.81}$$

Consequently,

$$f - d_L = 20 + 10.3790424 = 30.3790424 > 0 \tag{X.7.82}$$

and

$$f - d_R = 20 - 10.6165777 = 9.3834224 > 0. \tag{X.7.83}$$

We conclude that for this doublet there are physical/positive values of D_L and D_R for which any desired (negative) value of m can be achieved for the full system.

Selection of Magnification m and Determination of Lengths D_L and D_R

From (7.47) and (7.49) we see that

$$-D_L^e/f + 1 = 1/m \tag{X.7.84}$$

from which it follows that

$$D_L = f[1 - (1/m)] - d_L. \tag{X.7.85}$$

And from (7.50) we see that

$$D_R = f(1 - m) - d_R. \tag{X.7.86}$$

Let us now select some value for m. For example, suppose we select the value

$$m = -0.5 = -1/2. \tag{X.7.87}$$

Then D_L and D_R have the values

$$D_L = 20[1 + 2] + 10.3790424 = 70.3790424 \tag{X.7.88}$$

and

$$D_R = 20(3/2) - 10.6165777 = 19.3834223. \tag{X.7.89}$$

And, for the full system consisting of the device plus leading and trailing drifts, we find that the matrix R associated with the full linear map $\mathcal{R}$ has the related matrix G given by

$$G = \begin{pmatrix} -5.00000E-01 & 0.00000E+00 \\ -5.00000E-02 & -2.00000E+00 \end{pmatrix}. \tag{X.7.90}$$

Evidently, the full map $\mathcal{M}$ is imaging in paraxial approximation because $G_{12} = 0$, and the magnification is

$$m = G_{11} = -0.50, \tag{X.7.91}$$

as desired.

Vanishing of Petzval

What can be said about the third-order aberrations of this system consisting of the doublet plus leading and trailing drifts ? The f_4 entries for $\mathcal{M}$ are listed below.

```
Exhibit *.  Nonzero elements in generating polynomial assuming spherical lenses :

f( 33)=f( 20 00 01 )=-9.65181400681940E-02
f( 38)=f( 11 00 01 )=  1.9303628013639
f( 53)=f( 02 00 01 )= -18.561180782345
f( 67)=f( 00 20 01 )=-9.65181400681940E-02
f( 70)=f( 00 11 01 )=  1.9303628013639
f( 76)=f( 00 02 01 )= -18.561180782345
f( 83)=f( 00 00 03 )= -19.536145932398
f( 84)=f( 40 00 00 )=-2.58550347222222E-03
f( 85)=f( 31 00 00 )= 0.26342013888889
f( 90)=f( 22 00 00 )= -10.101302083333
f( 95)=f( 20 20 00 )=-5.17100694444444E-03
f( 96)=f( 20 11 00 )= 0.26342013888889
f( 99)=f( 20 02 00 )= -3.3671006944444
f(104)=f( 20 00 02 )=-0.13135810207182
f(105)=f( 13 00 00 )=  172.00868055556
f(110)=f( 11 20 00 )= 0.26342013888889
f(111)=f( 11 11 00 )= -13.468402777778
f(114)=f( 11 02 00 )=  172.00868055556
f(119)=f( 11 00 02 )=  4.3469930621546
f(140)=f( 04 00 00 )= -1098.2855902778
f(145)=f( 02 20 00 )= -3.3671006944444
f(146)=f( 02 11 00 )=  172.00868055556
f(149)=f( 02 02 00 )= -2196.5711805556
f(154)=f( 02 00 02 )= -42.459483682532
f(175)=f( 00 40 00 )=-2.58550347222222E-03
f(176)=f( 00 31 00 )= 0.26342013888889
f(179)=f( 00 22 00 )= -10.101302083333
f(184)=f( 00 20 02 )=-0.13135810207182
f(185)=f( 00 13 00 )=  172.00868055556
f(190)=f( 00 11 02 )=  4.3469930621546
```

```
f(195)=f( 00 04 00 )= -1098.2855902778
f(200)=f( 00 02 02 )= -42.459483682532
f(209)=f( 00 00 04 )= -24.897680899171
```

Calculation shows that for this f_4

$$\langle {}^4\chi_0^0\,,\, f_4\rangle = 0, \tag{X.7.92}$$

thereby indicating that the Petzval aberration indeed vanishes as desired. Alternatively, using (2.42), (2.43), and (3.23), we expect that

$$0 = (2C - 4D) = f(1,1,1,1) - 4f(0,2,2,0) = 0 \Leftrightarrow f(111) - 4f(145) = 0. \tag{X.7.93}$$

Examination of the values for $f(111)$ and $f(145)$ in the list of f_4 values above shows that the relation on the right side of (7.87) is indeed satisfied.

Evidently many of the f_4 entries listed above are nonzero. Calculation shows that there are the results

$$\langle {}^4\chi_2^2\,,\, f_4\rangle = *, \tag{X.7.94}$$

$$\langle {}^4\chi_1^2\,,\, f_4\rangle = *, \tag{X.7.95}$$

$$\langle {}^4\chi_0^2\,,\, f_4\rangle = *, \tag{X.7.96}$$

$$\langle {}^4\chi_{-1}^2\,,\, f_4\rangle = *, \tag{X.7.97}$$

$$\langle {}^4\chi_{-2}^2\,,\, f_4\rangle = *. \tag{X.7.98}$$

In view of (3.15) the scalar product results (7.92) and (7.94) through (7.98) specify f_4 completely because of the assumption/imposition of axial symmetry.

We have examined a particular system which is specified by the parameter values given by (7.73) through (7.75), (7.8), (7.9), (7.11) and the requirements (7.76) [which led to (7.78)] and (7.91). For this system we have verified that the third-order Petzval aberration vanishes, as desired. But upon reflection we see that, no matter what parameter values and requirements are imposed, the third-order Petzval aberration will vanish as long as (7.6) is satisfied.

Elimination of Remaining Third-Order Aberrations

Ideally we would like to have vanishing values for *all* the coefficients A through E appearing in (2.30) through (2.34). There are five such coefficients, but we have already caused the Petzval combination $(C-2D)$ to vanish thereby leaving four more goals to be met. We also observe that there are four surfaces S^1 through S^4 for which values β_4^1 through β_4^4 can be assigned. Can they be set in such a way that all the f_4 save for the F terms vanish? We will find that the answer is *yes*, but the matter is subtle.

For a spherical surface of radius r there are the relations

$$\beta_2 = 1/(2r) \tag{X.7.99}$$

and

$$\beta_4 = 1/(8r^3) \tag{X.7.100}$$

so that

$$\beta_4 = (\beta_2)^3. \tag{X.7.101}$$

For the calculation that produced the f_4 in Exhibit *, the β_4^j values were set in such a way that the relation (7.100) was satisfied for all four surfaces. Equivalently, the surfaces were assumed to be spherical.

But what happens if the β_4^j values are instead set to zero? In this case, because all surfaces are now parabolas of revolution through fourth order, one might hope that third-order aberrations would be reduced. Below are the relevant scalar products for the f_4 found in this case:

$$\langle {}^4\chi_0^0\,,\ f_4\rangle = 0, \tag{X.7.102}$$

$$\langle {}^4\chi_2^2\,,\ f_4\rangle = *, \tag{X.7.103}$$

$$\langle {}^4\chi_1^2\,,\ f_4\rangle = *, \tag{X.7.104}$$

$$\langle {}^4\chi_0^2\,,\ f_4\rangle = *, \tag{X.7.105}$$

$$\langle {}^4\chi_{-1}^2\,,\ f_4\rangle = *, \tag{X.7.106}$$

$$\langle {}^4\chi_{-2}^2\,,\ f_4\rangle = *. \tag{X.7.107}$$

Looking at (7.102), we see that the Petzval still vanishes as before. But this is not surprising since we know that the Petzval is independent of β_4, and the condition (7.6) is still met because the paraxial parameters have not been changed. What about the remaining entries? Comparison of the entries in (7.94) through (7.98) with those in (7.103) through (7.107) shows that the latter are still sizable despite all surfaces being parabolic. Why is this? First of all, surface maps are not the only source of aberrations. As we see from *, transfer maps for drifts involve ${}^4\chi_2^2$, ${}^6\chi_3^3\cdots$ generators. And these generators can turn into ${}^4\chi_m^2$, ${}^6\chi_m^3\cdots$ generators under the action of $\exp(:f_2:)$ maps that occur/act in the course of concatenation. Second, inspection of * for example, shows that for a surface map f_4 generator there is the expansion

$$f_4 = *{}^4\chi_0^0 + *{}^4\chi_0^2 + *{}^4\chi_{-1}^2 + *{}^4\chi_{-2}^2, \tag{X.7.108}$$

and only the ${}^4\chi_{-2}^2$ term depends on β_4 but does not vanish when $\beta_4 = 0$. See Exercise *. Evidently there are non-vanishing ${}^4\chi_m^2$ terms even when $\beta_4 = 0$.

What to do now that a simple strategy has been explored and found wanting? Since there are four goals to be achieved and four β_4^j parameters to be set, we might try varying the β_4^j to meet the goals

$$\langle {}^4\chi_m^2\,,\ f_4\rangle = 0 \text{ for } m = 2,1,0,-1. \tag{X.7.109}$$

Experience shows that this strategy succeeds. Below are values for the f_4 when the β_4^j are optimally set. Evidently all entries vanish save for the coefficients of q_x^4, $q_x^2q_y^2$, and q_y^4. All third-order aberrations that affect image formation have been caused to vanish. This was accomplished by selecting the values

$$\beta_4^1 =, \tag{X.7.110}$$

$$\beta_4^2 =, \tag{X.7.111}$$

$$\beta_4^3 =, \tag{X.7.112}$$

$$\beta_4^4 = . \tag{X.7.113}$$

It can be shown that the calculation leading to the satisfaction of (7.109), after the Petzval has already been eliminated, amounts to a linear fitting operation, and consequently the values (7.110) through (7.113) are unique.

Entries in f_4 when the β_4^j are given the values (7.110) through (7.113).

```
nonzero elements in generating polynomial are :

 f( 33)=f( 20 00 01 )=-9.65181400681940E-02
 f( 38)=f( 11 00 01 )=  1.9303628013639
 f( 53)=f( 02 00 01 )= -18.561180782345
 f( 67)=f( 00 20 01 )=-9.65181400681940E-02
 f( 70)=f( 00 11 01 )=  1.9303628013639
 f( 76)=f( 00 02 01 )= -18.561180782345
 f( 83)=f( 00 00 03 )= -19.536145932398
 f( 84)=f( 40 00 00 )=-2.58550347222222E-03
 f( 85)=f( 31 00 00 )= 0.26342013888889
 f( 90)=f( 22 00 00 )= -10.101302083333
 f( 95)=f( 20 20 00 )=-5.17100694444444E-03
 f( 96)=f( 20 11 00 )= 0.26342013888889
 f( 99)=f( 20 02 00 )= -3.3671006944444
 f(104)=f( 20 00 02 )=-0.13135810207182
 f(105)=f( 13 00 00 )=  172.00868055556
 f(110)=f( 11 20 00 )= 0.26342013888889
 f(111)=f( 11 11 00 )= -13.468402777778
 f(114)=f( 11 02 00 )=  172.00868055556
 f(119)=f( 11 00 02 )=  4.3469930621546
 f(140)=f( 04 00 00 )= -1098.2855902778
 f(145)=f( 02 20 00 )= -3.3671006944444
 f(146)=f( 02 11 00 )=  172.00868055556
 f(149)=f( 02 02 00 )= -2196.5711805556
 f(154)=f( 02 00 02 )= -42.459483682532
 f(175)=f( 00 40 00 )=-2.58550347222222E-03
 f(176)=f( 00 31 00 )= 0.26342013888889
 f(179)=f( 00 22 00 )= -10.101302083333
 f(184)=f( 00 20 02 )=-0.13135810207182
 f(185)=f( 00 13 00 )=  172.00868055556
 f(190)=f( 00 11 02 )=  4.3469930621546
 f(195)=f( 00 04 00 )= -1098.2855902778
 f(200)=f( 00 02 02 )= -42.459483682532
 f(209)=f( 00 00 04 )= -24.897680899171
```

Elimination of First Four Leading Fifth-Order Aberrations

The values of the β_6^j are also available to be set thereby attempting to cause four fifth-order aberrations to vanish. On the assumption that it is the sixth-order monomials with the highest powers of p_x and p_y that are the most damaging for image formation, we may attempt to cause the coefficients of ${}^6\chi_3^3$, ${}^6\chi_2^3$, ${}^6\chi_1^3$, and ${}^6\chi_0^3$ in f_6 to vanish by a suitable choice of the β_6^j. This goal can also be achieved. Below are the entries in f_6 for two cases. In the first case the β_6^j are set to zero [and the β_4^j are set to the values (7.88) through (7.91)]. In the second case the β_6^j are set to cause the coefficients of ${}^6\chi_3^3$ through ${}^6\chi_0^3$ in f_6 to vanish. These β_6^j have the values

$$\beta_6^1 =, \tag{X.7.114}$$

$$\beta_6^2 =, \tag{X.7.115}$$

$$\beta_6^3 =, \tag{X.7.116}$$

$$\beta_6^4 = . \tag{X.7.117}$$

Entries in f_6 when the β_6^j =0.
Entries in f_6 when the β_6^j have the values (7.114) through (7.117).

Corrector Strengths Depend on Magnification

Ray Traces

X.7.2 Aberration Corrected Hubble Telescope

Exercises

X.8 Inclusion of Chromatic Effects

X.9 Possibly Complementary Approaches

X.9.1 The Constant Index Case

Suppose the lenses of an optical system are all made of constant (not graded) index materials. In that case computers can numerically trace a large number of rays through the system in a very short time simply by invoking Snell's law at each interface.

In that case fairly rapid ray traces can also be performed using Lie methods. The interface maps $\mathcal{S}$ described in the Technical Report *Foundations of a Lie* $\cdots$ can be computed in milliseconds using equations (7.46a) through (7.46d) of that paper and combined in further milliseconds with surrounding transit maps (6.10) to yield the f_m displayed in the first line of equation (2.44) below:

$$\begin{aligned} w_\alpha^f &= \mathcal{M} w_\alpha^i = \{\exp(:f_2:)\exp(:f_4:)\exp(:f_6:)\exp(:f_8:)\cdots\} w_\alpha^i \\ &= g_1^\alpha(\boldsymbol{w}^i) + g_3^\alpha(\boldsymbol{w}i) + g_5^\alpha(\boldsymbol{w}^i) + g_7^\alpha(\boldsymbol{w}^i)\cdots . \end{aligned} \tag{X.9.1}$$

Then the homogeneous polynomials g_m^α displayed in the second line of (2.44) can be found in a few milliseconds more. The terms $g_1^\alpha(\boldsymbol{w}^i)$ describe the paraxial behavior of the optical system, and the terms $g_3^\alpha(\boldsymbol{w}^i)$, $g_5^\alpha(\boldsymbol{w}^i)$, $\cdots$ describe ever higher degree departures from paraxial behavior. [Note that the g_m^α are *not* independent because of the symplectic condition (1.16). Therefore they are ill suited for use in optimization/fitting procedures. By contrast, the f_m are independent, and any choice for them is consistent with the symplectic condition.] Finally, these polynomials $g_m^\alpha(\boldsymbol{w}^i)$ can be evaluated rapidly for any collection of initial conditions $\boldsymbol{w}^i$ to find the associated finial conditions $\boldsymbol{w}^f$. And if ray coordinates are desired at intermediate positions, they may be found by performing ray traces at intermediate positions as the full end-to-end map $\mathcal{M}$ is being built up.

Of course, this use of Lie methods presumes that the series in the second line of (2.44) is convergent and that to good approximation terms beyond some degree can be neglected. But this can be checked using the Snell's law ray trace. For example, let $w_\alpha^{fs\ell\mathrm{rt}}$ denote the result of a Snell's law ray trace for some initial condition $\boldsymbol{w}^i$ and let $w_\alpha^{f[7]}$ be the associated through seventh order result

$$w_\alpha^{f[7]} = g_1^\alpha(\boldsymbol{w}^i) + g_3^\alpha(\boldsymbol{w}^i) + g_5^\alpha(\boldsymbol{w}^i) + g_7^\alpha(\boldsymbol{w}^i). \tag{X.9.2}$$

Then we expect the result

$$w_\alpha^{f[7]} = w_\alpha^{fs\ell\mathrm{rt}} + O(|\boldsymbol{w}^i|^9). \tag{X.9.3}$$

The validity of (2.46) can be verified by comparing $w_\alpha^{f[7]}$ and $w_\alpha^{fs\ell\mathrm{rt}}$ for a variety of initial conditions $\boldsymbol{w}^i$ as $\boldsymbol{w}^i \to 0$.

Assuming that the series in the second line of (2.44) is convergent and that to good approximation terms beyond some degree can be neglected, it might be illuminating/interesting to monitor the Lie generators f_m during the course of a ray-trace-driven design/optimization process to observe, from a Lie perspective, what is being accomplished during the process. If it is observed, for example, that some particular f_m or set of f_m is being driven to zero, then one might experiment with including their values as part of a merit function.

In some settings, at least in the context of magnetic optics, it is useful to replace an optimization process by a *fitting* process is which several parameters are varied to drive several or all "offensive" Lie generators to zero. (To verify that some set of f_m is offensive, one can perform Lie algebraic ray traces with some of the f_m set to zero to see what effect that has on the $\boldsymbol{w}^f$ so computed. Note that so doing does not violate the symplectic condition.) For example, it is possible to design a complete third-order achromat in which the strengths of three quadrupoles, three sextuples, and eight octupoles are varied to set/remove various f_m in a particular basis, and all remaining f_m are cancelled by repetitive symmetry. (In magnetic optics parlance, an achromat bends a charged-particle beam, but otherwise acts as the identity map.) In so doing, 203 conditions are met. (Magnetic optical systems generally do not have axial symmetry and, therefore, there are many more aberrations to be corrected.) It is unlikely that this goal could have been achieved with an optimization program.

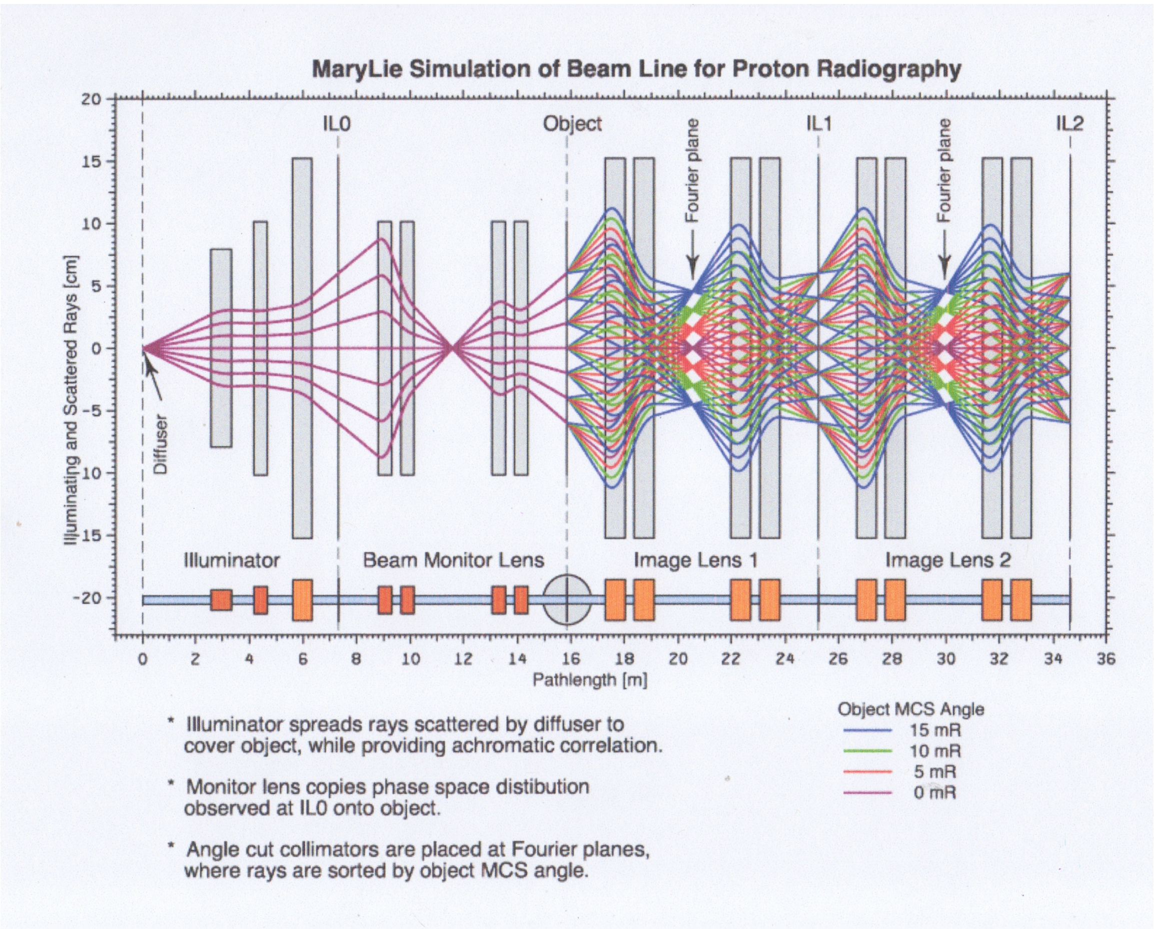
It is also possible to carry out procedures in which fitting loops are inside an optimization loop. This was done in connection with some octupole-corrected Los Alamos charged-particle beam projects.

X.9.2 The Graded Index Case

Suppose one wishes to employ lenses made of graded index material. (Something analogous is always the case in magnetic optics since magnetic fields are position dependent.) In that case one ray-tracing possibility is to employ direct numerical integration of the equations of motion (1.34) and (1.35) associated with H, perhaps with the aid of symplectic integrators to ensure maintenance of the symplectic condition. But this process is slow if many rays are to be traced with high accuracy. Moreover, extraction of aberration data from ray data, if desired, is subject to the numerical errors associated with high-order numerical differentiation.

A second option is to employ Lie methods. Suppose all the graded-index lenses have *flat faces*. In that case there are equations of motion for the Lie generators f_m that can be integrated to yield a Lie representation for the end-to-end $\mathcal{M}$. (See Chapter 10 of *Lie Methods for Nonlinear* $\cdots$.) And, once $\mathcal{M}$ is found in Lie form, numerous ray traces can be carried out rapidly. (And fitting/optimization can be carried out using both the values of the Lie generators and ray-trace results.) Figures 2 and 3 illustrate ray traces for two charged-particle beam devices carried out in this fashion.

The treatment of graded-index lenses with curved faces is more complicated, but some progress has also been made in handling this problem. And again, once $\mathcal{M}$ is found in Lie form, numerous ray traces can be carried out rapidly. And fitting/optimization can again be carried out using both the values of the Lie generators and ray-trace results.



* Illuminator spreads rays scattered by diffuser to cover object, while providing achromatic correlation.

* Monitor lens copies phase space distibution observed at IL0 onto object.

* Angle cut collimators are placed at Fourier planes, where rays are sorted by object MCS angle.

Figure X.9.1: Lie Algebraically Designed Magnetic Optical System for Fast Dynamic (Nanosecond) Imaging of Dense Objects Using High-Energy Proton Beams.

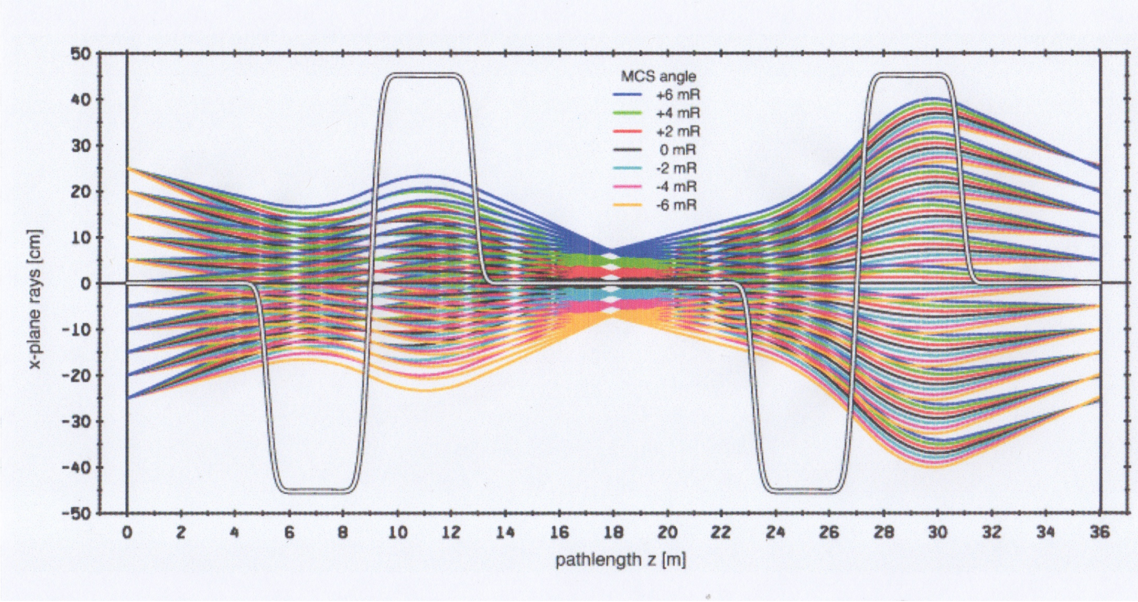


Figure X.9.2: Ray Trace of Soft-Edge Lie Algebraically Designed Super Lens for 50 GeV protons.

Bibliography

[1] D. Dilworth, *Lens Design*, IOP Publishing Ltd (2018). See the Web site http://iopscience.iop.org/book/978-0-7503-1611-8

[2] A. Dragt, "Lie algebraic theory of geometrical optics and optical aberrations", *J. Opt. Soc. Am.* **72** p 372 (1982).

[3] A. Dragt and E. Forest, *Foundations of a Lie Algebraic Theory of Geometrical Optics*, University of Maryland Physics Department Technical Report (1985).

[4] A. Dragt, E. Forest, and K. B. Wolf, "Foundations of a Lie algebraic theory of geometrical optics", published in *Lie Mehods in Optics*, p 105, J. S. Mondragón and K. B. Wolf, edit., Springer-Verlag (1986).

[5] A. Torre, *Linear Ray and Wave Optics in Phase Space: Bridging Ray and Wave Optics via the Wigner Phase-Space Picture*, Elsevier Science (2005).

Appendix Y

Relation between the Classical Poisson Bracket Lie Algebra and the Quantum Commutator-Based Lie Algebra

Overview

This appendix explores the relation between the classical Poisson bracket Lie algebra and the quantum commutator-based Lie algebra. Lie methods are used to construct bases for each. It is found that the basis in the quantum case coincides with the Weyl basis. Next a natural correspondence is set up between the classical and quantum bases. Finally, these bases are used to determine the structure constants for the classical and quantum Lie algebras. It is found that many, *but not all*, of the structure constants for the two Lie algebras are the same. In particular, it is found that the classical Lie algebra is a contraction of the quantum Lie algebra. Conversely, the quantum Lie algebra is a deformation of the classical Lie algebra.

Y.1 Classical Polynomial Basis

For introductory simplicity, work with a two-dimensional phase space with variables q and p. (Higher-dimensional cases can be treated in a similar manner.) Let f and g be any functions of the phase-space variables. Introduce a *classical mechanical* Lie product $[*,*]_{\rm cm}$ among such functions by use of the Poisson bracket,

$$[f,g]_{\rm cm} = (\partial f/\partial q)(\partial g/\partial p) - (\partial f/\partial p)(\partial g/\partial q). \tag{Y.1.1}$$

Given any phase-space function f, define an associated *Lie operator*, denoted by $:f:$, by the rule

$$:f:g = [f,g]_{\rm cm}. \tag{Y.1.2}$$

In view of (1.1) and (1.2), this is the usual definition of $:f:$, but presented with a slightly different notation.

Introduce basis monomials $a_{r-s,s}$ by the rule

$$a_{r-s,s}(q,p) = [(r-s)!/r!] : -p^2/2 :^s q^r. \tag{Y.1.3}$$

So doing yields, for the first few monomials, the results

$$a_{00} = 1; \tag{Y.1.4}$$

$$a_{10} = q, \tag{Y.1.5}$$

$$a_{01} = p; \tag{Y.1.6}$$

$$a_{20} = q^2, \tag{Y.1.7}$$

$$a_{11} = qp, \tag{Y.1.8}$$

$$a_{02} = p^2; \tag{Y.1.9}$$

$$a_{30} = q^3, \tag{Y.1.10}$$

$$a_{21} = q^2p, \tag{Y.1.11}$$

$$a_{12} = qp^2, \tag{Y.1.12}$$

$$a_{03} = p^3; \tag{Y.1.13}$$

$$a_{40} = q^4, \tag{Y.1.14}$$

$$a_{31} = q^3p, \tag{Y.1.15}$$

$$a_{22} = q^2p^2, \tag{Y.1.16}$$

$$a_{13} = qp^3, \tag{Y.1.17}$$

$$a_{04} = p^4. \tag{Y.1.18}$$

The degree of a monomial a_{rs} is given by the sum $(r+s)$, and we refer to the monomials of a fixed degree as a forming a *ladder*. Within a ladder and up to multiplicative factors, $: -p^2/2 :$ acts on a_{rs} as an operator that lowers r and raises s. Indeed, there is the recursion relation

$$: -p^2/2 : a_{r,s} = r a_{r-1,s+1}. \tag{Y.1.19}$$

Similarly, within a ladder and up to multiplicative factors, $: q^2/2 :$ acts as an operator that raises r and lowers s.

Y.2 Quantum Polynomial Basis

Let Q and P be the quantum-mechanical counterparts of q and p. They obey the commutation rule

$$\{Q,P\} = QP - PQ = i\hbar I \tag{Y.2.1}$$

where $\hbar$ is the reduced Planck's constant $h/(2\pi)$, and which we will eventually view as an adjustable parameter in Section 4. Suppose $F(Q,P)$ and $G(Q,P)$ are any polynomial functions of Q and P with some ordering rule for products of Q's and P's. Given these functions, define a *quantum mechanical* Lie product $[*,*]_{\rm qm}$ by the rule

$$[F,G]_{\rm qm} = (i\hbar)^{-1}\{F,G\}. \tag{Y.2.2}$$

Here we note two facts: First, $[*,*]_{\rm qm}$ is a Lie product because the commutator is a Lie product. Second, if F and G are Hermitian, $[F,G]_{\rm qm}$ will also be Hermitian. Indeed, from the definition (2.2) there is the relation

$$[F,G]^\dagger_{\rm qm} = -(i\hbar)^{-1}\{F,G\}^\dagger. \tag{Y.2.3}$$

But, there is also the relation

$$\{F,G\}^\dagger = (FG-GF)^\dagger = (FG)^\dagger - (GF)^\dagger = G^\dagger F^\dagger - F^\dagger G^\dagger = GF - FG = -\{F,G\}. \tag{Y.2.4}$$

Consequently, there is the advertised result

$$[F,G]^\dagger_{\rm qm} = [F,G]_{\rm qm}. \tag{Y.2.5}$$

Within the quantum mechanical context, define an operator $: -P^2/2 :$ by the rule

$$: -P^2/2 : G = [-P^2/2, G]_{\rm qm}. \tag{Y.2.6}$$

Powers of $: -P^2/2 :$ are defined by the rules

$$: -P^2/2 :^0 G = G, \tag{Y.2.7}$$

$$: -P^2/2 :^2 G = [-P^2/2, [-P^2/2, G]_{\rm qm}]_{\rm qm},\ \text{etc.} \tag{Y.2.8}$$

Next, in analogy to the construction (1.3), use $: -P^2/2 :$ to define polynomials $A_{r-s,s}$ by the rule

$$A_{r-s,s}(Q,P) = [(r-s)!/r!] : -P^2/2 :^s Q^r. \tag{Y.2.9}$$

Here use is to be made of the relation (2.1) to evaluate the commutators that occur, and we presume that

$$Q^0 = I. \tag{Y.2.10}$$

Doing so gives, for the first few values of r and s, the results

$$A_{00} = I; \tag{Y.2.11}$$

$$A_{10} = Q, \tag{Y.2.12}$$

$$A_{01} = P; \tag{Y.2.13}$$

$$A_{20} = Q^2, \tag{Y.2.14}$$

$$A_{11} = (QP + PQ)/2, \tag{Y.2.15}$$

$$A_{02} = P^2; \tag{Y.2.16}$$

$$A_{30} = Q^3, \tag{Y.2.17}$$

$$A_{21} = (Q^2P + PQ^2)/2, \tag{Y.2.18}$$

$$A_{12} = (QP^2 + P^2Q)/2, \tag{Y.2.19}$$

$$A_{03} = P^3; \tag{Y.2.20}$$

$$A_{40} = Q^4, \tag{Y.2.21}$$

$$A_{31} = (Q^3P + PQ^3)/2, \tag{Y.2.22}$$

$$A_{22} = (QP + PQ)^2/6 + (Q^2P^2 + P^2Q^2)/6, \tag{Y.2.23}$$

$$A_{13} = (QP^3 + P^3Q/)2, \tag{Y.2.24}$$

$$A_{04} = P^4. \tag{Y.2.25}$$

Note that the A_{rs} are Hermitian, as expected. Also, the A_{rs} are polynomials in the *Weyl* basis. That is, products of Q's and P's are Weyl ordered. Moreover, this ordering of constituents has not been achieved by demanding/imposing permutation symmetry, but instead arises *naturally* from a Lie algebraic procedure.

The degree of a polynomial A_{rs} is again given by the sum $(r+s)$, and we again refer to the polynomials of a fixed degree as a forming a ladder. Within a ladder and up to multiplicative factors, $: -P^2/2 :$ acts on A_{rs} as an operator that lowers r and raises s. Indeed, there is the recursion relation

$$: -P^2/2 : A_{r,s} = rA_{r-1,s+1}. \tag{Y.2.26}$$

Similarly, within a ladder and up to multiplicative factors, $: Q^2/2 :$ acts as an operator that raises r and lowers s.

Y.3 A Natural Correspondence between Classical and Quantum Bases

How can we set up a natural correspondence between the classical and quantum bases? In so doing, by linearity, we will also set up a natural correspondence between the classical Lie algebra of phase-space functions with Lie product $[*,*]_{\rm cm}$ and the quantum Lie algebra of polynomials in Q and P with Lie product $[*,*]_{\rm qm}$. We will call these Lie algebras $\mathcal{L}_{\rm cm}$ and $\mathcal{L}_{\rm qm}$.

First, it is natural to set up the correspondences

$$1 \leftrightarrow I, \tag{Y.3.1}$$

$$q^n \leftrightarrow Q^n, \tag{Y.3.2}$$

$$p^n \leftrightarrow P^n. \tag{Y.3.3}$$

But then, because of the Lie algebraic similarity of the definitions (1.3) and (2.9), it is also natural to set up the correspondences

$$a_{rs}(q,p) \leftrightarrow A_{rs}(Q,P), \tag{Y.3.4}$$

for which (3.1) through (3.3) are special cases.

Y.4 Relation between the Lie Algebras $\mathcal{L}_{\rm cm}$ and $\mathcal{L}_{\rm qm}$

The Lie algebra $\mathcal{L}_{\rm cm}$ has structure constants $c^{tu}_{jk;rs}$ defined by the relation

$$[a_{jk}, a_{rs}]_{\rm cm} = \sum_{tu} c^{tu}_{jk;rs}\, a_{tu}. \tag{Y.4.1}$$

Similarly, the Lie algebra $\mathcal{L}_{\rm qm}$ has structure constants $C^{tu}_{jk;rs}$ defined by the relation

$$[A_{jk}, A_{rs}]_{\rm qm} = \sum_{tu} C^{tu}_{jk;rs}\, A_{tu}. \tag{Y.4.2}$$

How do the structure constants $c^{tu}_{jk;rs}$ and $C^{tu}_{jk;rs}$ compare? They are not all the same, and therefore the two Lie algebras $\mathcal{L}_{\rm cm}$ and $\mathcal{L}_{\rm qm}$ are not manifestly the same.[1] However, *many* of the structure constants are the same. In particular, there are the equalities

$$C^{tu}_{jk;rs} = c^{tu}_{jk;rs} \text{ for } j+k \le 2. \tag{Y.4.3}$$

One consequence of (4.3) is that the subalgebras generated by classical and quantum polynomials of degree ≤ 2 are identical, and are in fact the Lie algebra $isp(2,\mathbb{R})$, the Lie algebra of the *inhomogeneous* symplectic group in two dimensions over the real field.[2]

Another consequence is the special role played by the a_{jk} and the A_{jk} with $j+k=2$. Both generate the Lie algebra for $sp(2,\mathbb{R})$. We have already seen that p^2 and q^2, and their quantum counterparts, act as raising and lowering operators within ladders. Recall (1.19)

[1] In the very early editions of Dirac's classic text *The Principles of Quantum Mechanics* he proceeded as if these two Lie algebras were the same. This misconception was removed by more careful wording in later editions.

[2] In light wave optics the quantity $\lambda/(2\pi)$, where λ is the wavelength, plays the role of $\hbar$. Consequently, Fourier optics results can be read off from paraxial ray optics results. We also remark that although the subalgebras are identical, the underling groups are not identical. In the classical case the group is the inhomogeneous symplectic group in two dimensions over the real field. In the quantum case the group is the inhomogeneous metaplectic group in two dimensions over the real field, which is a two-fold cover of the inhomogeneous symplectic group in two dimensions over the real field.

and (2.26). What can be said about the action of a_{11} and its quantum counterpart A_{11}? Simple calculation gives the result

$$: a_{11} : a_{rs} = [qp, q^r p^s]_{\text{cm}} = (s-r)a_{rs}. \tag{Y.4.4}$$

Thus the a_{rs} are eigenfunctions of the operator $: a_{11} :$ with eigenvalues $(s-r)$. It follows from (4.3) that there are the analogous quantum results

$$: A_{11} : A_{rs} = (s-r)A_{rs}. \tag{Y.4.5}$$

The A_{rs} are eigenfunctions of the operator $: A_{11} :$ with eigenvalues $(s-r)$.

The first differences between $\mathcal{L}_{\text{cm}}$ and $\mathcal{L}_{\text{qm}}$ results occur at degree 4. There are the classical relations

$$[a_{03}, a_{30}]_{\text{cm}} = -9a_{22} \tag{Y.4.6}$$

and

$$[a_{12}, a_{21}]_{\text{cm}} = -3a_{22}. \tag{Y.4.7}$$

By contrast, there are the quantum relations

$$[A_{03}, A_{30}]_{\text{qm}} = -9A_{22} + (3/2)\hbar^2 A_{00}, \tag{Y.4.8}$$

and

$$[A_{12}, A_{21}]_{\text{qm}} = -3A_{22} - (1/2)\hbar^2 A_{00}. \tag{Y.4.9}$$

The Lie products of all other degree 3 polynomials yield degree 4 results that are identical in the classical and quantum cases.

It can be verified that there are the relations

$$\lim_{\hbar \to 0} C^{tu}_{jk;rs} = c^{tu}_{jk;rs}, \tag{Y.4.10}$$

and consequently there is the limiting correspondence

$$[a_{jk}, a_{rs}]_{\text{cm}} \leftrightarrow \lim_{\hbar \to 0} [A_{jk}, A_{rs}]_{\text{qm}}. \tag{Y.4.11}$$

Thus, $\mathcal{L}_{\text{cm}}$ is a *contraction* of $\mathcal{L}_{\text{qm}}$ and, conversely, $\mathcal{L}_{\text{qm}}$ is a *deformation* of $\mathcal{L}_{\text{cm}}$.

It is notable that the differences between the classical and quantum cases actually involve $\hbar^2$ and not $\hbar$ itself.[3] Thus, the relation (4.10) may be replaced by the stronger result

$$C^{tu}_{jk;rs} = c^{tu}_{jk;rs} + O(\hbar^2). \tag{Y.4.12}$$

Moreover note, as inspection of (4.8) and (4.9) illustrates, that the $O(\hbar^2)$ terms involve only lower-degree polynomials than those that occur classically.

We also observe that the nature of the terms that can occur in a Lie product can be inferred from the Clebsch-Gordan series for the symplectic group. We have already seen that in the case of a two-dimensional phase space the basis elements A_{jk} have well-defined

[3] It follows that wave-optics aberration results, up to corrections of order $[\lambda/(2\pi)]^2$, can be read off from ray-optics aberration results.

transformation properties under the action of the symplectic group Lie algebra $sp(2, \mathbb{R})$. Consequently, their Lie products must also have well-defined transformation properties under the action of $sp(2, \mathbb{R})$. For example, the A_{jk} with $j + k = d$ belong to an irreducible representation of $sp(2, \mathbb{R})$ that behaves like "spin" $d/2$.[4] In the cases of (4.8) and (4.9), the ingredients of the Lie products on the left sides carry the representation $3/2$. According to Clebsch and Gordan, two spin $3/2$ objects can combine to produce objects of spins $3, 2, 1$, and 0. Since the Lie product is antisymmetric under the interchange of its ingredients, the spins 3 and 1 are ruled out by symmetry considerations. What possibly remain are spins 2 and 0. The two terms that occur on the right sides of (4.8) and (4.9), namely A_{22} and A_{00}, have spins 2 and 0, respectively.

Exercises

Y.4.1. Verify the results (1.4) through (1.19) and the results (2.11) through (2.26).

Y.4.2. Verify the results (4.3) through (4.11).

Y.4.3. Determine the effect of $: q^2/2 :$ on a_{rs} and the effect of $: Q^2/2 :$ on A_{rs}.

[4] Here we use the fact that the Lie algebras $sp(2, \mathbb{R})$ and $su(2)$ are equivalent over the complex field, and therefore their Clebsch-Gordan series are essentially the same. In the case of phase spaces of dimension $2n$, one must know the Clebsch-Gordan series for $sp(2n, \mathbb{R})$.

Bibliography

[1] B. Hall, *Quantum Theory for Mathematicians*, Springer (2013).

[2] T. Curtright, D. Fairlie, and C. Zachos, *A Concise Treatise on Quantum Mechanics in Phase Space*, World Scientific (2014).