DERIVATIONS OF APPLIED MATHEMATICS

VOLUME ONE OF TWO, INCLUDING PARTS I AND II

Derivations of Applied Mathematics

Thaddeus H. Black

Revised 27 February 2023

Thaddeus H. Black, 1967–.
Derivations of Applied Mathematics.
U.S. Library of Congress class QA401.

This book is published by the Debian Project [46]. Besides in Debian GNU/Linux and its derivatives [22], you can fetch the book's latest open-source printing at *derivations.org* [23].

The book was last revised 27 February 2023.

Contents

List of tables

List of figures

Preface

I never meant to write this book. It emerged unheralded, unexpectedly.

The book began in 1983 when a high-school classmate challenged me to prove the Pythagorean theorem on the spot. I lost the dare, but looking the proof up later I recorded it on loose leaves, adding to it the derivations of a few other theorems of interest to me. From such a kernel the notes grew over time, until family and friends suggested that the notes might make the material for the book you hold.

The book is neither a tutorial on the one hand nor a bald reference on the other. The book is rather a *study reference.* In this book, you can look up some particular result directly, or you can begin on page one and read—with toil and commensurate profit—straight through to the end of the last chapter.

The book as a whole surveys the general mathematical methods common to engineering and the physical sciences. As such, the book serves as a marshal or guide. It concisely arrays and ambitiously reprises the mathematics of the scientist and the engineer, deriving the mathematics it reprises, filling gaps in one's knowledge while extending one's mathematical reach.

Its focus on derivations is what principally distinguishes this book from the few others[1] of its class. No result is presented here but that it is justified in a style engineers, scientists and other applied mathematicians will recognize—not indeed in the contrived style of today's professional mathematician, which serves other needs; but in the longer-established natural style of applications.

Plan

Following its introduction in chapter 1 the book comes in three parts. The first part begins with a brief review of elementary algebra and geometry and develops thence the *calculus* of a single complex variable, this calculus

[1] Other books of the class include [38][25][88][6].

being the axle as it were about which higher mathematics turns. The second part laboriously constructs the broadly useful mathematics of *matrices* and *vectors,* without which so many modern applications (to the fresh incredulity of each generation of college undergraduates) remain analytically intractable—the jewel of this second part being the *eigenvalue* of chapter 14. The third and final part, the most interesting but also the most advanced, introduces the mathematics of the *Fourier transform, probability* and the *wave equation*—each of which is enhanced by the use of *special functions,* the third part's unifying theme.

Thus, the book's overall plan, though extensive enough to take several hundred pages to execute, is straightforward enough to describe in a single sentence. The plan is to derive as many mathematical results, useful to scientists, engineers and the like, as possible in a coherent train, recording and presenting the derivations together in an orderly manner in as few volumes as possible.[2] What constitutes "useful" or "orderly" is a matter of perspective and judgment, of course. My own peculiar, heterogeneous background in military service, electrical engineering, building construction, electromagnetic analysis, automotive manufacturing and software development, my nativity, residence and citizenship in the United States, undoubtedly bias the selection and presentation to some degree. How other authors go about writing their books, I do not know, but I suppose that what is true for me is true for many of them also: we begin by organizing notes for our own use, then observe that the same notes might prove useful to others, and then undertake to revise the notes and to bring them into a form which actually is useful to others.

Notation

The book deviates from—or cautiously improves, if you will endorse the characterization—the conventional notation of applied mathematics in one conspicuous respect which, I think, requires some defense here. The book employs hexadecimal numerals.

Why not decimal only? There is nothing wrong with decimal numerals as such. I am for them, whether in the Roman or the Arabic style. Decimal numerals are well in history and anthropology (man has ten fingers), finance and accounting (dollars, cents, pounds, shillings, pence: the base hardly matters), law and engineering (the physical units are arbitrary anyway); but they are merely serviceable in mathematical theory, never aesthetic.

[2] The book's nonelectronic print edition comes in two volumes.

Custom is not always defaced, but sometimes adorned, by the respectful attendance of a prudent discrimination. It is in this spirit that hexadecimal numerals are given place here.

Admittedly, one might judge the last to be more excuse than cause. Yet, though a dreary train of sophists down the years, impatient of experience, eager to innovate, has indisputably abused such causes—in ways which the mature reader of a certain cast of mind will find all too familiar—such causes would hardly merit abuse did they not sometimes hide a latent measure of justice. It is to the justice, or at least to the aesthetic, rather than to the sophistry that I affect to appeal here.

There unfortunately really is no gradual way to bridge the gap to hexadecimal (shifting to base eleven, thence to twelve, etc., is no use). If one wishes to reach hexadecimal ground then one must leap. Forty years of keeping my own private notes in hex have persuaded me that the leap justifies the risk. In other matters, by contrast, the book leaps seldom. The book in general walks a tolerably conventional applied mathematical line.

Audience

Besides those who have opened this book only to look up some particular result (a numerous and honorable clan, but likely not reading this preface), the book's intended readers arrive in two principle corps. First come the engineers and physical scientists who seek ideas toward, and logic to back, their analytical modeling of physical systems. Second come those ambitious students of calculus that want a broader, demand a deeper, and venture a terser treatment of the discipline than calculus textbooks usually afford.

There are also some others. In a third corps come the economist and his brethren, who may find the book a little long on physics and, comparatively, slightly short on statistics, but still edifying perhaps. Whether a few students of pure mathematics make a fourth corps, hunting sketches to elaborate, I shall not speculate.

Publication

The book belongs to the emerging tradition of open-source software where at the time of this writing it fills a void. Nevertheless it is a *book,* not a program. Lore among open-source developers holds that open development inherently leads to superior work. Well, maybe. Often it does in fact. Personally with regard to my own work, I should rather not make too many claims. It would be vain to deny that professional editing and formal peer review, neither of which the book enjoys, had substantial value. On the other hand, it does not do to despise the *amateur* (literally, one who does for the love of it: not such a bad motive, after all[3]) on principle, either—unless one would on the same principle despise a Cincinnatus or a Socrates. Open source has a spirit to it which leads readers to be more generous with their feedback than ever could be the case with an ordinary, proprietary book. Such readers, among whom a surprising concentration of talent and expertise are found, enrich the work freely. This has value, too.

The book's open-source publication implies that it can neither go out of print nor grow hard to find. It also implies that citations to the book can easily be followed hither. If desired you could, if expedient you should, copy, archive and distribute the book yourself, without further permission than the book's license already grants[4]—though as a courtesy to your own readers and to this writer you might publish the book's electronic address, *derivations.org,* along with the book.

Having addressed plan, notation, audience and publication, the preface will next touch points of edition, philosophy and reliance before concluding in acknowledgements and, thence, giving way to chapter 1.

Edition

A few marks of edition want note.

First, the book is extensively footnoted. Some of the footnotes unremarkably cite sources but many are discursive in nature, offering nonessential material which, though edifying, coheres insufficiently well to join the book's main narrative. The footnote is an imperfect messenger, of course. Catching the reader's eye, it can break the flow of otherwise good prose. Modern publishing promotes various alternatives to the footnote—numbered examples, sidebars, special fonts, colored inks, etc. Some of these are merely trendy. Others, like numbered examples, really do help the right kind of book; but for this book the humble footnote, long sanctioned by an earlier era of publishing, extensively employed by such sages as Gibbon [64] and Shirer [150], seems the most able messenger. In this book it shall have many messages to bear.

Second, in typical science/engineering style, the book numbers its sections, tables, figures and formulas, but not its theorems, the last of which it generally sets in italic type. Within the book, a theorem is referenced by the number of the section that states it.

[3] The expression is derived from an observation I seem to recall George F. Will making.

[4] [60]

Third, the book subjoins an alphabetical index as a standard convenience. Even so, the canny reader will avoid using the index (of this and other books), which alone of the book's pages is not to be regarded as a proper part of the book. Such a reader will tend rather to consult the book's table of contents which is a proper part.

Fourth, the book includes a bibliography listing works to which I have referred while writing. Mathematics can promote queer bibliographies, though; for it is derivation rather than authority that establishes mathematical methods and truths. The bibliography of the book you are reading less appeals to the works it lists than merely affords them due credit.

Regarding the last matter, not every point in the book is backed by a bibliographic citation of any kind. Some of the book consists of common mathematical knowledge or even of proofs I have worked out with my own pencil from various ideas gleaned—who knows from where?—over the years. The latter proofs are perhaps original or semi-original from my personal point of view but it is unlikely that many if any of them are truly new. To the initiated, the mathematics itself often tends to suggest the form of the proof: if to me, then surely also to others who came before; and even where a proof is new the idea proved is probably not.

Philosophy

Speaking of ideas and proofs: an idea is one thing, but what precisely constitutes a *proof?*

Modern pure mathematics tends to make one shy of the question. To me at least, a mathematical proof remains what an earlier era once unsuspectingly held it to be: it remains a morally convincing appeal to man's faculty of logic, geometry and number (though I study not to focus the reader's attention, unprofitably to the book's purpose, on any such metadefinition). Neither in this book nor elsewhere do I wish to deconstruct, reconstruct, lay bare, replace, supersede or explain away the faculty so named. Indeed, towering figures like Kant, Weierstrass and Hilbert notwithstanding, C. S. Lewis (1898–1963) speaks for me when he writes:

> You cannot go on "seeing through" things for ever. The whole point of seeing through something is to see something through it. It is good that the window should be transparent, because the street or garden beyond it is opaque. How if you saw through the garden too? It is no use trying to "see through" first principles. If you see through everything, then everything is transparent.

> But a wholly transparent world is an invisible world. To "see through" all things is the same as not to see. [108, chapter 3]

Such are my sympathies.

Would the Kantian era in which we live countenance it, the book should sooner merely have let pure mathematics' abstract foundations lie undisturbed. However, since the era probably will not countenance it, chapter 1 engages the question briefly but soberly, after which other chapters touch the question as necessary. When philosophically put to it, the book tends less to follow Kant, Weierstrass or Hilbert in spirit than Plato, Frege, Weyl and Gödel[5] (though there remains the peculiar matter of "the Courant-Hilbert-Shilov perspective," of which more will be said). That is in spirit. In method, with little further apology, the book follows the time-honored practice of the working physical scientist and engineer.

Reliance

I hope that the book harbors no more errors than do other books of the kind. I hope that the book harbors fewer. Having revised the book's manuscript (or the notes from which the manuscript is drawn) over a period of 40 years, I believe that the book's results are correct in the main.

Nevertheless, the book gives reasons the reader can evaluate. The book details steps the reader can check. The book illuminates patterns the reader can study. The book teaches principles the reader can absorb. To look up a result in the book without evaluating, checking, studying or absorbing might not always be an unreasonable risk to run when stakes are small and time is short, but application of the book's results must remain the responsibility of the applicationist.

[5] Readers who know the subject well may note the omission of the important name of Richard Dedekind (1831–1916) from these two lists. However, in which of the two would you name Dedekind? It is no easy question—nor is it a question this book will tackle. As respectable as Dedekind is, this book does not especially follow him, anyway. [151][166]

One could further mention several other respectable names—Georg Cantor's (1845–1918), Bertrand Russell's (1872–1970) and L. E. J. Brouwer's (1881–1966), for instance, after the name of the great Carl Friedrich Gauss (1777-1855)—and one could bring the lists generally more up to date, but we will leave the matter there.

To date the names listed: Plato (428–348 B.C.); Immanuel Kant (1724–1804); Karl Weierstrass (1815–1897); Gottlob Frege (1848–1925); David Hilbert (1862–1943); Hermann Weyl (1885–1955); Richard Courant (1888–1972); Kurt Gödel (1906–1978); Georgi E. Shilov (1917–1975).

Acknowledgements

The styles and insights of mentors including Profs. J. L. Young [189], D. M. Sullivan [160], R. J. Baker [9] and especially G. S. Brown [30] and of associates including Mr. D. E. Davis [41] implicitly touch the book in many places and in many ways.

Steady encouragement from my wife and children contribute to the book in ways only an author can properly appreciate.

For a preface to state that the book's shortcomings remain the author's own seems *de rigueur.* The statement remains true, too, though, so I am glad for this chance to state it.

More and much earlier than to mentors, associates, wife or children, the book owes a debt to my mother and, separately, to my father, without either of whom the book would never have come to be. Admittedly, any author of any book might say as much in a certain respect, but it is no office of a mathematical book's preface to burden readers with an author's expressions of filial piety. No, it is in entirely another respect that I lay the matter here. My mother taught me at her own hand most of the mathematics I ever learned as a child, patiently building a foundation that cannot but be said to undergird the whole book today. My father generously financed much of my formal education but—more than this—one would have had to grow up with my brother and me in my father's home to appreciate the grand sweep of the man's curiosity, the general depth of his knowledge, the profound wisdom of his practicality and the enduring example of his love of excellence.

May the book deserve such a heritage.

THB

Chapter 1

Introduction

The Pythagorean theorem holds that

$$a^2 + b^2 = c^2, \tag{1.1}$$

where a, b and c are the lengths of the legs and diagonal of a right triangle as in Fig. 1.1. Many proofs of the theorem are known.

One such proof posits a square of side length $a+b$ with a tilted square of side length c inscribed as in Fig. 1.2. The area of each of the four triangles in the figure is evidently $ab/2$. The area of the tilted inner square is c^2. The area of the large outer square is $(a+b)^2$. But the large outer square is comprised of the tilted inner square plus the four triangles, so the area of the large outer square equals the area of the tilted inner square plus the

Figure 1.1: A right triangle.

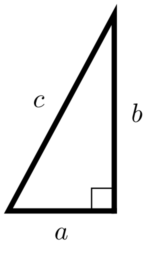
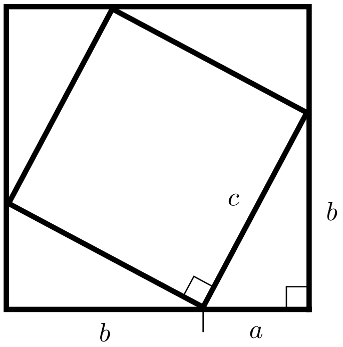


Figure 1.2: The Pythagorean theorem.



areas of the four triangles. In mathematical symbols, this is that

$$(a+b)^2 = c^2 + 4\left(\frac{ab}{2}\right),$$

which simplifies directly to (1.1).

If the foregoing appeals to you then you might read this book.

This book is a book of applied mathematical proofs. When you have seen a mathematical result somewhere, if you want to know why the result is so, then you can look for the proof here.

The book's purpose is to convey the essential ideas underlying the derivations of a large number of mathematical results useful in the modeling of physical systems. To this end, the book emphasizes main threads of mathematical argument and the motivation underlying the main threads, deëmphasizing formal mathematical rigor. It derives mathematical results from the applied perspective of the scientist and the engineer.

The book's chapters are topical. This first chapter explains the book's philosophy and otherwise treats a few introductory matters of general interest.

1.1 Applied mathematics

What is applied mathematics?

> Applied mathematics is a branch of mathematics that concerns itself with the application of mathematical knowledge to other domains. . . . The question of what is applied mathematics does not answer to logical classification so much as to the sociology of professionals who use mathematics. [103]

That is about right, on both counts. In this book we shall define *applied mathematics* to be correct mathematics useful to engineers, physical scientists and the like; proceeding not from reduced, well-defined sets of axioms but rather directly from a nebulous mass of natural arithmetical, geometrical and algebraic idealizations of physical systems; demonstrable but lacking the abstracted foundational rigor of the pure, professional mathematician.

1.2 Rigor

Applied and pure mathematics differ principally and essentially in the layer of abstract definitions the latter subimposes beneath the physical ideas the former seeks to model. That subimposed layer, the disciplined use of it, and the formal arithmetic associated with it may together be said to institute pure mathematical *rigor.*

Such pure mathematical rigor tends to dwell more comfortably in lone reaches of the professional mathematician's realm than among the hills and plats of applications, where it does not always travel so gracefully. If this book will be a book of mathematical derivations, then it might speak a little of rigor here at the start.

1.2.1 Axiom and definition

Whether explicitly or implicitly, the professional mathematician usually founds his rigor upon what he calls the *axiomatic method*—an *axiom*, according to Webster, being "a self-evident and necessary truth, or a proposition whose truth is so evident at first sight that no reasoning or demonstration can make it plainer; a proposition which it is necessary to take for granted."[1]

For example, the following could be an axiom: "For every set A and every set B, $A = B$ if and only if for every set x, x is a member of A if and only if x is a member of B."[2]

[1] [134]

[2] The source quoted is [185, § 2.1], which however uses the symbol $\in$ for "is a member of."

Axioms undergird the work of the professional mathematician. Indeed, so fundamental are axioms to the professional mathematician's work that—*ideally and at least in principle*—it may be that the professional will derive nothing until he has first declared the axioms upon which his derivations will rely; that is, until he has stated the least premises upon which he will argue. Conversely, aiming deeper—promoting the latter of Webster's two readings—the professional can illuminate the wisdom latent in his very axioms by their use in a suitable derivation.[3] Irreducibility is a prime aesthetic on either level: at best, no axiom should overlap the others or be specifiable in terms of the others. Nonaxiomatic geometrical argument—proof by sketch if you like, as the Pythagorean with its figures at the head of this chapter—is distrusted.[4] The professional mathematical literature at its best discourages undue pedantry indeed, but its readers still implicitly demand a convincing assurance that its writers *could* derive results in pedantic logical-arithmetical detail if called upon to do so. Precise definition here is critically important, which is why the professional mathematician tends not to accept blithe statements such as that

$$\frac{1}{0} = \infty,$$

among others, without first inquiring as to exactly what is meant by symbols like 0 and ∞.

The applied mathematician begins from a different base. His ideal lies not in precise definition or irreducible axiom, but rather in the elegant modeling of the essential features of some physical system. Here, mathematical definitions tend to be made up *ad hoc* along the way, based on previous experience solving similar problems, adapted implicitly to suit the model at hand. If you ask the applied mathematician exactly what his axioms are, which symbolic algebra he is using, he usually does not know; what he knows is that the physical system he is analyzing, describing or planning—say, a bridge—is to be founded in certain soils with observed tolerances, is to suffer such-and-such a wind load, and so on. To avoid error, the applied mathematician relies not on abstract formalism but rather on a thorough mental grasp of the essential physical features of the phenomenon he is trying to model. An equation like

$$\frac{1}{0} = \infty$$

may make perfect sense without further explanation to an applied mathematical readership, depending on the physical context in which the equation

[3] [78, Einleitung][77, chapter 1]

[4] [175, chapters 1 and 2]

is introduced. Nonaxiomatic geometrical argument—proof by sketch—is not only trusted but treasured. Abstract definitions are wanted only insofar as they smooth the analysis of the particular physical problem at hand; such definitions are seldom promoted for their own sakes.

The irascible Oliver Heaviside (1850–1925), responsible for the applied mathematical technique of phasor analysis,[5] once said,

> It is shocking that young people should be addling their brains over mere logical subtleties, trying to understand the proof of one obvious fact in terms of something equally ... obvious. [122]

Exaggeration, perhaps, but from the applied mathematical perspective Heaviside nevertheless had a point. The professional mathematicians Richard Courant (1888–1972) and David Hilbert (1862–1943) put it more soberly in 1924 when they wrote,

> Since the seventeenth century, physical intuition has served as a vital source for mathematical problems and methods. Recent trends and fashions have, however, weakened the connection between mathematics and physics; mathematicians, turning away from the roots of mathematics in intuition, have concentrated on refinement and emphasized the postulational side of mathematics, and at times have overlooked the unity of their science with physics and other fields. In many cases, physicists have ceased to appreciate the attitudes of mathematicians. [38, Preface]

And what are these "attitudes" of which Courant and Hilbert speak? To the mathematician Charles C. Pinter, they are not attitudes, but principles:

> Since the middle of the nineteenth century, the axiomatic method has been accepted as the only correct way of organizing mathematical knowledge. [131, chapter 1]

But accepted by whom? The mathematician Georgi E. Shilov, somewhat less enthusiastic than Pinter for the axiomatic method, is not so sure:

> There are other approaches to the theory ... where things I take as axioms are proved. ... Both treatments have a key deficiency, namely the absence of a proof of the compatibility of

[5] This book lacks occasion to treat phasor analysis as such but see chapter 5, which introduces the complex exponential function upon which the phasor is based, and chapter 18, which extends and generalizes phasors (without however mentioning the word) under the umbrella of the Fourier transform.

> the axioms. . . . The whole question, far from being a mere technicality, involves the very foundations of mathematical thought. In any event, this being the case, it is not very important where one starts a general treatment. . . . [147, Preface]

Although the present book responds to "the attitudes of mathematicians" with greater deference than some of Courant's and Hilbert's unnamed 1924 physicists might have done, though Shilov himself admittedly is more rigorous than his own, seemingly casual words let on, still, Courant and Hilbert could have been speaking for the engineers and other applied mathematicians of our own day as well as for the physicists of theirs; and still, Shilov like Heaviside has a point. To the applied mathematician, the mathematics is not principally meant to be developed and appreciated for its own sake; it is meant to be *used.* This book adopts the Courant-Hilbert-Shilov perspective.[6] for this reason.

But why? Is the Courant-Hilbert-Shilov perspective really necessary, after all? If unnecessary, is it desirable? Indeed, since the book you are reading is a book of derivations, would it not be a more elegant book if it began from the most primitive, pure mathematical fundamentals, and proceeded to applications thence?

If Heaviside was so irascible, then wasn't he just plain wrong?

[6] It is acknowledged that Hilbert at other times took what seems to be the opposite perspective; and that there remains the historically important matter of what the early twentieth century knew as "Hilbert's program," a subject this book will not address. Hilbert however, perhaps the greatest of the mathematical formalists [53, chapter 1], was a broad thinker, able to survey philosophical questions seriously from each of multiple points of view. What Hilbert's ultimate opinion might have been, and whether the words quoted more nearly represent Hilbert's own conviction or his student Courant's, and how the views of either had evolved before or would evolve after, are biographical questions this book will not try to treat. The book would accept the particular passage recited rather on its face.

Regarding Shilov, his formal mathematical rigor is easy and fluent, and his book [147] makes a good read even for an engineer. The book you now hold however adopts not Shilov's methods—for one can read Shilov's book for those—but only his perspective, as expressed in the passage recited.

For a taste of Shilov's actual methods, try this, the very first proof in his book: "Theorem. The system [of real numbers] contains a unique zero element. Proof. Suppose [that the system] contains two zero elements 0_1 and 0_2. Then it follows from [the axioms of commutation and identity] that $0_2 = 0_2 + 0_1 = 0_1 + 0_2 = 0_1$. Q.E.D." [147, § 1.31].

1.2.2 Mathematical Platonism

To appreciate the depth of the trouble in which the applied mathematician may soon find himself mired, should he too casually reject the Courant-Hilbert-Shilov perspective, consider John L. Bell's and Herbert Korté's difficult anecdote regarding Hilbert's brilliant student and later cordial rival, Hermann Weyl (1885–1955):

> Weyl . . . considers the experience of seeing a pencil lying on a table before him throughout a certain time interval. The position of the pencil during this interval may be taken as a function of the time, and Weyl takes it as a fact of observation that during the time interval in question this function is continuous and that its values fall within a definite range. And so, he says, "This observation entitles me to assert that during a certain period this pencil was on the table; and even if my right to do so is not absolute, it is nevertheless reasonable and well-grounded. It is obviously absurd to suppose that this right can be undermined by 'an expansion of our principles of definition'—as if new moments of time, overlooked by my intuition, could be added to this interval; moments in which the pencil was, perhaps, in the vicinity of Sirius or who knows where. . . ." [17]

In Weyl's mild irony lies a significant point, maybe, yet how should applied mathematics advocate such a point, or dispute it? Is applied mathematics even suited to such debates? What of engineering questions of a more mundane cast—such as, for instance, how likely Weyl's pencil might be to roll off if his knee bumped the table? After all, there is a pencil, and there is a table, and Sirius seems to have little immediately to do with either; and whether the pencil rolls off might concern us, irrespective of any particular substruction pure mathematics sought to build to support the technique we had used to model and analyze the case. Indeed, Weyl himself—a great mathematician, a consistent friend of the engineer and of the scientist, and a wise man—warns,

> The ultimate foundations and the ultimate meaning of mathematics remain an open problem; we do not know in what direction it will find its solution, nor even whether a final objective answer can be expected at all. [180]

Just so. Fascinating as it is, we shall not try to answer Weyl's deep question of mathematical philosophy in this book.

To the extent to which a professional mathematician classified this book as a work of applied mathematics, he might call it a *Platonist* work—and that is a fine adjective, is it not? The adjective is most subtle, most lofty; and perhaps the author had better not be too eager to adorn his own work with it; yet let us listen to what the professional mathematician Reuben Hersh has to say:

> Most writers on the subject seem to agree that the typical "working mathematician" is a Platonist on weekdays and a formalist on Sundays. That is, when he is doing mathematics, he is convinced that he is dealing with an objective reality whose properties he is attempting to determine. But then, when challenged to give a philosophical account of this reality, he finds it easiest to pretend that he does not believe in it after all. . . .
>
> The basis for Platonism is the awareness we all have that the problems and concepts of mathematics exist independently of us as individuals. The zeroes of the zeta function[7] are where they are, regardless of what I may think or know on the subject. . . .[8] [76]

Your author inclines to Platonism[9] all days of the week, yet even readers who do not so incline should find the book edifying on the working terms of which Hersh speaks.

Hersh goes on with tact and feeling at some length in the article from which his words are quoted, and it is fair to say that he probably would not endorse the present writer's approach in every respect. Notwithstanding, the philosopher Thomas Tymoczko—who unlike Hersh but like the present writer might fairly be described as a Platonist[10]—later writes of Hersh's article,

> . . . In so far as [the working philosophy of the professional mathematician] is restricted to the usual mix of foundational ideas,

[7] This book stops short of treating Riemann's zeta function, a rather interesting special function that however seems to be of even greater interest in pure than in applied mathematics. If you want to know, the zeta function is $\zeta(z) \equiv \sum_{k=1}^{\infty} 1/k^z$. [153, chapter 10]

[8] Hersh, who has thus so empathetically sketched mathematical Platonism, goes on tactfully to confess that he believes mathematical Platonism a myth, and to report (admittedly probably correctly) that most professional mathematicians also believe as he does on this point. The present writer however accepts the sketch, appreciates the tact, and believes *in* the myth, for the reasons outlined in this introduction among others.

[9] [55, chapter 2]

[10] Tymoczko's preferred term is not "Platonist" but "quasiëmpiricist," a word Tymoczko lends a subtly different emphasis. [169]

> Hersh charges, [this philosophy] is generally inconsistent, always irrelevant and sometimes harmful in practice and teaching.
>
> ... Hersh suggests [that] the best explanation of foundational concerns is in terms of the historical development of mathematics.... [H]e isolates some of the basic presuppositions of foundation studies: "that mathematics must be provided with an absolutely reliable foundation" and "that mathematics must be a source of indubitable truth." Hersh's point is that it is one thing to accept the assumption when, like Frege, Russell or Hilbert, we feel that the foundation is nearly attained. But it is quite another to go on accepting it, to go on letting it shape our philosophy, *long after*[11] we've abandoned any hope of attaining that goal.... [168]

The applied mathematician who rejects the Courant-Hilbert-Shilov perspective and inserts himself into *this* debate[12] may live to regret it. As the mathematician Ludwig Wittgenstein illustrates,

> [Bertrand] Russell [coäuthor of *Principia Mathematica* and arch-exponent of one of the chief schools of pure mathematical thought][181] gives us a calculus here. How this calculus of Russell's is to be *extended* you wouldn't know for your life, unless you had ordinary arithmetic in your bones. Russell doesn't even prove $10 \times 100 = 1000$.
>
> What you're doing is constantly taking for granted a particular interpretation. You have mathematics and you have Russell; you think mathematics is all right, and Russell is all right—more so; but isn't this a put-up job?[13] That you can correlate them in a way, is clear—not that one throws light on the other. [186, lecture XVI]

The book you hold will not correlate them but will (except in some inessential side commentary) confine its attention to the applied mathematician's

[11] Emphasis in the original.

[12] See also, in no particular order, [8][10][16][33][38][53][54][56][61][62][70][71][77][78][81][98][100][102][107][112][131][136][139][147][148][151][161][162][166][174][175][176][179][181][185].

[13] For readers whose first language is other than English, colloquially, a "put-up job" is a conspiratorially prearranged deception. The colloquialism is now somewhat outdated but, judging by this quote, was apparently current at Cambridge in 1939.

honorable, chief interest—which is to describe, quantify, model, plan and analyze particular physical phenomena of concern; and to understand topically why the specific mathematical techniques brought to bear on such phenomena should prosper; but not to place these techniques in the context of a larger ontological or epistemological dispute—a dispute that, though important in itself, does not directly move the applied mathematician's interest one way or the other.

To conclude this subsection's glance upon mathematical Platonism[14] we may well quote Plato himself:

> Then this is a kind of knowledge which legislation may fitly prescribe; and we must endeavour to persuade those who are to be the principal men of our State to go and learn arithmetic, not as amateurs, but they must carry on the study until they see the nature of numbers with the mind only; nor again, like merchants or retail-traders, with a view to buying or selling, but for the sake of their military use, and of the soul herself; and because this will be the easiest way for her to pass from becoming to truth and being. . . . I must add how charming the science is! . . . [A]rithmetic has a very great and elevating effect, compelling the soul to reason about abstract number, and rebelling against the introduction of visible or tangible objects into the argument. . . . [T]his knowledge may be truly called necessary, necessitating as it clearly does the use of the pure intelligence in the attainment of pure truth. . . .
>
> And next, shall we enquire whether the kindred science [of geometry] also concerns us? . . . [T]he question relates . . . to the greater and more advanced part of geometry—whether that tends in any degree to make more easy the vision of the idea of good; and thither, as I was saying, all things tend which compel the soul to turn her gaze towards that place, where is the full perfection of being, which she ought, by all means, to behold. . . . [T]he knowledge at which geometry aims is knowledge of the eternal, and not of aught perishing and transient. [G]eometry will draw the soul towards truth, and create the spirit of philosophy, and raise up that which is now unhappily allowed to fall down. [133, book VII].

[14] [10]

The vanity of modern man may affect to smile upon the ancient; but such vanity less indulges the ancient, who hardly needs indulgence, than indicts the modern.

Plato is not less right today than he was in the fourth century B.C.

1.2.3 Other foundational schools

Other foundational schools of mathematical thought than Platonism exist.[15] They go by such names as logicism, intuitionism, formalism, predicativism, naturalism, structuralism, nominalism, empiricism, rationalism, Kantianism, pluralism and computationalism. There is even a school called fictionalism, a fact which should caution us that few if any of the schools can be appraised by their names alone.

We shall not explore all these schools here. We certainly shall not undertake to judge them! Instead, we shall heed philosopher Alva Noë's words when he warns,

> [T]here is no stable or deeply understood account of how these autonomous domains fit together. The fact that we are getting along with business as if there were such an account is, well, a political or sociological fact about us that should do little to reassure. [121]

Noë is writing here about the nature of consciousness but could as well, with equal justice and for similar reasons, be writing about our problem of mathematical foundations.

1.2.4 Methods, notations, propositions and premises

The book purposely overlooks, and thus omits, several of the modern mathematics profession's methodological advances and some of its more recondite notations, largely unsuited to (or at any rate unintended for) applied use. Most notably, the book overlooks and omits the methods and notations of the Zermelo-Fraenkel and Choice set theory (ZFC)[16] and its consequents. Even were it practical for the book to develop its results on a ZFC basis it would not do so, for it inclines rather toward Weyl's view:

> [A set-theoretic approach] contradicts the essence of the continuum, which by its very nature cannot be battered into a set of

[15] [81]

[16] See [161][185][40]. The ZFC is a near descendant of the work of the aforementioned Bertrand Russell and, before Russell, of Georg Cantor.

> separated elements. Not the relationship of an element to a set, but that of a part to a whole should serve as the basis. . . . [143]

The years have brought many mathematical developments since Weyl wrote these words in 1925 but the present author still tends to think as Weyl does.

Besides ZFC, the book also omits measure theory.[17]

Such omissions do not of course mean that the book or its author intended to peddle nonintuitive mathematical propositions, unsupported, as fact. The book could hardly call itself a book of derivations if it did. What it does mean is that the book can assume without further foundation or explication—and without extremes of formal definition—for example, that a rotated square remains square; that a number like $\sqrt{3}/2$ occupies a definite spot in a comprehensible continuum;[18] that no numbers in the continuum other than 0 and 1 enjoy these two numbers' respective identifying properties;[19] that a continuous, differentiable,[20] real function's average slope over a real interval equals the function's instantaneous slope at at least one point;[21] and so on (and if you did not understand all of that, that is all right, for to explain such things is what the rest of the book's chapters are for). It also means that neither today's advances in pure mathematics nor the "recent trends and fashions" of which Courant and Hilbert in § 1.2.1 have spoken are suffered to discourage the use of older, more physically appealing modes of reason in this book.

The Pythagorean theorem at the chapter's head examples the book's

[17] [149]

[18] [179]

[19] See footnote 6.

[20] Are burdensome adjectives like "continuous, differentiable" necessary? Are they helpful? Do they sufficiently illuminate one's understanding that they should be suffered to clutter the text so?

Maybe they are indeed necessary. Maybe they are even helpful but, even if so, does the discerning reader want them? Does the *nondiscerning* reader want them? If neither, then whom do they serve? If the only answer is that they serve the investigator of foundational concerns, then what does this tell us about the wisdom of injecting foundational concerns into applications?

Regarding continuity and differentiability: the applicationist is inclined to wait until a specific problem arises in which a particular, concrete discontinuity or undifferentiability looms, when he will *work around* the discontinuity or undifferentiability as needed—whether by handling it as a parameterized limit or by addressing it in some other convenient way. None of this has much to do with foundations.

To treat every such point as a fundamental challenge to one's principles of definition just takes too long and anyway does not much help. The scientist or engineer wants to save that time to wrestle with physical materialities.

[21] [147, § 7.4]

approach. The theorem is proved briefly, without excessive abstraction, working upon the *implied, unstated premise* that a rotated square remains square.

If you can accept the premise, then you can accept the proof. If you can accept the *kind* of premise, then you can accept the book.

1.2.5 Rigor to forestall error

Aside from foundational concerns—whatever the ontological or epistemological merits of formal mathematical rigor may be—some will laud such rigor too for forestalling error, even in applications.[22] Does the rigor deserve this praise? Well, perhaps it does. Still, though the writer would not deny the praise's decorum in every instance, he does judge such praise to have been oversold by a few.

Notwithstanding, formal mathematical rigor serves two, distinct programs. On the one hand, it embodies the pure mathematician's noble attempt to establish, illuminate or discover the means and meaning of *truth.*[23] On the other hand, it cross-checks intuitive *logic* in a nonintuitive way. Neither hand holds absolute metaphysical guarantees; yet, even if the sole use of formal mathematical rigor were to draw the mathematician's attention systematically to certain species of questionable reasoning for further examination, such rigor would merit the applicationist's respect. As the mathematician Richard W. Hamming writes,

> When you yourself are responsible for some new application of mathematics in your chosen field, then your reputation, possibly millions of dollars and long delays in the work, and possibly even human lives, may depend on the results you predict. It is then that the *need* for mathematical rigor will become painfully

[22] To cite a specific source here might be more captious than helpful, so the reader is free to disregard the assertion as unsourced. However, if you have heard the argument—perhaps in conjunction with the example of a conditionally convergent sum or the like—then the writer has heard it, too.

[23] As courtesy to the reader, I should confess my own opinion in the matter, which is that it is probably, fundamentally not given to mortal man to lay bare the ultimate foundations of truth, as it is not given to the beasts, say, to grasp mathematics. Like the beasts, we too operate within the ontological constraints of our nature.

That this should be my opinion will not perhaps surprise readers who have read the preface and the chapter to this point. As far as I know, Aristotle, Aquinas and Gödel were right. However, be that as it may, my opinion in the matter is not very relevant to this book's purpose. I do not peddle it but mention it only to preclude misunderstanding regarding the sympathies and biases, such as they are, of the book's author. —THB—

> obvious to you. Before this time, mathematical rigor will often seem to be needless pedantry. . . . [70, § 1.6]

Sobering words. Nevertheless, Hamming's point is not a point this book will pursue. The working scientist or engineer uses so many mathematical results in his work that, if he pursued Hamming's method for all of them, he might never gain acquaintance with the reasons for any of them! He lacks time, and the book is already long enough. Nevertheless, an applications-level justification for a given formula, leveraging one's physical intuition, unburdened by excessive concern for recondite requirements of the mathematics profession, is not too much for the working scientist or engineer to ask. If this book conveys such a justification then it will have done what it set out to do.

The introduction you are reading is not the right venue for a full essay on why both kinds of mathematics, applied and pure, are needed at any rate. Each kind has its place; and though it is a stylistic error to mix the two indiscriminately, clearly the two have much to do with one another. However that may be, this book is a book of derivations of applied mathematics. The derivations here proceed by an applied approach.

1.3 Mathematical extension

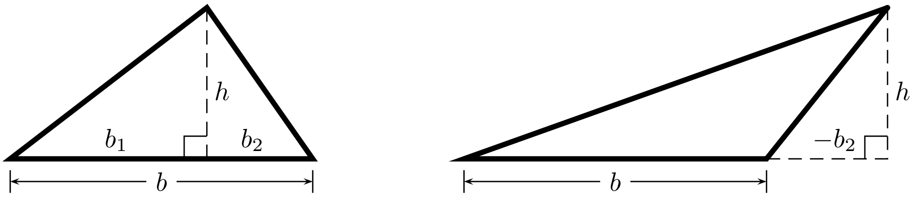
Profound results in mathematics occasionally are achieved simply by extending results already known. For example, the negative integers and their properties can be discovered by counting backward—3, 2, 1, 0—and then asking what follows (precedes?) 0 in the countdown and what properties this new, negative integer must have to interact smoothly with the already known positives. The astonishing Euler's formula (§ 5.4) is discovered by a similar but more sophisticated mathematical extension.

More often however, the results achieved by extension are unsurprising and not very interesting in themselves. Such extended results are the faithful servants of mathematical rigor. Consider for example the triangle on the left of Fig. 1.3. This triangle is evidently composed of two right triangles of areas

$$\begin{aligned} A_1 &= \frac{b_1 h}{2}, \\ A_2 &= \frac{b_2 h}{2} \end{aligned}$$

(each right triangle being exactly half a rectangle). Hence, the main trian-

Figure 1.3: Two triangles.



gle's area is

$$A = A_1 + A_2 = \frac{(b_1 + b_2)h}{2} = \frac{bh}{2}.$$

Very well. What about the triangle on the right? Its b_1 is not shown on the figure, and what is that $-b_2$, anyway? Answer: the triangle is composed of the *difference* of two right triangles, with b_1 the base of the larger, overall one: $b_1 = b + (-b_2)$. The b_2 is negative (whereby $-b_2$ is positive) because the sense of the small right triangle's area in the proof is negative: the small area is subtracted from the large rather than added. By extension on this basis, the main triangle's area is seen again to be $A = bh/2$. The proof is the same. In fact, once the central idea of adding two right triangles is grasped, the extension is really rather obvious—too obvious to be allowed to burden such a book as this.

Excepting the uncommon cases in which extension reveals something interesting or new, this book generally leaves the mere extension of proofs—including the validation of edge cases and over-the-edge cases—as an exercise to the interested reader.

Having mentioned rigor in § 1.2 and extension in § 1.3, the book shall proceed to treat derivations of applied algebra and geometry in chapter 2 presently. However, before it does, a few remarks regarding deduction, adduction, the complex variable and the book's text are in order.

1.4 Deduction, adduction and the complex variable

The English language derives from the Latin a nice counterpart to the transitive verb *to deduce,* a verb whose classical root means "to lead away." The counterpart is *to adduce,* "to lead toward." Adduction, as the word is here used, subtly reverses the sense of deduction: it establishes premises from necessary conclusions rather than the other way around.[24]

Applied mathematics sometimes prefers adduction to deduction. Attention is drawn to this preference because the preference governs the book's approach to the *complex number* and the *complex variable.* The book will speak much soon of the complex number and the complex variable, but we mention them now for the following reason.

An overall view of relevant analytical methods—including complex methods—and of their use in the modeling of physical systems, marks the applied mathematician more than does the abstract mastery of any received program of pure logic.[25] A *feel* for the math is the great thing. Formal definitions, axioms, symbolic algebras and the like, though often useful, are esteemed less as primary objects than as secondary supports. The book's adductive, rapidly staged development of the complex number and the complex variable is planned on this sensibility.

Sections 2.11, 3.10, 3.11, 4.3.3, 4.4, 6.2, 9.6 and 9.7, plus all of chapters 5 and 8, constitute the book's principal stages of complex development. In these sections and throughout the book, the reader comes to appreciate that most mathematical properties which apply for real numbers apply equally for complex, that few properties concern real numbers alone.

Pure mathematics develops its own, beautiful, abstract theory of the complex variable,[26] a theory whose arc regrettably takes off too late and flies too far from applications for such a book as this. Less beautiful, less abstract, more practical, nonaxiomatic paths to the topic exist,[27] and this book leads the reader along one of these.

[24] As § 1.2.1 has discussed, pure mathematics can occasionally, implicitly illuminate its *axioms* in the light of necessary conclusions. Since axioms are by definition a restricted kind of premise, one might arguably regard the illumination named as an elevated species of adduction. However, that is not what this section is about.

[25] Pure logic is a worthy endeavor, though whether such logic is more properly *received* or rather *illuminated* is a matter of long dispute. Besides Hilbert, of whom we have already spoken, see also Frege [62], the primer [162], and the commentary [176].

[26] [7][147][58][153][79]

[27] See chapter 8's footnote 8.

For supplemental reference, a bare sketch of the abstract theory of the complex variable is found in appendix C.

1.5 On the text

The book gives numerals in hexadecimal. It denotes variables in Greek letters as well as Roman. Readers unfamiliar with the hexadecimal notation will find a brief orientation thereto in appendix A. Readers unfamiliar with the Greek alphabet will find it in appendix B.

Licensed to the public under the GNU General Public License [60], version 2, this book meets the Debian Free Software Guidelines [47] and the Open Source Definition [123].

By its nature, a book of mathematical derivations can make strait, colorless reading. To delineate logic as it were in black and white is the book's duty. What then to tint? Is naught to be warm nor cool, naught ruddy nor blue? Though mathematics at its best should serve the demands not only of deduction but equally of insight, by the latter of which alone mathematics derives either feeling or use, neither every sequence of equations nor every conjunction of figures is susceptible to an apparent hue the writer can openly paint upon it—but only to that abeyant hue, that luster which reveals or reflects the fire of the reader's own mathematical imagination, which color remains otherwise unobserved.

The book begins by developing the calculus of a single variable.

Part I

The calculus of a single variable

Chapter 2

Elementary algebra and geometry

Probably every book must suppose something of its reader. This book supposes, or affects to suppose, little other than that its reader reads English and has a strong aptitude for mathematics, but it does assume that the reader has learned and absorbed the simplest elements of elementary arithmetic, algebra and geometry from his youth: that $1+1=2$; why $(2)(3)=6$; the reliability and rational repetition of the long-division algorithm; the digital representation in the real continuum of an arbitrary, perhaps irrational number; what it means when a letter like x stands in the place of an unspecified number; the technique to solve, say, that $3x-2=7$; how to read the functional notation $f(x)$; which quantity a square root $\sqrt{x}$ is; what to make of the several congruent angles that attend a line when the line intersects some parallels; and so on. Even so, some basic points of algebra and geometry seem worth touching briefly here. The book starts fast with these.

2.1 Basic arithmetic relationships

This section states some arithmetical rules.

2.1.1 Commutivity, associativity, distributivity, identity and inversion

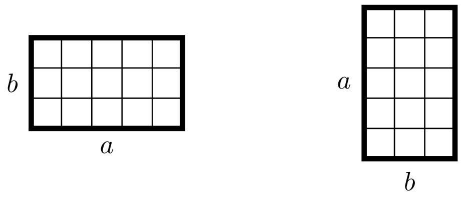
Table 2.1 lists several arithmetical rules,[1] each of which applies not only to real numbers but equally to the complex numbers of § 2.11. Most of the

[1] [147, § 1.2][153, chapter 1]

Table 2.1: Basic properties of arithmetic.

$a+b$	$=$	$b+a$	Additive commutivity
$a+(b+c)$	$=$	$(a+b)+c$	Additive associativity
$a+0=0+a$	$=$	a	Additive identity
$a+(-a)$	$=$	0	Additive inversion
ab	$=$	ba	Multiplicative commutivity
$(a)(bc)$	$=$	$(ab)(c)$	Multiplicative associativity
$(a)(1)=(1)(a)$	$=$	a	Multiplicative identity
$(a)(1/a)$	$=$	1	Multiplicative inversion
$(a)(b+c)$	$=$	$ab+ac$	Distributivity

Figure 2.1: Multiplicative commutivity.



rules are appreciated at once if the meaning of the symbols is understood. In the case of multiplicative commutivity, one imagines a rectangle with sides of lengths a and b, then the same rectangle turned on its side, as in Fig. 2.1: since the area of the rectangle is the same in either case, and since the area is the length times the width in either case (the area is more or less a matter of counting the little squares), evidently multiplicative commutivity holds. A similar argument validates multiplicative associativity, except that here one computes the *volume* of a three-dimensional rectangular box, which box one turns various ways.

Multiplicative inversion lacks an obvious interpretation when $a = 0$.

Loosely,

$$\frac{1}{0} = \infty.$$

But since $3/0 = \infty$ also, surely either the zero or the infinity, or both, somehow differ in the latter case.[2]

Looking ahead in the book, we note that the multiplicative properties do not always hold for more general linear transformations. For example, matrix multiplication is not commutative and vector cross-multiplication is not associative. Where associativity does not hold and parentheses do not otherwise group, right-to-left association is notationally implicit:[3,4]

$$\mathbf{A} \times \mathbf{B} \times \mathbf{C} = \mathbf{A} \times (\mathbf{B} \times \mathbf{C}).$$

The sense of it is that the thing on the left ($\mathbf{A} \times$) *operates* on the thing on the right ($\mathbf{B} \times \mathbf{C}$). (In the rare case in which the question arises, you may wish to use parentheses anyway.)

2.1.2 Negative numbers

Consider that

$$\begin{aligned}
(+a)(+b) &= +ab, \\
(+a)(-b) &= -ab, \\
(-a)(+b) &= -ab, \\
(-a)(-b) &= +ab.
\end{aligned}$$

The first three of the four equations probably are unsurprising, but the last is interesting. Why would a negative count $-a$ of a negative quantity $-b$

[2] Weierstrass, Kronecker, Dedekind and Frege, among others, spent much of the nineteenth century intensely debating the implications of this very question. The applied book you are reading however will treat the matter in a more relaxed manner than did these mathematical titans of yesteryear.

[3] The important C and C++ programming languages unfortunately are stuck with the reverse order of association, along with division inharmoniously on the same level of syntactic precedence as multiplication. Standard mathematical notation is more elegant:

$$abc/uvw = \frac{(a)(bc)}{(u)(vw)}.$$

[4] The nonassociative *cross product* $\mathbf{B} \times \mathbf{C}$ is introduced in § 15.2.2.

come to a positive product $+ab$? To see why, consider the progression

$$\begin{array}{rcr} & \vdots & \\ (+3)(-b) & = & -3b, \\ (+2)(-b) & = & -2b, \\ (+1)(-b) & = & -1b, \\ (0)(-b) & = & 0b, \\ (-1)(-b) & = & +1b, \\ (-2)(-b) & = & +2b, \\ (-3)(-b) & = & +3b, \\ & \vdots & \end{array}$$

The logic of arithmetic demands that the product of two negative numbers be positive for this reason.

2.1.3 Inequality

If[5]

$$a < b,$$

then necessarily

$$a + x < b + x.$$

However, the relationship between ua and ub depends on the sign of u:

$$\begin{array}{ll} ua < ub & \text{if } u > 0; \\ ua > ub & \text{if } u < 0. \end{array}$$

Also,

$$\begin{array}{ll} \frac{1}{a} > \frac{1}{b} & \text{if } a > 0 \ \text{ or } \ b < 0; \\ \frac{1}{a} < \frac{1}{b} & \text{if } a < 0 \text{ and } b > 0. \end{array}$$

See further § 2.5.6.

[5] Few readers attempting this book will need to be reminded that $<$ means "is less than," that $>$ means "is greater than," or that $\leq$ and $\geq$ respectively mean "is less than or equal to" and "is greater than or equal to."

2.1.4 The change of variable

The mathematician often finds it convenient *to change variables,* introducing new symbols to stand in place of old. For this we have the *change of variable* or *assignment* notation[6]

$$Q \leftarrow P,$$

which means, "in place of P, put Q"; or, "let Q now equal P." For example, if $a^2 + b^2 = c^2$, then the change of variable $2\mu \leftarrow a$ yields the new form $(2\mu)^2 + b^2 = c^2$.

Similar to the change of variable notation is the *definition* notation

$$Q \equiv P.$$

This means, "let the new symbol Q represent P."[7]

The two notations logically mean about the same thing. Subjectively, $Q \equiv P$ identifies a quantity P sufficiently interesting to be given a permanent name Q, whereas $Q \leftarrow P$ implies nothing especially interesting about P or Q but just introduces a (perhaps temporary) new symbol Q to ease the algebra. These concepts grow clearer as examples of the usage arise in the book.

2.2 Quadratics

Differences and sums of squares are conveniently factored as

$$\begin{aligned}
u^2 - v^2 &= (u+v)(u-v),\\
u^2 + v^2 &= (u+iv)(u-iv),\\
u^2 - 2uv + v^2 &= (u-v)^2,\\
u^2 + 2uv + v^2 &= (u+v)^2
\end{aligned} \tag{2.1}$$

[6] There appears to exist no broadly established standard applied mathematical notation for the change of variable, other than the = equal sign, which regrettably does not always fill the role well. One can indeed use the equal sign, but then what does the change of variable $k = k + 1$ mean? It looks like an impossible assertion that k and $k + 1$ were the same. The notation $k \leftarrow k + 1$ by contrast is unambiguous, incrementing k by one. Nevertheless, admittedly, the latter notation has seen only scattered use in the literature.

The C and C++ programming languages use `==` for equality and `=` for assignment (change of variable), as the reader may be aware.

[7] One would never write, $k \equiv k + 1$. Even $k \leftarrow k + 1$ can confuse readers inasmuch as it appears to imply two different values for the same symbol k, but the latter notation is sometimes used anyway when new symbols are unwanted or because more precise alternatives (like $k_n = k_{n-1} + 1$) seem overwrought. Still, usually it is better to introduce a new symbol, as in $j \leftarrow k + 1$.

In some books, $\equiv$ is printed as $\triangleq$.

(where i is the *imaginary unit,* a number defined such that $i^2 = -1$, introduced in § 2.11 below). Useful as these four forms are, however, none of them can directly factor the more general quadratic[8] expression[9]

$$z^2 - 2\beta z + \gamma^2.$$

To factor this, we *complete the square,* writing,

$$\begin{aligned} z^2 - 2\beta z + \gamma^2 &= z^2 - 2\beta z + \gamma^2 + (\beta^2 - \gamma^2) - (\beta^2 - \gamma^2) \\ &= z^2 - 2\beta z + \beta^2 - (\beta^2 - \gamma^2) \\ &= (z - \beta)^2 - (\beta^2 - \gamma^2). \end{aligned}$$

The expression evidently has *roots*—that is, it has values of z that null the expression—where

$$(z - \beta)^2 = (\beta^2 - \gamma^2),$$

or in other words where[10]

$$z = \beta \pm \sqrt{\beta^2 - \gamma^2}. \tag{2.2}$$

This suggests the factoring that

$$z^2 - 2\beta z + \gamma^2 = (z - z_1)(z - z_2), \tag{2.3}$$

where z_1 and z_2 are the two values of z given by (2.2).[11] Substituting into the equation the values of z_1 and z_2 and simplifying proves the suggestion correct.

It follows that the two solutions of the quadratic equation

$$z^2 = 2\beta z - \gamma^2 \tag{2.4}$$

[8] The adjective *quadratic* refers to the algebra of expressions in which no term has greater than second order. Examples of quadratic expressions include x^2, $2x^2 - 7x + 3$ and $x^2 + 2xy + y^2$. By contrast, the expressions $x^3 - 1$ and $5x^2 y$ are *cubic* not quadratic because they contain third-order terms. First-order expressions like $x + 1$ are *linear;* zeroth-order expressions like 3 are *constant.* Expressions of fourth and fifth order are *quartic* and *quintic,* respectively. (If not already clear from the context, *order* basically refers to the number of variables multiplied together in a term. The term $5x^2 y = 5[x][x][y]$ is of third order, for instance.)

[9] The β and γ are Greek letters, the full roster of which you can find in appendix B.

[10] The symbol ± means "+ or −." In conjunction with this symbol, the alternate symbol ∓ occasionally also appears, meaning "− or +"—which is the same thing except that, where the two symbols appear together, $(\pm z) + (\mp z) = 0$.

[11] It suggests it because the expressions on the left and right sides of (2.3) are each quadratic and because the two expressions appear to share the same roots.

are those given by (2.2), which is called *the quadratic formula.* (*Cubic* and *quartic formulas* also exist to extract the roots of polynomials respectively of third and fourth order, but they are much harder. See chapter 10 and its Tables 10.1 and 10.2.)

Nothing prevents the number γ^2 from being negative, incidentally. Section 2.11 will explain how to compute γ in that case, but until then you can regard γ^2 as it were a single symbol. You can replace it by the letter G if this helps.

Introducing the symbols a, b and c, assigning the value $-b/2a$ to β and the value c/a to γ^2, and rearranging terms and factors, one can write (2.2) and (2.4) respectively as

$$\begin{aligned} z &= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, \\ 0 &= az^2 + bz + c. \end{aligned} \tag{2.5}$$

Equation (2.4) is perhaps easier to remember but, depending on the form a quadratic expression takes, (2.5) might be more convenient than (2.4) to use.

Besides (2.1)'s *complex conjugate* form

$$u^2 + v^2 = (u + iv)(u - iv)$$

(whose meaning and use § 2.11 will clarify), there are also the *quadratic conjugate* forms

$$\begin{aligned} u^2 - s &= (u + \sqrt{s})(u - \sqrt{s}), \\ s - u^2 &= (\sqrt{s} + u)(\sqrt{s} - u), \end{aligned} \tag{2.6}$$

useful among other situations where an inconvenient square root appears in a ratio's denominator, as for example

$$\begin{aligned} \frac{\sqrt{3} - 1}{\sqrt{3} + 1} &= \left(\frac{\sqrt{3} - 1}{\sqrt{3} + 1}\right)\left(\frac{\sqrt{3} - 1}{\sqrt{3} - 1}\right) \\ &= \frac{\left(\sqrt{3} - 1\right)^2}{\left(\sqrt{3} + 1\right)\left(\sqrt{3} - 1\right)} \\ &= \frac{4 - 2\sqrt{3}}{2} = 2 - \sqrt{3}. \end{aligned}$$

2.3 Integer and series notation

Sums and products of series arise so often in mathematical work that one finds it convenient to define terse notations to express them. The summation notation

$$\sum_{k=a}^{b} f(k)$$

means to let k equal each of the integers $a, a+1, a+2, \ldots, b$ in turn, evaluating the function $f(k)$ at each k and then adding up the several $f(k)$. For example,[12]

$$\sum_{k=3}^{6} k^2 = 3^2 + 4^2 + 5^2 + 6^2 = \mathrm{0x56}.$$

The similar multiplication notation

$$\prod_{j=a}^{b} f(j)$$

means *to multiply* the several $f(j)$ rather than to add them. The symbols $\sum$ and $\prod$ are respectively the Greek letters for S and P, writ large, and may be regarded as standing for "Sum" and "Product." The j or k is a *dummy variable, index of summation* or *loop counter*—a variable with no independent existence, used only to facilitate the addition or multiplication of the series.[13] (Nothing prevents one from writing $\prod_k$ rather than $\prod_j$, incidentally. For a dummy variable, one can use any letter one likes. However, the general habit of writing $\sum_k$ and $\prod_j$ proves convenient at least in § 4.5.2 and chapter 8, so we start now.)

The product shorthand

$$\begin{aligned} n! &\equiv \prod_{j=1}^{n} j, \\ n!/m! &\equiv \prod_{j=m+1}^{n} j, \end{aligned}$$

[12] What's that 0x56? Answer: it is a *hexadecimal numeral* that represents the same number the familiar, decimal numeral 86 represents. It is an eighty-six. The book's preface explains why the book gives such numbers in hexadecimal. Appendix A tells how to read the numerals, if you do not already know.

[13] Section 7.3 speaks further of the dummy variable.

is very frequently used. The notation $n!$ is pronounced "n factorial." Regarding the notation $n!/m!$, this can of course be regarded correctly as $n!$ divided by $m!$, but it usually proves more amenable to regard the notation as a single unit.[14]

Because multiplication in its more general sense as linear transformation (§ 11.1.1) is not always commutative, we specify that

$$\prod_{j=a}^{b} f(j) = [f(b)][f(b-1)][f(b-2)]\cdots[f(a+2)][f(a+1)][f(a)]$$

rather than the reverse order of multiplication.[15] Multiplication proceeds from right to left. In the event that the reverse order of multiplication is needed, we will use the notation

$$\coprod_{j=a}^{b} f(j) = [f(a)][f(a+1)][f(a+2)]\cdots[f(b-2)][f(b-1)][f(b)].$$

Note that for the sake of definitional consistency,

$$\sum_{k=N+1}^{N} f(k) = 0 + \sum_{k=N+1}^{N} f(k) = 0,$$
$$\prod_{j=N+1}^{N} f(j) = (1) \prod_{j=N+1}^{N} f(j) = 1.$$

This means among other things that

$$0! = 1. \tag{2.7}$$

Context tends to make the notation

$$N, j, k \in \mathbb{Z}$$

unnecessary, but if used (as here and in § 2.5) it states explicitly that N, j and k are integers. (The symbol $\mathbb{Z}$ represents[16] the set of all integers:

[14] One reason among others for this is that factorials rapidly multiply to extremely large sizes, overflowing computer registers during numerical computation. If you can avoid unnecessary multiplication by regarding $n!/m!$ as a single unit, it helps.

[15] The extant mathematical literature seems to lack an established standard on the order of multiplication implied by the "$\prod$" symbol, but this is the order we will use in this book.

[16] The letter $\mathbb{Z}$ recalls the transitive and intransitive German verb *zählen,* "to count."

$\mathbb{Z} \equiv \{\ldots, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, \ldots\}$. The symbol $\in$ means "belongs to" or "is a member of" and § 6.1.4 tells more about it. Integers conventionally get the letters[17] i, j, k, m, n, M and N when available—though i sometimes is avoided because the same letter represents the imaginary unit of § 2.11. Where additional letters are needed ℓ, p and q, plus the capitals of these and the earlier listed letters, can be pressed into service, occasionally joined even by r and s. Greek letters are avoided, as—curiously, in light of the symbol $\mathbb{Z}$—are the Roman letters x, y and z. Refer to appendix B.)

On first encounter, the $\sum$ and $\prod$ notation seems a bit overwrought, whether or not the $\in \mathbb{Z}$ notation also is used. Admittedly it is easier for the beginner to read "$f(1)+f(2)+\cdots+f(N)$" than "$\sum_{k=1}^{N} f(k)$." However, experience shows the latter notation to be extremely useful in expressing more sophisticated mathematical ideas. We will use such notation extensively in this book.

2.4 The arithmetic series

A simple yet useful application of the series sum of § 2.3 is the *arithmetic*[18] *series*

$$\sum_{k=a}^{b} k = a + (a+1) + (a+2) + \cdots + b.$$

Pairing a with b, then $a+1$ with $b-1$, then $a+2$ with $b-2$, etc., the average of each pair is $[a+b]/2$; thus the average of the entire series is $[a+b]/2$. (The pairing may or may not leave an unpaired element at the series midpoint $k = [a+b]/2$, but this changes nothing.) The series has $b - a + 1$ terms. Hence,

$$\sum_{k=a}^{b} k = (b - a + 1)\frac{a+b}{2}. \tag{2.8}$$

Success with this arithmetic series leads one to wonder about the *geometric series* $\sum_{k=0}^{\infty} z^k$. Section 2.6.4 addresses that point.

[17] Though Fortran is less widely used a computer programming language than it once was, it dominated applied-mathematical computer programming for decades, during which the standard way to declare an integral variable to the Fortran compiler was simply to let its name begin with `I`, `J`, `K`, `L`, `M` or `N`; so, this alphabetical convention is fairly well cemented in practice.

[18] As an adjective, the word is pronounced "arithMETic."

Table 2.2: Power properties and definitions.

$$
\begin{aligned}
z^n &\equiv \prod_{j=1}^{n} z, \quad n \ge 0 \\
z &= (z^{1/n})^n = (z^n)^{1/n} \\
\sqrt{z} &\equiv z^{1/2} \\
(uv)^a &= u^a v^a \\
z^{p/q} &= (z^{1/q})^p = (z^p)^{1/q} \\
z^{ab} &= (z^a)^b = (z^b)^a \\
z^{a+b} &= z^a z^b \\
z^{a-b} &= \frac{z^a}{z^b} \\
z^{-b} &= \frac{1}{z^b} \\
j, n, p, q &\in \mathbb{Z}
\end{aligned}
$$

2.5 Powers and roots

This necessarily tedious section discusses powers and roots. It offers no surprises. Table 2.2 summarizes its definitions and results. Readers seeking more rewarding reading may prefer just to glance at the table and then to skip directly to the start of the next section.

In this section, the exponents

$$j, k, m, n, p, q, r, s \in \mathbb{Z}$$

are integers, but the exponents a and b are arbitrary real numbers. (What is a *real number?* Section 2.11 will explain; but, meanwhile, you can think of a real number as just a number, like 4, $-5/2$ or $\sqrt{3}$. There are also *complex numbers* like $7 + i4$, for which this section's results—and, indeed, most of the chapter's and book's results—turn out to be equally valid; but except in eqn. 2.1 we have not met this i yet, so you need not worry about it for now.)

2.5.1 Notation and integral powers

The power notation

$$z^n$$

indicates the number z, multiplied by itself n times. More formally, when the *exponent* n is a nonnegative integer,[19]

$$z^n \equiv \prod_{j=1}^{n} z. \tag{2.9}$$

For example,[20]

$$\begin{aligned} z^3 &= (z)(z)(z), \\ z^2 &= (z)(z), \\ z^1 &= z, \\ z^0 &= 1. \end{aligned}$$

Notice that in general,

$$z^{n-1} = \frac{z^n}{z}.$$

This leads us to extend the definition to negative integral powers with

$$z^{-n} = \frac{1}{z^n}. \tag{2.10}$$

From the foregoing it is plain that

$$\begin{aligned} z^{m+n} &= z^m z^n, \\ z^{m-n} &= \frac{z^m}{z^n}, \end{aligned} \tag{2.11}$$

for any integral m and n. For similar reasons,

$$z^{mn} = (z^m)^n = (z^n)^m. \tag{2.12}$$

On the other hand, from multiplicative associativity and commutivity,

$$(uv)^n = u^n v^n. \tag{2.13}$$

[19] The symbol "$\equiv$" means "$=$", but further usually indicates that the expression on its right serves to define the expression on its left. Refer to § 2.1.4.

[20] The case 0^0 is interesting because it lacks an obvious interpretation. The specific interpretation depends on the nature and meaning of the two zeros. For interest, if $E \equiv 1/\epsilon$, then

$$\lim_{\epsilon\to 0^+} \epsilon^\epsilon = \lim_{E\to\infty} \left(\frac{1}{E}\right)^{1/E} = \lim_{E\to\infty} E^{-1/E} = \lim_{E\to\infty} e^{-(\ln E)/E} = e^0 = 1.$$

2.5.2 Roots

Fractional powers are not something we have defined yet, so for consistency with (2.12) we let

$$(u^{1/n})^n = u.$$

This has $u^{1/n}$ as the number which, raised to the nth power, yields u. Setting

$$v = u^{1/n},$$

it follows by successive steps that

$$\begin{aligned} v^n &= u, \\ (v^n)^{1/n} &= u^{1/n}, \\ (v^n)^{1/n} &= v. \end{aligned}$$

Taking the u and v formulas together, then,

$$(z^{1/n})^n = z = (z^n)^{1/n} \tag{2.14}$$

for any z and integral n.

The number $z^{1/n}$ is called the *nth root* of z—or in the very common case that $n = 2$, the *square root* of z, often written as

$$\sqrt{z}.$$

When z is real and nonnegative, the last notation usually implicitly is taken to mean the real, nonnegative square root. In any case, the power and root operations mutually invert one another.

What about powers expressible neither as n nor as $1/n$, such as the $3/2$ power? If z and w are numbers related by

$$w^q = z,$$

then

$$w^{pq} = z^p.$$

Taking the qth root,

$$w^p = (z^p)^{1/q}.$$

But $w = z^{1/q}$, so this has that

$$(z^{1/q})^p = (z^p)^{1/q},$$

which is to say that it does not matter whether one applies the power or the root first: the result is the same. Extending (2.12) therefore, we define $z^{p/q}$ such that

$$(z^{1/q})^p = z^{p/q} = (z^p)^{1/q}. \tag{2.15}$$

Since one can arbitrarily closely approximate any real number by a ratio of integers, (2.15) implies a power definition for all real exponents.

Equation (2.15) is this subsection's main result. However, § 2.5.3 will find it useful if we can also show here that

$$(z^{1/q})^{1/s} = z^{1/qs} = (z^{1/s})^{1/q}. \tag{2.16}$$

The proof is straightforward. If

$$w \equiv z^{1/qs},$$

then raising to the qs power yields that

$$(w^s)^q = z.$$

Successively taking the qth and sth roots gives that

$$w = (z^{1/q})^{1/s}.$$

By identical reasoning,

$$w = (z^{1/s})^{1/q}.$$

But since $w \equiv z^{1/qs}$, the last two equations imply (2.16), as we have sought.

2.5.3 Powers of products and powers of powers

Per (2.13),

$$(uv)^p = u^p v^p.$$

Raising this equation to the $1/q$ power, we have that

$$\begin{aligned}
(uv)^{p/q} &= [u^p v^p]^{1/q} \\
&= \left[(u^p)^{q/q}(v^p)^{q/q}\right]^{1/q} \\
&= \left[(u^{p/q})^q(v^{p/q})^q\right]^{1/q} \\
&= \left[(u^{p/q})(v^{p/q})\right]^{q/q} \\
&= u^{p/q}v^{p/q}.
\end{aligned}$$

But as argued already in § 2.5.2, some ratio p/q of integers exists to approach any real number a with arbitrary precision, so the last means that

$$(uv)^a = u^a v^a \tag{2.17}$$

for any real a.

On the other hand, per (2.12),

$$z^{pr} = (z^p)^r.$$

Raising this equation to the $1/qs$ power and applying (2.12), (2.15) and (2.16) to reorder the powers, we have that

$$z^{(p/q)(r/s)} = \left(z^{p/q}\right)^{r/s}.$$

By identical reasoning,

$$z^{(p/q)(r/s)} = \left(z^{r/s}\right)^{p/q}.$$

Again as before, p/q and r/s approximate real numbers, so

$$(z^a)^b = z^{ab} = (z^b)^a \tag{2.18}$$

for any real a and b.

2.5.4 Sums of exponents

With (2.11), (2.17) and (2.18), one can reason that

$$z^{(p/q)+(r/s)} = \left(z^{ps+rq}\right)^{1/qs} = \left(z^{ps}z^{rq}\right)^{1/qs} = z^{p/q}z^{r/s},$$

or in other words that

$$z^{a+b} = z^a z^b. \tag{2.19}$$

In the case that $a = -b$,

$$1 = z^{-b+b} = z^{-b}z^b,$$

which implies that

$$z^{-b} = \frac{1}{z^b}. \tag{2.20}$$

But then replacing $-b \leftarrow b$ in (2.19) leads to

$$z^{a-b} = z^a z^{-b},$$

which according to (2.20) is that

$$z^{a-b} = \frac{z^a}{z^b}. \tag{2.21}$$

2.5.5 Summary and remarks

Table 2.2 on page 31 summarizes the section's definitions and results.

Looking ahead to § 2.11, § 3.11 and chapter 5, we observe that nothing in the foregoing analysis requires the base variables z, w, u and v to be real numbers; if complex (§ 2.11), the formulas remain valid. Still, the analysis does imply that the various exponents m, n, p/q, a, b and so on are real numbers. We shall remove this restriction later, purposely defining the action of a complex exponent to comport with the results found here. With such a definition the results apply not only for all bases but also for all exponents, real or complex.

2.5.6 Power-related inequality

If

$$0 < x < y$$

are real numbers (for this subsection alone of the section does not apply to the complex numbers of § 2.11), then inductively—since $0 < (x)(x) < (y)(y)$, $0 < (x)(x^2) < (y)(y^2)$, and so on—we have that $0 < x^p < y^p$ for positive, real, integral p. Moreover, the implication is reversible, so $0 < x^{1/q} < y^{1/q}$, too. Combining these with $a = p/q$ and recalling § 2.1.3,

$$\begin{aligned}
0 < x^a < y^a \qquad & \text{if } a > 0, \\
0 < x^a = y^a \qquad & \text{if } a = 0, \\
0 < y^a < x^a \qquad & \text{if } a < 0.
\end{aligned}$$

Similar reasoning has further that

$$\begin{aligned}
1 < x < x^a \qquad & \text{if } x > 1 \qquad \text{and } a > 1, \\
1 < x^a < x \qquad & \text{if } x > 1 \qquad \text{and } 0 < a < 1, \\
0 < x^a < x < 1 \qquad & \text{if } 0 < x < 1 \text{ and } a > 1, \\
0 < x < x^a < 1 \qquad & \text{if } 0 < x < 1 \text{ and } 0 < a < 1,
\end{aligned}$$

among others.

2.6 Multiplying and dividing power series

A *power series*[21] is a weighted sum of integral powers:

$$A(z) = \sum_{k=-\infty}^{\infty} a_k z^k, \tag{2.22}$$

in which the several weights a_k are arbitrary constants. This section discusses the multiplication and division of power series.

[21] Another name for the *power series* is *polynomial.* The word "polynomial" usually connotes a power series with a finite number of terms, but the two names in fact refer to essentially the same thing.

Professional mathematicians use the terms more precisely. Equation (2.22), they call—or at any rate some of them call—a "power series" only if $a_k = 0$ for all $k < 0$—in other words, technically, not if it includes negative powers of z. They call it a "polynomial" only if it is a "power series" with a finite number of terms. They call (2.22) in general a *Laurent series.*

The name "Laurent series" is a name we shall meet again in § 8.14. In the meantime however we admit that the professionals have vaguely daunted us by adding to the name some pretty sophisticated connotations, to the point that we applied mathematicians (at least in the author's country) seem to feel somehow unlicensed actually to use the name. We tend to call (2.22) a "power series with negative powers," or just "a power series."

This book follows the last usage. You however can call (2.22) a *Laurent series* if you prefer (and if you pronounce it right: "lor-ON"). That is after all exactly what it is. Nevertheless, if you do use the name "Laurent series," be prepared for some people subjectively—for no particular reason—to expect you to establish complex radii of convergence, to sketch some annulus in the Argand plane, and/or to engage in other maybe unnecessary formalities. If that is not what you seek, then you may find it better just to call the thing by the less lofty name of "power series"—or better, if it has a finite number of terms, by the even humbler name of "polynomial."

Semantics. All these names mean about the same thing, but one is expected most carefully always to give the right name in the right place. What a bother! (Someone once told the writer that the Japanese language can give different names to the same object, depending on whether the *speaker* is male or female. The power-series terminology seems to share a spirit of that kin.) If you seek just one word for the thing, the writer recommends that you call it a "power series" and then not worry too much about it until someone objects. When someone does object, you can snow him with the big word "Laurent series," instead.

The experienced scientist or engineer may notice that the above vocabulary omits the name "Taylor series." The vocabulary omits the name because that name fortunately remains unconfused in usage—it means quite specifically a power series without negative powers and tends to connote a representation of some particular function of interest—as we shall see in chapter 8.

2.6.1 Multiplying power series

Given two power series

$$\begin{aligned} A(z) &= \sum_{k=-\infty}^{\infty} a_k z^k, \\ B(z) &= \sum_{k=-\infty}^{\infty} b_k z^k, \end{aligned} \tag{2.23}$$

the product of the two series is evidently

$$P(z) \equiv A(z)B(z) = \sum_{k=-\infty}^{\infty} \left[\left(\sum_{j=-\infty}^{\infty} a_j b_{k-j} \right) z^k \right]. \tag{2.24}$$

2.6.2 Dividing power series

The quotient $Q(z) = B(z)/A(z)$ of two power series is a little harder to calculate, and there are at least two ways to do it. Section 2.6.3 below will do it by matching coefficients, but this subsection does it by long division. For example,

$$\begin{aligned} \frac{2z^2 - 3z + 3}{z-2} &= \frac{2z^2 - 4z}{z-2} + \frac{z+3}{z-2} = 2z + \frac{z+3}{z-2} \\ &= 2z + \frac{z-2}{z-2} + \frac{5}{z-2} = 2z + 1 + \frac{5}{z-2}. \end{aligned}$$

The strategy is to take the dividend[22] $B(z)$ piece by piece, purposely choosing pieces easily divided by $A(z)$.

If you feel that you understand the example, then that is really all there is to it, and you can skip over several pages of thick notation straight to § 2.6.4 if you like. Indeed, to skip is recommended to many or most readers—though, if you do skip, you might nonetheless glance along the way at Tables 2.3 and 2.4, which summarize and formalize the procedure the example has used and which also include the clever, alternate procedure of § 2.6.3.

Formally, we prepare the long division $B(z)/A(z)$ by writing,

$$B(z) = A(z)Q_n(z) + R_n(z), \tag{2.25}$$

[22] If $Q(z)$ is a *quotient* and $R(z)$ a *remainder,* then $B(z)$ is a *dividend* (or *numerator*) and $A(z)$ a *divisor* (or *denominator*). Such are the Latin-derived names of the parts of a long division.

where $R_n(z)$ is a *remainder* (being the part of $B[z]$ *remaining* to be divided); and

$$\begin{aligned}
A(z) &= \sum_{k=-\infty}^{K} a_k z^k, \quad a_K \neq 0, \\
B(z) &= \sum_{k=-\infty}^{N} b_k z^k, \\
R_N(z) &= B(z), \\
Q_N(z) &= 0, \\
R_n(z) &= \sum_{k=-\infty}^{n} r_{nk} z^k, \\
Q_n(z) &= \sum_{k=n-K+1}^{N-K} q_k z^k,
\end{aligned} \tag{2.26}$$

where K and N identify the greatest orders k of z^k present in $A(z)$ and $B(z)$, respectively.

Well, that is a lot of symbology. What does it mean? The key to understanding it lies in understanding (2.25), which is not one but several equations—one equation for each value of n, where $n = N, N-1, N-2, \ldots$. The dividend $B(z)$ and the divisor $A(z)$ stay the same from one n to the next, but the quotient $Q_n(z)$ and the remainder $R_n(z)$ change. At start, $Q_N(z) = 0$ while $R_N(z) = B(z)$, but the thrust of the long division process is to build $Q_n(z)$ up by wearing $R_n(z)$ down. The goal is to grind $R_n(z)$ away to nothing, to make it disappear as $n \to -\infty$.

As in the example, we pursue the goal by choosing from $R_n(z)$ an easily divisible piece containing the whole high-order term of $R_n(z)$. The piece we choose is $(r_{nn}/a_K)z^{n-K}A(z)$, which we add and subtract from (2.25) to obtain the form

$$B(z) = A(z)\left[Q_n(z) + \frac{r_{nn}}{a_K}z^{n-K}\right] + \left[R_n(z) - \frac{r_{nn}}{a_K}z^{n-K}A(z)\right].$$

Matching this equation against the desired iterate

$$B(z) = A(z)Q_{n-1}(z) + R_{n-1}(z)$$

and observing from the definition of $Q_n(z)$ that $Q_{n-1}(z) = Q_n(z) +$

$q_{n-K}z^{n-K}$, we find that

$$\begin{aligned} q_{n-K} &= \frac{r_{nn}}{a_K}, \\ R_{n-1}(z) &= R_n(z) - q_{n-K}z^{n-K}A(z), \end{aligned} \tag{2.27}$$

where no term remains in $R_{n-1}(z)$ higher than a z^{n-1} term.

To begin the actual long division, we initialize

$$R_N(z) = B(z),$$

for which (2.25) is trivially true if $Q_N(z) = 0$. Then we iterate per (2.27) as many times as desired. If an infinite number of times, then so long as $R_n(z)$ tends to vanish as $n \to -\infty$, it follows from (2.25) that

$$\frac{B(z)}{A(z)} = Q_{-\infty}(z). \tag{2.28}$$

Iterating only a finite number of times leaves a remainder,

$$\frac{B(z)}{A(z)} = Q_n(z) + \frac{R_n(z)}{A(z)}, \tag{2.29}$$

except that it may happen that $R_n(z) = 0$ for sufficiently small n.

Table 2.3 summarizes the long-division procedure.[23] In its q_{n-K} equation, the table includes also the result of § 2.6.3 below.

The foregoing algebra is probably punishing enough; but if not, then one can further observe in light of Table 2.3 that if[24]

$$\begin{aligned} A(z) &= \sum_{k=K_o}^{K} a_k z^k, \\ B(z) &= \sum_{k=N_o}^{N} b_k z^k, \end{aligned}$$

then

$$R_n(z) = \sum_{k=n-(K-K_o)+1}^{n} r_{nk}z^k \quad \text{for all } n < N_o + (K - K_o). \tag{2.30}$$

[23] [156, § 3.2]

[24] The notations K_o, a_k and z^k are usually pronounced, respectively, as "K naught," "a sub k" and "z to the k" (or, more fully, "z to the kth power")—at least in the author's country.

Table 2.3: Dividing power series through successively smaller powers.

$$\begin{aligned}
B(z) &= A(z)Q_n(z) + R_n(z) \\
A(z) &= \sum_{k=-\infty}^{K} a_k z^k, \quad a_K \neq 0 \\
B(z) &= \sum_{k=-\infty}^{N} b_k z^k \\
R_N(z) &= B(z) \\
Q_N(z) &= 0 \\
R_n(z) &= \sum_{k=-\infty}^{n} r_{nk} z^k \\
Q_n(z) &= \sum_{k=n-K+1}^{N-K} q_k z^k \\
q_{n-K} &= \frac{r_{nn}}{a_K} = \frac{1}{a_K}\left(b_n - \sum_{k=n-K+1}^{N-K} a_{n-k} q_k\right) \\
R_{n-1}(z) &= R_n(z) - q_{n-K} z^{n-K} A(z) \\
\frac{B(z)}{A(z)} &= Q_{-\infty}(z)
\end{aligned}$$

That is, the remainder has residual order one less than the divisor has. The reason for this, of course, is that we have strategically planned the long-division iteration precisely to cause the divisor's leading term to cancel the remainder's leading term at each step.[25] (If not clear from the context, a polynomial's *residual order* is the difference between the least and greatest orders of its several terms. For example, the residual order of $9x^5 - 7x^4 + 6x^3$ is two because $5 - 3 = 2$—or, if you prefer, because $9x^5 - 7x^4 + 6x^3 = [x^3][9x^2 - 7x + 6]$, where $9x^2 - 7x + 6$ is of second order.[26])

The long-division procedure of Table 2.3 extends the quotient $Q_n(z)$ through successively smaller powers of z. Often, however, one prefers to extend the quotient through successively *larger* powers of z, where a z^K term is $A(z)$'s term of least rather than greatest order. In this case, the long division goes by the complementary rules of Table 2.4.

2.6.3 Dividing power series by matching coefficients

There is another, sometimes quicker way to divide power series than by the long division of § 2.6.2. One can divide them by matching coefficients.[27] If

$$Q_\infty(z) = \frac{B(z)}{A(z)}, \tag{2.31}$$

[25] If a more formal demonstration of (2.30) is wanted, then consider per (2.27) that

$$R_{m-1}(z) = R_m(z) - \frac{r_{mm}}{a_K} z^{m-K} A(z).$$

If the least-order term of $R_m(z)$ is a z^{N_o} term (as clearly is the case at least for the initial remainder $R_N[z] = B[z]$), then according to the equation so also must the least-order term of $R_{m-1}(z)$ be a z^{N_o} term, unless an even lower-order term is contributed by the product $z^{m-K}A(z)$. But that very product's term of least order is a $z^{m-(K-K_o)}$ term. Under these conditions, evidently the least-order term of $R_{m-1}(z)$ is a $z^{m-(K-K_o)}$ term when $m - (K - K_o) \le N_o$; otherwise a z^{N_o} term. This is better stated after the change of variable $n + 1 \leftarrow m$: the least-order term of $R_n(z)$ is a $z^{n-(K-K_o)+1}$ term when $n < N_o + (K - K_o)$; otherwise a z^{N_o} term.

The greatest-order term of $R_n(z)$ is by definition a z^n term. So, in summary, when $n < N_o + (K - K_o)$, the terms of $R_n(z)$ run from $z^{n-(K-K_o)+1}$ through z^n, which is exactly the claim (2.30) makes.

[26] But what of $0x^5 - 7x^4 + 6x^3$ with its leading null coefficient? Is *this* polynomial's residual order also two?

Answer: that depends on what you mean. The strictly semantic question of what a mere phrase ought to signify is not always very interesting. After all, an infinite number of practically irrelevant semantic distinctions *could* be drawn. The applied mathematician lacks the time.

Anyway, whatever semantics might eventually be settled upon, at least (2.30) and Table 2.3 remain unambiguous.

[27] [101][58, § 2.5]

Table 2.4: Dividing power series through successively larger powers.

$$
\begin{aligned}
B(z) &= A(z)Q_n(z) + R_n(z) \\
A(z) &= \sum_{k=K}^{\infty} a_k z^k, \quad a_K \neq 0 \\
B(z) &= \sum_{k=N}^{\infty} b_k z^k \\
R_N(z) &= B(z) \\
Q_N(z) &= 0 \\
R_n(z) &= \sum_{k=n}^{\infty} r_{nk} z^k \\
Q_n(z) &= \sum_{k=N-K}^{n-K-1} q_k z^k \\
q_{n-K} &= \frac{r_{nn}}{a_K} = \frac{1}{a_K}\left(b_n - \sum_{k=N-K}^{n-K-1} a_{n-k} q_k\right) \\
R_{n+1}(z) &= R_n(z) - q_{n-K} z^{n-K} A(z) \\
\frac{B(z)}{A(z)} &= Q_\infty(z)
\end{aligned}
$$

where

$$\begin{aligned}A(z) &= \sum_{k=K}^{\infty} a_k z^k, \quad a_K \neq 0,\\ B(z) &= \sum_{k=N}^{\infty} b_k z^k\end{aligned}$$

are known and

$$Q_\infty(z) = \sum_{k=N-K}^{\infty} q_k z^k$$

is to be calculated, then one can rearrange (2.31) as that

$$A(z)Q_\infty(z) = B(z).$$

Expanding the rearranged equation's left side according to (2.24) and changing indices suitably on both sides,

$$\sum_{n=N}^{\infty}\left[\left(\sum_{k=N-K}^{n-K} a_{n-k}q_k\right)z^n\right] = \sum_{n=N}^{\infty} b_n z^n.$$

But for this to hold for all z, the coefficients must match for each n:

$$\sum_{k=N-K}^{n-K} a_{n-k}q_k = b_n, \quad n \geq N.$$

Transferring all terms but $a_K q_{n-K}$ to the equation's right side and dividing by a_K, we have that

$$q_{n-K} = \frac{1}{a_K}\left(b_n - \sum_{k=N-K}^{n-K-1} a_{n-k}q_k\right), \quad n \geq N. \tag{2.32}$$

Equation (2.32) computes the coefficients of $Q(z)$, each coefficient depending not on any remainder but directly on the coefficients earlier computed.

The coefficient-matching technique of this subsection is easily adapted to the division of series in decreasing, rather than increasing, powers of z. Tables 2.3 and 2.4 incorporate the technique both ways.

Admittedly, the fact that (2.32) yields a sequence of coefficients does not necessarily mean that the resulting power series $Q_\infty(z)$ converges to some definite value over a given domain. Consider for instance (2.36), which diverges when[28] $|z| > 1$, even though all its coefficients are known. At least (2.32) is correct when $Q_\infty(z)$ does converge. Even when $Q_\infty(z)$ as such does not converge, however, often what interest us are only the series' first several terms

$$Q_n(z) = \sum_{k=N-K}^{n-K-1} q_k z^k.$$

In this case, in light of (2.25),

$$Q_\infty(z) = \frac{B(z)}{A(z)} = Q_n(z) + \frac{R_n(z)}{A(z)} \tag{2.33}$$

and convergence is not an issue. Solving (2.33) or (2.25) for $R_n(z)$,

$$R_n(z) = B(z) - A(z)Q_n(z). \tag{2.34}$$

2.6.4 Common power-series quotients and the geometric series

Frequently encountered power-series quotients, calculated by the long division of § 2.6.2, computed by the coefficient matching of § 2.6.3, and/or verified by multiplying, include[29]

$$\frac{1}{1 \pm z} = \begin{cases} \displaystyle\sum_{k=0}^{\infty} (\mp z)^k, & |z| < 1; \\ \displaystyle -\sum_{k=-\infty}^{-1} (\mp z)^k, & |z| > 1. \end{cases} \tag{2.35}$$

Equation (2.35) almost incidentally answers a question which has arisen in § 2.4 and which often arises in practice: to what total does the infinite *geometric series* $\sum_{k=0}^{\infty} z^k$, $|z| < 1$, sum? Answer: it sums exactly to $1/(1-z)$. However, there is a simpler, more aesthetic, more instructive way to demonstrate the same thing, as follows. Let

$$S_0 \equiv \sum_{k=0}^{\infty} z^k, \quad |z| < 1.$$

[28] See footnote 29.

[29] The notation $|z|$ represents the magnitude of z. For example, $|5| = 5$ and $|8| = 8$, but also $|-5| = 5$ and $|-8| = 8$.

Multiplying by z yields that

$$zS_0 = \sum_{k=1}^{\infty} z^k.$$

Subtracting the latter equation from the former leaves that

$$(1-z)S_0 = 1,$$

which, after dividing by $1-z$, implies that

$$S_0 \equiv \sum_{k=0}^{\infty} z^k = \frac{1}{1-z}, \quad |z| < 1, \tag{2.36}$$

as was to be demonstrated.

2.6.5 Variations on the geometric series

Besides being more aesthetic than the long division of § 2.6.2, the difference technique of § 2.6.4 permits one to extend the basic geometric series in several ways. For instance, one can compute the sum

$$S_1 \equiv \sum_{k=0}^{\infty} kz^k, \quad |z| < 1$$

(which arises in, among others, Planck's quantum blackbody radiation calculation[30]) as follows. Multiply the unknown S_1 by z, producing

$$zS_1 = \sum_{k=0}^{\infty} kz^{k+1} = \sum_{k=1}^{\infty} (k-1)z^k.$$

Subtract zS_1 from S_1, leaving

$$(1-z)S_1 = \sum_{k=0}^{\infty} kz^k - \sum_{k=1}^{\infty} (k-1)z^k = \sum_{k=1}^{\infty} z^k = z\sum_{k=0}^{\infty} z^k = \frac{z}{1-z},$$

where we have used (2.36) to collapse the last sum. Dividing by $1-z$,

$$S_1 \equiv \sum_{k=0}^{\infty} kz^k = \frac{z}{(1-z)^2}, \quad |z| < 1, \tag{2.37}$$

[30] [116]

which was to be found.

Further series of the kind, such as $\sum_k k^2 z^k$, can be calculated in like manner as the need for them arises. Introducing the derivative, though, chapter 4 does it better:[31]

$$S_n \equiv \sum_{k=0}^{\infty} k^n z^k = z\frac{dS_{n-1}}{dz}, \quad n \in \mathbb{Z},\ n > 0; \tag{2.38}$$

except that you must first read chapter 4 or otherwise know about derivatives to understand this.[32] See also § 8.1.

2.7 Indeterminate constants, independent variables and dependent variables

Mathematical models use *indeterminate constants, independent variables* and *dependent variables.* The three are best illustrated by example as follows. Consider the time t a sound needs to travel from its source to a distant listener:

$$t = \frac{\Delta r}{v_{\mathrm{sound}}},$$

where Δr is the distance from source to listener and v_{sound} is the speed of sound. Here, v_{sound} is an indeterminate constant (given particular atmospheric conditions, it does not vary), Δr is an independent variable, and t is a dependent variable. The model gives t as a function of Δr; so, if you tell the model how far the listener sits from the sound source, then the model returns the time the sound needs to propagate from one to the other. Regarding the third quantity, the indeterminate constant v_{sound}, one conceives of this as having a definite, fixed value; yet, oddly, notwithstanding that the value is (or is thought of as) fixed, the model's abstract validity may not depend on whether one actually knows what the value is (if I tell you that sound goes at 350 m/s, but later you find out that the real figure is 331 m/s, this probably does not ruin the theoretical part of your analysis; you may only have to recalculate numerically). Knowing the value is not the point. The point is that conceptually there preëxists some correct figure for the

[31] It does not really matter, but you can regard k^n to be unity—that is, $k^n = 1$—when $n = 0$ and $k = 0$, though $n = 0$ technically lies outside the domain of (2.38) as expressed. See also footnote 20.

[32] This of course is a forward reference. Logically, (2.38) belongs in or after chapter 4, but none of the earlier chapters use it, so it is kept here with the rest of the geometric-series math. See chapter 4's footnote 34.

indeterminate constant; that sound goes at some constant speed—whatever it is—and that one can calculate the delay in terms of this.[33]

Though the three kinds of quantity remain in some sense distinct, still, which particular quantity one regards as an indeterminate constant, as an independent variable, or as a dependent variable may depend less upon any property of the quantity itself—or of the thing the quantity quantifies—than upon the mathematician's point of view. Moreover, the mathematician's point of view can waver. The same model in the example would remain valid if atmospheric conditions were changing (v_{sound} would then be an independent variable) or if the model were used in designing a musical concert hall[34] to suffer a maximum acceptable sound time lag from the stage to the hall's back row (t would then be an independent variable; Δr, dependent). Occasionally one goes so far as deliberately to shift one's point of view in mid-analysis—now regarding as an independent variable, for instance, that which one a moment ago had regarded as an indeterminate constant (a typ-

[33] Besides the quantities themselves, there is also the manner in which, or pattern by which, the quantities relate to one another. The philosophy that attends this distinction lies mostly beyond the book's scope, but it still seems worth a footnote to quote Bertrand Russell (1872–1970) on the subject:

> Given any propositional concept, or any unity . . . , which may in the limit be simple, its constituents are in general of two sorts: (1) those which may be replaced by anything else whatever without destroying the unity of the whole; (2) those which have not this property. Thus in "the death of Caesar," anything else may be substituted for Caesar, but a proper name must not be substituted for *death,* and hardly anything can be substituted for *of.* Of the unity in question, the former class of constituents will be called *terms,* the latter *concepts.* . . . [Emphases in the original.][139, appendix A, § 482]

Sections 2.10 and 2.11, and a few later ones, glance upon the matter.

[34] As a child, were you ever let to read one of those trendy, second-rate arithmetic textbooks that had you calculate such as the quantity of air in an astronaut's round helmet? One could have calculated the quantity of water in a kitchen's mixing bowl just as well, but astronauts' helmets are so much more interesting than bowls, you see. (Whether you will have endured the condescending frivolity specifically of the ersatz astronaut's textbook depends largely on when and where you were raised. Trends will come and go, but maybe you will have met another year's version of the same kind of thing.)

So, what of the concert hall? The chance that the typical reader will ever specify the dimensions of a real musical concert hall is of course vanishingly small. However, it is the idea of the example that matters here, because the chance that the typical reader will ever specify *something* technical is quite large. Although sophisticated models with many factors and terms do indeed play a large role in engineering, the great majority of practical engineering calculations—for quick, day-to-day decisions where small sums of money and negligible risk to life are at stake, or for proëmial or exploratory analysis—are done with models hardly more sophisticated than the one shown here. So, maybe the concert-hall example is not so unreasonable, after all.

ical case of such a shift arising in the solution of differential equations by the method of unknown coefficients, § 9.5).

It matters which symbol represents which of the three kinds of quantity in part because, in calculus, one analyzes how change in independent variables affects dependent variables as indeterminate constants remain fixed.

(Section 2.3 has introduced the dummy variable, which the present section's threefold taxonomy seems to exclude. However, in fact, most dummy variables are just independent variables—a few are dependent variables—whose scope is restricted to a particular expression. Such a dummy variable does not seem very "independent," of course; but its dependence is on the operator controlling the expression, not on some other variable within the expression. Within the expression, the dummy variable fills the role of an independent variable; without, it fills no role because logically it does not exist there. Refer to §§ 2.3 and 7.3.)

2.8 Exponentials and logarithms

In § 2.5 we have considered the power operation z^a, where (in § 2.7's language) the independent variable z is the base and the indeterminate constant a is the exponent. There is another way to view the power operation, however. One can view it as the *exponential* operation

$$a^z,$$

where the variable z is the exponent and the constant a is the base.

2.8.1 The logarithm

The exponential operation follows the same laws the power operation follows; but, because the variable of interest is now the exponent rather than the base, the inverse operation is not the root but rather the *logarithm:*

$$\log_a(a^z) = z. \tag{2.39}$$

For example, $\log_2 8 = 3$. The logarithm $\log_a w$ answers the question, "To what power must I raise a to get w?"

Raising a to the power of the last equation, we have that

$$a^{\log_a(a^z)} = a^z.$$

With the change of variable $w \leftarrow a^z$, this is that

$$a^{\log_a w} = w. \tag{2.40}$$

Thus, the exponential and logarithmic operations mutually invert one another.

2.8.2 Properties of the logarithm

The basic properties of the logarithm include that

$$\begin{aligned}
\log_a uv &= \log_a u + \log_a v, &(2.41)\\
\log_a \frac{u}{v} &= \log_a u - \log_a v, &(2.42)\\
\log_a(w^z) &= z\log_a w, &(2.43)\\
w^z &= a^{z\log_a w}, &(2.44)\\
\log_b w &= \frac{\log_a w}{\log_a b}. &(2.45)
\end{aligned}$$

Of these, (2.41) follows from the steps

$$\begin{aligned}
(uv) &= (u)(v),\\
(a^{\log_a uv}) &= (a^{\log_a u})(a^{\log_a v}),\\
a^{\log_a uv} &= a^{\log_a u + \log_a v};
\end{aligned}$$

and (2.42) follows by similar reasoning. Equations (2.43) and (2.44) follow from the steps

$$\begin{aligned}
w^z &= \quad (w^z) = (w)^z,\\
w^z &= a^{\log_a(w^z)} = (a^{\log_a w})^z,\\
w^z &= a^{\log_a(w^z)} = a^{z\log_a w}.
\end{aligned}$$

Equation (2.45) follows from the steps

$$\begin{aligned}
w &= b^{\log_b w},\\
\log_a w &= \log_a(b^{\log_b w}),\\
\log_a w &= \log_b w \log_a b.
\end{aligned}$$

Among other purposes, (2.41) through (2.45) serve respectively to transform products to sums, quotients to differences, powers to products, exponentials to differently based exponentials, and logarithms to differently based logarithms. Table 2.5 repeats the equations along with (2.39) and (2.40) (which also emerge as restricted forms of eqns. 2.43 and 2.44), thus summarizing the general properties of the logarithm.

Table 2.5: General properties of the logarithm.

$$\begin{aligned}
\log_a uv &= \log_a u + \log_a v \\
\log_a \frac{u}{v} &= \log_a u - \log_a v \\
\log_a(w^z) &= z \log_a w \\
w^z &= a^{z \log_a w} \\
\log_b w &= \frac{\log_a w}{\log_a b} \\
\log_a(a^z) &= z \\
w &= a^{\log_a w}
\end{aligned}$$

2.9 The triangle

This section develops several facts about the triangle.[35]

2.9.1 Area

The area of a right triangle[36] is half the area of the corresponding rectangle. This is seen by splitting a rectangle down its diagonal into a pair of right triangles of equal size. The fact that *any* triangle's area is half its base length times its height is seen by dropping a perpendicular from one point of the triangle to the opposite side (see Fig. 1.3 on page 15), thereby dividing the triangle into two right triangles, for each of which the fact is true. In algebraic symbols,

$$A = \frac{bh}{2}, \tag{2.46}$$

where A stands for area, b for base length, and h for perpendicular height.

[35] Fashion seems to ask a writer to burden the plain word "triangle" with various accurate but not-very-helpful adjectives like "planar" and "Euclidean." We like planes and Euclid (§ 2.9.5) but would resist the fashion. Readers already know what a triangle is.

[36] As the reader likely knows, a *right triangle* is a triangle, one of whose three angles is perfectly square.

2.9.2 The triangle inequalities

Any two sides of a triangle together are longer than the third alone, which itself is longer than the difference between the two. In symbols,

$$|a - b| < c < a + b, \tag{2.47}$$

where a, b and c are the lengths of a triangle's three sides. These are the *triangle inequalities.* The truth of the sum inequality, that $c < a + b$, is seen by sketching some triangle on a sheet of paper and asking: if c is the direct route between two points and $a + b$ is an indirect route, then how can $a + b$ not be longer? Of course the sum inequality is equally good on any of the triangle's three sides, so one can write that $a < c + b$ and $b < c + a$ just as well as that $c < a + b$. Rearranging the a and b inequalities, we have that $a - b < c$ and $b - a < c$, which together say that $|a - b| < c$. The last is the difference inequality, completing (2.47)'s proof.[37]

2.9.3 The sum of interior angles

A triangle's three interior angles[38] sum to $2\pi/2$. One way to see the truth of this fact is to imagine a small car rolling along one of the triangle's sides. Reaching the corner, the car turns to travel along the next side, and so on round all three corners to complete a circuit, returning to the start. Since the car again faces the original direction, we reason that it has turned a total of 2π, a full revolution. But the angle ϕ the car turns at a corner and the triangle's inner angle ψ there together form the straight angle $2\pi/2$ (the sharper the inner angle, the more the car turns: see Fig. 2.2). In mathematical notation,

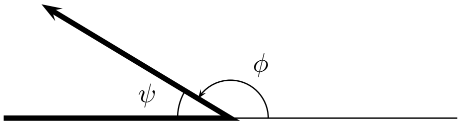
$$\begin{aligned} \phi_1 + \phi_2 + \phi_3 &= 2\pi; \\ \phi_k + \psi_k &= \frac{2\pi}{2}, \quad k \in \{1, 2, 3\}; \end{aligned}$$

[37] Section 13.9 proves the triangle inequalities more generally, though regrettably without recourse to this subsection's properly picturesque geometrical argument.

[38] Most readers will already know the notation 2π and its meaning as the angle of full revolution. The notation is properly introduced in §§ 3.1, 3.6 and 8.11 at any rate. Briefly nevertheless, the symbol 2π represents a complete turn, a full circle, a spin to face the same direction as before. Hence (for instance) $2\pi/4$ represents a square turn or right angle.

You may be used to the notation 360° in place of 2π; but, for the reasons explained in appendix A and in footnote 17 of chapter 3, this book tends to avoid the notation 360°.

Figure 2.2: The sum of a triangle's inner angles: turning at the corner.



where ψ_k and ϕ_k are respectively the triangle's inner angles and the angles through which the car turns; and where the *set notation*[39] $k \in \{1, 2, 3\}$, met in § 2.3, means that the last equation holds whether $k = 1$, 2 or 3. Solving the latter equation for ϕ_k and substituting into the former yields that

$$\psi_1 + \psi_2 + \psi_3 = \frac{2\pi}{2}, \tag{2.48}$$

which was to be demonstrated.

Extending the same technique to the case of an n-sided polygon,

$$\begin{aligned} \sum_{k=1}^{n} \phi_k &= 2\pi, \\ \phi_k + \psi_k &= \frac{2\pi}{2}. \end{aligned}$$

Solving the latter equation for ϕ_k and substituting into the former, we have that

$$\sum_{k=1}^{n} \left(\frac{2\pi}{2} - \psi_k \right) = 2\pi,$$

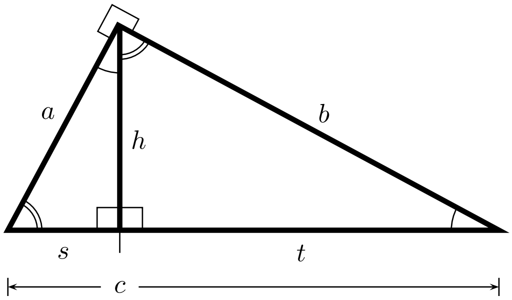
or in other words that

$$\sum_{k=1}^{n} \psi_k = (n-2)\frac{2\pi}{2}. \tag{2.49}$$

Equation (2.48) is then seen to be a special case of (2.49) with $n = 3$.

[39] Applied mathematicians tend to less enthusiasm than professional mathematicians do over set notation like the membership symbol $\in$, but such notation still finds use in applications now and again as, for example, it does in this instance.

Figure 2.3: A right triangle subdivided.



2.9.4 The Pythagorean theorem

Along with Euler's formula (5.12), the fundamental theorem of calculus (7.2), Cauchy's integral formula (8.29), Fourier's equation (18.1) and maybe a few others, the Pythagorean theorem is one of the most famous results in all of mathematics.

The introduction to chapter 1 has proved the theorem.[40] Alternately and less concisely, Fig. 2.3 too can prove it as follows. In the figure are three right triangles: a small one *sha* on the left; a medium one *htb* on the right; and, if the small and medium are joined together, a large one *abc* overall. The small triangle shares an angle *sa* with the large which, in light of (2.48) and of that both are right triangles, means that *all three angles* of the small triangle equal the corresponding three angles of the large, which in turn means that the small and large triangles are *similar* to one another; that is, the small triangle is merely a scaled, rotated, possibly reflected version of the large. The medium triangle shares an angle *tb* with the large which likewise means that the medium and large triangles are similar.

Insofar as triangles are similar (§ 2.9.5), corresponding ratios of their

[40] The elegant proof of chapter 1 is simpler than the one famously given by the ancient geometer Euclid, yet more appealing than alternate proofs often found in print. Whether Euclid was acquainted with either of the two Pythagorean proofs the book you are reading gives, or indeed was acquainted with both, this writer does not know; but it is possible [182, "Pythagorean theorem," 02:32, 31 March 2006] that Euclid chose his proof because it comported better with the restricted set of geometrical elements with which he permitted himself to work. Be that as it may, the present writer encountered the proof of chapter 1 somewhere years ago and has never seen it in print since (but has not looked for it, either), so can claim no credit for originating it. Unfortunately the citation is now long lost. A current, electronic source for the proof of chapter 1 is [182] as cited earlier in this footnote.

sides must be equal:

$$\frac{a}{c} = \frac{s}{a}; \quad \frac{b}{c} = \frac{t}{b}.$$

Rearranging factors,

$$\frac{a^2}{c} = s; \quad \frac{b^2}{c} = t.$$

The sum of the last line's two equations is that

$$\frac{a^2 + b^2}{c} = s + t;$$

or rather, because $s + t = c$, that

$$\frac{a^2 + b^2}{c} = c,$$

which, when multiplied by c, yields the Pythagorean theorem (1.1),

$$a^2 + b^2 = c^2.$$

This paragraph's proof has found no new conclusion but one is glad to discover that the Pythagorean theorem seems to remain true no matter how one reaches it.[41]

The Pythagorean theorem is readily extended into three dimensions as

$$a^2 + b^2 + h^2 = r^2, \tag{2.50}$$

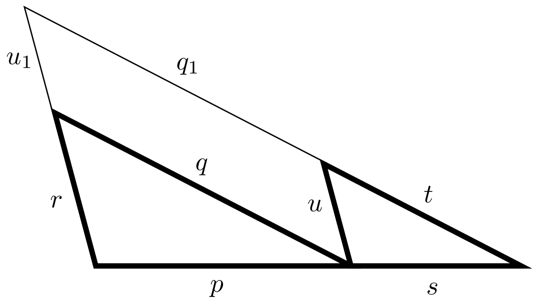
where h is an altitude perpendicular to both a and b and thus also to c; and where r is the corresponding three-dimensional diagonal—the diagonal of the right triangle whose legs are c and h. Inasmuch as (1.1) applies to any right triangle, it follows that $c^2 + h^2 = r^2$, which equation expands directly to yield (2.50).

2.9.5 Congruence and Euclid's similar proportionality

The careful reader might have noticed the presence of an unproved assertion in the last subsection. The assertion was this: that, if each of a triangle's angles equals the corresponding angle of another triangle, then the one triangle is a scaled, rotated, possibly reflected version of the other; or, stated

[41] The source of this proof is an otherwise forgotten book that stood on a shelf in the writer's high-school library in 1983. The writer still remembers the proof but has long since lost the citation.

Figure 2.4: Proportionality of similar triangles (or the AA criterion).



another way, that, if the respective angles are equal, then the respective sides though maybe not equal are at least proportionate.

Does such an assertion not require proof?

Whether such an assertion requires proof depends on the vividness of one's geometrical intuition and/or on one's attitude toward the kind of assertion the assertion is. If one's attitude is an applied attitude and if intuition finds the assertion obvious enough, then the assertion may not require proof.

Even given an applied attitude, though, it's a borderline case, isn't it? Fortunately, the ancient geometer Euclid has supplied us a proof in book VI of his *Elements* [52]. Though we will not fully detail Euclid's proof we can digest it as follows.

Fig. 2.4 arranges two triangles *pqr* and *stu* whose respective angles are equal. Extending the indicated sides and exploiting the equal angles, the figure extrudes a parallelogram from the two triangles. Being opposite sides of a parallelogram,

$$q_1 = q; \quad u_1 = u. \tag{2.51}$$

Because[42] q runs parallel to $q_1 + t$,

$$\frac{r}{u_1} = \frac{p}{s}.$$

[42] The pure mathematician who loves Euclid can abhor such shortcuts! "Because"?! Where is the proof that the lines are parallel, the purist wants to know? Where is the proof that opposite sides of a parallelogram are equal in length? Notwithstanding, busily rushing ahead, we applicationists fail to notice the purist's consternation.

Applying (2.51),

$$\frac{r}{u} = \frac{p}{s}.$$

Rearranging factors,

$$\frac{r}{p} = \frac{u}{s}. \tag{2.52}$$

By the same kind of argument,

$$\frac{t}{s} = \frac{q}{p},$$

or, reversed and inverted,

$$\frac{p}{q} = \frac{s}{t}. \tag{2.53}$$

Multiplying (2.52) by (2.53),

$$\frac{r}{q} = \frac{u}{t}. \tag{2.54}$$

Together, (2.52), (2.53) and (2.54) complete Euclid's proof. Proceeding nevertheless if you wish, one may further rearrange the three equations respectively to read,

$$\frac{r}{u} = \frac{p}{s}, \quad \frac{p}{s} = \frac{q}{t}, \quad \frac{q}{t} = \frac{r}{u};$$

or, in other words,

$$\frac{p}{s} = \frac{q}{t} = \frac{r}{u}, \tag{2.55}$$

which, using this book's modern applied notation (which Euclid lacked) condenses Euclid's elegant finding into a single line.

The finding is indeed elegant. However, as earlier asked, to an applicationist, is the finding's proof *necessary?* Could one not instead just look at Fig. 2.4 and *see* that the triangles are proportionate?[43] (By *see,* we do not precisely mean, "see with the eyes." The eyes do help but what we here mean is *to perceive mentally* as in chapter 1.)

After all, if these triangles were not proportionate, then how else should they be?

[43] The modern professional mathematician might answer, "Sure, if you are an engineer or the like, then go ahead. However, if you wish to be a proper analyst, then no. We have a theory of *metric spaces.* We have theories of other things, too. You needn't just skip all that."

This book, a book of applied mathematics, will not go the route of metric spaces. However, [147] (which though not open source has long been available in an inexpensive paperback edition) offers the interested reader a comparatively accessible account of such.

Until Euclid's proof had been given, would any mathematician (whether applied or professional) who had sketched a triangle on a sheet of paper really have conjectured a different result? Was the investigation's outcome ever in doubt? To put it another way: suppose that the investigation had found that the triangles were disproportionate; would its finding have been believed? Or is the proof merely a salutary logical exercise, a noble sport whose object is to hunt down and annihilate every premise one can possibly do without?

Effort and mental energy are not unlimited resources. Time spent in speculative foundational investigations may be regretted, if that time is as a consequence later unavailable to devote to the acquisition of more advanced, more practical mathematical techniques. Mathematics is after all far too vast for a single mathematician to master the whole of it during a single lifetime. One must prioritize.

The question is left for the reader to contemplate.

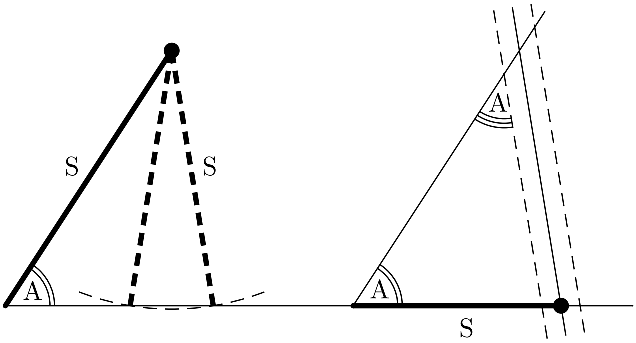
Meanwhile, a pair of definitions: as § 2.9.4 has already suggested, two triangles are *similar* if their respective angles are equal, as, for example, in Fig. 2.4; and two triangles are *congruent* (this definition is new) if they are similar and are equal in size. Note that every triangle is congruent to its own reflection (as in a mirror). Regarding similarity, the three triangles of Fig. 2.3 are mutually similar but not congruent.

Two triangles are congruent to one another if they meet any of the following criteria:

- SSS (side-side-side);
- SAS (side-angle-side);
- ASA (angle-side-angle);
- AAS (angle-angle-side).

The SSS criterion requires that each of the triangles have three sides as long as the other triangle's (for example, a triangle whose sides are 5, 6 and 8 cm long and a triangle whose sides are 6, 5 and 8 cm long together satisfy the criterion). The SAS criterion requires that each of the triangles have two sides as long as the other triangle's and, where a triangle's two sides meet, one equal angle (for example, two triangles, each of whose sides of 6 and 8 cm meet to form a two-hour or 30° angle, together satisfy the criterion). The ASA criterion requires that each of the triangles have two equal angles joined by one equal side. The AAS criterion requires that each

Figure 2.5: The SSA and AAS criteria.



of the triangles have two equal angles plus one equal side, this side however being another than the side that joins the two angles.

As Euclid has proved earlier in the subsection, two triangles are similar (though not necessarily congruent) if they meet this criterion:

- AA (angle-angle).

Why is the criterion not AAA? Answer: because AA suffices, for, according to (2.48), AA implies AAA.

The criterion

- SSA (side-side-angle),

depicted in Fig. 2.4, is tricky: it does not quite establish congruence in and of itself, but if *three* triangles mutually meet the SSA criterion then at least two of them are congruent.

Besides SSA, Fig. 2.4 also illustrates the AAS criterion. Illustration of criteria besides SSA and AAS is left as an exercise.

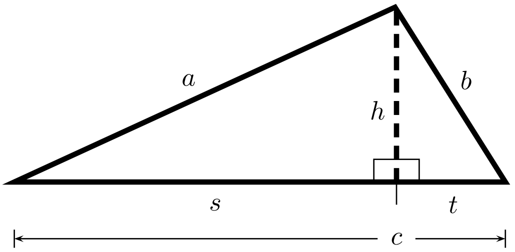
See also § 3.7.

2.9.6 Altitude and Heron's rule

A triangle's *semiperimeter*

$$s \equiv \frac{a+b+c}{2} \tag{2.56}$$

Figure 2.6: Altitude and Heron's rule.



being half the sum of the lengths of the triangle's three sides as in Fig. 2.6, *Heron's rule*[44] finds the triangle's area to be

$$A = \sqrt{(s)(s-a)(s-b)(s-c)}. \tag{2.57}$$

To prove Heron's rule, one may begin from the Pythagorean theorem (1.1), according to which

$$h^2 = a^2 - s^2 = b^2 - t^2.$$

Rearranging terms,

$$s^2 - t^2 = a^2 - b^2.$$

Factoring the left side per (2.1) and observing in the figure that $s + t = c$,

$$(c)(s-t) = a^2 - b^2.$$

Expanding each of s and t according to Pythagoras,

$$(c)\left(\sqrt{a^2-h^2} - \sqrt{b^2-h^2}\right) = a^2 - b^2. \tag{2.58}$$

Squaring and rearranging,

$$\left(2c^2\right)\sqrt{\left(a^2-h^2\right)\left(b^2-h^2\right)} = a^2c^2 + b^2c^2 + 2a^2b^2 - a^4 - b^4 - 2c^2h^2.$$

Squaring again, combining like terms, and isolating on the left all terms containing h,

$$\begin{aligned}\left(a^4 - 2a^2b^2 + b^4\right)4c^2h^2 = &-\left(a^4 - 2a^2b^2 + b^4\right)c^4 \\ &+\left(a^6 - a^4b^2 - a^2b^4 + b^6\right)2c^2 \\ &-\left(a^8 - 4a^6b^2 + 6a^4b^4 - 4a^2b^6 + b^8\right).\end{aligned}$$

[44] Also known as "Heron's formula," as for instance in [95][182, "Heron's formula," 21:30, 2 Dec. 2022], the sources for (2.57)'s semiperimeter-based formulation.

To recognize that the sum $a^4 - 2a^2b^2 + b^4$, which appears twice in the last equation, can be expressed as $(a^2 - b^2)^2$ is not too hard, after which one may notice that the sum $a^8 - 4a^6b^2 + 6a^4b^4 - 4a^2b^6 + b^8$ similarly can be expressed as $(a^2 - b^2)^4$. (If you did not notice it, then wait till §§ 4.2 and 4.3 and their Fig. 4.2 arrive. Those will help.) The sum $a^6 - a^4b^2 - a^2b^4 + b^6$ is less obvious, perhaps; but, after some experimentation, one eventually discovers the factorization $(a^2 - b^2)(a^4 - b^4)$ which, after the factor $a^4 - b^4$ is further expanded, comes to $(a^2 - b^2)^2(a^2 + b^2)$. Applying all these factorizations, we have that

$$\left(a^2 - b^2\right)^2 4c^2h^2 = -(a^2 - b^2)^2c^4 + \left(a^2 - b^2\right)^2 \left(a^2 + b^2\right) 2c^2 - (a^2 - b^2)^4.$$

Dividing by $(a^2 - b^2)^2$,

$$4c^2h^2 = -c^4 + \left(a^2 + b^2\right) 2c^2 - (a^2 - b^2)^2.$$

The triangle's area, as already expressed in different symbols by (2.46), is

$$A = \frac{ch}{2} \tag{2.59}$$

which, when applied to the preceding equation, implies that

$$(\mathrm{0x10})A^2 = -c^4 + \left(a^2 + b^2\right) 2c^2 - (a^2 - b^2)^2.$$

Distributing factors, one concludes that

$$(\mathrm{0x10})A^2 = 2a^2b^2 + 2a^2c^2 + 2b^2c^2 - a^4 - b^4 - c^4. \tag{2.60}$$

Equation (2.60) is a form of Heron's rule. One may use it directly. However, as long ago as the first century A.D., Heron of Alexandria appears cleverly to have rewritten (2.60) in terms of a triangle's semiperimeter (2.56), the result being the elegant (2.57)—as one may verify by substituting (2.56) into (2.57) and multiplying out the product that results. (The writer is unacquainted with the inspiration for Heron's cleverness. History knows Heron to have been a sometime surveyor,[45] though, so one may conjecture that Heron was used to stretching a primitive surveyor's cord[46] about stakes at the three points of a triangular plot of land. After tying the cord's ends to form a taut loop about the triangle and then removing the loop from the stakes, Heron might during the ordinary course of his work have collapsed

[45] [32, page 53]

[46] [109, § 1A][126, § 67]

and straightened the then-slack loop—with ends still tied—to semiperimeter's length; so maybe the idea of working in terms of semiperimeters came naturally to Heron. As for a factor like $s-a$, that would come if the collapsed slack loop were left caught about one stake and turned sideways, unspread, about a second. Such a story is plausible at any rate. An ancient Egyptian surveyor could tell us if the story were true. In the surveyor's absence, we merely observe that mathematical inspiration arrives in many guises.)

A triangle's altitude h may be calculated via Heron's rule by rearranging (2.60) to read

$$h = \frac{2A}{c} \tag{2.61}$$

and then using (2.57) to get the area A. Indeed, a triangle, having three angles, has three altitudes, h_a, h_b and h_c, the last of which is the h of Fig. 2.6. Heron's rule conveniently yields all three altitudes at one go:

$$h_a = \frac{2A}{a}; \quad h_b = \frac{2A}{b}; \quad h_c = \frac{2A}{c}. \tag{2.62}$$

The fastidious reader may have noticed that the derivation has divided by $(a^2-b^2)^2$, a quantity that could be zero. However, this is not a real trouble and may be dealt with in various ways, for instance by lengthening a and shortening b ever so slightly to evade the zero without substantially altering the triangle's area (a technique chapter 4 will generalize); or instead by leaving a and b as they are but reworking the derivation to exploit the geometrical symmetry the two lengths' equality implies. Details are left to the interested reader as an exercise.

The other objection one might reasonably raise regards a negative s or t like the negative b_2 of page 15's Fig. 1.3. Though § 1.3 has generally dismissed objections of this kind, let us nevertheless briefly consider the objection here. A negative t (say) would have caused (2.58) to read,

$$(c)\left(\sqrt{a^2-h^2}+\sqrt{b^2-h^2}\right) = a^2 - b^2,$$

with a $+$ sign on the left in place of (2.58)'s $-$ sign. Fortunately, the derivation's repeated squaring operation would have made the change of sign irrelevant a few lines later, as the objector's review of the derivation would reveal.

2.10 Functions

Briefly, a *function* is a mapping from one number (or vector of several numbers, §§ 11.1 and 8.16) to another. This book is not the place for a gentle introduction to the concept of the function; but as an example, $f(x) \equiv x^2 - 1$ is a function which maps 1 to 0 and -3 to 8, among others.

When discussing functions, one often speaks of *domains* and *ranges.* The *domain* of a function is the set of numbers one can put into it. The *range* of a function is the corresponding set of numbers one can get out of it. In the example, if the domain is restricted to real x such that $|x| \le 3$—that is, such that $-3 \le x \le 3$—then the corresponding range is $-1 \le f(x) \le 8$.

If $y = f(x)$, then the notation $f^{-1}(\cdot)$ indicates the *inverse* of the function $f(\cdot)$ such that

$$\begin{aligned} f^{-1}[f(x)] &= x, \\ f[f^{-1}(y)] &= y, \end{aligned} \tag{2.63}$$

thus swapping the function's range with its domain. In the example, $f(-3) = 8$, so $f^{-1}(8) = -3$. Unfortunately, $f^{-1}(8) = 3$ as well, making the example's function $f(x)$ strictly *noninvertible* over its given domain; but, fortunately, various attendant considerations (whether these be theoretical considerations, as in § 8.4, or merely practical considerations) tend in concrete circumstances to distinguish between candidates like the example's -3 and 3, rendering many such functions as $f(x)$ effectively invertible in context. Therefore, the applied mathematician may not always have to worry too much about strict invertibility.

Inconsistently, inversion's notation $f^{-1}(\cdot)$ clashes with the similar-looking but different-meaning notation $f^2(\cdot) \equiv [f(\cdot)]^2$, whereas $f^{-1}(\cdot) \ne [f(\cdot)]^{-1}$. Both notations are conventional and both are used in this book.

Other terms that arise when discussing functions are *root* (or *zero*), *singularity* and *pole.* A *root* (or *zero*) of a function is a domain point at which the function evaluates to zero (the example has roots at $x = \pm 1$). A *singularity* of a function is a domain point at which the function's output *diverges;* that is, where the function's output goes infinite.[47] A *pole* is a singularity that behaves locally like $1/x$ (rather than, for example, like $1/\sqrt{x}$). A singularity that behaves as $1/x^N$ is a *multiple pole*, which (§ 9.7.2) can be thought of as N poles. The example's function $f(x)$ has no singularities for finite x; however, the function $h(x) \equiv 1/(x^2 - 1)$ has poles at $x = \pm 1$.

[47] Here is one example of the book's deliberate lack of formal mathematical rigor. A more precise formalism to say that "the function's output goes infinite" might be that

$$\lim_{x \to x_o} |f(x)| = \infty,$$

and yet preciser formalisms than this are conceivable, and occasionally even are useful. Be that as it may, the applied mathematician tends to avoid such formalisms *where there seems no immediate need for them.*

(Besides the root, the singularity and the pole, there is also the troublesome *branch point,* an infamous example of which is $z = 0$ in the function $g[z] \equiv \sqrt{z}$. Branch points are important, but the book must lay a more extensive foundation before introducing them properly in § 8.5.[48])

2.11 Complex numbers (introduction)

Section 2.5.2 has introduced square roots. What it has not done is to tell us how to regard a quantity like $\sqrt{-1}$. Since there exists no real number i such that

$$i^2 = -1 \tag{2.64}$$

and since the quantity i thus defined is found to be critically important across broad domains of higher mathematics, we accept (2.64) as the definition of a fundamentally new kind of quantity, *the imaginary number.*[49]

Imaginary numbers are given their own number line, plotted at right angles to the familiar real number line as in Fig. 2.7. The sum of a real number x and an imaginary number iy is the *complex number*

$$z = x + iy.$$

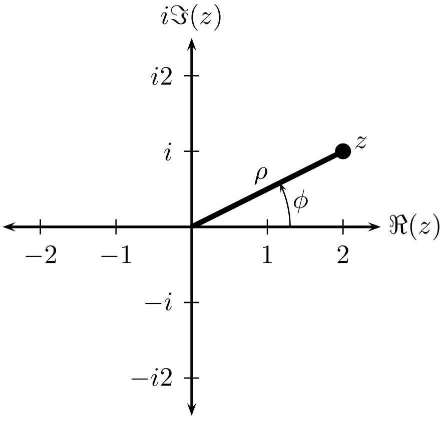
[48] There is further the *essential singularity,* an example of which is $z = 0$ in $p(z) \equiv \exp(1/z)$, but typically the best way to handle so unreasonable a singularity is to change a variable, as $w \leftarrow 1/z$, or otherwise to frame the problem such that one need not approach the singularity. Except tangentially much later when it treats asymptotic series, this book will have little to say of the essential singularity.

[49] The English word *imaginary* is evocative, but perhaps not of quite the right concept in this usage. Imaginary numbers are not to mathematics as, say, imaginary elfs are to the physical world. In the physical world, imaginary elfs are (presumably) not substantial objects. However, in the mathematical realm, imaginary numbers *are* substantial. The word *imaginary* in the mathematical sense is thus more of a technical term than a descriptive adjective.

The number i is just a concept, of course, but then so is the number 1 (though you and I have often met one *of something*—one apple, one chair, one summer afternoon, etc.—neither of us has ever met just 1). In Platonic [55, chapter 2] or Fregean [62] terms, i is literally no less valid a *form* than 1 is. The reason imaginary numbers have been called "imaginary" probably has to do with the fact that they emerge from mathematical operations only, never directly from counting things. Notice, however, that the number 1/2 never emerges directly from counting things, either. If for some reason the $i$year were offered as a unit of time, then the period separating your fourteenth and twenty-first birthdays could have been measured as $-i7$ $i$years. Madness? No, let us not call it that; let us call it a useful formalism, rather.

The unpersuaded reader is asked to suspend judgment a while. He will soon see the use.

Figure 2.7: The complex (or Argand) plane, and a complex number $z = 2+i1$ therein.



The *conjugate* z^* of this complex number is defined to be[50]

$$z^* = x - iy.$$

The *magnitude* (or *modulus,* or *absolute value*) $|z|$ of the complex number is defined to be the length ρ in Fig. 2.7, which per the Pythagorean theorem (§ 2.9.4) is such that

$$|z|^2 = x^2 + y^2. \tag{2.65}$$

The *phase* $\arg z$ of the complex number is defined to be the angle ϕ in the figure, which in terms of the trigonometric functions of § 3.1[51] is such that

$$\tan(\arg z) = \frac{y}{x}. \tag{2.66}$$

[50] For some inscrutable reason, in the author's country at least, professional mathematicians seem universally to write $\overline{z}$ instead of z^*, whereas rising engineers take the mathematicians' courses at school and then, having passed those courses, promptly start writing z^* for the rest of their lives. The writer has his preference between the two notations and this book reflects it, but the curiously absolute character of the notational split is interesting as a social phenomenon.

[51] This is a forward reference, returned by footnote 5 of chapter 3. If the equation does not make sense to you yet for this reason, skip it for now. The important point is that $\arg z$ is the angle ϕ in the figure.

Specifically to extract the real and imaginary parts of a complex number, the notations

$$\begin{aligned}\Re(z) &= x,\\ \Im(z) &= y,\end{aligned} \tag{2.67}$$

are conventionally recognized (although often the symbols $\Re[\cdot]$ and $\Im[\cdot]$ are written $\mathrm{Re}[\cdot]$ and $\mathrm{Im}[\cdot]$, particularly when printed by hand). For example, if $z = 2 + i1$, then $\Re(z) = 2$, $\Im(z) = 1$, $|z| = \sqrt{5}$, and $\arg z = \arctan(1/2)$.

2.11.1 Multiplication and division of complex numbers in rectangular form

Several elementary properties of complex numbers are readily seen if the fact that $i^2 = -1$ is kept in mind, including that

$$z_1 z_2 = (x_1 x_2 - y_1 y_2) + i(y_1 x_2 + x_1 y_2), \tag{2.68}$$

$$\begin{aligned}\frac{z_1}{z_2} &= \frac{x_1 + iy_1}{x_2 + iy_2} = \left(\frac{x_2 - iy_2}{x_2 - iy_2}\right)\frac{x_1 + iy_1}{x_2 + iy_2}\\ &= \frac{(x_1 x_2 + y_1 y_2) + i(y_1 x_2 - x_1 y_2)}{x_2^2 + y_2^2}.\end{aligned} \tag{2.69}$$

It is a curious fact that

$$\frac{1}{i} = -i. \tag{2.70}$$

It is a useful fact that

$$z^* z = x^2 + y^2 = |z|^2 \tag{2.71}$$

(the curious fact, eqn. 2.70, is useful, too). Sometimes convenient are the forms

$$\begin{aligned}\Re(z) &= \frac{z + z^*}{2},\\ \Im(z) &= \frac{z - z^*}{i2},\end{aligned} \tag{2.72}$$

trivially proved. Surprising and significant is that, though $\sqrt{-1} = i$ has revealed a previously unknown class of quantity in i, nevertheless, $\sqrt{i} = (1 + i)/\sqrt{2}$ reveals no further new class of quantity, but only a quantity expressible[52] in terms of 1 and i, as (2.68) verifies (and upon which § 3.11 will shed a clearer light).

[52] [57, § I:22-5]

2.11.2 Complex conjugation

An important property of complex numbers descends subtly from the fact that

$$i^2 = -1 = (-i)^2.$$

If one defined some number $j \equiv -i$, asserting that j not i were the true imaginary unit,[53] then one would find that

$$(-j)^2 = -1 = j^2,$$

and thus that all the basic properties of complex numbers in the j system held just as well as they did in the i system. The units i and j would differ indeed, but would perfectly mirror one another in every respect.

That is the basic idea. To establish it symbolically needs a page or so of slightly abstract algebra as follows, the goal of which will be to show that $[f(z)]^* = f(z^*)$ for some unspecified function $f(z)$ with specified properties. To begin with, if

$$z = x + iy,$$

then

$$z^* = x - iy$$

by definition. Proposing that $(z^{k-1})^* = (z^*)^{k-1}$ (which may or may not be true but for the moment we assume it), we can write,

$$\begin{aligned} z^{k-1} &= s_{k-1} + it_{k-1}, \\ (z^*)^{k-1} &= s_{k-1} - it_{k-1}, \end{aligned}$$

where s_{k-1} and t_{k-1} are symbols introduced respectively to represent the real and imaginary parts of z^{k-1}. Multiplying the former equation by $z = x + iy$ and the latter by $z^* = x - iy$, we have that

$$\begin{aligned} z^{k} &= (xs_{k-1} - yt_{k-1}) + i(ys_{k-1} + xt_{k-1}), \\ (z^*)^{k} &= (xs_{k-1} - yt_{k-1}) - i(ys_{k-1} + xt_{k-1}). \end{aligned}$$

With the definitions that $s_k \equiv xs_{k-1} - yt_{k-1}$ and $t_k \equiv ys_{k-1} + xt_{k-1}$, this is written more succinctly,

$$\begin{aligned} z^{k} &= s_k + it_k, \\ (z^*)^{k} &= s_k - it_k. \end{aligned}$$

[53] [45][57, § I:22-5]

In other words, if $(z^{k-1})^* = (z^*)^{k-1}$, then it necessarily follows that $(z^k)^* = (z^*)^k$. Solving the definitions of s_k and t_k for s_{k-1} and t_{k-1} yields the reverse definitions that $s_{k-1} = (xs_k + yt_k)/(x^2+y^2)$ and $t_{k-1} = (-ys_k + xt_k)/(x^2+y^2)$. Therefore, except when $z = x+iy$ happens to be null or infinite, the implication is reversible by reverse reasoning, so by mathematical induction[54] we have that

$$(z^k)^* = (z^*)^k \tag{2.73}$$

for all integral k. We have also from (2.68) that

$$(z_1z_2)^* = z_1^*z_2^* \tag{2.74}$$

for any complex z_1 and z_2.

Consequences of (2.73) and (2.74) include that if

$$\begin{aligned} f(z) &\equiv \sum_{k=-\infty}^{\infty} (a_k + ib_k)(z - z_o)^k, & (2.75)\\ f^*(z) &\equiv \sum_{k=-\infty}^{\infty} (a_k - ib_k)(z - z_o^*)^k, & (2.76) \end{aligned}$$

where a_k and b_k are real and imaginary parts of the coefficients peculiar to the function $f(\cdot)$, then

$$[f(z)]^* = f^*(z^*). \tag{2.77}$$

In the common case in which $b_k = 0$ for all k and in which $z_o = x_o$ is a real number, $f(\cdot)$ and $f^*(\cdot)$ are the same function, so (2.77) reduces to the desired form

$$[f(z)]^* = f(z^*), \tag{2.78}$$

which says that the effect of conjugating the function's input is merely to conjugate its output.

Equation (2.78) expresses a significant, general rule of complex numbers and complex variables which is better explained in words than in mathematical symbols. The rule is this: for most equations and systems of equations used to model physical systems, *one can produce an equally valid alternate model simply by simultaneously conjugating all the complex quantities present.*[55]

[54] *Mathematical induction* is an elegant old technique for the construction of mathematical proofs. Section 8.1 elaborates on the technique and offers a more extensive example. Beyond the present book, a very good introduction to mathematical induction is found in [70].

[55] [70][153]

2.11.3 Power series and analytic functions (preview)

Equation (2.75) expresses a general power series[56] in $z - z_o$. Such power series have broad application.[57] It happens in practice that most functions of interest in modeling physical phenomena can conveniently be constructed (at least over a local domain) as power series with suitable choices of a_k, b_k and z_o.[58]

The property (2.77) applies to all such functions, with (2.78) also applying to those for which $b_k = 0$ and $z_o = x_o$. The property the two equations represent is called the *conjugation property.* Basically, it says that if one replaces all the i in some mathematical model with $-i$, then the resulting conjugate model is equally as valid as the original.[59]

Such functions, whether $b_k = 0$ and $z_o = x_o$ or not, are *analytic functions* (§ 8.4). In the formal mathematical definition, a function is analytic which is infinitely differentiable (chapter 4) in the immediate domain neighborhood of interest. However, for applications a fair working definition of the analytic function might be "a function expressible as a power series." Chapter 8 elaborates. All power series are infinitely differentiable except at their poles.

There nevertheless exist one common group of functions which cannot be constructed as power series. These all have to do with the parts of complex numbers and have been introduced in this very section: the magnitude $|\cdot|$; the phase $\arg(\cdot)$; the conjugate $(\cdot)^*$; and the real and imaginary parts $\Re(\cdot)$ and $\Im(\cdot)$. These functions are not analytic and do not in general obey the conjugation property. Also not analytic are the Heaviside unit step $u(t)$ and the Dirac delta $\delta(t)$ (§ 7.7), used to model discontinuities explicitly.

We shall have more to say about analytic functions in chapter 8. We shall have more to say about complex numbers in §§ 3.11, 4.3.3, and 4.4,

[56] [79, § 10.8]

[57] That is a pretty impressive-sounding statement: "Such power series have broad application." However, molecules, air and words also have "broad application"; merely stating the fact does not tell us much. In fact the general power series is a sort of one-size-fits-all mathematical latex glove, which can be stretched to fit around almost any function of interest. Admittedly nevertheless, what grips one's attention here is not so much in the general form (2.75) of the series as it is in the specific choice of a_k and b_k, which this section does not discuss.

Observe that the Taylor series (which this section also does not discuss: see § 8.3) is a power series with $a_k = b_k = 0$ for $k < 0$.

[58] But see also the Fourier series of chapter 17 which, by a different approach, can construct many functions over *nonlocal* domains.

[59] To illustrate, from the fact that $(1 + i2)(2 + i3) + (1 - i) = -3 + i6$, the conjugation property infers immediately that $(1 - i2)(2 - i3) + (1 + i) = -3 - i6$. Observe however that no such property holds for the real parts: $(-1 + i2)(-2 + i3) + (-1 - i) \neq 3 + i6$.

and much more yet in chapter 5.

Chapter 3

Trigonometry

Trigonometry is that branch of mathematics which relates angles to lengths. This chapter introduces the functions of trigonometry and derives several of these functions' properties.

3.1 Definitions

Consider the circle-inscribed right triangle of Fig. 3.1.

In considering the triangle and its circle in the figure, we shall find some terminology useful. The *angle* ϕ in the figure is measured in *radians,* where a radian is that angle which, when centered in a unit circle, describes or intercepts an arc of unit length as measured—not in a straight line—but along the curve of the circle's perimeter. A *unit circle* is a circle whose radius is $\rho = 1$. Similarly, a *unit length* is a length of 1 (not one centimeter or one mile, or anything like that, but just an abstract 1). In the more general circle of radius ρ (where the *radius* is the distance from the circle's center to its perimeter) as in the figure, an angle ϕ describes or intercepts an arc of length $\rho\phi$ along the curve.

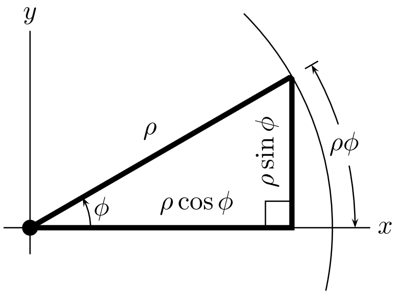
An angle in radians is a dimensionless number, so one need not write, "$\phi = 2\pi/4$ radians"; but it suffices to write, "$\phi = 2\pi/4$." In mathematical theory, we express angles in radians.

The angle of full revolution is given the symbol 2π—which thus is the unit circle's circumference.[1] A quarter revolution, $2\pi/4$, is then the *right angle* or *square angle.*

The trigonometric functions $\sin\phi$ and $\cos\phi$, the *sine* and *cosine* of ϕ, relate the angle ϕ to the lengths in the figure. The tangent function is then

[1] Section 8.11 computes the numerical value of 2π.

Figure 3.1: The sine and the cosine (shown on a circle-inscribed right triangle, with the circle centered at the triangle's point).



defined to be

$$\tan\phi \equiv \frac{\sin\phi}{\cos\phi}, \tag{3.1}$$

which is the vertical rise per unit horizontal run, this ratio being the *slope,* of the triangle's diagonal.[2] Inverses of the three trigonometric functions can also be defined:

$$\begin{aligned}\arcsin\,(\sin\phi) &= \phi;\\ \arccos\,(\cos\phi) &= \phi;\\ \arctan\,(\tan\phi) &= \phi.\end{aligned}$$

When the last of these is written,

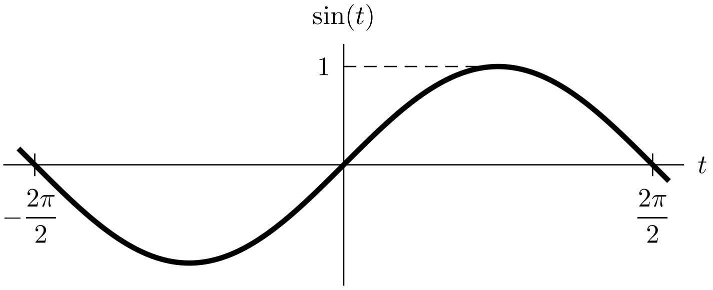
$$\arctan\frac{y}{x},$$

it normally implied that x and y are to be interpreted as rectangular coordinates[3] and that the arctan function is to return ϕ in the correct quad-

[2] Often seen in print is the additional notation $\sec\phi \equiv 1/\cos\phi$, $\csc\phi \equiv 1/\sin\phi$ and $\cot\phi \equiv 1/\tan\phi$; respectively the "secant," "cosecant" and "cotangent." This book does not use that notation.

[3] *Rectangular coordinates* are pairs of numbers (x, y) which uniquely specify points in a plane. Conventionally, the x coordinate indicates distance eastward as it were; the y coordinate, northward. For instance, the coordinates $(3, -4)$ mean the point three units eastward and four units southward (that is, -4 units northward) of the *origin* $(0, 0)$. When needed, a third rectangular coordinate can moreover be added—(x, y, z)—the z indicating distance upward, above the plane of the x and y. (Consistency with the book's general style should have suggested the spelling "coördinates" except that no mathematics book this writer knows spells the word that way, nor even as "co-ordinates" which is how the word after all is pronounced.)

Figure 3.2: The sine function.



rant $-\pi < \phi \leq \pi$ (for example, $\arctan[1/(-1)] = [+3/8][2\pi]$, whereas $\arctan[(-1)/1] = [-1/8][2\pi]$). This is similarly the usual interpretation when an equation like

$$\tan\phi = \frac{y}{x}$$

is written.

By the Pythagorean theorem (§ 2.9.4 and the introduction to chapter 1), it is seen that[4]

$$\cos^2\phi + \sin^2\phi = 1 \tag{3.2}$$

for any ϕ.

Fig. 3.2 plots the sine function. The shape in the plot is called a *sinusoid.*[5]

3.2 Simple properties

Inspecting Fig. 3.1 and observing (3.1) and (3.2), one discovers the trigonometric properties of Table 3.1. (Observe in the table as usually elsewhere in the book that $n \in \mathbb{Z}$, § 2.3.)

[4] The notation $\cos^2\phi$ means $(\cos\phi)^2$.

[5] This section completes the forward reference of § 2.11. See chapter 2's footnote 51.

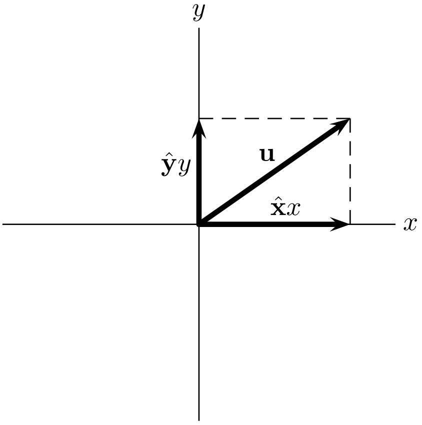
Table 3.1: Simple properties of the trigonometric functions.

$$
\begin{array}{rcl@{\qquad}rcl}
\sin(-\phi) &=& -\sin\phi & \cos(-\phi) &=& +\cos\phi \\
\sin(2\pi/4-\phi) &=& +\cos\phi & \cos(2\pi/4-\phi) &=& +\sin\phi \\
\sin(2\pi/2-\phi) &=& +\sin\phi & \cos(2\pi/2-\phi) &=& -\cos\phi \\
\sin(\phi\pm 2\pi/4) &=& \pm\cos\phi & \cos(\phi\pm 2\pi/4) &=& \mp\sin\phi \\
\sin(\phi\pm 2\pi/2) &=& -\sin\phi & \cos(\phi\pm 2\pi/2) &=& -\cos\phi \\
\sin(\phi+n2\pi) &=& \sin\phi & \cos(\phi+n2\pi) &=& \cos\phi
\end{array}
$$

$$
\begin{array}{rcl}
\tan(-\phi) &=& -\tan\phi \\
\tan(2\pi/4-\phi) &=& +1/\tan\phi \\
\tan(2\pi/2-\phi) &=& -\tan\phi \\
\tan(\phi\pm 2\pi/4) &=& -1/\tan\phi \\
\tan(\phi\pm 2\pi/2) &=& +\tan\phi \\
\tan(\phi+n2\pi) &=& \tan\phi
\end{array}
$$

$$
\begin{array}{rcl}
\dfrac{\sin\phi}{\cos\phi} &=& \tan\phi \\
\cos^2\phi+\sin^2\phi &=& 1 \\
1+\tan^2\phi &=& \dfrac{1}{\cos^2\phi} \\
1+\dfrac{1}{\tan^2\phi} &=& \dfrac{1}{\sin^2\phi}
\end{array}
$$

Figure 3.3: A two-dimensional vector $\mathbf{u} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y$, shown with its rectangular components.



3.3 Scalars, vectors, and vector notation

In applied mathematics, a *geometrical vector*—whose name the narrative will hereinafter usually abbreviate as *vector*—is an amplitude of some kind coupled with a direction.[6,7] For example, "55 miles per hour northwestward" is a vector, as is the entity $\mathbf{u}$ depicted in Fig. 3.3. The entity $\mathbf{v}$ depicted in Fig. 3.4 is also a vector, in this case a three-dimensional one.

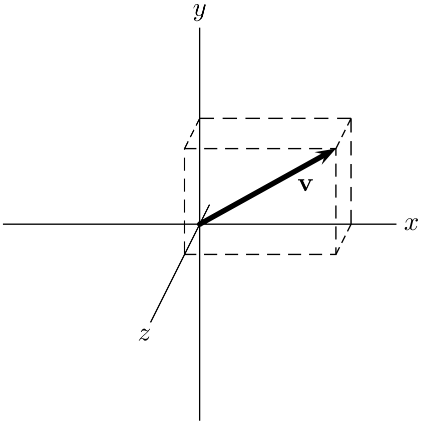
Many readers will already find the basic vector concept familiar, but for

[6] The same word *vector* also is used to indicate an ordered set of N scalars (§ 8.16) or an $N \times 1$ matrix (chapter 11), but those are not the uses of the word meant here. See also the introduction to chapter 15.

[7] The word "amplitude" is sometimes used interchangeably with "magnitude," even by this writer. However, used more precisely, an *amplitude* unlike a magnitude can be negative or positive, as for example the A in $f(t) \equiv A \cos \omega t$. Indeed, this A makes a typical example of the usage, for the word "amplitude" is especially applied to measure the extent to which a wave deviates from its neutral position.

In practice, the last usage typically connotes that the quantity whose amplitude the amplitude is is a *linear* quantity like displacement, force, electric current or electric tension (voltage). A *squared* quantity—that is, a quantity developed as the product of two other, linear quantities—is not usually said to possess an amplitude. Especially, energy and power are not usually said to possess amplitudes, whereas displacement and force (whose product is an energy) are each amplitudinous; and, likewise, current and tension (whose product is a power) are each amplitudinous.

Figure 3.4: A three-dimensional vector $\mathbf{v} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y + \hat{\mathbf{z}}z$.



those who do not, a brief review: vectors such as the

$$\begin{aligned}\mathbf{u} &= \hat{\mathbf{x}}x + \hat{\mathbf{y}}y,\\ \mathbf{v} &= \hat{\mathbf{x}}x + \hat{\mathbf{y}}y + \hat{\mathbf{z}}z,\end{aligned}$$

of the figures are composed of multiples of the *unit basis vectors* $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$, which are themselves vectors of unit length pointing in the cardinal directions their respective symbols suggest.[8] Any vector $\mathbf{a}$ can be factored into an *amplitude* a and a *unit vector* $\hat{\mathbf{a}}$, as

$$\mathbf{a} = \hat{\mathbf{a}}a = \hat{\mathbf{a}}\left|\mathbf{a}\right|,$$

where the $\hat{\mathbf{a}}$ represents direction only and has unit magnitude by definition, and where the a or $|\mathbf{a}|$ represents amplitude only and carries the physical units if any.[9] For example, $a = 55$ miles per hour, $\hat{\mathbf{a}} =$ northwestward. The

[8] Printing by hand, one customarily writes a general vector like $\mathbf{u}$ as "$\vec{u}$" or just "$\overline{u}$", and a unit vector like $\hat{\mathbf{x}}$ as "$\hat{x}$".

[9] The word "unit" here is unfortunately overloaded. As an adjective in mathematics, or in its nounal form "unity," it refers to the number one (1)—not one mile per hour, one kilogram, one Japanese yen or anything like that; just an abstract 1. The word "unit" itself as a noun however usually signifies a physical or financial reference quantity of measure, like a mile per hour, a kilogram or even a Japanese yen. There is no inherent logical unity to 1 mile per hour (otherwise known as 0.447 meters per second, among other names). By contrast, a "unitless 1"—a 1 with no physical unit attached, also known as a

unit vector $\hat{\mathbf{a}}$ itself can be expressed in terms of the unit basis vectors; for example, if $\hat{\mathbf{x}}$ points east and $\hat{\mathbf{y}}$ points north, then $\hat{\mathbf{a}} = -\hat{\mathbf{x}}(1/\sqrt{2})+\hat{\mathbf{y}}(1/\sqrt{2})$ means "northwestward," where per the Pythagorean theorem $(-1/\sqrt{2})^2 + (1/\sqrt{2})^2 = 1^2$.

A single number which is not a vector or a matrix (chapter 11) is called a *scalar.* In the example, $a = 55$ miles per hour is a scalar. Though the scalar a in the example happens to be real, scalars can be complex, too—which might surprise one, since scalars by definition lack direction and the Argand phase ϕ of Fig. 2.7 so strongly resembles a direction. However, phase is not an actual direction in the vector sense (the real number line in the Argand plane cannot be said to run west-to-east, or anything like that). The x, y and z of Fig. 3.4 are each (possibly complex) scalars; $\mathbf{v} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y + \hat{\mathbf{z}}z$ is a vector. If x, y and z are complex, then[10]

$$\begin{aligned} |\mathbf{v}|^2 &= |x|^2 + |y|^2 + |z|^2 = x^*x + y^*y + z^*z \\ &= [\Re(x)]^2 + [\Im(x)]^2 + [\Re(y)]^2 + [\Im(y)]^2 \\ &\quad + [\Re(z)]^2 + [\Im(z)]^2 . \end{aligned} \tag{3.3}$$

A point is often identified by the vector expressing its distance from and direction relative to the origin $(0, 0)$ of a coordinate system. That is, the point (x, y) can be, and frequently is, identified with the vector $\hat{\mathbf{x}}x + \hat{\mathbf{y}}y$. [Likewise, the origin $(0, 0)$ itself can be identified with the null vector $\hat{\mathbf{x}}0 + \hat{\mathbf{y}}0 = 0$.] However, not every vector need be associated with an origin, for vectors in the more general sense represent *relative* distances and directions, whether the thing to which they are relative happens to be an origin or something else.

Notice incidentally our mathematical use of the word "distance." A mathematical distance may or may not represent a distance in the literal,

"dimensionless 1"—does represent a logical unity.

Consider the ratio $r = h_1/h_o$ of your height h_1 to my height h_o. Maybe you are taller than I am and so $r = 1.05$ (not 1.05 cm or 1.05 feet, just 1.05). Now consider the ratio h_1/h_1 of your height to your own height. That ratio is of course unity, exactly 1.

There is nothing ephemeral in the concept of mathematical unity, nor in the concept of unitless quantities in general. The concept is quite straightforward and is entirely practical. That $r > 1$ means neither more nor less than that you are taller than I am. In applications, one often puts physical quantities in ratio precisely to strip the physical units from them, comparing the ratio to unity without regard to physical units.

Incidentally, a more general term to comprise physical units, financial units and other such quantities is *units of measure.*

[10] Some books print $|\mathbf{v}|$ as $\|\mathbf{v}\|$ or even $\|\mathbf{v}\|_2$ to emphasize that it represents the real, scalar magnitude of a complex vector. The reason the last notation subscripts a numeral 2 is obscure, having to do with the professional mathematician's generalized definition of a thing he calls the "norm." This book just renders it $|\mathbf{v}|$.

physical sense, but it will at least represent a quantity one might proportionably diagram as though it were a distance. If such a quantity also has direction, then it can be a vector. This is why "55 miles per hour northwestward" is a vector despite that the mile per hour is a unit of speed rather than of distance. This is also why "55 kilograms," which lacks direction, is not a vector.

Observe the relative orientation of the axes of Fig. 3.4. The axes are oriented such that if you point your flat right hand in the x direction, bend your fingers in the y direction, and extend your thumb, then the thumb points in the z direction. This is orientation by the *right-hand rule.* A left-handed orientation is equally possible, of course, but since neither orientation owns a natural advantage over the other, mathematicians arbitrarily but conventionally accept the right-handed one as standard.[11]

Sections 3.4 and 3.9 and chapters 15 and 16 will speak further of the geometrical vector.

3.4 Rotation

A fundamental problem in trigonometry arises when a vector

$$\mathbf{u} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y \tag{3.4}$$

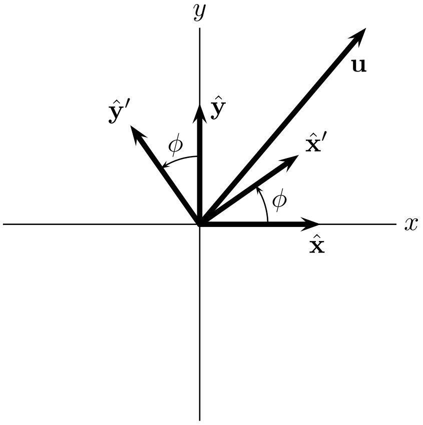
must be expressed in terms of alternate unit vectors $\hat{\mathbf{x}}'$ and $\hat{\mathbf{y}}'$, where $\hat{\mathbf{x}}'$ and $\hat{\mathbf{y}}'$ stand at right angles to one another and lie in the plane[12] of $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$, but are rotated from the latter pair by an angle ϕ as depicted in Fig. 3.5.[13] In terms of the trigonometric functions of § 3.1, evidently,

[11] The writer does not know the etymology for certain, but verbal lore in American engineering has it that the name "right-handed" comes from experience with a standard right-handed wood screw or machine screw. If you hold the screwdriver in your right hand and turn the screw in the natural manner clockwise, turning the screw slot from the x orientation toward the y, then the screw advances away from you in the z direction into the bore. If somehow you came across a left-handed screw, you'd probably find it easier to drive that screw with the screwdriver in your left hand.

[12] A *plane,* as the reader on this tier undoubtedly knows, is a flat (but not necessarily level) surface, infinite in extent unless otherwise specified. Space is three-dimensional. A plane is two-dimensional. A line is one-dimensional. A point is zero-dimensional. The plane belongs to this geometrical hierarchy.

[13] The "$'$" mark is pronounced "prime" or "primed" (for no especially good reason of which the author is aware, but anyway, that's how it's pronounced). Mathematical writing employs the mark for a variety of purposes. Here, the mark merely distinguishes the new unit vector $\hat{\mathbf{x}}'$ from the old $\hat{\mathbf{x}}$.

Figure 3.5: Vector basis rotation.



$$\begin{aligned}\hat{\mathbf{x}}' &= +\hat{\mathbf{x}}\cos\phi + \hat{\mathbf{y}}\sin\phi,\\ \hat{\mathbf{y}}' &= -\hat{\mathbf{x}}\sin\phi + \hat{\mathbf{y}}\cos\phi;\end{aligned} \tag{3.5}$$

and by appeal to symmetry it stands to reason that

$$\begin{aligned}\hat{\mathbf{x}} &= +\hat{\mathbf{x}}'\cos\phi - \hat{\mathbf{y}}'\sin\phi,\\ \hat{\mathbf{y}} &= +\hat{\mathbf{x}}'\sin\phi + \hat{\mathbf{y}}'\cos\phi.\end{aligned} \tag{3.6}$$

Substituting (3.6) into (3.4) yields that

$$\mathbf{u} = \hat{\mathbf{x}}'(x\cos\phi + y\sin\phi) + \hat{\mathbf{y}}'(-x\sin\phi + y\cos\phi), \tag{3.7}$$

which was to be derived.

Equation (3.7) finds general application where rotations in rectangular coordinates are involved. If the question is asked, “what happens if I rotate not the unit basis vectors but rather the vector **u** instead?” the answer is that it amounts to the same thing, except that the sense of the rotation is reversed:

$$\mathbf{u}' = \hat{\mathbf{x}}(x\cos\phi - y\sin\phi) + \hat{\mathbf{y}}(x\sin\phi + y\cos\phi). \tag{3.8}$$

Whether it is the basis or the vector which rotates thus depends on your point of view.[14]

[14] This is only true, of course, with respect to the vectors themselves. When one actually rotates a physical body, the body experiences forces during rotation which might or might not change the body internally in some relevant way.

Much later in the book, § 15.1 will extend rotation in two dimensions to reorientation in three dimensions.

3.5 Trigonometric functions of sums and differences of angles

With the results of § 3.4 in hand, we now stand in a position to consider trigonometric functions of sums and differences of angles. Let

$$\begin{aligned} \hat{\mathbf{a}} &\equiv \hat{\mathbf{x}}\cos\alpha + \hat{\mathbf{y}}\sin\alpha, \\ \hat{\mathbf{b}} &\equiv \hat{\mathbf{x}}\cos\beta + \hat{\mathbf{y}}\sin\beta, \end{aligned}$$

be vectors of unit length in the xy plane, respectively at angles α and β from the x axis. If we wanted $\hat{\mathbf{b}}$ to coïncide with $\hat{\mathbf{a}}$, we would have to rotate it by $\phi = \alpha - \beta$. According to (3.8) and the definition of $\hat{\mathbf{b}}$, if we did this we would obtain that

$$\begin{aligned} \hat{\mathbf{b}}' = \hat{\mathbf{x}}&[\cos\beta\cos(\alpha-\beta) - \sin\beta\sin(\alpha-\beta)] \\ &+ \hat{\mathbf{y}}[\cos\beta\sin(\alpha-\beta) + \sin\beta\cos(\alpha-\beta)]. \end{aligned}$$

Since we have deliberately chosen the angle of rotation such that $\hat{\mathbf{b}}' = \hat{\mathbf{a}}$, we can separately equate the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ terms in the expressions for $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}'$ to obtain the pair of equations

$$\begin{aligned} \cos\alpha &= \cos\beta\cos(\alpha-\beta) - \sin\beta\sin(\alpha-\beta), \\ \sin\alpha &= \cos\beta\sin(\alpha-\beta) + \sin\beta\cos(\alpha-\beta). \end{aligned}$$

Solving the last pair simultaneously[15] for $\sin(\alpha-\beta)$ and $\cos(\alpha-\beta)$ and observing that $\sin^2(\cdot) + \cos^2(\cdot) = 1$ yields that

$$\begin{aligned} \sin(\alpha-\beta) &= \sin\alpha\cos\beta - \cos\alpha\sin\beta, \\ \cos(\alpha-\beta) &= \cos\alpha\cos\beta + \sin\alpha\sin\beta. \end{aligned} \tag{3.9}$$

[15] The easy way to do this is

- to subtract $\sin\beta$ times the first equation from $\cos\beta$ times the second and then to solve the result for $\sin(\alpha-\beta)$;
- to add $\cos\beta$ times the first equation to $\sin\beta$ times the second and then to solve the result for $\cos(\alpha-\beta)$.

This shortcut technique for solving a pair of equations simultaneously for a pair of variables is well worth mastering. In this book alone, it proves useful many times.

With the change of variable $\beta \leftarrow -\beta$ and the observations from Table 3.1 that $\sin(-\phi) = -\sin\phi$ and $\cos(-\phi) = +\cos(\phi)$, eqns. (3.9) become

$$\begin{aligned} \sin(\alpha+\beta) &= \sin\alpha\cos\beta + \cos\alpha\sin\beta, \\ \cos(\alpha+\beta) &= \cos\alpha\cos\beta - \sin\alpha\sin\beta. \end{aligned} \tag{3.10}$$

Equations (3.9) and (3.10) are the basic formulas for trigonometric functions of sums and differences of angles.

3.5.1 Variations on the sums and differences

Several useful variations on (3.9) and (3.10) are achieved by combining the equations in various straightforward ways.[16] These include that

$$\begin{aligned} \sin\alpha\sin\beta &= \frac{\cos(\alpha-\beta)-\cos(\alpha+\beta)}{2}, \\ \sin\alpha\cos\beta &= \frac{\sin(\alpha-\beta)+\sin(\alpha+\beta)}{2}, \\ \cos\alpha\cos\beta &= \frac{\cos(\alpha-\beta)+\cos(\alpha+\beta)}{2}. \end{aligned} \tag{3.11}$$

With the change of variables $\delta \leftarrow \alpha - \beta$ and $\gamma \leftarrow \alpha + \beta$, (3.9) and (3.10) become

$$\begin{aligned} \sin\delta &= \sin\left(\frac{\gamma+\delta}{2}\right)\cos\left(\frac{\gamma-\delta}{2}\right) - \cos\left(\frac{\gamma+\delta}{2}\right)\sin\left(\frac{\gamma-\delta}{2}\right), \\ \cos\delta &= \cos\left(\frac{\gamma+\delta}{2}\right)\cos\left(\frac{\gamma-\delta}{2}\right) + \sin\left(\frac{\gamma+\delta}{2}\right)\sin\left(\frac{\gamma-\delta}{2}\right), \\ \sin\gamma &= \sin\left(\frac{\gamma+\delta}{2}\right)\cos\left(\frac{\gamma-\delta}{2}\right) + \cos\left(\frac{\gamma+\delta}{2}\right)\sin\left(\frac{\gamma-\delta}{2}\right), \\ \cos\gamma &= \cos\left(\frac{\gamma+\delta}{2}\right)\cos\left(\frac{\gamma-\delta}{2}\right) - \sin\left(\frac{\gamma+\delta}{2}\right)\sin\left(\frac{\gamma-\delta}{2}\right). \end{aligned}$$

[16] Refer to footnote 15 above for the technique.

Combining these in various ways, we have that

$$\begin{aligned}
\sin\gamma + \sin\delta &= 2\sin\left(\frac{\gamma+\delta}{2}\right)\cos\left(\frac{\gamma-\delta}{2}\right),\\
\sin\gamma - \sin\delta &= 2\cos\left(\frac{\gamma+\delta}{2}\right)\sin\left(\frac{\gamma-\delta}{2}\right),\\
\cos\delta + \cos\gamma &= 2\cos\left(\frac{\gamma+\delta}{2}\right)\cos\left(\frac{\gamma-\delta}{2}\right),\\
\cos\delta - \cos\gamma &= 2\sin\left(\frac{\gamma+\delta}{2}\right)\sin\left(\frac{\gamma-\delta}{2}\right).
\end{aligned} \tag{3.12}$$

3.5.2 Trigonometric functions of double and half angles

If $\alpha = \beta$, then eqns. (3.10) become the *double-angle formulas*

$$\begin{aligned}
\sin 2\alpha &= 2\sin\alpha\cos\alpha,\\
\cos 2\alpha &= 2\cos^2\alpha - 1 = \cos^2\alpha - \sin^2\alpha = 1 - 2\sin^2\alpha.
\end{aligned} \tag{3.13}$$

Solving (3.13) for $\sin^2\alpha$ and $\cos^2\alpha$ yields the *half-angle formulas*

$$\begin{aligned}
\sin^2\alpha &= \frac{1-\cos 2\alpha}{2},\\
\cos^2\alpha &= \frac{1+\cos 2\alpha}{2}.
\end{aligned} \tag{3.14}$$

3.6 Trigonometric functions of the hour angles

In general, one uses the Taylor series of chapter 8 to calculate trigonometric functions of specific angles. We're not ready for that yet. However, for angles which happen to be integral multiples of an *hour*—there being twenty-four or 0x18 hours in a circle, just as there are twenty-four or 0x18 hours in a day[17]—for such angles simpler expressions exist. Figure 3.6 shows the angles. Since such angles arise very frequently in practice, it seems worth our while to study them specially.

Table 3.2 tabulates the trigonometric functions of these *hour angles.* To see how the values in the table are calculated, look at the square and the

[17] Hence an hour is 15°, but you weren't going to write your angles in such inelegant conventional notation as "15°," were you? Well, if you were, you're in good company.

The author is fully aware of the barrier the unfamiliar notation poses for most first-time readers of the book. The barrier is erected neither lightly nor disrespectfully. Consider:

Figure 3.6: The 0x18 (twenty-four) hours in a circle.

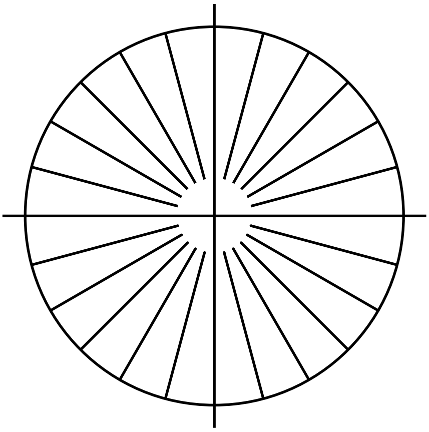
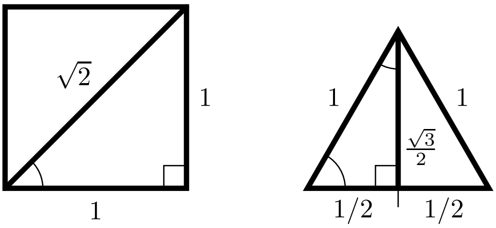


Table 3.2: Trigonometric functions of the hour angles.

ANGLE ϕ				
[radians]	[hours]	$\sin\phi$	$\tan\phi$	$\cos\phi$
0	0	0	0	1
$\frac{2\pi}{\mathrm{0x18}}$	1	$\frac{\sqrt{3}-1}{2\sqrt{2}}$	$\frac{\sqrt{3}-1}{\sqrt{3}+1}$	$\frac{\sqrt{3}+1}{2\sqrt{2}}$
$\frac{2\pi}{\mathrm{0x10}}$	$\frac{3}{2}$	$\frac{\sqrt{2-\sqrt{2}}}{2}$	$\sqrt{\frac{2-\sqrt{2}}{2+\sqrt{2}}}$	$\frac{\sqrt{2+\sqrt{2}}}{2}$
$\frac{2\pi}{\mathrm{0xC}}$	2	$\frac{1}{2}$	$\frac{1}{\sqrt{3}}$	$\frac{\sqrt{3}}{2}$
$\frac{2\pi}{8}$	3	$\frac{1}{\sqrt{2}}$	1	$\frac{1}{\sqrt{2}}$
$\frac{2\pi}{6}$	4	$\frac{\sqrt{3}}{2}$	$\sqrt{3}$	$\frac{1}{2}$
$\frac{(3)(2\pi)}{\mathrm{0x10}}$	$\frac{9}{2}$	$\frac{\sqrt{2+\sqrt{2}}}{2}$	$\sqrt{\frac{2+\sqrt{2}}{2-\sqrt{2}}}$	$\frac{\sqrt{2-\sqrt{2}}}{2}$
$\frac{(5)(2\pi)}{\mathrm{0x18}}$	5	$\frac{\sqrt{3}+1}{2\sqrt{2}}$	$\frac{\sqrt{3}+1}{\sqrt{3}-1}$	$\frac{\sqrt{3}-1}{2\sqrt{2}}$
$\frac{2\pi}{4}$	6	1	∞	0

$$\frac{\sqrt{3}\pm 1}{\sqrt{3}\mp 1} = 2 \pm \sqrt{3} \qquad \frac{2\pm\sqrt{2}}{2\mp\sqrt{2}} = 3 \pm 2\sqrt{2}$$

Figure 3.7: A square and an equilateral triangle for calculating trigonometric functions of the hour angles.



equilateral triangle[18] of Fig. 3.7. Each of the square's four angles naturally measures six hours; and, since a triangle's angles always total twelve hours (§ 2.9.3), by symmetry each of the angles of the equilateral triangle in the figure measures four. Also by symmetry, the perpendicular splits the triangle's top angle into equal halves of two hours each and its bottom leg into equal segments of length 1/2 each; and the diagonal splits the square's corner into equal halves of three hours each. The Pythagorean theorem (§ 2.9.4

- There are 0x18 hours in a circle.
- There are 360 degrees in a circle.

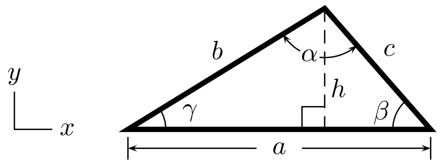
Both sentences say the same thing, don't they? But even though the "0x" hex prefix is a bit clumsy, the first sentence nevertheless says the thing rather better. The reader is urged to invest the attention and effort to master the notation.

There is a psychological trap regarding the hour. The familiar, standard clock face shows only twelve hours not twenty-four, so the angle between eleven o'clock and twelve *on the clock face* is not an hour of arc! That angle is two hours of arc. This is so because the clock face's geometry is artificial. If you have ever been to the Old Royal Observatory at Greenwich, England, you may have seen the big clock face there with all twenty-four hours on it. It'd be a bit hard to read the time from such a crowded clock face were it not so big, but anyway, the angle between hours on the Greenwich clock is indeed an honest hour of arc. [21]

The hex and hour notations are recommended mostly only for theoretical math work. It is not claimed that they offered much benefit in most technical work of the less theoretical kinds. If you wrote an engineering memorandum describing a survey angle as 0x1.80 hours instead of 22.5 degrees, for example, you'd probably not like the reception the memo got. Nonetheless, the improved notation fits a book of this kind so well that the author hazards it. It is hoped that, after trying the notation a while, the reader will approve the choice.

[18] An *equilateral* triangle is, as the name and the figure suggest, a triangle whose three sides all have the same length.

Figure 3.8: The laws of sines and cosines.



and the introduction to chapter 1) then supplies the various other lengths in the figure, after which we observe from Fig. 3.1 that

- the sine of a non-right angle in a right triangle is the opposite leg's length divided by the diagonal's,

- the tangent is the opposite leg's length divided by the adjacent leg's, and

- the cosine is the adjacent leg's length divided by the diagonal's.

With this observation and the lengths in the figure, one can calculate the sine, tangent and cosine of angles of two, three and four hours.

The values for one and five hours are found by applying (3.9) and (3.10) against the values for two and three hours just calculated. The values for zero and six hours are, of course, seen by inspection.[19]

Though quarters of a right angle are half-integral multiples of an hour, quarters of a right angle arise often enough in engineering practice that the table mentions them, too. The values for the quarters are found by applying (3.14) against the values for three and nine hours.

3.7 The laws of sines and cosines

Refer to the triangle of Fig. 3.8. By the definition of the sine function, one can write that

$$c \sin \beta = h = b \sin \gamma,$$

[19] The creative reader may notice that he can extend the table to any angle by repeated application of the various sum, difference and half-angle formulas from the preceding sections to the values already in the table. However, the Taylor series (§ 8.9) offers a cleaner, quicker way to calculate trigonometrics of non-hour angles.

or in other words that

$$\frac{\sin\beta}{b} = \frac{\sin\gamma}{c}.$$

But there is nothing special about β and γ; what is true for them must be true for α, too.[20] Hence,

$$\frac{\sin\alpha}{a} = \frac{\sin\beta}{b} = \frac{\sin\gamma}{c}. \tag{3.15}$$

This equation is known as *the law of sines.*

On the other hand, if one expresses a and b as vectors emanating from the point γ,[21]

$$\begin{aligned}\mathbf{a} &= \hat{\mathbf{x}}a,\\ \mathbf{b} &= \hat{\mathbf{x}}b\cos\gamma + \hat{\mathbf{y}}b\sin\gamma,\end{aligned}$$

then

$$\begin{aligned}c^2 &= |\mathbf{b}-\mathbf{a}|^2\\ &= (b\cos\gamma - a)^2 + (b\sin\gamma)^2\\ &= a^2 + (b^2)(\cos^2\gamma + \sin^2\gamma) - 2ab\cos\gamma.\end{aligned}$$

Since $\cos^2(\cdot) + \sin^2(\cdot) = 1$, this is that

$$c^2 = a^2 + b^2 - 2ab\cos\gamma, \tag{3.16}$$

known as *the law of cosines.*

3.8 Summary of properties

Table 3.2 on page 84 has listed the values of trigonometric functions of the hour angles. Table 3.1 on page 74 has summarized simple properties of the trigonometric functions. Table 3.3 on page 88 summarizes further properties, gathering them from §§ 3.4, 3.5 and 3.7.

[20] "But," it is objected, "there *is* something special about α. The perpendicular h drops from it."

True. However, the h is just a utility variable to help us to manipulate the equation into the desired form; we're not interested in h itself. Nothing prevents us from dropping additional perpendiculars h_β and h_γ from the other two corners and using those as utility variables, too, if we like. We can use any utility variables we want.

[21] Here is another example of the book's judicious relaxation of formal rigor, or at any rate of formal nomenclature. Of course there is no "point γ"; γ is an angle not a point. However, the writer suspects in light of Fig. 3.8 that few readers will be confused as to which point is meant. The skillful applied mathematician does not multiply labels without need!

Table 3.3: Further properties of the trigonometric functions.

$$
\begin{aligned}
\mathbf{u} &= \hat{\mathbf{x}}'(x\cos\phi + y\sin\phi) + \hat{\mathbf{y}}'(-x\sin\phi + y\cos\phi) \\
\sin(\alpha \pm \beta) &= \sin\alpha\cos\beta \pm \cos\alpha\sin\beta \\
\cos(\alpha \pm \beta) &= \cos\alpha\cos\beta \mp \sin\alpha\sin\beta \\
\sin\alpha\sin\beta &= \frac{\cos(\alpha-\beta) - \cos(\alpha+\beta)}{2} \\
\sin\alpha\cos\beta &= \frac{\sin(\alpha-\beta) + \sin(\alpha+\beta)}{2} \\
\cos\alpha\cos\beta &= \frac{\cos(\alpha-\beta) + \cos(\alpha+\beta)}{2} \\
\sin\gamma + \sin\delta &= 2\sin\left(\frac{\gamma+\delta}{2}\right)\cos\left(\frac{\gamma-\delta}{2}\right) \\
\sin\gamma - \sin\delta &= 2\cos\left(\frac{\gamma+\delta}{2}\right)\sin\left(\frac{\gamma-\delta}{2}\right) \\
\cos\delta + \cos\gamma &= 2\cos\left(\frac{\gamma+\delta}{2}\right)\cos\left(\frac{\gamma-\delta}{2}\right) \\
\cos\delta - \cos\gamma &= 2\sin\left(\frac{\gamma+\delta}{2}\right)\sin\left(\frac{\gamma-\delta}{2}\right) \\
\sin 2\alpha &= 2\sin\alpha\cos\alpha \\
\cos 2\alpha &= 2\cos^2\alpha - 1 = \cos^2\alpha - \sin^2\alpha = 1 - 2\sin^2\alpha \\
\sin^2\alpha &= \frac{1-\cos 2\alpha}{2} \\
\cos^2\alpha &= \frac{1+\cos 2\alpha}{2} \\
\frac{\sin\gamma}{c} &= \frac{\sin\alpha}{a} = \frac{\sin\beta}{b} \\
c^2 &= a^2 + b^2 - 2ab\cos\gamma
\end{aligned}
$$

3.9 Cylindrical and spherical coordinates

Section 3.3 has introduced the concept of the vector

$$\mathbf{v} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y + \hat{\mathbf{z}}z.$$

The coefficients (x, y, z) on the equation's right side are *coordinates*—specifically, *rectangular coordinates*—which given a specific, orthonormal[22] set of unit basis vectors $[\hat{\mathbf{x}}\ \hat{\mathbf{y}}\ \hat{\mathbf{z}}]$ uniquely identify a point (see Fig. 3.4 on page 76; and also, much later in the book, refer to § 15.3). Such rectangular coordinates are simple and general, and are convenient for many purposes. However, there are at least two broad classes of conceptually simple problems for which rectangular coordinates tend to be inconvenient: problems in which an axis or a point dominates. Consider for example an electric wire's magnetic field, whose intensity varies with distance from the wire (an axis); or the illumination a lamp sheds on a printed page of this book, which depends on the book's distance from the lamp (a point).

To attack a problem dominated by an axis, the *cylindrical coordinates* $(\rho; \phi, z)$ can be used instead of the rectangular coordinates (x, y, z). To attack a problem dominated by a point, the *spherical coordinates* $(r; \theta; \phi)$ can be used.[23] Refer to Fig. 3.9. Such coordinates are related to one another and to the rectangular coordinates by the formulas of Table 3.4.

Cylindrical and spherical coordinates can greatly simplify the analyses of the kinds of problems they respectively fit, but they come at a price. There are no constant unit basis vectors to match them. That is,

$$\mathbf{v} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y + \hat{\mathbf{z}}z \neq \hat{\boldsymbol{\rho}}\rho + \hat{\boldsymbol{\phi}}\phi + \hat{\mathbf{z}}z \neq \hat{\mathbf{r}}r + \hat{\boldsymbol{\theta}}\theta + \hat{\boldsymbol{\phi}}\phi.$$

It doesn't work that way. Nevertheless, *variable* unit basis vectors are defined:

$$\begin{aligned}
\hat{\boldsymbol{\rho}} &\equiv +\hat{\mathbf{x}}\cos\phi + \hat{\mathbf{y}}\sin\phi, \\
\hat{\boldsymbol{\phi}} &\equiv -\hat{\mathbf{x}}\sin\phi + \hat{\mathbf{y}}\cos\phi, \\
\hat{\mathbf{r}} &\equiv +\hat{\mathbf{z}}\cos\theta + \hat{\boldsymbol{\rho}}\sin\theta, \\
\hat{\boldsymbol{\theta}} &\equiv -\hat{\mathbf{z}}\sin\theta + \hat{\boldsymbol{\rho}}\cos\theta;
\end{aligned} \tag{3.17}$$

[22] *Orthonormal* in this context means "of unit length and at right angles to the other vectors in the set." [182, "Orthonormality," 14:19, 7 May 2006]

[23] Notice that the ϕ is conventionally written second in cylindrical $(\rho; \phi, z)$ but third in spherical $(r; \theta; \phi)$ coordinates. This odd-seeming convention is to maintain proper right-handed coordinate rotation. (The explanation will seem clearer once chapters 15 and 16 are read.)

Figure 3.9: A point on a sphere, in spherical $(r; \theta; \phi)$ and cylindrical $(\rho; \phi, z)$ coordinates. (The axis labels bear circumflexes in this figure only to disambiguate the $\hat{z}$ axis from the cylindrical coordinate z.)

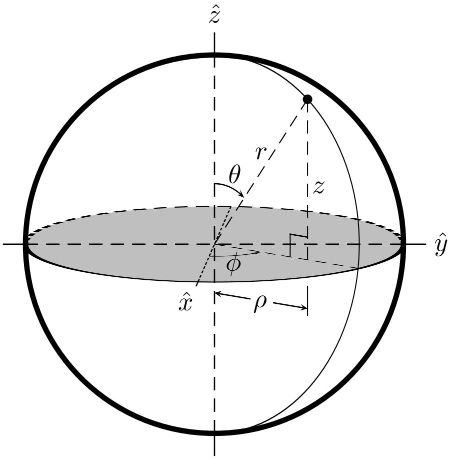


Table 3.4: Relations among the rectangular, cylindrical and spherical coordinates.

$$
\begin{aligned}
\rho^2 &= x^2 + y^2 \\
r^2 &= \rho^2 + z^2 = x^2 + y^2 + z^2 \\
\tan\theta &= \frac{\rho}{z} \\
\tan\phi &= \frac{y}{x} \\
z &= r\cos\theta \\
\rho &= r\sin\theta \\
x &= \rho\cos\phi = r\sin\theta\cos\phi \\
y &= \rho\sin\phi = r\sin\theta\sin\phi
\end{aligned}
$$

or, substituting identities from the table,

$$
\begin{aligned}
\hat{\boldsymbol{\rho}} &= \frac{\hat{\mathbf{x}}x + \hat{\mathbf{y}}y}{\rho}, \\
\hat{\boldsymbol{\phi}} &= \frac{-\hat{\mathbf{x}}y + \hat{\mathbf{y}}x}{\rho}, \\
\hat{\mathbf{r}} &= \frac{\hat{\mathbf{z}}z + \hat{\boldsymbol{\rho}}\rho}{r} = \frac{\hat{\mathbf{x}}x + \hat{\mathbf{y}}y + \hat{\mathbf{z}}z}{r}, \\
\hat{\boldsymbol{\theta}} &= \frac{-\hat{\mathbf{z}}\rho + \hat{\boldsymbol{\rho}}z}{r}.
\end{aligned}
\tag{3.18}
$$

Such variable unit basis vectors point locally in the directions in which their respective coordinates advance.

Combining pairs of (3.17)'s equations appropriately, we have also that

$$
\begin{aligned}
\hat{\mathbf{x}} &= +\hat{\boldsymbol{\rho}}\cos\phi - \hat{\boldsymbol{\phi}}\sin\phi, \\
\hat{\mathbf{y}} &= +\hat{\boldsymbol{\rho}}\sin\phi + \hat{\boldsymbol{\phi}}\cos\phi, \\
\hat{\mathbf{z}} &= +\hat{\mathbf{r}}\cos\theta - \hat{\boldsymbol{\theta}}\sin\theta, \\
\hat{\boldsymbol{\rho}} &= +\hat{\mathbf{r}}\sin\theta + \hat{\boldsymbol{\theta}}\cos\theta.
\end{aligned}
\tag{3.19}
$$

Convention usually orients $\hat{\mathbf{z}}$ in the direction of a problem's axis. Occasionally however a problem arises in which it is more convenient to orient $\hat{\mathbf{x}}$ or $\hat{\mathbf{y}}$ in the direction of the problem's axis (usually because $\hat{\mathbf{z}}$ has already been established in the direction of some other pertinent axis). Changing the meanings of known symbols like ρ, θ and ϕ would probably not be a good idea, but you can use symbols like

$$
\begin{aligned}
(\rho^x)^2 &= y^2 + z^2, & (\rho^y)^2 &= z^2 + x^2, \\
\tan\theta^x &= \frac{\rho^x}{x}, & \tan\theta^y &= \frac{\rho^y}{y}, \\
\tan\phi^x &= \frac{z}{y}, & \tan\phi^y &= \frac{x}{z},
\end{aligned}
\tag{3.20}
$$

instead if needed.[24]

[24] Symbols like ρ^x are logical but, as far as this writer is aware, not standard. The writer is not aware of any conventionally established symbols for quantities like these, but § 15.6 at least will use the ρ^x-style symbology.

3.10 The complex triangle inequalities

If the real, two-dimensional vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ represent the three sides of a triangle such that $\mathbf{a} + \mathbf{b} + \mathbf{c} = 0$, then per (2.47)

$$|\mathbf{a}| - |\mathbf{b}| \le |\mathbf{a} + \mathbf{b}| \le |\mathbf{a}| + |\mathbf{b}| .$$

These are just the triangle inequalities of § 2.9.2 in vector notation.[25] But if the triangle inequalities hold for real vectors in a plane, then why not equally for complex scalars? Consider the geometric interpretation of the Argand plane of Fig. 2.7 on page 65. Evidently,

$$|z_1| - |z_2| \le |z_1 + z_2| \le |z_1| + |z_2| \tag{3.21}$$

for any two complex numbers z_1 and z_2. Extending the sum inequality, we have that

$$\left| \sum_k z_k \right| \le \sum_k |z_k| . \tag{3.22}$$

(Naturally, the inequalities 3.21 and 3.22 hold as well for real numbers as for complex. One may find the latter inequality useful for sums of real numbers, for example, when some of the numbers summed are positive and others negative.)[26]

An important consequence of (3.22) is that if $\sum |z_k|$ converges, then $\sum z_k$ also converges. Such a consequence is important because mathematical derivations sometimes need the convergence of $\sum z_k$ established, which can be hard to do directly. Convergence of $\sum |z_k|$, which per (3.22) implies convergence of $\sum z_k$, is often easier to establish.

See also (9.19). Equation (3.22) will find use among other places in § 8.10.3.

3.11 De Moivre's theorem

Compare the Argand-plotted complex number of Fig. 2.7 (page 65) against the vector of Fig. 3.3 (page 75). Although complex numbers are scalars not vectors, the figures do suggest an analogy between complex phase and vector direction. With reference to Fig. 2.7 we can write,

$$z = (\rho)(\cos\phi + i\sin\phi) = \rho \operatorname{cis}\phi, \tag{3.23}$$

[25] Reading closely, one might note that § 2.9.2 uses the "<" sign rather than the "≤," but that's all right. See § 1.3.

[26] Section 13.9 proves the triangle inequalities more generally.

where

$$\operatorname{cis}\phi \equiv \cos\phi + i\sin\phi. \tag{3.24}$$

If $z = x + iy$, then evidently

$$\begin{aligned} x &= \rho\cos\phi, \\ y &= \rho\sin\phi. \end{aligned} \tag{3.25}$$

Per (2.68),

$$z_1 z_2 = (x_1 x_2 - y_1 y_2) + i(y_1 x_2 + x_1 y_2).$$

Applying (3.25) to this equation yields that

$$\frac{z_1 z_2}{\rho_1 \rho_2} = (\cos\phi_1 \cos\phi_2 - \sin\phi_1 \sin\phi_2) + i(\sin\phi_1 \cos\phi_2 + \cos\phi_1 \sin\phi_2).$$

But according to (3.10), this is just that

$$\frac{z_1 z_2}{\rho_1 \rho_2} = \cos(\phi_1 + \phi_2) + i\sin(\phi_1 + \phi_2),$$

or in other words that

$$z_1 z_2 = \rho_1 \rho_2 \operatorname{cis}(\phi_1 + \phi_2). \tag{3.26}$$

Equation (3.26) is a significant result. It says that if you want to multiply complex numbers, it suffices

- to multiply their magnitudes and
- to add their phases.

It follows by parallel reasoning (or by extension) that

$$\frac{z_1}{z_2} = \frac{\rho_1}{\rho_2} \operatorname{cis}(\phi_1 - \phi_2) \tag{3.27}$$

and by extension that

$$z^a = \rho^a \operatorname{cis} a\phi. \tag{3.28}$$

Equations (3.26), (3.27) and (3.28) are known as *de Moivre's theorem.*[27,28]

[27] Also called *de Moivre's formula.* Some authors apply the name of de Moivre directly only to (3.28), or to some variation thereof; but, since the three equations express essentially the same idea, if you refer to any of them as *de Moivre's theorem* then you are unlikely to be misunderstood.

[28] [153][182]

We have not shown yet, but will in § 5.4, that

$$\operatorname{cis}\phi = \exp i\phi = e^{i\phi},$$

where $\exp(\cdot)$ is the natural exponential function and e is the natural logarithmic base, both defined in chapter 5. De Moivre's theorem is most useful in this light.

Section 5.5 will revisit the derivation of de Moivre's theorem.

Chapter 4

The derivative and its incidents

The mathematics of *calculus* concerns a complementary pair of questions:[1]

- Given some function $f(t)$, what is the function's instantaneous rate of change, or *derivative,* $f'(t)$?
- Interpreting some function $f'(t)$ as an instantaneous rate of change, what is the corresponding accretion, or *integral,* $f(t)$?

This chapter builds toward a basic understanding of the first question.

[1] Although once grasped the concept is relatively simple, to understand this pair of questions—so briefly stated—is no trivial thing. They are the pair which eluded or confounded the most brilliant mathematical minds of the ancient world.

The greatest conceptual hurdle—the stroke of brilliance—probably lies in simply stating the pair of questions clearly. Sir Isaac Newton and G. W. Leibnitz cleared this hurdle for us in the seventeenth century, so now at least we know the right pair of questions to ask. With the pair in hand, the calculus beginner's first task is quantitatively to understand the pair's interrelationship, generality and significance. Such an understanding constitutes the basic calculus concept.

It cannot be the role of a book like this to lead the beginner gently toward an apprehension of the basic calculus concept. Once grasped, the concept is simple and briefly stated. Therefore, in this book, we shall take the concept as simple, briefly state it, and then move along; whereas, of course, you would find it hard first to learn calculus like that.

Many instructional textbooks have been written to lead the beginner gently. Worthy examples of such a textbook include [70].

4.1 Infinitesimals and limits

Calculus systematically treats numbers so large and so small, they lie beyond the reach of our mundane number system.

4.1.1 The infinitesimal

A number ϵ is an *infinitesimal* if it is so small that

$$0 < |\epsilon| < a$$

for all possible, mundane positive numbers a.

This is somewhat a difficult concept, so if it is not immediately clear then let us approach the matter colloquially. Let me propose to you that I have an infinitesimal.

"How big is your infinitesimal?" you ask.

"Very, very small," I reply.

"How small?"

"Very small."

"Smaller than 0x0.01?"

"Smaller than what?"

"Than 2^{-8}. You said that we should use hexadecimal notation in this book, remember?"

"Sorry. Yes, right, smaller than 0x0.01."

"What about 0x0.0001? Is it smaller than that?"

"Much smaller."

"Smaller than 0x0.0000 0000 0000 0001?"

"Smaller."

"Smaller than $2^{-\mathrm{0x1\,0000\,0000\,0000\,0000}}$?"

"Now *that* is an impressively small number. Nevertheless, my infinitesimal is smaller still."

"Zero, then."

"Oh, no. Bigger than that. My infinitesimal is definitely bigger than zero."

This is the idea of the infinitesimal. It is a definite number of a certain nonzero magnitude, but its smallness conceptually lies beyond the reach of our mundane number system.

If ϵ is an infinitesimal, then $1/\epsilon$ can be regarded as an *infinity:* a very large number much larger than any mundane number one can name.[2]

The principal advantage of using symbols like ϵ rather than 0 for infinitesimals is in that it permits us conveniently to compare one infinitesimal against another, to add them together, to divide them, etc. For instance, if $\delta = 3\epsilon$ is another infinitesimal, then the quotient δ/ϵ is not some unfathomable $0/0$; rather it is $\delta/\epsilon = 3$. In physical applications, the infinitesimals often are not true mathematical infinitesimals but rather relatively very small quantities such as the mass of a wood screw compared to the mass of a wooden house frame, or the audio power of your voice compared to that of a jet engine. The additional cost of inviting one more guest to the wedding may or may not be infinitesimal, depending upon your point of view. The key point is that the infinitesimal quantity be negligible by comparison, whatever "negligible" might mean in the context.[3]

The second-order infinitesimal $\epsilon^2 = (\epsilon)(\epsilon)$ is so small on the scale of the common, first-order infinitesimal ϵ that the even latter cannot measure it. The ϵ^2 is an infinitesimal to the infinitesimals. Third- and higher-order infinitesimals likewise are possible.

The notation $u \ll v$, or $v \gg u$, indicates that u is much less than v, typically such that one can regard the quantity u/v to be an infinitesimal. In fact, one common way to specify that ϵ be infinitesimal is to write that $\epsilon \ll 1$.

4.1.2 Limits

The notation $\lim_{z\rightarrow z_o}$ indicates that z draws as near to z_o as it possibly can. When written as $\lim_{z\rightarrow z_o^+}$, the implication is that z draws toward z_o from

[2] Some professional mathematicians have deprecated talk of this sort. However, their reasons are abstruse and appear to bear little on applications. See § 1.2.

The literature in this and related matters is vast, deep and inconclusive. It includes [19][38][40][58][62][71][76][77][78][131][136][139][147][151][166][168][175][176][179][180][185] among others.

[3] Among scientists and engineers who study wave phenomena, there is an old rule of thumb that sinusoidal waveforms be discretized not less finely than ten points per wavelength. In keeping with this book's adecimal theme (appendix A) and the concept of the hour of arc (§ 3.6), we should probably render the rule as *twelve* points per wavelength here. In any case, even very roughly speaking, a quantity greater than 1/0xC of the principal to which it compares probably cannot rightly be regarded as infinitesimal. On the other hand, a quantity less than 1/0x10000 of the principal is indeed infinitesimal for most practical purposes (but not all: for example, positions of spacecraft and concentrations of chemical impurities must sometimes be accounted more precisely). For quantities between 1/0xC and 1/0x10000, it depends on the accuracy one seeks.

the positive side such that $z > z_o$. Similarly, when written as $\lim_{z\to z_o^-}$, the implication is that z draws toward z_o from the negative side.

The reason for the notation is to provide an orderly way to handle expressions like

$$\frac{f(z)}{g(z)}$$

as z vanishes or approaches some otherwise difficult value. For example, if $f(z) \equiv 3z + 5z^3$ and $g(z) \equiv 2z + z^2$, then

$$\lim_{z\to 0} \frac{f(z)}{g(z)} = \lim_{z\to 0} \frac{3z + 5z^3}{2z + z^2} = \lim_{z\to 0} \frac{3 + 5z^2}{2 + z} = \frac{3+0}{2+0} = \frac{3}{2},$$

which is preferable to writing naïvely that $f(z)/g(z)|_{z=0} = 0/0$ (the "$|_{z=0}$" meaning, "given that, or evaluated when, $z = 0$"). The symbol "$\lim_Q$" is short for "in the limit as Q," so "$\lim_{z\to 0}$" says, "in the limit as z approaches 0."

Observe that lim is not a function like log or sin. Rather, it is a mere reminder. It is a reminder that a quantity like z approaches some value, used when saying that the quantity *equaled* the value would be inconvenient or confusing.

4.2 Combinatorics

In its general form, the problem of selecting k specific items out of a set of n available items belongs to probability theory (chapter 20). In its basic form however, the same problem also applies to the handling of polynomials or power series. This section treats the problem in its basic form.[4]

4.2.1 Combinations and permutations

Consider the following scenario. I have several small, wooden blocks of various shapes and sizes, painted different colors so that you can readily tell each block from the others. If I offer you the blocks and you are free to take all, some or none of them at your option, if you can take whichever blocks you like, then how many distinct choices of blocks confront you? Answer: the symbol n representing the number of blocks I have, a total of 2^n distinct choices confront you, for you can accept or reject the first block, then accept or reject the second, then the third, and so on.

[4] [70]

Now, suppose that what you want are exactly k blocks, neither more nor fewer. Desiring exactly k blocks, you select your favorite block first: there are n options for this. Then you select your second favorite: for this, there are $n-1$ options (why not n options? because you have already taken one block from me; I have only $n-1$ blocks left). Then you select your third favorite—for this there are $n-2$ options—and so on until you have k blocks. There are evidently

$$P\binom{n}{k} \equiv n!/(n-k)! \tag{4.1}$$

ordered ways, or *permutations,* available for you to select exactly k blocks.

However, some of these distinct permutations would put exactly the same *combination* of blocks in your hand; for instance, the permutations red-green-blue and green-red-blue constitute the same combination, whereas red-white-blue is a different combination entirely. For a single combination of k blocks (red, green, blue), evidently $k!$ permutations are possible (red-green-blue, red-blue-green, green-red-blue, green-blue-red, blue-red-green, blue-green-red). Thus, dividing the number of permutations (4.1) by $k!$ yields the number of combinations

$$\binom{n}{k} \equiv \frac{n!/(n-k)!}{k!}. \tag{4.2}$$

Table 4.1 repeats the definitions (4.1) and (4.2), and then proceeds to list several properties of the number $\binom{n}{k}$ of combinations. Among the several properties, the property of the table's third line results from changing the variable $k \leftarrow n-k$ in (4.2). The property of the table's fourth line is seen when an nth block—let us say that it is a black block—is added to an existing set of $n-1$ blocks: to choose k blocks then, you can choose either k from the original set, or the black block plus $k-1$ from the original set. The next four lines come directly from the definition (4.2); they relate combinatoric coefficients to their neighbors in Pascal's triangle (§ 4.2.2). The last line merely observes, again as at the head of this section, that 2^n total combinations are possible if any k is allowed.

Because one can choose neither fewer than zero nor more than n from n blocks,

$$\binom{n}{k} = 0 \quad \text{unless } 0 \le k \le n. \tag{4.3}$$

For $\binom{n}{k}$ when $n < 0$, there is no obvious definition.[5]

[5] So, does that mean that $\binom{n}{k}$ is not allowed when $n < 0$? Answer: probably. After

Table 4.1: Combinatorical properties.

$$\begin{aligned}
P\binom{n}{k} &\equiv n!/(n-k)! \\
\binom{n}{k} &\equiv \frac{n!/(n-k)!}{k!} = \frac{1}{k!}P\binom{n}{k} \\
&= \binom{n}{n-k} \\
&= \binom{n-1}{k-1} + \binom{n-1}{k} \\
&= \frac{n-k+1}{k}\binom{n}{k-1} \\
&= \frac{k+1}{n-k}\binom{n}{k+1} \\
&= \frac{n}{k}\binom{n-1}{k-1} \\
&= \frac{n}{n-k}\binom{n-1}{k} \\
\sum_{k=0}^{n}\binom{n}{k} &= 2^n
\end{aligned}$$

Figure 4.1: The plan for Pascal's triangle.

$$\begin{array}{c}
\binom{0}{0} \\
\binom{1}{0}\binom{1}{1} \\
\binom{2}{0}\binom{2}{1}\binom{2}{2} \\
\binom{3}{0}\binom{3}{1}\binom{3}{2}\binom{3}{3} \\
\binom{4}{0}\binom{4}{1}\binom{4}{2}\binom{4}{3}\binom{4}{4} \\
\binom{5}{0}\binom{5}{1}\binom{5}{2}\binom{5}{3}\binom{5}{4}\binom{5}{5} \\
\vdots
\end{array}$$

4.2.2 Pascal's triangle

Consider the triangular layout in Fig. 4.1 of the various possible $\binom{n}{k}$. Evaluated, this yields Fig. 4.2, *Pascal's triangle.* Notice that each entry in the triangle is the sum of the two entries immediately above, as Table 4.1 predicts. (In fact, this is the easy way to fill out Pascal's triangle: for each entry, just add the two entries above.)

4.3 The binomial theorem

This section presents the binomial theorem and one of its significant consequences.

all, it seems hard to imagine how one could allow such a quantity while retaining internal consistency within Table 4.1, for a division by zero seems to be implied. However, the question may not be the sort of question the applied mathematician is even likely to ask. He is likely to ask, rather, what $\binom{n}{k}$, $n < 0$, would mean—if anything—*in light of a particular physical problem of interest.* Only once the latter question has been answered will the applied mathematician consider whether or how to treat the quantity.

Figure 4.2: Pascal's triangle (in hexadecimal notation).

$$\begin{array}{ccccccccccccccccccc}
&&&&&&&&&1&&&&&&&&& \\
&&&&&&&&1&&1&&&&&&&& \\
&&&&&&&1&&2&&1&&&&&&& \\
&&&&&&1&&3&&3&&1&&&&&& \\
&&&&&1&&4&&6&&4&&1&&&&& \\
&&&&1&&5&&\mathrm{A}&&\mathrm{A}&&5&&1&&&& \\
&&&1&&6&&\mathrm{F}&&14&&\mathrm{F}&&6&&1&&& \\
&&1&&7&&15&&23&&23&&15&&7&&1&& \\
&1&&8&&1\mathrm{C}&&38&&46&&38&&1\mathrm{C}&&8&&1& \\
1&&9&&24&&54&&7\mathrm{E}&&7\mathrm{E}&&54&&24&&9&&1 \\
&&&&&&&&&\vdots&&&&&&&&&
\end{array}$$

Figure 4.3: Pascal's triangle (in decimal notation).

$$\begin{array}{ccccccccccccccccccc}
&&&&&&&&&1&&&&&&&&& \\
&&&&&&&&1&&1&&&&&&&& \\
&&&&&&&1&&2&&1&&&&&&& \\
&&&&&&1&&3&&3&&1&&&&&& \\
&&&&&1&&4&&6&&4&&1&&&&& \\
&&&&1&&5&&10&&10&&5&&1&&&& \\
&&&1&&6&&15&&20&&15&&6&&1&&& \\
&&1&&7&&21&&35&&35&&21&&7&&1&& \\
&1&&8&&28&&56&&70&&56&&28&&8&&1& \\
1&&9&&36&&84&&126&&126&&84&&36&&9&&1 \\
&&&&&&&&&\vdots&&&&&&&&&
\end{array}$$

4.3.1 Expanding the binomial

The *binomial theorem* holds that[6]

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k. \tag{4.4}$$

In the common case that $a = 1$, $b = \epsilon$, $|\epsilon| \ll 1$, this is that

$$(1+\epsilon)^n = \sum_{k=0}^{n} \binom{n}{k} \epsilon^k \tag{4.5}$$

(actually, eqn. 4.5 holds for any ϵ, small or large; but the typical case of interest has that $|\epsilon| \ll 1$). In either form, the binomial theorem is a direct consequence of the combinatorics of § 4.2. Since

$$(a+b)^n = (a+b)(a+b)\cdots(a+b)(a+b),$$

each $(a+b)$ factor corresponds to one of the "wooden blocks," where a means rejecting the block and b, accepting it.

4.3.2 Powers of numbers near unity

Since $\binom{n}{0} = 1$ and $\binom{n}{1} = n$, it follows from (4.5) for

$$(m,n) \in \mathbb{Z},\ m > 0,\ n \geq 0,\ |\delta| \ll 1,\ |\epsilon| \ll 1,\ |\epsilon_o| \ll 1,$$

that[7]

$$1 + m\epsilon_o \approx (1+\epsilon_o)^m$$

to arbitrary precision as long as ϵ_o is small enough. Furthermore, raising the equation to the $1/m$ power then changing $\delta \leftarrow m\epsilon_o$, we have that

$$(1+\delta)^{1/m} \approx 1 + \frac{\delta}{m}.$$

[6] The author is given to understand that, by an heroic derivational effort, (4.4) can be extended directly to nonintegral n. However, we shall have no immediate need for such an extension. Later, in Table 8.1, we will compute the Taylor series for $(1+z)^{a-1}$, anyway, which indirectly amounts to much the same thing as the extension, and has a more elegant form to boot, and moreover (at least in the author's experience) arises much more often in applications.

[7] The symbol $\approx$ means "approximately equals."

Changing $1+\delta \leftarrow (1+\epsilon)^n$ and observing from the $(1+\epsilon_o)^m$ equation above that this implies that $\delta \approx n\epsilon$, we have that

$$(1+\epsilon)^{n/m} \approx 1 + \frac{n}{m}\epsilon.$$

Inverting this equation yields that

$$(1+\epsilon)^{-n/m} \approx \frac{1}{1+(n/m)\epsilon} = \frac{[1-(n/m)\epsilon]}{[1-(n/m)\epsilon][1+(n/m)\epsilon]} \approx 1 - \frac{n}{m}\epsilon.$$

Taken together, the last two equations imply that

$$(1+\epsilon)^x \approx 1 + x\epsilon \tag{4.6}$$

for any real x.

Obscurely, (4.6) is called the *first-order Taylor expansion.* The reason the equation is called by such an unwieldy name will be explained in chapter 8, but howsoever the equation may be called, it is an important result. The equation offers a simple, accurate way of approximating any real power of numbers in the near neighborhood of 1.

4.3.3 Complex powers of numbers near unity

Equation (4.6) is fine as far as it goes, but its very form suggests the question: what if ϵ or x, or both, is complex? Changing the symbol $z \leftarrow x$ and observing that the infinitesimal ϵ also may be complex, one wants to know whether

$$(1+\epsilon)^z \approx 1 + z\epsilon \tag{4.7}$$

still holds. No work we have yet done in the book answers the question, because though a complex infinitesimal ϵ poses no particular problem, the action of a complex power z remains undefined. Still, for consistency's sake, one would like (4.7) to hold. In fact nothing prevents us from *defining* the action of a complex power such that (4.7) does hold, which we now do, logically extending the known result (4.6) into the new domain.

But we cannot just define that, can we? Surely we cannot glibly assert that "nothing prevents us" and then go to define whatever we like!

Can we?

Actually, yes, in this case we can. Consider that, insofar as (4.7) holds,

$$\begin{aligned}
(1+\epsilon)^{z_1+z_2} &= (1+\epsilon)^{z_1}(1+\epsilon)^{z_2} \approx (1+z_1\epsilon)(1+z_2\epsilon) \\
&= 1 + z_1\epsilon + z_2\epsilon + z_1z_2\epsilon^2 \approx 1 + (z_1+z_2)\epsilon; \\
(1+\epsilon)^{z_1z_2} &= [(1+\epsilon)^{z_1}]^{z_2} \approx [1+z_1\epsilon]^{z_2} \approx 1 + z_1z_2\epsilon;
\end{aligned}$$

and so on. These alone do not of course conclusively prove that our new definition is destined to behave well in every circumstance of future interest. Experience will tell. Notwithstanding, in the meantime, since we seem unable for the moment to identify a relevant circumstance in which our new definition misbehaves, since our definition does seem a natural extension of (4.6), since it does not seem to contradict anything we already know, and since no obvious alternative presents itself, let us provisionally accept the definition and find out to what results it leads.

Section 5.4 will investigate the extremely interesting effects which arise when $\Re(\epsilon) = 0$ and the power z in (4.7) grows large, but for the moment we shall use the equation in a more ordinary manner to develop the concept and basic application of the derivative, as follows.

4.4 The derivative

With (4.7) at least provisionally in hand, we can now turn to the chapter's subject proper, the derivative.

What is the derivative? The *derivative* is the instantaneous rate or slope of a function. In mathematical symbols and for the moment using real numbers,[8]

$$f'(t) \equiv \lim_{\epsilon \to 0^+} \frac{f(t+\epsilon/2) - f(t-\epsilon/2)}{\epsilon}. \tag{4.8}$$

Alternately,

$$f'(t) \equiv \lim_{\epsilon \to 0^+} \frac{f(t+\epsilon) - f(t)}{\epsilon}. \tag{4.9}$$

Because ϵ is infinitesimal, either the balanced definition (4.8) or the unbalanced definition (4.9) should in theory yield the same result (where it does not, you have a problem: the derivative does not exist at that value of t; for example, given $f[t] = 1/t$, $f'[t]_{t=0}$ does not exist despite that it exists at other values of t). Both definitions have their uses but applied mathematicians tend to prefer the balanced (4.8) because it yields comparatively accurate results in practical approximations in which ϵ, though small, is not actually infinitesimal.[9] Except where otherwise stated, this book will prefer the balanced (4.8)—or rather, as we shall eventually see, will prefer its generalized form, the balanced (4.13).

[8] Professional mathematicians tend to prefer another, more self-contained definition. Section 4.4.9 will briefly present it. See too eqns. (4.13) and (4.14).

[9] [57, §§ I:9.6 and I:9.7][39, § 4.3.4]

(Note: from this section through § 4.7, the mathematical notation necessarily grows a little thick. This cannot be helped, so if you are reading straight through, be prepared for a bit of a hard slog.)

4.4.1 The derivative of the power series

In the very common case that $f(t)$ is the power series

$$f(t) = \sum_{k=-\infty}^{\infty} c_k t^k, \tag{4.10}$$

where the c_k are in general complex coefficients, (4.8) says that

$$\begin{aligned} f'(t) &= \sum_{k=-\infty}^{\infty} \lim_{\epsilon\to 0^+} \frac{(c_k)(t+\epsilon/2)^k - (c_k)(t-\epsilon/2)^k}{\epsilon} \\ &= \sum_{k=-\infty}^{\infty} \lim_{\epsilon\to 0^+} c_k t^k \frac{(1+\epsilon/2t)^k - (1-\epsilon/2t)^k}{\epsilon}. \end{aligned}$$

Applying (4.7),

$$f'(t) = \sum_{k=-\infty}^{\infty} \lim_{\epsilon\to 0^+} c_k t^k \frac{(1+k\epsilon/2t) - (1-k\epsilon/2t)}{\epsilon},$$

which simplifies to

$$f'(t) = \sum_{k=-\infty}^{\infty} c_k k t^{k-1}, \tag{4.11}$$

assuming of course that the sum converges.[10] Equation (4.11) gives the general derivative of the power series.[11]

4.4.2 The Leibnitz notation

The $f'(t)$ notation used above for the derivative is due to Sir Isaac Newton, and is easier to start with. Usually better on the whole, however (but see

[10] The book will seldom again draw attention to such caveats of abstract rigor, even in passing. For most *concrete* series to which one is likely to apply (4.11) in practice, the series' convergence or nonconvergence will be plain enough on its face, as abstract considerations of theoretical sumworthiness fade into an expedient irrelevance. (For a closer applied consideration of sumworthiness nevertheless, see [3].)

[11] Equation (4.11) has not admittedly, explicitly considered what happens when the real t becomes the complex z, but § 4.4.7 will remedy the oversight.

appendix C), is G. W. Leibnitz's notation,

$$\begin{aligned} dt &= \epsilon, \\ df &= f(t + dt/2) - f(t - dt/2), \end{aligned}$$

such that per (4.8),

$$f'(t) = \frac{df}{dt}. \tag{4.12}$$

Here, dt is the infinitesimal, and df is a dependent infinitesimal whose size *relative to dt* depends on the independent variable t.

Conceptually, one can choose any sufficiently small size ϵ for the independent infinitesimal dt; and, actually, though we have called dt "independent," what we really mean is that the variable t with which dt is associated is independent. The size of dt may be constant (this is typically easiest) but may instead depend on t as $dt = \epsilon(t)$. Fortunately, one seldom needs to say, or care, what the size of an independent infinitesimal like dt is. All one normally needs to worry about are the sizes of other infinitesimals in proportion to dt.

As an example of the infinitesimal's use, if $f(t) \equiv 3t^3 - 5$, then $f(t \pm dt/2) = 3(t \pm dt/2)^3 - 5 = 3t^3 \pm (9/2)t^2\,dt + (9/4)t\,dt^2 \pm (3/8)\,dt^3 - 5$, whence $df = f(t + dt/2) - f(t - dt/2) = 9t^2\,dt + (3/4)\,dt^3$, and thus $df/dt = 9t^2 + (3/4)\,dt^2$—which has that $df/dt = 9t^2$ in the limit as dt tends to vanish. The example is easier if (4.7) is used to approximate that $f[(t)(1 \pm dt/2t)] \approx 3t^3 \pm (9/2)t^2\,dt - 5$, the details of which are left as an exercise.

Where two or more independent variables are simultaneously in play, say s and t, the mathematician can have two, distinct independent infinitesimals ds and dt—or, as one often styles them in such cases, ∂s and ∂t. The size of ∂s may be constant but may depend on s, t, or both, as $\partial s = \delta(s, t)$ where the δ is like ϵ an infinitesimal; and, likewise, the size of ∂t may be constant but may depend on s, t, or both, as $\partial t = \epsilon(s, t)$. Fortunately, as before, one seldom needs to say or care what the sizes are.

An applied mathematician ought to acquire, develop and retain a clear, lively, flexible mental image of Leibnitz's infinitesimal.

4.4.3 Considerations of the Leibnitz notation

The precise meaning of Leibnitz's letter d subtly depends on its context. In (4.12), the meaning is clear enough: $d(\cdot)$ signifies the amount by which $(\cdot)$ changes while the independent variable t is increasing by dt. Indeed, so essential is this point to the calculus concept that it bears repeating for emphasis!

> Insofar as $(\cdot)$ depends on t, the notation $d(\cdot)$ signifies the amount by which $(\cdot)$ changes while t is increasing by dt.

The following notational anomaly intrudes to complicate the matter. Where two or more independent variables are at work in the same equation or model, for instance s and t, convention warps Leibnitz's letter d into the shape of Carl Jacobi's letter ∂ (already seen in § 4.4.2). Convention warps the letter, not for any especially logical reason, but as a visual reminder that multiple independents are in play. For example, if $f(s,t) \equiv s^2 + 3st^2 + t^4$, then $\partial_s f = (2s + 3t^2)\partial s$ [which represents the change f undergoes while s is increasing by an increment ∂s and t is held constant] but $\partial_t f = (6st + 4t^3)\partial t$ [which represents the change f undergoes while t is increasing by an increment ∂t and s is held constant].

In practice, the style of $\partial_s f$ and $\partial_t f$ is usually overelaborate. Usually, one abbreviates each as ∂f. Context normally clarifies.

A derivative like $\partial f/\partial s$ or $\partial f/\partial t$ (that is, like $\partial_s f/\partial s$ or $\partial_t f/\partial t$) is called a *partial derivative* because in it, only one of two or more independent variables is varying. An equation containing derivatives (whether partial or otherwise), or containing infinitesimals like df or ∂f that represent the change a dependent variable like f undergoes, is called a *differential equation.* A differential equation whose derivatives or infinitesimals track more than one independent variable is called a *partial differential equation.*[12] A differential equation whose derivatives or infinitesimals track only one independent variable is called an *ordinary differential equation.*

Observe incidentally that the notation $\partial_s f$ is nonstandard. For obscure reasons (§§ 4.4.4 and 4.4.5), the style usually instead seen in print is that of $(\partial f/\partial s)\,ds$, rather.

The symbol ∂ is merely a warped letter d. Chapter 7 will use the warped letter a little, as will §§ 8.16 and 13.7. Chapter 16 will use the warped letter a lot.

We have mentioned equations with two or more independent variables. However, some equations with infinitesimals, such as the potential-kinetic energy equation that $ma\,dx = mv\,dv$, do not explicitly include or refer to any independent variable at all.[13] Context can sometimes supply an independent the equation does not mention, like t, upon which x and v

[12] Chapter 16 gives many examples of partial differential equations, for instance (16.27).

[13] The m stands for mass, the x for position, the v for speed, and the a for acceleration. The model's independent variable would probably be t for time but that variable does not happen to appear in this equation.

both depend; but it may be that the equation speaks only to how x and v change conjointly, without suggesting that either change caused the other and without explicit reference to an independent of any kind. Another example of the sort would be the economist's demand-elasticity equation, $e\,dP/P = dQ/Q$, which speaks to how P and Q change conjointly.[14] This is all right. Moreover, even in the rare application in which the lack of an independent does pose some trouble, one can often remedy the trouble by introducing a purely formal parameter to serve as it were an independent.

Convention sports at least one other notational wrinkle we should mention, a wrinkle that comes into view from chapter 7. Convention writes $\int_0^t f(\tau)\,d\tau$ rather than $\int_0^t f(\tau)\,\partial\tau$, which is to say that it eschews the warped ∂ when writing chapter 7's infinitesimal factor of integration. One could explain this by observing that the true independent t acts as a constant within the dummy's scope, and thus that the dummy sees itself as a lone independent within that scope; but whatever the explanation, that is how mathematicians will write the thing.

4.4.4 Remarks on the Leibnitz notation

A deep mystery is implicated. None has wholly plumbed it. Perhaps none ever will.

During antiquity, Thales, Anaximander, Anaximenes, Heraclitus, Parmenides, Zeno of Elea, Melissus, Anaxagoras, Leucippus, Democritus, Eudoxus, Euclid, Archimedes, Epicurus, Zeno of Cition, Chrysippus, Plato and Aristotle[15] long debated—under various forms—whether material reality and, more pertinently, immaterial reality[16] are essentially *continuous,* as geometry; or essentially *discrete,* as arithmetic. (Here we use "arithmetic" in the ancient sense of the word.) We still do not know. Indeed, we do not even know whether this is the right question to ask.

One feels obliged to salute the erudition of the professional mathematician's ongoing effort to find the question and give the answer; and yet, after twenty-five centuries, when the best efforts to give the answer seem to have

[14] The e (stated as a unitless negative number) stands for demand elasticity, the P for price, and the Q for quantity demanded. Refer to [91, chapter 4].

[15] These names and others are marshaled and accounted by [15].

[16] An influential school of thought asserts that immaterial reality does not exist, or that it might as well not exist. The school is acknowledged but the writer makes no further comment except that that is not what this paragraph is about.

Meanwhile, mathematical ontology is something we can discuss but general ontology lies beyond the writer's expertise.

succeeded chiefly to the extent to which they have *tailored the question*[17] to suit whatever answer is currently known and deemed best, why, the applicationist's interest in the matter may waver.

What the applicationist knows or believes is this: that the continuous and the discrete—whether each separately or both together—appeal directly to the mathematical intuition. If the mind's eye already sees both, and indeed sees them together in the same mental scene, then one may feel little need further to deconstruct the two.[18]

One might, of course, inadvertently obscure the mental scene by elaborating certain definitions, innocently having meant to place the whole scene upon an indisputably unified basis; but, if one does elaborate certain definitions *and this act indeed obscures the scene,* then one might ask: is it not the definitions which are suspect? Nowhere after all is it written that any indisputably unified basis shall, even in principle, be accessible to the mind of man.[19] Such a basis might be accessible, or it might not. The search is honorable and worthwhile, but it can utterly fail. We are not required to abandon mathematics if it does.

To the applicationist meanwhile, during the modeling of a given physical system, the choice of whether to employ the continuous or the discrete will chiefly depend not on abstract considerations but rather on the idiosyncratic demands of the problem at hand.

Such is the applicationist's creed.

4.4.5 The professional's creed; further remarks

Yet, what of the *professional's* creed? May the professional not also be heard? And, anyway, what do any of these things have to do with the Leibnitz notation?

The professional may indeed be heard, and more eloquently elsewhere than in this book; but, for this book's purpose, the answers are a matter of

[17] Is this adverse criticism? No. Indeed, one can hardly see what else the best efforts might have done, given the Herculean task those efforts had set for themselves. A task may, however, be too large, or be inherently impossible, even for Hercules.

[18] The mind's eye may be deceived where pure analysis does not err, of course, insofar as pure analysis relies upon disciplined formalisms. This is not denied. What the mind's eye *is,* and how it is able to perceive the abstract, are great questions of epistemology beyond the book's scope.

[19] "But ZFC is such a basis!" comes the objection.

However, whether ZFC is truly a basis or is rather a clever contraption recently bolted onto the side of preëxisting, even primordial mathematics is a question one can debate. Wittgenstein debated it. See § 1.2.2.

perspective. The professional mathematician L. E. J. Brouwer has memorably described the truths of mathematics as "fascinating by their immovability, but horrifying by their lifelessness, like stones from barren mountains of disconsolate infinity."[20] Brouwer's stones have not always appeared to the professional to countenance Leibnitz's infinitesimal.[21] Though the professional may tolerate, though the professional may even indulge, the applicationist's use of Leibnitz's infinitesimal as an expedient, the professional may find himself unable to summon greater enthusiasm for the infinitesimal than this.

Indeed, he may be even less enthusiastic. As the professional mathematician Bertrand Russell succinctly judges, "[I]nfinitesimals as explaining continuity must be regarded as unnecessary, erroneous, and self-contradictory."[22]

The professional mathematician Georg Cantor recounts:

> [A] great quarrel arose among the philosophers, of whom some followed Aristotle, others Epicurus; still others, in order to remain aloof from this quarrel, declared with Thomas Aquinas that the continuum consisted neither of infinitely many nor of a finite number of parts, but of absolutely no parts. This last opinion seems to me to contain less an explanation of the facts than a tacit confession that one has not got to the bottom of the matter and prefers to get genteelly out of its way.[23]

(The present writer has no opinion on St. Thomas' declaration or on Cantor's interpretation thereof but, were it possible, would concur in Thomas' preference to get out of the way. Alas!)

On the opposite side, you have professional mathematicians like Hermann Weyl (as quoted in §§ 1.2.2 and 1.2.4) and Abraham Robinson, along maybe with Brouwer himself,[24] who have seemed to suggest that professional mathematics might rearrange—or at least search out a more congenial perspective upon—Brouwer's immovable stones to countenance the infinitesimal nevertheless.[25] All the while, scientists and engineers have cheerfully kept plying the infinitesimal. Scientists and engineers appear to have been obtaining good results, too: bridges; weather forecasts; integrated circuits;

[20] Source: [8]. Brouwer later changed his mind.

[21] [16]

[22] [15]

[23] *Ibid.*

[24] See footnote 20.

[25] [179][143][136][15]

space shots; etc. It seems that whether one approves the infinitesimal depends chiefly on whether one's focus is in applications or in foundations.

The present book's focus is of course in applications. Fortunately, if you are a Platonist as the author is, or even if you are an intuitionist as Brouwer was, then this particular collision of foundations against applications need not much disturb you. See § 1.2.2.

So, what of the infinitesimal, this concept which has exercised the great philosophical minds of the ages? After all this, how shall the student of applied mathematics now approach it?

Following twenty-five centuries of spirited debate and vacillating consensus, the prudent student will remain skeptical of *whatever* philosophy the latest consensus (or this book) might press him to adopt. Otherwise, at the undergraduate level at any rate, many students though having learned a little calculus have not quite learned how to approach the infinitesimal at all. That is, they have never developed a clear intuition as to what Leibnitz elements like df, dt, ∂g and ∂x individually might mean—especially when these students have seen such elements heretofore only in certain, specific combinations like df/dt, $\partial g/\partial x$ and $\int f(t)\,dt$. Often, these students have developed positive misunderstandings regarding such elements. The vacillation of consensus may be blamed for this.

Therefore, in this book, having acknowledged that an opposing, meritorious school of thought exists, let us not hesitate between the two schools. Though the writer acknowledges the cleverness of certain of Cantor's results (as for example in [147, §§ 1.8 and 2.4]) and allows Cantor's ambition regarding the continuum its due, the author is an applicationist and thus his book (you are reading it) esteems Leibnitz more highly than it does Cantor. The charge could be leveled that the author does not grasp Cantor and there would be some truth in this charge, but the book's esteem for Leibnitz is more than a mere matter of preferred style. A mystery of mathematical philosophy is involved and deeply entwined, touched upon in § 1.2 and again in this section, as Cantor himself would undoubtedly have been the first to insist. Also, remember, Weyl disputed Cantor, too.

Even setting aside foundational mysteries and the formidable Cantor, there is at any rate significant practical benefit in learning how to handle the Leibnitz notation correctly. Physicists, chemists, engineers and economists have long been used to handling Leibnitz elements individually. For all these reasons among others, the present section seeks to present each Leibnitz element in its proper, individual light.

The chief recent source to which professional mathematicians seem to turn to bridge the questions this section ponders is Robinson's 1966 book

Nonstandard Analysis [136]. To conclude the subsection, therefore, we may hear Robinson's words:

> Suppose that we ask a well-trained mathematician for the meaning of [the derivative]
>
> $$\lim_{x \to x_o} \frac{f(x) - f(x_o)}{x - x_o} = a.$$
>
> Then we may rely on it that [he will explain it as § 4.4.9 below].
>
> Let us now ask our mathematician whether he would not accept the following more direct interpretation. . . .
>
> For any x in the interval of definition of $f(x)$ such that $dx = x - x_o$ is *infinitely close* to 0 but not equal to 0, the ratio df/dx, where
>
> $$df = f(x) - f(x_o),$$
>
> is *infinitely close* to a.
>
> To this question we may expect the answer that our definition may be simpler in appearance but unfortunately it is also meaningless. If we then try to explain that two numbers are infinitely close to one another if their distance . . . is *infinitely small*, . . . we shall probably be faced with the rejoinder that this is possible only if the numbers coïncide. And, so we may be told charitably, this obviously is not what we meant since it would make our explanation trivially wrong.
>
> However, in spite of this shattering rebuttal, the idea of infinitely small or *infinitesimal* quantities seems to appeal naturally to our intuition. At any rate, the use of infinitesimals was widespread during the formative stages of the Differential and Integral Calculus. . . . [136, § 1.1].

Your author is an engineer. John Derbyshire quips[26] that we "[e]ngineers were a semi-civilized tribe on an adjacent island, beery oafs who played hard rugby and never listened to concert music." Thus it seems fitting—whether your author be an oaf or not!—that in the book you are reading the "shattering rebuttal" Robinson's hypothetical interlocutor has delivered shall not unduly discomfit us.

The book will henceforth ply the Leibnitz notation and its infinitesimals vigorously, with little further hedge, cavil or equivocation.

[26] Derbyshire is author of [48] but this quip came, not written, but spoken by him in 2015.

4.4.6 Higher-order derivatives

Conventional shorthand for $d(df)$ is d^2f; for $(dt)^2$, dt^2; so

$$\frac{d(df/dt)}{dt} = \frac{d^2f}{dt^2}$$

is a derivative of a derivative, or *second derivative.* By extension, the notation

$$\frac{d^kf}{dt^k}$$

represents the kth derivative. For example, if $k = 3$, then $d[d(df)] = d^3f$, by which one writes d^3f/dt^3 for the derivative of a derivative of a derivative, or third derivative. So, if $f(t) \equiv t^4/8$, then $df/dt = t^3/2$, $d^2f/dt^2 = 3t^2/2$ and $d^3f/dt^3 = 3t$.

4.4.7 The derivative of a function of a complex variable

For (4.8) to be robust, one should like its ratio to approach a single, common value for all sufficiently small ϵ, for only when ϵ grows beyond infinitesimal size should the ratio of (4.8) become inexact. However, (4.8) considers only real, positive ϵ. What if ϵ were not positive? Indeed, what if ϵ were not even real?

This turns out to be an important question, so let us now revise (4.8) to establish the slightly more general form

$$\frac{df}{dz} \equiv \lim_{\epsilon\to 0} \frac{f(z+\epsilon/2) - f(z-\epsilon/2)}{\epsilon} \tag{4.13}$$

and let us incidentally revise (4.9), also, to establish the corresponding unbalanced form

$$\frac{df}{dz} \equiv \lim_{\epsilon\to 0} \frac{f(z+\epsilon) - f(z)}{\epsilon}, \tag{4.14}$$

where as in the section's introduction so here too applications tend to prefer the balanced (4.13) over the unbalanced (4.14).

As in (4.8), so too in (4.13) one should like the ratio to approach a single, common value[27] for all sufficiently small ϵ. However, in (4.13) one must consider not only the magnitude $|\epsilon|$ of the referential infinitesimal but

[27] One can construct apparent exceptions like $f(z) = \sin(1/z)$. If feeling obstreperous, one can construct far more unreasonable exceptions such as the one found toward the end of § 8.4. The applied mathematician can hardly be asked to expand his definitions to accommodate all such mischief! He hasn't the time.

When an apparent exception of the less unreasonable kinds arises in the context of a also its phase $\arg\epsilon$ (§ 2.11). For example, supposing that the symbol δ represented some positive, real infinitesimal, it should be equally valid to let $\epsilon = \delta$, $\epsilon = -\delta$, $\epsilon = i\delta$, $\epsilon = -i\delta$, $\epsilon = (4 - i3)\delta$, or any other infinitesimal value. The ratio df/dt of (4.13) ought to come out the same for all these. In fact, for the sake of robustness, one normally demands that the ratio does come out the same; and (4.13) rather than (4.8) is the definition we normally use for the derivative for this reason. Specifically, where the limit (4.13) or even (4.14) is sensitive to $\arg\epsilon$, there we normally say that the derivative does not exist.

Where the derivative (4.13) does exist—that is, where the derivative is finite and is insensitive to our choice of a complex, infinitesimal value for ϵ—there we say that the function $f(z)$ is *differentiable.*

Excepting the nonanalytic parts of complex numbers ($|\cdot|$, $\arg[\cdot]$, $[\cdot]^*$, $\Re[\cdot]$ and $\Im[\cdot]$; see § 2.11.3), plus the Heaviside unit step $u(t)$ and the Dirac delta $\delta(t)$ (§ 7.7), most functions encountered in applications do meet the criterion (4.13) except at isolated nonanalytic points (like $z = 0$ in $h[z] \equiv 1/z$ or $g[z] \equiv \sqrt{z}$). Meeting the criterion, such functions are fully differentiable except at their poles (where the derivative goes infinite in any case) and other nonanalytic points. Particularly, the key formula (4.7), written here as

$$(1+\epsilon)^w \approx 1 + w\epsilon,$$

works without modification when ϵ is complex; so the derivative (4.11) of the general power series,

$$\frac{d}{dz}\sum_{k=-\infty}^{\infty} c_k z^k = \sum_{k=-\infty}^{\infty} c_k k z^{k-1} \tag{4.15}$$

holds equally well for complex z as for real (but see also the next subsection).

4.4.8 The derivative of z^a

Inspection of the logic of § 4.4.1 in light of (4.7) reveals that nothing prevents us from replacing the real t, real ϵ and integral k of that section with

particular physical model, rather than attempting to accommodate the exception under the roof of an abstruse, universal rule, the applicationist is more likely to cook up a way to work around the exception in the specific context of the physical model at hand (as for example in the so-called Hadamard finite part of [5]).

arbitrary, complex z, ϵ and a. That is,

$$\begin{aligned}\frac{d(z^a)}{dz} &= \lim_{\epsilon\to 0}\frac{(z+\epsilon/2)^a-(z-\epsilon/2)^a}{\epsilon}\\ &= \lim_{\epsilon\to 0} z^a\frac{(1+\epsilon/2z)^a-(1-\epsilon/2z)^a}{\epsilon}\\ &= \lim_{\epsilon\to 0} z^a\frac{(1+a\epsilon/2z)-(1-a\epsilon/2z)}{\epsilon},\end{aligned}$$

which simplifies to

$$\frac{d(z^a)}{dz} = az^{a-1} \tag{4.16}$$

for any complex z and a.

How exactly to evaluate z^a or z^{a-1} when a is complex is another matter, treated in § 5.4 and its (5.13); but in any case you can use (4.16) for real a right now.

4.4.9 An alternate definition of the derivative

Professional mathematicians tend to prefer a less picturesque, alternate definition of the derivative (4.13) or (4.14): "For any positive number ϵ there exists a positive number δ such that

$$\left|\frac{f(z)-f(z_o)}{z-z_o}-a\right| < \epsilon \tag{4.17}$$

for all z ... for which

$$0 < |z-z_o| < \delta, \tag{4.18}$$

[the quantity a being the *derivative* df/dz]."[28]

Equations (4.17) and (4.18) bring few practical advantages to applications but are at least more self-contained than (4.13) or (4.14) is. However that may be, the derivative is such a pillar of mathematics that it behooves the applied mathematician to learn at least to recognize the professional's preferred definition of it. See also §§ 4.4.4 and 4.4.5.

4.4.10 The logarithmic derivative

Sometimes one is more interested to know the rate of $f(t)$ *in proportion to the value of* $f(t)$ than to know the absolute rate itself. For example, if

[28] The quoted original is [136, § 1.1], from which the notation has been adapted to this book's usage.

you inform me that you earn $ 1000 a year on a bond you hold, then I may commend you vaguely for your thrift but otherwise the information does not tell me much. However, if you inform me instead that you earn 10 percent a year on the same bond, then I might want to invest. The latter figure is a *proportional rate* or *logarithmic derivative,*

$$\frac{df/dt}{f(t)} = \frac{d}{dt}\ln f(t). \tag{4.19}$$

The investment principal grows at the absolute rate df/dt but the bond's proportional rate, also called (in the case of a bond) its *interest rate,* is $(df/dt)/f(t)$.

The natural logarithmic notation $\ln f(t)$ may not mean much to you yet, for we'll not introduce it formally until § 5.2, so you can ignore the right side of (4.19) for the moment; but the equation's left side at least should make sense to you. It expresses the significant concept of a proportional rate, like 10 percent annual interest on a bond.

4.5 Basic manipulation of the derivative

This section introduces the derivative chain and product rules.

4.5.1 The derivative chain rule

If f is a function of w, which itself is a function of z, then[29]

$$\frac{df}{dz} = \left(\frac{df}{dw}\right)\left(\frac{dw}{dz}\right). \tag{4.20}$$

Equation (4.20) is the *derivative chain rule.*[30]

4.5.2 The derivative product rule

In general per (4.13),

$$d\left[\prod_j f_j(z)\right] = \prod_j f_j\left(z+\frac{dz}{2}\right) - \prod_j f_j\left(z-\frac{dz}{2}\right).$$

But to first order,

$$f_j\left(z\pm\frac{dz}{2}\right) \approx f_j(z) \pm \left(\frac{df_j}{dz}\right)\left(\frac{dz}{2}\right) = f_j(z) \pm \frac{df_j}{2};$$

[29] For example, one can rewrite

$$f(z) = \sqrt{3z^2-1}$$

in the form

$$\begin{aligned} f(w) &= w^{1/2}, \\ w(z) &= 3z^2-1. \end{aligned}$$

Then

$$\begin{aligned} \frac{df}{dw} &= \frac{1}{2w^{1/2}} = \frac{1}{2\sqrt{3z^2-1}}, \\ \frac{dw}{dz} &= 6z, \end{aligned}$$

so by (4.20),

$$\frac{df}{dz} = \left(\frac{df}{dw}\right)\left(\frac{dw}{dz}\right) = \frac{6z}{2\sqrt{3z^2-1}} = \frac{3z}{\sqrt{3z^2-1}}.$$

[30] It bears emphasizing to readers who may inadvertently have picked up unhelpful ideas about the Leibnitz notation in the past: the dw factor in the denominator cancels the dw factor in the numerator, and a thing divided by itself is 1. On an applied level, this more or less is all there is to it (but see § 4.4). Other than maybe in degenerate cases like $dw = 0$, cases the applied mathematician will treat individually as they come, there is hardly more to the applied proof of the derivative chain rule than this (but see [153, Prob. 3.39]).

so, in the limit,

$$d\left[\prod_j f_j(z)\right] = \prod_j \left(f_j(z) + \frac{df_j}{2}\right) - \prod_j \left(f_j(z) - \frac{df_j}{2}\right).$$

Since the product of two or more df_j is negligible compared to the first-order infinitesimals to which they are here added,[31] this simplifies to

$$d\left[\prod_j f_j(z)\right] = \left[\prod_j f_j(z)\right]\left[\sum_k \frac{df_k}{2f_k(z)}\right] - \left[\prod_j f_j(z)\right]\left[\sum_k \frac{-df_k}{2f_k(z)}\right],$$

or in other words

$$d\prod_j f_j = \left[\prod_j f_j\right]\left[\sum_k \frac{df_k}{f_k}\right]. \tag{4.21}$$

In the common case of only two f_j, this comes to

$$d(f_1 f_2) = f_2\,df_1 + f_1\,df_2. \tag{4.22}$$

On the other hand, if $f_1(z) = f(z)$ and $f_2(z) = 1/g(z)$, then by the derivative chain rule (4.20), $df_2 = -dg/g^2$; so,

$$d\left(\frac{f}{g}\right) = \frac{g\,df - f\,dg}{g^2}, \tag{4.23}$$

and indeed

$$d\left(\frac{f^a}{g^b}\right) = \frac{f^{a-1}}{g^{b+1}}(ag\,df - bf\,dg). \tag{4.24}$$

Similarly,

$$\begin{aligned} d\left(f_1^{a_1} f_2^{a_2}\right) &= \left(f_1^{a_1-1} f_2^{a_2-1}\right)(a_1 f_2\,df_1 + a_2 f_1\,df_2) \\ &= \left(f_1^{a_1} f_2^{a_2}\right)\left(\frac{a_1\,df_1}{f_1} + \frac{a_2\,df_2}{f_2}\right). \end{aligned} \tag{4.25}$$

Equation (4.21) is the *derivative product rule.*

[31] Unless $df_j \approx 0$ to first order, in which case it contributes nothing to the derivative, anyway.

After studying the complex exponential in chapter 5, we shall stand in a position to write (4.21) in the slightly specialized but often useful form[32]

$$\begin{aligned} d\left[\prod_j g_j^{a_j} \prod_j e^{b_j h_j} \prod_j \ln c_j p_j\right] &\\ = \left[\prod_j g_j^{a_j} \prod_j e^{b_j h_j} \prod_j \ln c_j p_j\right]& \\ \times \left[\sum_k a_k \frac{dg_k}{g_k} + \sum_k b_k\,dh_k + \sum_k \frac{dp_k}{p_k \ln c_k p_k}\right]&. \end{aligned} \tag{4.26}$$

where the a_k, b_k and c_k are arbitrary complex coefficients and the g_k, h_k and p_k are arbitrary functions.[33]

4.5.3 A derivative product pattern

According to (4.22) and (4.16), the derivative of the product $z^a f(z)$ with respect to its independent variable z is

$$\frac{d}{dz}[z^a f(z)] = z^a \frac{df}{dz} + az^{a-1} f(z).$$

Swapping the equation's left and right sides then dividing through by z^a yields that

$$\frac{df}{dz} + a\frac{f}{z} = \frac{d(z^a f)}{z^a\,dz}, \tag{4.27}$$

a pattern worth committing to memory, emerging among other places in § 16.9.[34]

4.6 Extrema and higher derivatives

One problem which arises very frequently in applied mathematics is the problem of finding a local *extremum*—that is, a local minimum or maximum—of a real-valued function $f(x)$. Refer to Fig. 4.4. The almost dis-

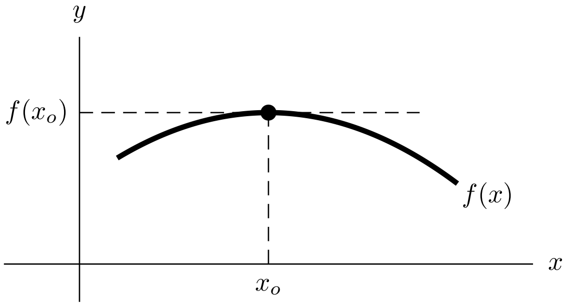
[32] This paragraph is extra. You can skip it for now if you prefer.

[33] The subsection is sufficiently abstract that it is a little hard to understand unless one already knows what it means. An example may help:

$$d\left[\frac{u^2v^3}{z}e^{-5t}\ln 7s\right] = \left[\frac{u^2v^3}{z}e^{-5t}\ln 7s\right]\left[2\frac{du}{u} + 3\frac{dv}{v} - \frac{dz}{z} - 5\,dt + \frac{ds}{s\ln 7s}\right].$$

[34] This section completes the forward reference of § 2.6.5. See chapter 2's footnote 32.

Figure 4.4: A local extremum.



tinctive characteristic of the extremum $f(x_o)$ is that[35]

$$\left.\frac{df}{dx}\right|_{x=x_o} = 0. \tag{4.28}$$

At the extremum, the slope is zero. The curve momentarily runs level there. One solves (4.28) to find the extremum.

Whether the extremum be a minimum or a maximum depends on whether the curve turn from a downward slope to an upward, or from an upward slope to a downward, respectively. If from downward to upward, then the derivative of the slope is evidently positive; if from upward to downward, then negative. But the derivative of the slope is just the derivative of the derivative, or second derivative. Hence if $df/dx = 0$ at $x = x_o$, then

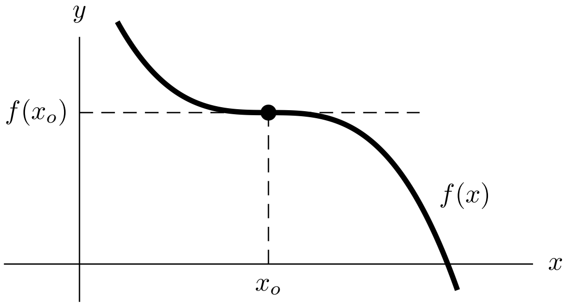
$$\begin{aligned}
\left.\frac{d^2f}{dx^2}\right|_{x=x_o} &> 0 \quad \text{implies a local minimum at } x_o;\\
\left.\frac{d^2f}{dx^2}\right|_{x=x_o} &< 0 \quad \text{implies a local maximum at } x_o.
\end{aligned}$$

Regarding the case

$$\left.\frac{d^2f}{dx^2}\right|_{x=x_o} = 0,$$

[35] The notation $P|_Q$ means "P when Q," "P, given Q," or "P evaluated at Q." Sometimes it is alternately written $P|Q$ or $[P]_Q$.

Figure 4.5: A level inflection.



this might be either a minimum or a maximum but more probably is neither, being rather a *level inflection point* as depicted in Fig. 4.5.[36] (In general the term *inflection point* signifies a point at which the second derivative is zero. The inflection point of Fig. 4.5 is *level* because its first derivative is zero, too.)

4.7 L'Hôpital's rule

If $z = z_o$ is a root of both $f(z)$ and $g(z)$, or alternately if $z = z_o$ is a pole of both functions—that is, if both functions go to zero or infinity together at $z = z_o$—then *l'Hôpital's rule* holds that

$$\lim_{z \to z_o} \frac{f(z)}{g(z)} = \left. \frac{df/dz}{dg/dz} \right|_{z=z_o} . \tag{4.29}$$

[36] Of course if the first and second derivatives are zero not just at $x = x_o$ but everywhere, then $f(x) = y_o$ is just a level straight line, but you knew that already. Whether one chooses to call some arbitrary point on a level straight line an inflection point or an extremum, or both or neither, would be a matter of definition, best established not by prescription but rather by the needs of the model at hand.

In the case in which $z = z_o$ is a root, l'Hôpital's rule is proved by reasoning[37]

$$\begin{aligned}\lim_{z\to z_o} \frac{f(z)}{g(z)} &= \lim_{z\to z_o} \frac{f(z)-0}{g(z)-0} \\ &= \lim_{z\to z_o} \frac{f(z)-f(z_o)}{g(z)-g(z_o)} = \lim_{z\to z_o} \frac{df}{dg} = \lim_{z\to z_o} \frac{df/dz}{dg/dz}.\end{aligned}$$

In the case in which $z = z_o$ is a pole, new functions $F(z) \equiv 1/f(z)$ and $G(z) \equiv 1/g(z)$ of which $z = z_o$ is a root are defined, with which

$$\lim_{z\to z_o} \frac{f(z)}{g(z)} = \lim_{z\to z_o} \frac{G(z)}{F(z)} = \lim_{z\to z_o} \frac{dG}{dF} = \lim_{z\to z_o} \frac{-dg/g^2}{-df/f^2},$$

where we have used the fact from (4.16) that $d(1/u) = -du/u^2$ for any u. Canceling the minus signs and multiplying by g^2/f^2, we have that

$$\lim_{z\to z_o} \frac{g(z)}{f(z)} = \lim_{z\to z_o} \frac{dg}{df}.$$

Inverting,

$$\lim_{z\to z_o} \frac{f(z)}{g(z)} = \lim_{z\to z_o} \frac{df}{dg} = \lim_{z\to z_o} \frac{df/dz}{dg/dz}.$$

And if z_o itself is infinite? Then, whether it represents a root or a pole, we define the new variable $Z \equiv 1/z$ and the new functions $\Phi(Z) \equiv f(1/Z) = f(z)$ and $\Gamma(Z) \equiv g(1/Z) = g(z)$, with which we apply l'Hôpital's rule for $Z \to 0$ to obtain

$$\begin{aligned}\lim_{z\to\infty} \frac{f(z)}{g(z)} &= \lim_{Z\to 0} \frac{\Phi(Z)}{\Gamma(Z)} = \lim_{Z\to 0} \frac{d\Phi/dZ}{d\Gamma/dZ} = \lim_{Z\to 0} \frac{df/dZ}{dg/dZ} \\ &= \lim_{\substack{z\to\infty,\\ Z\to 0}} \frac{(df/dz)(dz/dZ)}{(dg/dz)(dz/dZ)} = \lim_{z\to\infty} \frac{(df/dz)(-z^2)}{(dg/dz)(-z^2)} = \lim_{z\to\infty} \frac{df/dz}{dg/dz}.\end{aligned}$$

Nothing in the derivation requires that z or z_o be real. Nothing prevents one from applying l'Hôpital's rule recursively, should the occasion arise.[38]

[37] Partly with reference to [182, "L'Hopital's rule," 03:40, 5 April 2006].

[38] Consider for example the ratio $\lim_{x\to 0}(x^3+x)^2/x^2$, which is 0/0. The easier way to resolve this particular ratio would naturally be to cancel a factor of x^2 from it; but just to make the point let us apply l'Hôpital's rule instead, reducing the ratio to $\lim_{x\to 0} 2(x^3+x)(3x^2+1)/2x$, which is still 0/0. Applying l'Hôpital's rule again to the result yields $\lim_{x\to 0} 2[(3x^2+1)^2+(x^3+x)(6x)]/2 = 2/2 = 1$. Where expressions involving trigonometric functions (chapters 3 and 5) or special functions (mentioned in part III) appear in ratio, a recursive application of l'Hôpital's rule can be just the thing one needs.

Observe that one must stop applying l'Hôpital's rule once the ratio is no longer 0/0 or ∞/∞. In the example, applying the rule a third time would have ruined the result.

L'Hôpital's rule is used in evaluating indeterminate forms of the kinds $0/0$ and ∞/∞, plus related forms like $(0)(\infty)$ which can be recast into either of the two main forms. Good examples of the use require mathematics from chapter 5 and later, but if we may borrow from (5.8) the natural logarithmic function and its derivative,[39]

$$\frac{d}{dx}\ln x = \frac{1}{x},$$

then a typical l'Hôpital example is[40]

$$\lim_{x\to\infty}\frac{\ln x}{\sqrt{x}} = \lim_{x\to\infty}\frac{1/x}{1/2\sqrt{x}} = \lim_{x\to\infty}\frac{2}{\sqrt{x}} = 0.$$

The example incidentally shows that natural logarithms grow slower than square roots, an instance of a more general principle we shall meet in § 5.3.

Section 5.3 will put l'Hôpital's rule to work.

4.8 The Newton-Raphson iteration

The *Newton-Raphson iteration* is a powerful, fast converging, broadly applicable method for finding roots numerically. Given a function $f(z)$ of which the root is desired, the Newton-Raphson iteration is

$$z_{k+1} = \left[z - \frac{f(z)}{\frac{d}{dz}f(z)} \right]_{z=z_k}. \tag{4.30}$$

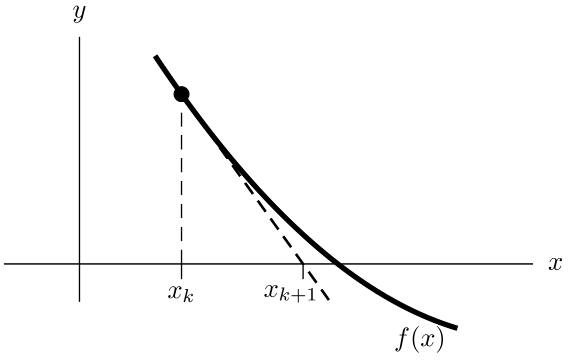
One begins the iteration by guessing the root and calling the guess z_0. Then z_1, z_2, z_3, etc., calculated in turn by the iteration (4.30), give successively better estimates of the true root z_∞.

To understand the Newton-Raphson iteration, consider the function $y = f(x)$ of Fig 4.6. The iteration approximates the curve $f(x)$ by its tangent

[39] This paragraph is optional reading for the moment. You can read chapter 5 first, then come back here and read the paragraph if you prefer.

[40] [146, § 10-2]

Figure 4.6: The Newton-Raphson iteration.



line[41] (shown as the dashed line in the figure):

$$\tilde{f}_k(x) = f(x_k) + \left[\frac{d}{dx}f(x)\right]_{x=x_k}(x - x_k).$$

It then approximates the root x_{k+1} as the point at which $\tilde{f}_k(x_{k+1}) = 0$:

$$\tilde{f}_k(x_{k+1}) = 0 = f(x_k) + \left[\frac{d}{dx}f(x)\right]_{x=x_k}(x_{k+1} - x_k).$$

Solving for x_{k+1}, we have that

$$x_{k+1} = \left[x_k - \frac{f(x)}{\frac{d}{dx}f(x)}\right]_{x=x_k},$$

which is (4.30) with $x \leftarrow z$.

Although the illustration uses real numbers, nothing forbids complex z and $f(z)$. The Newton-Raphson iteration works just as well for these.

[41] A *tangent* line, also just called a *tangent,* is the line which most nearly approximates a curve at a given point. The tangent touches the curve at the point, and in the neighborhood of the point it goes in the same direction the curve goes. The dashed line in Fig. 4.6 is a good example of a tangent line.

The relationship between the tangent line and the trigonometric tangent function of chapter 3 is slightly obscure, maybe more of linguistic interest than of mathematical. The trigonometric tangent function is named from a variation on Fig. 3.1 in which the triangle's bottom leg is extended to unit length, leaving the rightward leg tangent to the circle.

The principal limitation of the Newton-Raphson arises when the function has more than one root, as most interesting functions do. The iteration often converges on the root nearest the initial guess z_o but does not always, and in any case there is no guarantee that the root it finds is the one you wanted. The most straightforward way to beat this problem is to find *all* the roots: first you find some root α, then you remove that root (without affecting any of the other roots) by dividing $f(z)/(z-\alpha)$, then you find the next root by iterating on the new function $f(z)/(z-\alpha)$, and so on until you have found all the roots. If this procedure is not practical (perhaps because the function has a large or infinite number of roots), then you should probably take care to make a sufficiently accurate initial guess if you can.

A second limitation of the Newton-Raphson is that, if you happen to guess z_0 especially unfortunately, then the iteration might never converge at all. For example, the roots of $f(z) = z^2 + 2$ are $z = \pm i\sqrt{2}$, but if you guess that $z_0 = 1$ then the iteration has no way to leave the real number line, so it never converges[42] (and if you guess that $z_0 = \sqrt{2}$—well, try it with your pencil and see what z_2 comes out to be). You can fix the problem with a different, possibly complex initial guess.

A third limitation arises where there is a multiple root. In this case, the Newton-Raphson normally still converges, but relatively slowly. For instance, the Newton-Raphson converges relatively slowly on the triple root of $f(z) = z^3$. However, even the relatively slow convergence is still pretty fast and is usually adequate, even for calculations by hand.

Usually in practice, the Newton-Raphson iteration works very well. For most functions, once the Newton-Raphson finds the root's neighborhood, it converges on the actual root remarkably quickly. Figure 4.6 shows why: in the neighborhood, the curve hardly departs from the straight line.

The Newton-Raphson iteration is a champion square-root calculator, incidentally. Consider

$$f(x) = x^2 - p,$$

whose roots are

$$x = \pm\sqrt{p}.$$

Per (4.30), the Newton-Raphson iteration for this is

$$x_{k+1} = \frac{1}{2}\left[x_k + \frac{p}{x_k}\right]. \tag{4.31}$$

If you start by guessing

$$x_0 = 1$$

[42] It is entertaining to try this on a computer. Then try again with $z_0 = 1 + i2^{-0\text{x}10}$.

and iterate several times, the iteration (4.31) converges on $x_\infty = \sqrt{p}$ fast. To calculate the nth root $x = p^{1/n}$, let

$$f(x) = x^n - p$$

and iterate[43,44]

$$x_{k+1} = \frac{1}{n}\left[(n-1)x_k + \frac{p}{x_k^{n-1}}\right]. \tag{4.32}$$

Section 13.7 generalizes the Newton-Raphson iteration to handle vector-valued functions.

This concludes the chapter. Chapter 8, treating the Taylor series, will continue the general discussion of the derivative.

[43] Equations (4.31) and (4.32) work not only for real p but also usually for complex. Given $x_0 = 1$, however, they converge reliably and orderly only for real, nonnegative p. (To see why, sketch $f[x]$ in the fashion of Fig. 4.6.)

If reliable, orderly convergence is needed for complex $p = u + iv = \sigma \operatorname{cis} \psi$, $\sigma \geq 0$, you can decompose $p^{1/n}$ per de Moivre's theorem (3.28) as $p^{1/n} = \sigma^{1/n} \operatorname{cis}(\psi/n)$, in which $\operatorname{cis}(\psi/n) = \cos(\psi/n) + i\sin(\psi/n)$ is calculated by the Taylor series of Table 8.1. Then σ is real and nonnegative, upon which (4.32) reliably, orderly computes $\sigma^{1/n}$.

The Newton-Raphson iteration however excels as a *practical* root-finding technique, so it often pays to be a little less theoretically rigid in applying it. If so, then don't bother to decompose; seek $p^{1/n}$ directly, using complex z_k in place of the real x_k. In the uncommon event that the direct iteration does not seem to converge, start over again with some randomly chosen complex z_0. This saves effort and usually works.

[44] [146, § 4-9][119, § 6.1.1][178]

Chapter 5

The complex exponential

The complex exponential, especially the complex natural exponential, is ubiquitous in higher mathematics. There seems hardly a corner of calculus, basic or advanced, in which the complex exponential does not strongly impress itself and frequently arise. Because the complex exponential emerges (at least pedagogically) out of the real exponential—and, especially, because the complex *natural* exponential emerges out of the real natural exponential—this chapter introduces first the real natural exponential and its inverse, the real natural logarithm; and then proceeds to show how the two can operate on complex arguments.

This chapter develops the close relationship between the natural exponential and chapter 3's trigonometrics, showing that all these (including the complex natural exponential) belong to a single exponential/trigonometric family. After developing the relationship, the chapter works out the derivatives of the family's several functions. Also, the chapter treats the functions' inverses and works out the derivatives of these inverses—the derivatives of the inverses turning out to be particularly interesting.

5.1 The real exponential

Consider the factor

$$(1+\epsilon)^N.$$

This is the overall factor by which a quantity grows after N iterative rounds of multiplication[1] by $(1+\epsilon)$. What happens when ϵ is very small but N is

[1] For example, let a quantity A be multiplied by $(1+\epsilon)$, then by $(1+\epsilon)$ again, and then by $(1+\epsilon)$ a third time. The product $(1+\epsilon)^3 A$ results from these three rounds of

very large? The really interesting question is, what happens in the limit, as $\epsilon \to 0$ and $N \to \infty$, while $x = \epsilon N$ remains a finite number? The answer is that the factor becomes

$$\exp x \equiv \lim_{\epsilon \to 0} (1 + \epsilon)^{x/\epsilon}. \tag{5.1}$$

Equation (5.1) defines the *natural exponential function*—commonly, more briefly named the *exponential function.* Another way to write the same definition is

$$\begin{aligned} \exp x &= e^x, && (5.2) \\ e &\equiv \lim_{\epsilon \to 0} (1 + \epsilon)^{1/\epsilon}. && (5.3) \end{aligned}$$

In whichever form we write it, the question remains as to whether the limit actually exists; that is, whether $1 < e < \infty$; whether in fact we can put some concrete bound on e. To show that we can,[2] we observe per (4.13) and (4.6) that the derivative of the exponential function is

$$\begin{aligned} \frac{d}{dx} \exp x &= \lim_{\delta \to 0} \frac{\exp(x + \delta/2) - \exp(x - \delta/2)}{\delta} \\ &= \lim_{\delta, \epsilon \to 0} \frac{(1 + \epsilon)^{(x + \delta/2)/\epsilon} - (1 + \epsilon)^{(x - \delta/2)/\epsilon}}{\delta} \\ &= \lim_{\delta, \epsilon \to 0} (1 + \epsilon)^{x/\epsilon} \frac{(1 + \epsilon)^{+\delta/2\epsilon} - (1 + \epsilon)^{-\delta/2\epsilon}}{\delta} \\ &= \lim_{\delta, \epsilon \to 0} (1 + \epsilon)^{x/\epsilon} \frac{(1 + \delta/2) - (1 - \delta/2)}{\delta} \\ &= \lim_{\epsilon \to 0} (1 + \epsilon)^{x/\epsilon}, \end{aligned}$$

which is to say that

$$\frac{d}{dx} \exp x = \exp x. \tag{5.4}$$

This is a curious, important result: the derivative of the exponential function is the exponential function itself; the slope and height of the exponential

multiplication. Overall, such a product is $(1 + \epsilon)^3$ times the quantity A with which we started. In this example, $N = 3$.

[2] Excepting (5.4), the author would prefer to omit much of the rest of this section, but even at the applied level cannot think of a logically permissible way to do it. It seems nonobvious that the limit $\lim_{\epsilon \to 0} (1 + \epsilon)^{1/\epsilon}$ actually does exist. The rest of this section shows why it does.

function are everywhere equal. For the moment, however, what interests us is that

$$\left.\frac{d}{dx}\exp x\right|_{x=0} = \exp 0 = \lim_{\epsilon\to 0}(1+\epsilon)^0 = 1,$$

which says that the slope and height of the exponential function are both unity at $x = 0$, implying that the straight line which best approximates the exponential function in that neighborhood—the *tangent line,* which just grazes the curve—is

$$y(x) = 1 + x.$$

With the tangent line $y(x)$ found, the next step toward putting a concrete bound on e is to show that $y(x) \le \exp x$ for all real x, that the curve runs nowhere below the line. To show this, we observe per (5.1) that the essential action of the exponential function is to multiply repeatedly by $1+\epsilon$ as x increases, to divide repeatedly by $1+\epsilon$ as x decreases. Since $1+\epsilon > 1$, this action means for real x that

$$\exp x_1 \le \exp x_2 \quad \text{if} \quad x_1 \le x_2.$$

However, a positive number remains positive no matter how many times one multiplies or divides it by $1+\epsilon$, so the same action also means that

$$0 \le \exp x$$

for all real x. In light of (5.4), the last two equations imply further that

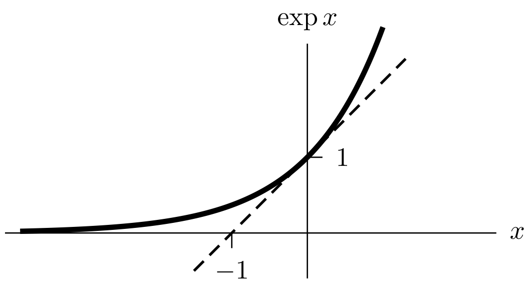
$$\begin{aligned}\left.\frac{d}{dx}\exp x\right|_{x=x_1} &\le \left.\frac{d}{dx}\exp x\right|_{x=x_2} \quad \text{if } x_1 \le x_2,\\ 0 &\le \frac{d}{dx}\exp x.\end{aligned}$$

But we have already established the tangent line $y(x) = 1 + x$ such that

$$\begin{aligned}\exp 0 = {} & \quad y(0) & = 1,\\ \left.\frac{d}{dx}\exp x\right|_{x=0} = {} & \left.\frac{dy}{dx}\right|_{x=0} & = 1;\end{aligned}$$

that is, such that the line just grazes the curve of $\exp x$ at $x = 0$. Rightward, at $x > 0$, the curve's slope evidently only increases, bending upward away from the line. Leftward, at $x < 0$, the curve's slope evidently only decreases,

Figure 5.1: The natural exponential.



again bending upward away from the line. Either way, the curve never crosses below the line for real x. In symbols,

$$y(x) \le \exp x.$$

Figure 5.1 depicts.

Evaluating the last inequality at $x = -1/2$ and $x = 1$, we have that

$$\begin{aligned}\frac{1}{2} &\le \exp\left(-\frac{1}{2}\right),\\ 2 &\le \exp(1).\end{aligned}$$

But per (5.2) $\exp x = e^x$, so

$$\begin{aligned}\frac{1}{2} &\le e^{-1/2},\\ 2 &\le e^1,\end{aligned}$$

or in other words,

$$2 \le e \le 4, \tag{5.5}$$

which in consideration of (5.2) puts the desired bound on the exponential function. The limit does exist.

Dividing (5.4) by $\exp x$ yields the *logarithmic derivative* (§ 4.4.10)

$$\frac{d(\exp x)}{(\exp x)\, dx} = 1, \tag{5.6}$$

a form which expresses or captures the deep curiosity of the natural exponential maybe even better than does (5.4).

By the Taylor series of Table 8.1, the value[3]

$$e \approx \text{0x2.B7E1}$$

can readily be calculated, but the derivation of that series does not come until chapter 8.

5.2 The natural logarithm

In the general exponential expression b^x one can choose any base b; for example, $b = 2$ is an interesting choice. As we shall see in § 5.4, however, it turns out that $b = e$, where e is the constant introduced in (5.3), is the most interesting choice of all. For this reason among others, the base-e logarithm is similarly interesting, such that we define for it the special notation

$$\ln(\cdot) = \log_e(\cdot)$$

and call it the *natural logarithm.* Just as for any other base b, so also for base $b = e$; thus the natural logarithm inverts the natural exponential and vice versa:

$$\begin{aligned} \ln \exp x = \ln e^x &= x, \\ \exp \ln x = e^{\ln x} &= x. \end{aligned} \tag{5.7}$$

Figure 5.2 plots the natural logarithm.

If

$$y = \ln x,$$

then

$$x = \exp y,$$

and per (5.4),

$$\frac{dx}{dy} = \exp y.$$

But this means that

$$\frac{dx}{dy} = x,$$

the inverse of which is

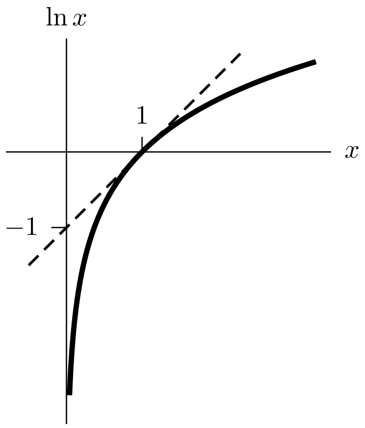
$$\frac{dy}{dx} = \frac{1}{x}.$$

In other words,

$$\frac{d}{dx} \ln x = \frac{1}{x}. \tag{5.8}$$

[3] [152, sequence A004593]

Figure 5.2: The natural logarithm.



Like many of the equations in these early chapters, here is another rather significant result.[4]

One can specialize Table 2.5's logarithmic base-conversion identity to read

$$\log_b w = \frac{\ln w}{\ln b}. \tag{5.9}$$

This equation converts any logarithm to a natural logarithm. Base $b = 2$ logarithms are interesting, so we note here that[5]

$$\ln 2 = -\ln\frac{1}{2} \approx \text{0x0.B172},$$

which chapter 8 and its Table 8.1 will show how to calculate.

5.3 Fast and slow functions

The exponential $\exp x$ is a *fast function.* The logarithm $\ln x$ is a *slow function.* These functions grow, diverge or decay respectively faster and slower than x^a.

[4] Besides the result itself, the technique which leads to the result is also interesting and is worth mastering. We will use the technique more than once in this book.

[5] [152, sequence A002162]

Such claims are proved by l'Hôpital's rule (4.29). Applying the rule, we have that

$$\begin{aligned}
\lim_{x\to\infty}\frac{\ln x}{x^a} &= \lim_{x\to\infty}\frac{-1}{ax^a} = \begin{cases} 0 & \text{if } a > 0, \\ +\infty & \text{if } a \le 0, \end{cases} \\
\lim_{x\to 0}\frac{\ln x}{x^a} &= \lim_{x\to 0}\frac{-1}{ax^a} \;\;= \begin{cases} -\infty & \text{if } a \ge 0, \\ 0 & \text{if } a < 0, \end{cases}
\end{aligned} \tag{5.10}$$

which reveals the logarithm to be a slow function. Since the $\exp(\cdot)$ and $\ln(\cdot)$ functions are mutual inverses, we can leverage (5.10) to show also that

$$\begin{aligned}
\lim_{x\to\infty}\frac{\exp(\pm x)}{x^a} &= \lim_{x\to\infty}\exp\left[\ln\frac{\exp(\pm x)}{x^a}\right] \\
&= \lim_{x\to\infty}\exp\left[\pm x - a\ln x\right] \\
&= \lim_{x\to\infty}\exp\left[(x)\left(\pm 1 - a\frac{\ln x}{x}\right)\right] \\
&= \lim_{x\to\infty}\exp\left[(x)\left(\pm 1 - 0\right)\right] \\
&= \lim_{x\to\infty}\exp\left[\pm x\right].
\end{aligned}$$

That is,

$$\begin{aligned}
\lim_{x\to\infty}\frac{\exp(+x)}{x^a} &= \infty, \\
\lim_{x\to\infty}\frac{\exp(-x)}{x^a} &= 0,
\end{aligned} \tag{5.11}$$

which reveals the exponential to be a fast function. Exponentials grow or decay faster than powers; logarithms diverge slower.

Such conclusions are extended to bases other than the natural base e simply by observing that $\log_b x = \ln x / \ln b$ and that $b^x = \exp(x\ln b)$. Thus exponentials generally are fast and logarithms generally are slow, regardless of the base.[6]

It is interesting and worthwhile to contrast the sequence

$$\ldots, -\frac{3!}{x^4}, \frac{2!}{x^3}, -\frac{1!}{x^2}, \frac{0!}{x^1}, \frac{x^0}{0!}, \frac{x^1}{1!}, \frac{x^2}{2!}, \frac{x^3}{3!}, \frac{x^4}{4!}, \ldots$$

[6] There are of course some degenerate edge cases like $b = 0$ and $b = 1$. The reader can detail these as the need arises.

against the sequence

$$\ldots, -\frac{3!}{x^4}, \frac{2!}{x^3}, -\frac{1!}{x^2}, \frac{0!}{x^1}, \ln x, \frac{x^1}{1!}, \frac{x^2}{2!}, \frac{x^3}{3!}, \frac{x^4}{4!}, \ldots$$

As $x \to +\infty$, each sequence increases in magnitude going rightward. Also, each term in each sequence is the derivative with respect to x of the term to its right—except left of the middle element in the first sequence and right of the middle element in the second. The exception is peculiar. What is going on here?

The answer is that x^0 (which is just a constant) and $\ln x$ *both are of zeroth order in x.* This seems strange at first because $\ln x$ diverges as $x \to \infty$ whereas x^0 does not, but the divergence of the former is extremely slow—so slow, in fact, that per (5.10) $\lim_{x\to\infty}(\ln x)/x^\epsilon = 0$ for any positive ϵ no matter how small.[7] Figure 5.2 has plotted $\ln x$ only for $x \sim 1$, but beyond the figure's window the curve (whose slope is $1/x$) flattens rapidly rightward, to the extent that it locally resembles the plot of a constant value; and indeed because

$$0 = \lim_{u\to\infty} \ln \frac{x+u}{u},$$

whence

$$\begin{aligned} 0 &= \lim_{u\to\infty} \ln(x+u) - \ln u, \\ \lim_{u\to\infty} \ln u &= \lim_{u\to\infty} \ln(x+u), \\ 1 &= \lim_{u\to\infty} \frac{\ln(x+u)}{\ln u}, \end{aligned}$$

one can write,

$$x^0 = \lim_{u\to\infty} \frac{\ln(x+u)}{\ln u},$$

which casts x^0 as a logarithm greatly shifted and moderately scaled. Admittedly, one ought not strain such logic too far, because $\ln x$ is not in fact a constant, but the point nevertheless remains that x^0 and $\ln x$ often play analogous roles in mathematics. The logarithm can in some situations profitably be thought of as a "diverging constant" of sorts.

[7] One does not grasp how truly slow the divergence is until one calculates a few concrete values. Consider for instance how far out x must run to make $\ln x = \mathrm{0x100}$. It's a long, long way. The natural logarithm does indeed eventually diverge to infinity, in the literal sense that there is no height it does not eventually reach, but it certainly does not hurry. As we have seen, it takes practically forever just to reach 0x100.

Less strange-seeming perhaps is the consequence of (5.11) that $\exp x$ is of infinite order in x, that x^∞ and $\exp x$ play analogous roles.

It befits an applied mathematician to internalize subjectively (5.10) and (5.11), to remember that $\ln x$ resembles x^0 and that $\exp x$ resembles x^∞. A qualitative sense that logarithms are slow and exponentials, fast, helps one to grasp mentally the essential features of many mathematical models one encounters in practice.

Now leaving aside fast and slow functions for the moment, we turn our attention in the next section to the highly important matter of the exponential of a complex argument.

5.4 Euler's formula

The result of § 5.1 leads to one of the central questions in all of mathematics. How can one evaluate

$$\exp i\theta = \lim_{\epsilon\to 0}(1+\epsilon)^{i\theta/\epsilon},$$

where $i^2 = -1$ is the imaginary unit introduced in § 2.11?

To begin, one can take advantage of (4.7) to write the last equation in the form

$$\exp i\theta = \lim_{\epsilon\to 0}(1+i\epsilon)^{\theta/\epsilon},$$

but from here it is not obvious where to go. The book's development up to the present point gives no obvious direction. In fact it appears that the interpretation of $\exp i\theta$ remains for us to define, if we can find a way to define it which fits sensibly with our existing notions of the real exponential. So, if we don't quite know where to go with this yet, what do we know?

One thing we know is that if $\theta = \epsilon$, then

$$\exp(i\epsilon) = (1+i\epsilon)^{\epsilon/\epsilon} = 1 + i\epsilon.$$

But per § 5.1, the essential operation of the exponential function is to multiply repeatedly by some factor, the factor being not quite exactly unity and, in this case, being $1 + i\epsilon$. With such thoughts in mind, let us multiply a complex number $z = x + iy$ by $1 + i\epsilon$, obtaining

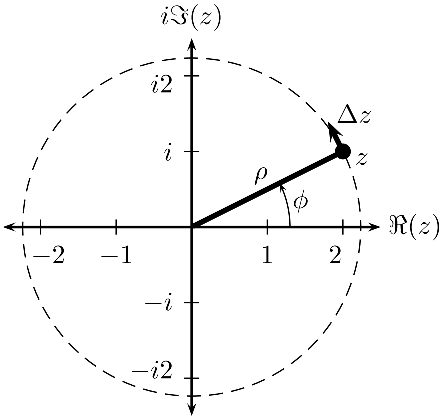
$$(1+i\epsilon)(x+iy) = (x-\epsilon y) + i(y+\epsilon x).$$

The resulting change in z is[8]

$$\Delta z = (1+i\epsilon)(x+iy) - (x+iy) = (\epsilon)(-y+ix),$$

[8] As the context implies, the notation Δz means "the change in z." We have briefly met such notation already in § 2.7.

Figure 5.3: The complex exponential and Euler's formula.



in which it is seen that

$$\begin{aligned} |\Delta z| &= (\epsilon)\sqrt{y^2+x^2} &&= \epsilon\rho, \\ \arg(\Delta z) &= \arctan\frac{x}{-y} &&= \phi + \frac{2\pi}{4}. \end{aligned}$$

The Δz, $\rho = |z|$ and $\phi = \arg z$ are as shown in Fig. 5.3. Whether in the figure or in the equations, *the change Δz is evidently proportional to the magnitude of z, but at a right angle to z's radial arm in the Argand plane.*

To travel about a circle wants motion always perpendicular to the circle's radial arm, which happens to be just the kind of motion Δz represents. Referring to the figure and the last equations, we have then that

$$\begin{aligned} \Delta\rho &\equiv |z+\Delta z| - |z| &&= 0, \\ \Delta\phi &\equiv \arg(z+\Delta z) - \arg z = \frac{|\Delta z|}{\rho} = \frac{\epsilon\rho}{\rho} &&= \epsilon, \end{aligned}$$

which results evidently are valid for infinitesimal $\epsilon \to 0$ and, importantly, stand independently of the value of ρ. (But does ρ not grow at least a little, as the last equations almost seem to suggest? The answer is no; or, if you prefer, the answer is that $\Delta\rho \approx \{[\sqrt{1+\epsilon^2}] - 1\}\rho \approx \epsilon^2\rho/2 \approx 0$, a second-order infinitesimal inconsequential on the scale of $\epsilon\rho$, utterly vanishing by comparison in the limit $\epsilon \to 0$. Remember that Δz has a *phase,* a direction in the Argand plane; and that, as the figure depicts, this phase points at a right angle to the phase of z. In mathematical symbols, $\arg[\Delta z] - \arg z = 2\pi/4$. Now, if the difference between these phases were greater than $2\pi/4$, that would mean that Δz pointed inward, which would cause $|z + \Delta z| < |z|$, wouldn't it? And if the difference were less than $2\pi/4$, that would mean that Δz pointed outward, which would cause $|z+\Delta z| > |z|$. So, what phase differential exactly causes $|z + \Delta z| = |z|$? Where indeed is the boundary between the inward and outward domains? Answer: $2\pi/4$. Such are the paradoxes of calculus!) With such results in hand, now let us recall from earlier in the section that—as we have asserted or defined—

$$\exp i\theta = \lim_{\epsilon\to 0}(1 + i\epsilon)^{\theta/\epsilon},$$

and that this remains so for arbitrary real θ. Yet what does such an equation do, mechanically, but to compute $\exp i\theta$ by multiplying 1 by $1+i\epsilon$ repeatedly, θ/ϵ times? The plain answer is that such an equation does precisely this and nothing else.[9] We have recently seen how each multiplication of the kind the equation suggests increments the phase ϕ by $\Delta\phi = \epsilon$ while not changing the magnitude ρ. Since the phase ϕ begins from $\arg 1 = 0$ it must become

$$\phi = \frac{\theta}{\epsilon}\epsilon = \theta$$

after θ/ϵ increments of ϵ each, while the magnitude must remain

$$\rho = 1.$$

Reversing the sequence of the last two equations and recalling that $\rho \equiv |\exp i\theta|$ and that $\phi \equiv \arg(\exp i\theta)$,

$$\begin{aligned} |\exp i\theta| &= 1, \\ \arg(\exp i\theta) &= \theta. \end{aligned}$$

Moreover, had we known that θ were just $\phi \equiv \arg(\exp i\theta)$, naturally we should have represented it by the symbol ϕ from the start. Changing $\phi \leftarrow \theta$ now, we have for real ϕ that

$$\begin{aligned} |\exp i\phi| &= 1, \\ \arg(\exp i\phi) &= \phi, \end{aligned}$$

which equations together say neither more nor less than that

$$\exp i\phi = \cos\phi + i\sin\phi = \operatorname{cis}\phi, \tag{5.12}$$

[9] See footnote 1.

where the notation cis(·) is as defined in § 3.11.

Along with the Pythagorean theorem (1.1), the fundamental theorem of calculus (7.2), Cauchy's integral formula (8.29) and Fourier's equation (18.1), eqn. (5.12) is one of the most famous results in all of mathematics. It is called *Euler's formula,*[10,11] and it opens the exponential domain fully to complex numbers, not just for the natural base e but for any base. How? Consider in light of Fig. 5.3 and (5.12) that one can express any complex number in the form

$$z = x + iy = \rho \exp i\phi.$$

If a complex base w is similarly expressed in the form

$$w = u + iv = \sigma \exp i\psi,$$

then it follows that

$$\begin{aligned} w^z &= \exp[\ln w^z] \\ &= \exp[z \ln w] \\ &= \exp[(x + iy)(i\psi + \ln \sigma)] \\ &= \exp[(x \ln \sigma - \psi y) + i(y \ln \sigma + \psi x)]. \end{aligned}$$

Since $\exp(\alpha + \beta) = e^{\alpha+\beta} = \exp \alpha \exp \beta$, the last equation is

$$w^z = \exp(x \ln \sigma - \psi y) \exp i(y \ln \sigma + \psi x), \tag{5.13}$$

where

$$\begin{aligned} x &= \rho \cos \phi, \\ y &= \rho \sin \phi, \\ \sigma &= \sqrt{u^2 + v^2}, \\ \tan \psi &= \frac{v}{u}. \end{aligned}$$

Equation (5.13) serves to raise any complex number to a complex power.

[10] For native English speakers who do not speak German, Leonhard Euler's name is pronounced as "oiler."

[11] An alternate derivation of Euler's formula (5.12)—less intuitive and requiring slightly more advanced mathematics, but briefer—constructs from Table 8.1 the Taylor series for $\exp i\phi$, $\cos \phi$ and $i \sin \phi$, then adds the latter two to show them equal to the first of the three. Such an alternate derivation lends little insight, perhaps, but at least it builds confidence that we actually knew what we were doing when we came up with the incredible (5.12).

Curious consequences of Euler's formula (5.12) include that[12]

$$\begin{aligned} e^{\pm i2\pi/4} &= \pm i; \\ e^{\pm i2\pi/2} &= -1; \\ e^{in2\pi} &= 1, \quad n \in \mathbb{Z}. \end{aligned} \tag{5.14}$$

For the natural logarithm of a complex number in light of Euler's formula, we have that

$$\ln w = \ln\left(\sigma e^{i\psi}\right) = \ln\sigma + i\psi. \tag{5.15}$$

5.5 Complex exponentials and de Moivre's theorem

Euler's formula (5.12) implies that complex numbers z_1 and z_2 can be written as

$$\begin{aligned} z_1 &= \rho_1 e^{i\phi_1}, \\ z_2 &= \rho_2 e^{i\phi_2}. \end{aligned} \tag{5.16}$$

By the basic power properties of Table 2.2, then,

$$\begin{array}{rclcl} z_1 z_2 &=& \rho_1\rho_2 e^{i(\phi_1+\phi_2)} &=& \rho_1\rho_2 \exp[i(\phi_1+\phi_2)], \\ \dfrac{z_1}{z_2} &=& \dfrac{\rho_1}{\rho_2} e^{i(\phi_1-\phi_2)} &=& \dfrac{\rho_1}{\rho_2} \exp[i(\phi_1-\phi_2)], \\ z^a &=& \rho^a e^{ia\phi} &=& \rho^a \exp[ia\phi]. \end{array} \tag{5.17}$$

This is de Moivre's theorem, introduced in § 3.11.

5.6 Complex trigonometrics

Applying Euler's formula (5.12) to $+\phi$ then to $-\phi$, we have that

$$\begin{aligned} \exp(+i\phi) &= \cos\phi + i\sin\phi, \\ \exp(-i\phi) &= \cos\phi - i\sin\phi. \end{aligned}$$

Adding the two equations and solving for $\cos\phi$ yields that

$$\cos\phi = \frac{\exp(+i\phi) + \exp(-i\phi)}{2}. \tag{5.18}$$

[12] Notes of the obvious, like $n \in \mathbb{Z}$, are sometimes omitted by this book because they clutter the page. However, the note is included in this instance.

Subtracting the second equation from the first and solving for $\sin\phi$ yields that

$$\sin\phi = \frac{\exp(+i\phi) - \exp(-i\phi)}{i2}. \tag{5.19}$$

Thus are the trigonometrics expressed in terms of complex exponentials.

5.6.1 The hyperbolic functions

The forms (5.18) and (5.19) suggest the definition of new functions

$$\begin{aligned} \cosh\phi &\equiv \frac{\exp(+\phi) + \exp(-\phi)}{2}, & (5.20) \\ \sinh\phi &\equiv \frac{\exp(+\phi) - \exp(-\phi)}{2}, & (5.21) \\ \tanh\phi &\equiv \frac{\sinh\phi}{\cosh\phi}. & (5.22) \end{aligned}$$

These are called the *hyperbolic functions.* Their inverses arccosh, etc., are defined in the obvious way. The Pythagorean theorem for trigonometrics (3.2) is that $\cos^2\phi + \sin^2\phi = 1$, verified by combining (5.18) and (5.19); and from (5.20) and (5.21) one can derive the hyperbolic analog:

$$\begin{aligned} \cos^2\phi + \sin^2\phi &= 1, \\ \cosh^2\phi - \sinh^2\phi &= 1. \end{aligned} \tag{5.23}$$

Although Fig. 5.3 has visualized only real ϕ, complex ϕ can be considered, too. Nothing prevents one from taking (5.18) through (5.21), as written, *to define* the trigonometrics of complex ϕ; so that's what we now do. From this it follows that (5.23) and others must likewise hold[13] for complex ϕ.

[13] Chapter 15 teaches that the *dot product* of a unit vector and its own conjugate is unity—$\hat{\mathbf{v}}^* \cdot \hat{\mathbf{v}} = 1$, in the notation of that chapter—which tempts one to suppose incorrectly by analogy that $(\cos\phi)^*\cos\phi + (\sin\phi)^*\sin\phi = 1$ and that $(\cosh\phi)^*\cosh\phi - (\sinh\phi)^*\sinh\phi = 1$ when the angle ϕ is complex. However, (5.18) through (5.21) can be generally true only if (5.23) holds exactly as written for complex ϕ as well as for real. Hence in fact $(\cos\phi)^*\cos\phi + (\sin\phi)^*\sin\phi \neq 1$ and $(\cosh\phi)^*\cosh\phi - (\sinh\phi)^*\sinh\phi \neq 1$.

Fig. 3.1 is quite handy for real ϕ but what if anything the figure means when ϕ is complex is not obvious. The ϕ of the figure cannot quite be understood to mean an actual direction or bearing in the east-north-west-south sense. Therefore, visual analogies between geometrical vectors like $\hat{\mathbf{v}}$, on the one hand, and Argand-plotted complex numbers, on the other, can analytically fail, especially in circumstances in which ϕ may be complex. (The professional mathematician might smile at this, gently prodding us that this is why one should rely on analysis rather than on mere geometrical intuition. If so, then we would acknowledge the prod [33] without further comment in this instance.)

The notation $\exp i(\cdot)$ or $e^{i(\cdot)}$ is sometimes felt to be too bulky. Although less commonly seen than the other two, the notation

$$\operatorname{cis}(\cdot) \equiv \exp i(\cdot) = \cos(\cdot) + i\sin(\cdot)$$

is also conventionally recognized, as earlier seen in § 3.11. Also conventionally recognized are $\sin^{-1}(\cdot)$ and occasionally $\operatorname{asin}(\cdot)$ for $\arcsin(\cdot)$, and likewise for the several other trigs.

Replacing $z \leftarrow \phi$ in this section's several equations implies a coherent definition for trigonometric functions of a complex variable. Then, comparing (5.18) and (5.19) respectively to (5.20) and (5.21), we have that

$$\begin{aligned} \cosh z &= \cos iz, \\ i\sinh z &= \sin iz, \\ i\tanh z &= \tan iz, \end{aligned} \tag{5.24}$$

by which one can immediately adapt the many trigonometric properties of Tables 3.1 and 3.3 to hyperbolic use.

At this point in the development one begins to notice that the cos, sin, exp, cis, cosh and sinh functions are each really just different facets of the same mathematical phenomenon. Likewise their respective inverses: arccos, arcsin, ln, $-i\ln$, arccosh and arcsinh. Conventional names for these two mutually inverse families of functions are unknown to the author, but one might call them the *natural exponential* and *natural logarithmic families.* Or, if the various tangent functions were included, then one might call them the *trigonometric* and *inverse trigonometric families.*

5.6.2 Inverse complex trigonometrics

Since one can express the several trigonometric functions in terms of complex exponentials one would like to know, complementarily, whether one cannot express the several inverse trigonometric functions in terms of complex logarithms. As it happens, one can.[14]

Let us consider the arccosine function, for instance. If per (5.18)

$$z = \cos w = \frac{e^{iw} + e^{-iw}}{2},$$

[14] [153, chapter 2]

then by successive steps

$$\begin{aligned} e^{iw} &= 2z - e^{-iw}, \\ \left[e^{iw}\right]^2 &= \left[e^{iw}\right]\left[2z - e^{-iw}\right] = 2z\left(e^{iw}\right) - 1, \\ e^{iw} &= z \pm \sqrt{z^2-1}, \end{aligned}$$

the last step of which has used the quadratic formula (2.2). Taking the logarithm, we have that

$$w = \frac{1}{i}\ln\left(z \pm i\sqrt{1-z^2}\right);$$

or, since by definition $z = \cos w$, that

$$\arccos z = \frac{1}{i}\ln\left(z \pm i\sqrt{1-z^2}\right). \tag{5.25}$$

Similarly,

$$\arcsin z = \frac{1}{i}\ln\left(iz \pm \sqrt{1-z^2}\right). \tag{5.26}$$

The arctangent goes only a little differently:

$$\begin{aligned} z = \tan w &= -i\frac{e^{iw} - e^{-iw}}{e^{iw} + e^{-iw}}, \\ ze^{iw} + ze^{-iw} &= -ie^{iw} + ie^{-iw}, \\ (i+z)e^{iw} &= (i-z)e^{-iw}, \\ e^{i2w} &= \frac{i-z}{i+z}, \end{aligned}$$

implying that

$$\arctan z = \frac{1}{i2}\ln\frac{i-z}{i+z}. \tag{5.27}$$

By the same means, one can work out the inverse hyperbolics to be

$$\begin{aligned} \operatorname{arccosh} z &= \ln\left(z \pm \sqrt{z^2-1}\right), \\ \operatorname{arcsinh} z &= \ln\left(z \pm \sqrt{z^2+1}\right), \\ \operatorname{arctanh} z &= \frac{1}{2}\ln\frac{1+z}{1-z}. \end{aligned} \tag{5.28}$$

5.7 Summary of properties

Table 5.1 gathers properties of the complex exponential from this chapter and from §§ 2.11, 3.11 and 4.4.

Table 5.1: Complex exponential properties.

$$
\begin{aligned}
i^2 &= -1 = (-i)^2 \\
\frac{1}{i} &= -i \\
e^{i\phi} &= \cos\phi + i\sin\phi \\
e^{iz} &= \cos z + i\sin z \\
z_1 z_2 &= \rho_1\rho_2 e^{i(\phi_1+\phi_2)} = (x_1x_2 - y_1y_2) + i(y_1x_2 + x_1y_2) \\
\frac{z_1}{z_2} &= \frac{\rho_1}{\rho_2} e^{i(\phi_1-\phi_2)} = \frac{(x_1x_2 + y_1y_2) + i(y_1x_2 - x_1y_2)}{x_2^2 + y_2^2} \\
z^a &= \rho^a e^{ia\phi} \\
w^z &= e^{x\ln\sigma - \psi y} e^{i(y\ln\sigma + \psi x)} \\
\ln w &= \ln\sigma + i\psi
\end{aligned}
$$

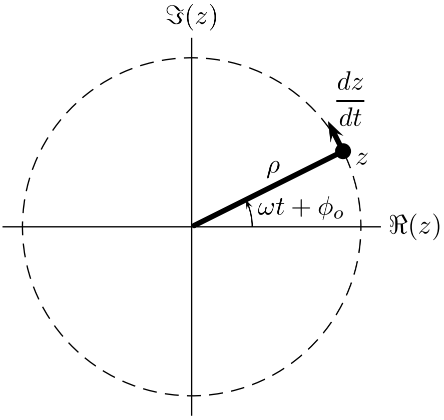
$$
\begin{aligned}
\sin z &= \frac{e^{iz} - e^{-iz}}{i2} & \sin iz &= i\sinh z & \sinh z &= \frac{e^{z} - e^{-z}}{2} \\
\cos z &= \frac{e^{iz} + e^{-iz}}{2} & \cos iz &= \cosh z & \cosh z &= \frac{e^{z} + e^{-z}}{2} \\
\tan z &= \frac{\sin z}{\cos z} & \tan iz &= i\tanh z & \tanh z &= \frac{\sinh z}{\cosh z}
\end{aligned}
$$

$$
\begin{aligned}
\arcsin z &= \frac{1}{i}\ln\left(iz \pm \sqrt{1-z^2}\right) & \operatorname{arcsinh} z &= \ln\left(z \pm \sqrt{z^2+1}\right) \\
\arccos z &= \frac{1}{i}\ln\left(z \pm i\sqrt{1-z^2}\right) & \operatorname{arccosh} z &= \ln\left(z \pm \sqrt{z^2-1}\right) \\
\arctan z &= \frac{1}{i2}\ln\frac{i-z}{i+z} & \operatorname{arctanh} z &= \frac{1}{2}\ln\frac{1+z}{1-z}
\end{aligned}
$$

$$\cos^2 z + \sin^2 z = 1 = \cosh^2 z - \sinh^2 z$$

$$
\begin{aligned}
z &\equiv x + iy = \rho e^{i\phi} & \frac{d}{dz}\exp z &= \exp z \\
w &\equiv u + iv = \sigma e^{i\psi} & \frac{d}{dw}\ln w &= \frac{1}{w} \\
\exp z &\equiv e^z & \frac{df/dz}{f(z)} &= \frac{d}{dz}\ln f(z) \\
\operatorname{cis} z &\equiv \cos z + i\sin z = e^{iz} & \log_b w &= \frac{\ln w}{\ln b}
\end{aligned}
$$

Figure 5.4: The derivatives of the sine and cosine functions.



5.8 Derivatives of complex exponentials

This section computes the derivatives of the various trigonometric and inverse trigonometric functions.

5.8.1 Derivatives of sine and cosine

One could compute derivatives of the sine and cosine functions from (5.18) and (5.19). To do so is left as an exercise. Meanwhile, however, another, more sporting way to find the derivatives is known: one can directly examine the circle from which the sine and cosine functions come.

Refer to Fig. 5.4. Suppose that the point z in the figure is not fixed but travels steadily about the circle such that[15]

$$z(t) = (\rho)\left[\cos(\omega t + \phi_o) + i\sin(\omega t + \phi_o)\right]. \tag{5.29}$$

How fast then is the rate dz/dt, and in what Argand direction? Answer:

$$\frac{dz}{dt} = (\rho)\left[\frac{d}{dt}\cos(\omega t + \phi_o) + i\frac{d}{dt}\sin(\omega t + \phi_o)\right]. \tag{5.30}$$

Evidently however, considering the figure,

[15] Observe the Greek letter ω, omega, which is not a Roman w. Refer to appendix B.

- the speed $|dz/dt|$ is also $(\rho)(d\phi/dt) = \rho\omega$;
- the direction is at right angles to the arm of ρ, which is to say that $\arg(dz/dt) = \phi + 2\pi/4$.

With these observations we can write that

$$\begin{aligned} \frac{dz}{dt} &= (\rho\omega)\left[\cos\left(\omega t + \phi_o + \frac{2\pi}{4}\right) + i\sin\left(\omega t + \phi_o + \frac{2\pi}{4}\right)\right] \\ &= (\rho\omega)\left[-\sin(\omega t + \phi_o) + i\cos(\omega t + \phi_o)\right]. \end{aligned} \tag{5.31}$$

Matching the real and imaginary parts of (5.30) against those of (5.31), we have that

$$\begin{aligned} \frac{d}{dt}\cos(\omega t + \phi_o) &= -\omega\sin(\omega t + \phi_o), \\ \frac{d}{dt}\sin(\omega t + \phi_o) &= +\omega\cos(\omega t + \phi_o). \end{aligned} \tag{5.32}$$

If $\omega = 1$ and $\phi_o = 0$, these are that

$$\begin{aligned} \frac{d}{dt}\cos t &= -\sin t, \\ \frac{d}{dt}\sin t &= +\cos t. \end{aligned} \tag{5.33}$$

5.8.2 Derivatives of the trigonometrics

Equations (5.4) and (5.33) give the derivatives of $\exp(\cdot)$, $\sin(\cdot)$ and $\cos(\cdot)$. From these, with the help of (5.23) and the derivative chain and product rules (§ 4.5), we can calculate the several derivatives of Table 5.2.[16]

5.8.3 Derivatives of the inverse trigonometrics

Observe the pair

$$\begin{aligned} \frac{d}{dz}\exp z &= \exp z, \\ \frac{d}{dw}\ln w &= \frac{1}{w}. \end{aligned}$$

The natural exponential $\exp z$ belongs to the trigonometric family of functions, as does its derivative. The natural logarithm $\ln w$, by contrast, belongs

[16] [146, back endpaper]

Table 5.2: Derivatives of the trigonometrics.

$$\begin{aligned}
\frac{d}{dz}\exp z &= +\exp z & \qquad \frac{d}{dz}\frac{1}{\exp z} &= -\frac{1}{\exp z} \\
\frac{d}{dz}\sin z &= +\cos z & \qquad \frac{d}{dz}\frac{1}{\sin z} &= -\frac{1}{\tan z\sin z} \\
\frac{d}{dz}\cos z &= -\sin z & \qquad \frac{d}{dz}\frac{1}{\cos z} &= +\frac{\tan z}{\cos z}
\end{aligned}$$

$$\begin{aligned}
\frac{d}{dz}\tan z &= +\left(1+\tan^2 z\right) &&= +\frac{1}{\cos^2 z} \\
\frac{d}{dz}\frac{1}{\tan z} &= -\left(1+\frac{1}{\tan^2 z}\right) &&= -\frac{1}{\sin^2 z}
\end{aligned}$$

$$\begin{aligned}
\frac{d}{dz}\sinh z &= +\cosh z & \qquad \frac{d}{dz}\frac{1}{\sinh z} &= -\frac{1}{\tanh z\sinh z} \\
\frac{d}{dz}\cosh z &= +\sinh z & \qquad \frac{d}{dz}\frac{1}{\cosh z} &= -\frac{\tanh z}{\cosh z}
\end{aligned}$$

$$\begin{aligned}
\frac{d}{dz}\tanh z &= 1-\tanh^2 z &&= +\frac{1}{\cosh^2 z} \\
\frac{d}{dz}\frac{1}{\tanh z} &= 1-\frac{1}{\tanh^2 z} &&= -\frac{1}{\sinh^2 z}
\end{aligned}$$

to the inverse trigonometric family of functions; but its derivative is simpler, not a trigonometric or inverse trigonometric function at all. In Table 5.2, one notices that all the trigonometrics have trigonometric derivatives. By analogy with the natural logarithm, do all the inverse trigonometrics have simpler derivatives?

It turns out that they do. Refer to the account of the natural logarithm's derivative in § 5.2. Following a similar procedure, we have by successive steps that

$$\begin{aligned}
\arcsin w &= z, \\
w &= \sin z, \\
\frac{dw}{dz} &= \cos z, \\
\frac{dw}{dz} &= \pm\sqrt{1-\sin^2 z}, \\
\frac{dw}{dz} &= \pm\sqrt{1-w^2}, \\
\frac{dz}{dw} &= \frac{\pm 1}{\sqrt{1-w^2}}, \\
\frac{d}{dw}\arcsin w &= \frac{\pm 1}{\sqrt{1-w^2}}. \qquad (5.34)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\arctan w &= z, \\
w &= \tan z, \\
\frac{dw}{dz} &= 1+\tan^2 z, \\
\frac{dw}{dz} &= 1+w^2, \\
\frac{dz}{dw} &= \frac{1}{1+w^2}, \\
\frac{d}{dw}\arctan w &= \frac{1}{1+w^2}. \qquad (5.35)
\end{aligned}$$

Derivatives of the other inverse trigonometrics are found in the same way. Table 5.3 summarizes.

Table 5.3 may prove useful when the integration technique of § 9.1 is applied.

Table 5.3: Derivatives of the inverse trigonometrics.

$$\begin{aligned}
\frac{d}{dw}\ln w &= \frac{1}{w} \\
\frac{d}{dw}\arcsin w &= \frac{\pm 1}{\sqrt{1-w^2}} \\
\frac{d}{dw}\arccos w &= \frac{\mp 1}{\sqrt{1-w^2}} \\
\frac{d}{dw}\arctan w &= \frac{1}{1+w^2} \\
\frac{d}{dw}\operatorname{arcsinh} w &= \frac{\pm 1}{\sqrt{w^2+1}} \\
\frac{d}{dw}\operatorname{arccosh} w &= \frac{\pm 1}{\sqrt{w^2-1}} \\
\frac{d}{dw}\operatorname{arctanh} w &= \frac{1}{1-w^2}
\end{aligned}$$

5.9 The actuality of complex quantities

Doing all this theoretically interesting complex mathematics, the applied mathematician can lose sight of some questions he probably ought to keep in mind: do complex quantities arise in nature? If they do not, then what physical systems do we mean to model with them? Hadn't we better avoid these complex quantities, leaving them to the professional mathematical theorists?

As developed by Oliver Heaviside in 1887,[17] the answer depends on your point of view. If I have 300 g of grapes and 100 g of grapes, then I have 400 g altogether. Alternately, if I have 500 g of grapes and -100 g of grapes, again I have 400 g altogether. (What does it mean to have -100 g of grapes? Maybe that I ate some!) But what if I have $200 + i100$ g of grapes and $200 - i100$ g of grapes? Answer: again, 400 g.

Probably you would not choose to think of $200 + i100$ g of grapes and $200 - i100$ g of grapes, but because of (5.18) and (5.19), one often describes wave phenomena as linear superpositions (sums) of countervailing

[17] [122]

complex exponentials. Consider for instance the propagating wave

$$A\cos[\omega t - kz] = \frac{A}{2}\exp[+i(\omega t - kz)] + \frac{A}{2}\exp[-i(\omega t - kz)].$$

The benefit of splitting the real cosine into two complex parts is that, while the magnitude of the cosine changes with time t, the magnitude of either exponential alone remains steady (see the circle in Fig. 5.3). It turns out to be easier to analyze two complex wave quantities of constant magnitude than to analyze one real wave quantity of varying magnitude. Better yet, since each complex wave quantity is the complex conjugate of the other, the analyses thereof are mutually conjugate, too (§ 2.11.2); so one normally need not actually analyze the second. The one analysis suffices for both.[18]

Some authors have gently denigrated the use of imaginary parts in physical applications as a mere mathematical trick, as though the parts were not actually there.[19] Well, that is one way to treat the matter, but it is not the way this book recommends. Nothing in the mathematics *requires* you to regard the imaginary parts as physically nonexistent. One need not abuse Ockham's razor! (Ockham's razor, "Do not multiply objects without necessity,"[20] is a sound philosophical indicator when properly used. However, the razor is overused in some circles, particularly in circles in which Aristotle[21] is believed—mistakenly, in this writer's view—to be vaguely outdated; or

[18] If the point is not immediately clear, an example: suppose that by the Newton-Raphson iteration (§ 4.8) you have found a root of the polynomial $x^3 + 2x^2 + 3x + 4$ at $x \approx -\mathrm{0x0.2D} + i\mathrm{0x1.8C}$. Where is there another root? Answer: there is a conjugate root at $x \approx -\mathrm{0x0.2D} - i\mathrm{0x1.8C}$. Because the polynomial's coefficients are real, one need not actually run the Newton-Raphson again to find the conjugate root.

Another example, this time with a wave: suppose that, when fed by a time-varying electric current of $(5.0\ \text{milliamps})\exp\{+i(60\ \text{sec}^{-1})2\pi t\}$, an electric capacitor develops a voltage—that is, develops an electric tension or potential—of $(40\ \text{volts})\exp\{+i[(60\ \text{sec}^{-1})2\pi t - 2\pi/4]\}$. It immediately follows, without further analysis, that the same capacitor, if fed by a time-varying electric current of $(5.0\ \text{milliamps})\exp\{-i(60\ \text{sec}^{-1})2\pi t\}$, would develop a voltage of $(40\ \text{volts})\exp\{-i[(60\ \text{sec}^{-1})2\pi t - 2\pi/4]\}$. The conjugate current gives rise to a conjugate voltage.

The reason to analyze an electric circuit in such a way is that, after analyzing it, one can sum the two complex currents to get a real a.c. current like the current an electric wall receptacle supplies. If one does this, then one can likewise sum the two complex voltages to compute the voltage the capacitor would develop. Indeed, this is how electrical engineers normally analyze a.c. systems (well, electrical engineers know some shortcuts, but this is the idea), because $\exp(\cdot)$ is so much easier a function to handle than $\cos(\cdot)$ or $\sin(\cdot)$ is.

[19] One who gently denigrates the use can nevertheless still apply the trick! They often do.

[20] [158, chapter 12]

[21] [55]

more likely in circles in which Aristotle has been altogether forgotten. More often than one likes to believe, the necessity to multiply objects remains hidden until one has ventured the multiplication, nor reveals itself to the one who wields the razor, whose hand humility should stay; for even Immanuel Kant—no Aristotelean he!—has cautioned, "*entium varietates non temere esse minuendas,*"[22] "the variety of beings ought not rashly be diminished.") It is true by Euler's formula (5.12) that a complex exponential $\exp i\phi$ can be decomposed into a sum of trigonometrics. However, it is equally true by the complex trigonometric formulas (5.18) and (5.19) that *a trigonometric can be decomposed into a sum of complex exponentials.* So, if each can be decomposed into the other, then which of the two is the true decomposition? Answer: that depends on your point of view. Experience seems to recommend viewing the complex exponential as the basic element—as the element of which the trigonometrics are composed—rather than the other way around. From this point of view, it is (5.18) and (5.19) which are the true decomposition. Euler's formula (5.12) itself could be viewed in this sense as secondary.

The complex exponential method of offsetting imaginary parts offers an elegant yet practical mathematical means to model physical wave phenomena. It may find other uses, too, so go ahead: regard the imaginary parts as actual. Aristotle would regard them so (or so the writer suspects). To regard the imaginary parts as actual hurts nothing, and it helps with the math.

[22] From Kant's [94, book III, chapter II, appendix]. Kant wrote in German but wrote this particular phrase in Latin, which is why it is printed in Latin here.

Chapter 6

Primes, roots and averages

This chapter gathers a few significant topics, each of whose treatment seems too brief (or in § 6.5 too idiosyncratic) for a chapter of its own.

6.1 Prime numbers

A *prime number*—or simply, a *prime*—is an integer greater than one, divisible only by one and itself. A *composite number* is an integer greater than one and not prime. A composite number can be composed as a product of two or more prime numbers. All positive integers greater than one are either composite or prime.

The mathematical study of prime numbers and their incidents constitutes *number theory,* and it is a deep area of mathematics. The deeper results of number theory seldom arise in applications,[1] however, so except in § 6.5 we will confine our study of number theory in this book to a handful of its simplest, most broadly interesting results.

6.1.1 The infinite supply of primes

The first primes are evidently $2, 3, 5, 7, \mathrm{0xB}, \ldots$. Is there a last prime?

To show that there is no last prime, suppose that there were. More precisely, suppose that there existed exactly N primes, with N finite, letting $p_1, p_2, \ldots, p_N$ represent these primes from least to greatest. Now consider

[1] The deeper results of number theory do arise in cryptography, or so the author has been led to understand. Although cryptography is literally an application of mathematics, its spirit is that of pure mathematics rather than of applied. If you seek cryptographic derivations, this book is probably not the one you want.

the product of all the primes,

$$C = \prod_{j=1}^{N} p_j.$$

What of $C + 1$? Since $p_1 = 2$ divides C, it cannot divide $C + 1$. Similarly, since $p_2 = 3$ divides C, it also cannot divide $C + 1$. The same goes for $p_3 = 5$, $p_4 = 7$, $p_5 = \mathrm{0xB}$, etc. Apparently none of the primes in the p_j series divides $C + 1$, which implies either that $C + 1$ itself is prime, or that $C + 1$ is composed of primes not in the series. But the latter is assumed impossible on the ground that the p_j series includes all primes; and the former is assumed impossible on the ground that $C + 1 > C > p_N$, with p_N the greatest prime. The contradiction proves false the assumption that gave rise to it. The false assumption: that there were a last prime.

Thus there is no last prime. No matter how great a prime number one finds, a greater can always be found. The supply of primes is infinite.[2]

Attributed to the ancient geometer Euclid, the foregoing proof is a classic example of mathematical *reductio ad absurdum,* or as usually styled in English, *proof by contradiction.*[3]

6.1.2 Compositional uniqueness

Occasionally in mathematics, plausible assumptions can hide subtle logical flaws. One such plausible assumption is the assumption that every positive integer has a unique *prime factorization.* It is readily seen that the first several positive integers—$1 = ()$, $2 = (2^1)$, $3 = (3^1)$, $4 = (2^2)$, $5 = (5^1)$, $6 = (2^1)(3^1)$, $7 = (7^1)$, $8 = (2^3)$, ...—each have unique prime factorizations, but is this necessarily true of all positive integers?

To show that it is true, suppose that it were not.[4] More precisely, suppose that there did exist positive integers factorable each in two or more distinct ways, with the symbol C representing the least such integer. Noting that C must be composite (prime numbers by definition are each factorable

[2] [155]

[3] [141, appendix 1][182, "Reductio ad absurdum," 02:36, 28 April 2006]

[4] Unfortunately the author knows no more elegant proof than this, yet cannot even cite this one properly. The author encountered the proof in some book in the library of Hayward State University, California, during the 1990s. The identity of that book is now long forgotten.

only one way, like $5 = [5^1]$), let

$$\begin{aligned}
C_p &\equiv \prod_{j=1}^{N_p} p_j, \\
C_q &\equiv \prod_{k=1}^{N_q} q_k, \\
C_p = C_q &= C, \\
p_j &\le p_{j+1}, \\
q_k &\le q_{k+1}, \\
p_1 &\le q_1, \\
N_p &> 1, \\
N_q &> 1,
\end{aligned}$$

where C_p and C_q represent two distinct prime factorizations of the same number C and where the p_j and q_k are the respective primes ordered from least to greatest. We see that

$$p_j \ne q_k$$

for any j and k—that is, that the same prime cannot appear in both factorizations—because if the same prime r did appear in both then C/r either would be prime (in which case both factorizations would be $[r][C/r]$, defying our assumption that the two differed) or would constitute an ambiguously factorable composite integer less than C when we had already defined C to represent the least such. Among other effects, the fact that $p_j \ne q_k$ strengthens the definition $p_1 \le q_1$ to read

$$p_1 < q_1.$$

Let us now rewrite the two factorizations in the form

$$\begin{aligned}
C_p &= p_1 A_p, \\
C_q &= q_1 A_q, \\
C_p = C_q &= C, \\
A_p &\equiv \prod_{j=2}^{N_p} p_j, \\
A_q &\equiv \prod_{k=2}^{N_q} q_k,
\end{aligned}$$

where p_1 and q_1 are the least primes in their respective factorizations. Since C is composite and since $p_1 < q_1$, we have that

$$1 < p_1 < q_1 \leq \sqrt{C} \leq A_q < A_p < C,$$

which implies that

$$p_1 q_1 < C.$$

The last inequality lets us compose the new positive integer

$$B = C - p_1 q_1,$$

which might be prime or composite (or unity), but which either way enjoys a unique prime factorization because $B < C$, with C the least positive integer factorable two ways. Observing that some integer s which divides C necessarily also divides $C \pm ns$, we note that each of p_1 and q_1 necessarily divides B. This means that B's unique factorization includes both p_1 and q_1, which further means that the product $p_1 q_1$ divides B. But if $p_1 q_1$ divides B, then it divides $B + p_1 q_1 = C$, also.

Let E represent the positive integer which results from dividing C by $p_1 q_1$:

$$E \equiv \frac{C}{p_1 q_1}.$$

Then,

$$\begin{aligned} Eq_1 &= \frac{C}{p_1} = A_p, \\ Ep_1 &= \frac{C}{q_1} = A_q. \end{aligned}$$

That $Eq_1 = A_p$ says that q_1 divides A_p. But $A_p < C$, so A_p's prime factorization is unique—and we see above that A_p's factorization *does not include any q_k,* not even q_1. The contradiction proves false the assumption that gave rise to it. The false assumption: that there existed a least composite number C prime-factorable in two distinct ways.

Thus no positive integer is ambiguously factorable. Prime factorizations are always unique.

This finding has by some been called[5] the *fundamental theorem of arithmetic.*

[5] The source [177] omits the proof but does list the name. It is not a name a scientist or engineer is likely often to encounter.

We have observed at the start of this subsection that plausible assumptions can hide subtle logical flaws. Indeed this is so. Interestingly however, the plausible assumption of the present subsection has turned out absolutely correct; we have just had to do some extra work to prove it. Such effects are typical on the shadowed frontier on which applied shades into pure mathematics: with sufficient experience and with a firm grasp of the model at hand, if you think that it's true, then it probably is. Judging when to delve into the mathematics anyway, seeking a more rigorous demonstration of a proposition one feels pretty sure is correct, is a matter of applied mathematical style. It depends on how sure one feels, and more importantly on whether the unsureness felt is true uncertainty or is just an unaccountable desire for more precise mathematical definition (if the latter, then unlike the author you may have the right temperament to become a professional mathematician). The author does judge the present subsection's proof to be worth the applied effort; but nevertheless, when one lets logical minutiae distract one to too great a degree, one admittedly begins to drift out of the applied mathematical realm that is the subject of this book.

6.1.3 Determination

The least composite divisible by no prime smaller[6] than p is evidently p^2. Therefore, every composite less than p^2 has at least one prime factor smaller than p. It follows that, *to determine whether an integer n within the domain $1 < n < p^2$ is prime, it suffices to try dividing n by each prime smaller than p.*

For example, let $n = \mathrm{0x77}$. Because $n < \mathrm{0xB}^2$, to try dividing n by each of 2, 3, 5 and 7 suffices. The 2, 3 and 5 are found not to divide n but the 7 is indeed found to divide n, so $n = \mathrm{0x77}$ is composite. In fact, $\mathrm{0x77} = (7)(\mathrm{0x77}/7) = (7)(\mathrm{0x11})$.

For another example, let $n = \mathrm{0x71}$. Again, because $n < \mathrm{0xB}^2$, to try dividing n by each of 2, 3, 5 and 7 suffices. All four trials fail, so $n = \mathrm{0x71}$ is prime.

For a third example, let $n = \mathrm{0x7F}$. This time, $n \geq \mathrm{0xB}^2$. However, $n < \mathrm{0xD}^2$, so to try dividing n by each of 2, 3, 5, 7 and 0xB suffices. All *five* trials fail, so $n = \mathrm{0x7F}$ is prime.[7]

[6] The phrasing is accurate but wants close reading: "divisible by *no prime smaller* than p." Divisibility by p and by larger primes is not by this particular phrase considered.

Divisibility by p and by larger primes is indeed considered by the rest of the sentence, though, so if the sentence's truth is not obvious then answer the following, listed questions.

1. What is the least composite divisible by 0xB? (Answer: [2][0xB] = 0x16.)
2. What is the least composite divisible by 0xB but not divisible by 2? (Answer: [3][0xB] = 0x21.)
3. What is the least composite divisible by 0xB but divisible neither by 2 nor by 3? (Answer: [5][0xB] = 0x37.)
4. What is the least composite divisible by 0xB but divisible neither by 2 nor by 3 nor by 5? (Answer: [7][0xB] = 0x4D.)
5. What is the least composite divisible by 0xB but divisible neither by 2 nor by 3 nor by 5 nor by 7? (Answer: [0xB][0xB] = 0x79.)
6. What is the least composite divisible by **0xD** but divisible neither by 2 nor by 3 nor by 5 nor by 7? (Answer: [0xB][0xD] = 0x8F.)
7. What is the least composite divisible by **0x11** but divisible neither by 2 nor by 3 nor by 5 nor by 7? (Answer: [0xB][0x11] = 0xBB.)
8. (And so on. . .)

Since every composite has at least two prime factors, exhaustion of possibilities along the listed questions' lines leads unavoidably to the conclusion that no composite less than $\mathrm{0xB}^2 = \mathrm{0x79}$ can be constructed except by use of at least one prime factor less than 0xB.

The foregoing series of questions hinges on 0xB but a similar series hinging on any prime p could equally well be asked.

[7] The reader might recall the schoolboy's trick to determine whether 3 divides a decimal

6.1.4 Rational and irrational numbers

A *rational number* is a finite real number expressible as a ratio of integers[8,9]

$$x = \frac{p}{q}, \quad (p, q) \in \mathbb{Z}, \; q > 0.$$

The ratio is *irreducible* or *fully reduced* if p and q have no prime factors in common. For instance, 4/6 is not fully reduced, whereas 2/3 is.

An *irrational number* is a finite real number which is not rational. For example, $\sqrt{2}$ is irrational. In fact any $x = \sqrt{n}$ is irrational unless integral; there is no such thing as a $\sqrt{n}$ which is not an integer but is rational.

To prove[10] the last point, suppose that there did exist a fully reduced

$$x = \frac{p}{q} = \sqrt{n}, \quad (n, p, q) \in \mathbb{Z}, \; n > 0, \; p > 0, \; q > 1.$$

Squaring the equation, we have that

$$\frac{p^2}{q^2} = n,$$

which form is evidently also fully reduced. But if $q > 1$, then the fully reduced $n = p^2/q^2$ is not an integer as we had assumed that it was. The contradiction proves false the assumption which gave rise to it. Hence there

numeral like, say 1581: $1 + 5 + 8 + 1 = 15$; 3 divides 15, so 3 divides 1581, too. The same trick works in hexadecimal, also; and, moreover, the trick works not only for 3 but for 5 in hex, as well. Take 0xD417, for example: $\mathrm{D} + 4 + 1 + 7 = \mathrm{0x19}$; 3 does not divide 0x19 but 5 does, so 3 does not divide 0xD417 but 5 does.

Since the trick is mentioned only in a footnote, the book leaves the proof as an exercise but here is a hint: $(\mathrm{0x10}^n - 1) \bmod 3 = 0$ and $(\mathrm{0x10}^n - 1) \bmod 5 = 0$ for any nonnegative integer n. Section 6.1.5 explains mod.

Because hexadecimal gathers bits in groups of four (appendix A), the trick does not work for 7. However, if bits are instead gathered in groups of *three* (this is called *octal* rather than hexadecimal) then the trick does indeed work for 7, and for the same reason.

[8] Section 2.3 explains the $\in \mathbb{Z}$ notation, but see also below in the present subsection.

[9] The letters p and q are used for a different purpose here, in the rest of the section, and in § 6.5 than in §§ 6.1.1 through 6.1.3 above, where they represented prime factors. Another prime factor is to appear in § 6.1.6, but there the letter h rather than the letter p will represent it. (To ease the reading, one tries to honor established English-language mathematical, scientific or engineering convention in the choice of letters; but there are only so many letters in the alphabet and, sometimes, conventions contradict. The writer is unaware of any convention that represents a prime factor by the letter h, but the present section and § 6.5 have already claimed all plausibly conventional alternate letters for conflicting purposes. The letter h being available, it must serve!)

[10] A proof somewhat like the one presented here is found in [141, appendix 1].

exists no rational, nonintegral $\sqrt{n}$, as was to be demonstrated. The proof is readily extended to show that any $x = n^{j/k}$ is irrational if nonintegral, the extension by writing that $p^k/q^k = n^j$ and then following similar steps as those this paragraph outlines.

For information, conventional mathematical notation broadly recognized even by scientists and engineers takes[11,12]

- $\mathbb{Z}$ to represent the integers,
- $\mathbb{Q}$ to represent the rational numbers (which include the integers),
- $\mathbb{R}$ to represent the real numbers (which include the rationals), and
- $\mathbb{C}$ to represent the complex numbers (which include the reals).

Thus, for example, that $n \in \mathbb{Q}$ states that n is a rational number. The four sets nest as

$$\mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C},$$

in which the operator $\subset$ is conventionally used instead of the operator $\in$ because it is the members of the set to the operator's left, rather than the set itself, that belong to the set to the operator's right.[13] Of the four

[11] [26, Introduction][27, chapters 5 through 7]

[12] Many books including [26] print **Z**, **Q**, **R** and **C**–or even, as [147] does, merely Z, Q, R and C—rather than $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$. However, the typeface aside, the same four letters are used by everyone as far as the author knows. The letters stand respectively—again as far as the author knows—for the German *Zahl* (or *zählen*), *Quotient* and *reele Zahl* and the English *complex number* or French *nombre complexe.* (The author hesitates to cite a source, for the sources he has read either conflict in certain details, or omit to state *their* pertinent sources, or else seem less than wholly reliable in adjacent matters. In English, the symbols seem often to have been treated as common mathematical lore, or quasi-lore, like the $\sum$ sign, not quite requiring citation. The symbology's proper, published source is presumably written in 19th-century German but this writer is unacquainted with it and has been unable to follow a chain of citations from his bookshelf back to it.)

Often listed alongside $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ is also $\mathbb{N}$, which comprises the positive integers and, depending on the source, often zero, as well. However, other than to mention $\mathbb{N}$ in this footnote the book does not use the symbol. Further listed in some sources can be $\mathbb{A}$ for the "algebraic numbers" (which, lacking a strong physical application as a set within the limits of your author's experience, lie beyond this book's scope), perhaps among other such capitals.

Whether ∞ belongs to any or all of the four sets $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ or, indeed, is a number at all is a definitional question the prudent applied mathematician will defer until a concrete need to answer arises, as it does for instance in [147, § 1.9].

[13] If the reader thinks the distinction between $\in$ and $\subset$ too fine for physical applications, the writer is inclined to agree. Notwithstanding, chiefly for professional, logicofoundational reasons, this is how the two symbols are conventionally used. [161, § 2.2]

symbols, $\mathbb{Z}$ is the only one the book often uses, but one should know how to read all four symbols nevertheless.

Regarding the four sets of numbers, Nicolas Bourbaki writes,[14]

> Every measure of size implies a diffuse notion of real numbers. From the point of view of mathematics, the origins of the theory of real numbers must go back to the progressive formation, in Babylonian science, of a system of numbering capable (in principle) of denoting values as near as required to every real number. The possession of such a system, and the confidence in numerical calculations which cannot but follow from it, terminated inevitably, in fact, with a naïve notion of real numbers, which hardly differs from that which one finds today . . . in elementary teaching or with physicists and engineers. Such a notion does not allow itself to be defined with exactness. . . . [27, chapter 12]

Bourbaki assumes, of course, that it were necessary to define such a notion at all. If one pursues the goal Bourbaki pursues then it may indeed be necessary, but meanwhile Bourbaki's "physicists and engineers"[15] may ask: hypothetically, if the real numbers of "Babylonian science" were found to conflict with the exact definitions Bourbaki seeks, then would the conflict call the real continuum into question? Or would it call the *exact definitions* into question? Even Bourbaki would surely have to admit that it would call only the latter. See also Hermann Weyl's advice in § 1.2.4.

Here the book circles back to a theme: the matter of whether the concepts of applied mathematics—which until the 19th century would have been called simply *the concepts of mathematics*—require justification in terms of a corresponding foundational program. Bourbaki might answer that it depends on one's goals, and if he did then your author would agree. Worth hearing also in the matter, though, are Richard W. Hamming in § 1.2.5, Ludwig Wittgenstein in § 1.2.2 and Felix Klein in § 22.5.

[14] The translator's punctuation, spacing, capitalization and exact English phrasing are not here reproduced.

[15] *Scare quotes,* which convert an honest mark of punctuation into a snide sign of disapprobation, are one of the most annoying affectations of subpar early 21st-century written American English, the affectation's unfortunate prevalence requiring an otherwise unnecessary footnote like this—a waste of good ink. This book's "·" are no scare quotes, of course. They're just quotes, Bourbaki here being quoted.

(If the writer disagrees with Bourbaki or anyone else, the writer will just say so.)

6.1.5 Remainders and the modulo notation

The *modulo* notation

$$p \bmod q$$

represents the remainder division of the arbitrary integer p by the positive integer q leaves. For example, $0\mathrm{xE3} \bmod 0\mathrm{x40} = 0\mathrm{x23}$. By definition,[16]

$$\begin{aligned} p \bmod q &\equiv p - rq, \\ 0 \le p \bmod q &< q, \\ (p, q, r) &\in \mathbb{Z}, \end{aligned} \tag{6.1}$$

the r, known as the *integral quotient,*[17]—picturesquely denoted in some books[18] $\lfloor p/q \rfloor$ (the picture being that of a *floor*[19])—being that unique integer that lets $p \bmod q$ fall within the specified range.

Too obvious to bother proving is that

$$gp \bmod gq = (g)(p \bmod q), \qquad g \in \mathbb{Z},\ g > 0. \tag{6.2}$$

6.1.6 Relative primes and the greatest common divisor

A pair of integers is *relatively prime* if no integer greater than 1 divides both. For example, the pair $0\mathrm{xE} = (2)(7)$ and $0\mathrm{xF} = (3)(5)$ is relatively prime whereas the pair $0\mathrm{xE} = (2)(7)$ and $0\mathrm{x15} = (3)(7)$ is not. A ratio is irreducible (§ 6.1.4) if and only if its dividend and divisor are relatively prime.

The *greatest common divisor* (GCD or gcd) of a pair of integers is the greatest integer that divides both. For example, $\gcd(0\mathrm{xE}, 0\mathrm{xF}) = 1$ and $\gcd(0\mathrm{xE}, 0\mathrm{x15}) = 7$. Therefore, if a and b are integers, then these propositions are equivalent:

- a and b are relatively prime;

[16] One could extend the definition to cover negative q, too, if need arose, but the need does not arise in this book. (Refer to footnote 12 regarding definitional deferral.)

[17] [66, § 10]

[18] The book [29] affords a typical example. The paper [66] footnote 17 has mentioned, however, a century older, employs the notation $I(p/q)$ for the same purpose, so the notation $\lfloor p/q \rfloor$ would appear either to be of comparatively recent provenance or, at least, not to be universally agreed upon.

[19] For information or interest, the standard library of the C programming language [92, § 7.12.9.2], incorporated also into the standard library of the C++ programming language [93, § 29.9.1] among others, gives the name `floor()` to the function that performs the indicated operation.

- $\gcd(a, b) = 1$.

Either both are true or both are false.

Plainly, $\gcd(b, a) = \gcd(a, b)$. Observe though that $\gcd(a, \pm a) = |a|$, $\gcd(a, \pm 1) = 1$ and $\gcd(a, 0) = |a|$ for every integer a, except that[20] $\gcd(0, 0) = \infty$.

Similarly to (6.2),

$$\gcd(ga, gb) = g \gcd(a, b), \qquad g \in \mathbb{Z},\ g > 0. \tag{6.3}$$

6.1.7 Euclid's algorithm

Euclid's algorithm[21] extracts the greatest common divisor (GCD or gcd) of a pair of integers—one positive, the other nonnegative—as follows:

$$\begin{aligned} &\gcd(\text{0x60}, \text{0x24}) = \gcd(\text{0x24}, \text{0x60} \bmod \text{0x24}) \\ &= \gcd(\text{0x24}, \text{0x18}) = \gcd(\text{0x18}, \text{0x24} \bmod \text{0x18}) \\ &= \gcd(\text{0x18},\ \text{0xC}) = \gcd(\ \text{0xC}, \text{0x18} \bmod\ \text{0xC}) \\ &= \gcd(\ \text{0xC},\quad 0) = \text{0xC}. \end{aligned}$$

Generally,

$$\begin{aligned} a_{k+1} &\equiv a_{k-1} \bmod a_k \quad \text{for } 0 < k < K, \\ a_K &= 0, \\ (k, K, a_{k-1}, a_k, a_{k+1}, a_K) &\in \mathbb{Z}, \qquad 0 < K, \\ 0 \le a_1 &\le a_0, \qquad 0 < a_0, \end{aligned} \tag{6.4}$$

the first line of which directs the algorithm's iteration and the second line of which, as soon the first line has discovered a null $a_{(\cdot)}$, determines K, whose value is not theretofore known. In the example, for instance, $a_0 = \text{0x60}$, $a_1 = \text{0x24}$, $a_2 = \text{0x18}$, $a_3 = \text{0xC}$, $a_4 = 0$ and, because $a_4 = 0$, $K = 4$.

Together, (6.1) and (6.4) imply that

$$0 \le a_{k+1} < a_k \quad \text{for } 0 < k < K. \tag{6.5}$$

[20] This use of the symbol ∞ probably does not follow the style and practice the mathematics profession would prefer, for it formally confuses pure analysis [147, §§ 1.9 and 10.47][148, §§ 1.23c and 1.45c, and Fig. 8] and admits esoteric questions such as whether, say, ∞ were an integer; but we don't mind. Actually, as can be seen in the just-cited [148], professional mathematicians do not necessarily much mind, either. For further commentary, refer to § 1.2.1 in the book you are now reading.

[21] [52, book VII, propositions 1 and 2]

Because according to (6.4) either $0 = a_1 < a_0$ or $0 < a_1 \le a_0$ and because $a_2 \equiv a_0 \bmod a_1$ if $K \ge 2$,

$$\begin{aligned}
0 = a_2 < a_1 \text{ and } K = 2 &\quad \text{if } a_1 = a_0,\\
0 = a_1 < a_0 &\quad \text{if } K = 1,\\
0 = a_2 < a_1 < a_0 &\quad \text{if } K = 2 \text{ and } a_1 \ne a_0,\\
0 = a_3 < a_2 < a_1 < a_0 &\quad \text{if } K = 3,\\
0 = a_4 < a_3 < a_2 < a_1 < a_0 &\quad \text{if } K = 4,\\
&\vdots\\
0 = a_K < a_{K-1} < \cdots < a_2 < a_1 < a_0, &
\end{aligned}$$

the sequence of the a_k strictly decreasing, except in the special case of $a_1 = a_0$, as the index k increases.

In view of the foregoing, Euclid asserts that

$$\gcd(a_k, a_{k+1}) = \gcd(a_{k-1}, a_k) \quad \text{for all } 0 < k < K \tag{6.6}$$

and, consequently by induction, that

$$\gcd(a_0, a_1) = \gcd(a_{K-1}, 0) = a_{K-1}. \tag{6.7}$$

Euclid's assertion (6.6) and its consequence (6.7) apparently prosper in the example's case, for gcd(0x60, 0x24) = gcd(0xC, 0) = 0xC indeed. The assertion and consequence look plausible in any case. Plausibility is not proof, though, so how can one be certain that Euclid's assertion (6.6) is in every case true?

The proof is harder and turns out to be slipperier than one might expect, partly due to the special case of $a_1 = a_0$ and partly for another reason we shall see; so, rather than grasping the full proof immediately, let us return to the example with which the subsection has begun, *reducing* it by dividing it through by 0xC and *restricting* it by excluding the special case, explicitly requiring that the pair of integers with which the example begins be unequal:

$$\begin{aligned}
&\gcd(8,3) = \gcd(3, 8 \bmod 3)\\
&= \gcd(3,2) = \gcd(2, 3 \bmod 2)\\
&= \gcd(2,1) = \gcd(1, 2 \bmod 1)\\
&= \gcd(1,0) = 1.
\end{aligned}$$

Generally,

$$
\begin{aligned}
b_{k+1} &\equiv b_{k-1} \bmod b_k \quad \text{for } 0 < k < K, \\
\gcd(b_0, b_1) &= 1, \\
b_K &= 0, \\
(k, K, b_{k-1}, b_k, b_{k+1}, b_K) &\in \mathbb{Z}, \quad 0 < K, \\
0 \le b_1 &< b_0,
\end{aligned}
\tag{6.8}
$$

which resembles (6.4) but with the letter b in place of the letter a, with the aforementioned restriction, and with the extra stipulation, on (6.8)'s second line, that b_0 and b_1 be relatively prime (§ 6.1.6).

At this point a warning against the prospect of circular reasoning can be raised: the reduced, restricted algorithm (6.8) presents itself for invocation, as we have just said, only on condition that b_0 and b_1 be relatively prime; yet to determine before invocation whether b_0 and b_1 are indeed relatively prime, does one not require the use of Euclid's algorithm (6.4)? For Euclid's algorithm remains unproven.

The warning however misses one point: Euclid's algorithm is, or once proven means to be, a *quick* way to compute a greatest common divisor and thus (incidentally) to determine whether two integers are relatively prime. It has never been the *only* way to do so. To the extent to which Euclid's algorithm has (because one is still in process of proving it) not yet become available, one can always just factor b_0 and b_1 each into their constituent primes—whether by trial divisions as in § 6.1.3 or by ruder[22] handling. If the

[22] Readers whose first language is a Germanic language other than English might not recognize the English adjective *rude,* which does not (as such readers might suspect) mean "advisable", unless such readers have specifically learned the English adjective. Modern German does possess the stockman's and biologist's technical adjective *rüde* which, though it has somewhat a different meaning, comes from the same root; but the word is seldom met in German outside its technical contexts and thus perhaps does not much help. In English at any rate, the adjective *rude* means "coarse," "unrefined."

Readers whose first language is English might wonder what this is all about until they consider the peculiarly English transitive and intransitive verb *read,* a verb no other language knows with that meaning. The verb comes of the Anglo-Saxon noun *rede,* "advice," "counsel," "speech." Unlike the verb *read,* the noun *rede* is known to all the Germanic languages and as far as the writer knows remains in everyday use in all of them except English (in which the use outside Scotland has become archaic), with only slight variations between languages in pronunciation and spelling. Given sounds and appearances, to suppose a connection between *rude* and *rede* would be natural, yet the two English words are properly unrelated. And yet sounds and appearances matter still. The two English words are phonically entangled down the centuries in a fascinating way.

To chase the following chain of words might interest readers that enjoy amateur philoltwo are thus found to share no primes in common then the two are relatively prime; if found to share then they are not. This observation dispenses with the warning.

Together, (6.1) and (6.8) imply that

$$0 \le b_{k+1} < b_k \quad \text{for } 0 \le k < K \tag{6.9}$$

(the index k having here a broader domain than in eqn. 6.5), whereupon one asserts that

$$\gcd(b_k, b_{k+1}) = 1 \quad \text{for all } 0 \le k < K. \tag{6.10}$$

This assertion is definitionally true for $k = 0$. To show by induction that the assertion is true for larger k, it wants only to be established that, if $\gcd(b_{k-1}, b_k) = 1$, then $\gcd(b_k, b_{k+1}) = 1$; that is, it wants only to be established that if b_{k-1} and b_k are relatively prime then b_k and b_{k+1} must be relatively prime, too, and that this remains so regardless of any specific value one might have chosen for k.

To establish it, let b_{k-1} and b_k be relatively prime. Applying (6.1) to (6.8)'s definition of b_{k+1},

$$b_{k+1} = b_{k-1} - rb_k.$$

ogy: *rubble; ruderibus* (Latin); *rudiment; rudis* (Latin); *rude; rüde* (technical German); *rugire* (Latin); *rut; ru* (Danish); *ruw* (Dutch); *rauh* (German); *rug* (archaic Swedish); *hrjúft* (Icelandic); *rûch* (Frisian); *rough; Rat* (German); *rede; read; lesan* (Anglo-Saxon); *läsa* (Swedish); *plocka* (Swedish); *pluck; assembler* (French); *assemble; sammeln* (German); *lesen* (German); *lesson; lecture; lectio* (Latin); *legere* (Latin). Precise details of the medieval interaction between *ru* and *rudis,* properly unrelated words presumably both known by the literate in the region of medieval England known as the Danelaw, are unknown to this writer; but one may surmise that it was an inadvertent phonic entanglement of the two words that lent *rude* the tenacity to enter the English lexicon. As for *lesan* and cognata, the contrast between their proper Germanic etymology (meaning not "to read" but "to assemble" or "to pluck") and their phonic Latin entanglement (with *legere* which does mean "to read")—the entanglement in this instance occurring earlier and on the Continent—is even more interesting. The footnote has already mentioned the modern German technical adjective *rüde,* which yet again seems phonically entangled with the properly unrelated English intransitive verb *rut,* though whether it has in this instance been the German that has influenced the English or the English, the German—or whether the entanglement comes indirectly by way of the Latin *rugire,* "to roar"—is unclear. The sources [117] and [50] seem to disagree on the last point but, whichever source is right, the cluster of words this footnote mentions is entangled all round: it's quite the knot. A footnote in a book like this is not the place to disentangle the cluster further and the present writer is unqualified further to conduct the disentanglement in any event, but see the aforementioned sources and also [67] and [163].

Insofar as b_{k-1} and b_k are relatively prime, none of the prime factors that compose b_k divides b_{k-1}. Does any of these prime factors divide b_{k+1}, instead, though?

One possibility is that $b_k = 1$, in which case b_k has no prime factors at all. If a larger b_k however had a prime factor—let us call it h—that divided b_{k+1} then one could write,

$$\begin{gathered} b_k = hc, \\ b_{k+1} = b_{k-1} - rch, \\ b_{k+1} \bmod h = b_k \bmod h = 0, \\ (b_{k-1}, b_k, b_{k+1}, r, c, h) \in \mathbb{Z}, \quad h \text{ being prime,} \end{gathered}$$

the first line of which introduces the integer $c \equiv b_k/h$ to be the product of prime factors that compose b_k other than h, the second line of which comes by substituting $ch \leftarrow b_k$ in the unnumbered equation the last paragraph has displayed.[23] Defining now another integer

$$m \equiv rc,$$

we have that

$$\begin{gathered} b_{k+1} = b_{k-1} - mh, \\ (b_{k-1}, b_k, b_{k+1}, m, h) \in \mathbb{Z}, \quad h \text{ being prime,} \end{gathered}$$

which implies that either

- h divides both b_{k+1} and b_{k-1} or
- h divides neither b_{k+1} nor b_{k-1}.

But we have already determined that h, being a prime factor of b_k, cannot divide b_{k-1}, for b_{k-1} and b_k are relatively prime. This determination forces us to conclude that h does not divide b_{k+1}, either.

In short, no prime factor h of b_k with the sought property exists. No prime factor of b_k divides b_{k+1}. But *this* means that b_{k+1} is composed solely of prime factors b_k lacks, a composition that makes the pair (b_k, b_{k+1}) relatively prime, thus according to § 6.1.6 proving the assertion (6.10) as was to be done.

[23] Remember that the section's introduction has defined a prime to be *an integer greater than one,* divisible only by one and itself. Primes are positive. Therefore, h is positive by definition.

Admittedly, applied mathematics is not always entirely consistent in its definitions and terminology [61], but in any case this subsection's letter h represents a positive number.

Some readers, though otherwise persuaded, might still doubt whether (6.10) holds during the algorithm's final iteration, during which $b_{k+1} = b_K = 0$. Even if such readers admit that nothing in the proof has precisely excluded the final iteration, they could still observe that the only way for b_{K-1} and $b_K = 0$ to be relatively prime is for

$$b_{K-1} = 1 \tag{6.11}$$

(the 1, divisible by no prime factor, being the sole positive integer to be prime relative to 0, divisible by every prime factor). Were $b_{K-1} > 1$ possible, it would invalidate (6.10)'s proof.

However, for $k = K - 1$, eqn. (6.8) has that $0 = b_{K-2} \bmod b_{K-1}$, thus finding b_{K-2} to be a multiple of b_{K-1}. Insofar as the proof of (6.10) is trusted at least for all $0 \le k < K - 1$, for $k = K - 2$ eqn. (6.10) finds b_{K-1} to be prime relative to b_{K-2}; whereas the only positive integer to be prime relative to a multiple of itself is 1, which is just what (6.11) delivers. Whether it was necessary thus to buttress (6.10)'s proof specially for the algorithm's final iteration can be debated but, if it was, there it is, banishing the doubt.

Now what about the pair (a_0, a_1), which (unlike the pair b_0, b_1) has not been constructed to be relatively prime? This pair is messy because per (6.4) it admits the special case of $a_1 = a_0$, but (6.4) makes $a_2 = 0$ in the special case and then (6.6) follows at once. In every other case $a_1 < a_0$, so defining

$$g \equiv \gcd(a_0, a_1) \tag{6.12}$$

to be the factor[24] common to a_0 and a_1 and setting

$$b_0 \equiv \frac{a_0}{g}, \qquad b_1 \equiv \frac{a_1}{g}, \tag{6.13}$$

the b_0 and b_1 being thus a_0 and a_1 with the common factor canceled, we have that $\gcd(b_0, b_1) = 1$, which allows (6.8) to be applied and from which (6.10) flows. Multiplying (6.10) then by g and distributing according to (6.3),

$$\gcd(gb_k, gb_{k+1}) = g \quad \text{for all } 0 \le k < K.$$

By (6.13), the last equation implies Euclid's assertion (6.6), completing the subsection's proof.

[24] Notice that it does not say, "the prime factor." According to § 6.1.6, the common factor g might be any positive integer, even 1 or, as in the example that leads the subsection, 0xC.

6.1.8 The least common multiple

Because the b_0, b_1 and $g \equiv \gcd(a_0, a_1)$ of § 6.1.7 are each prime relative to the others, the product $|b_0 b_1| g$—which one can alternately express either as $|b_0 a_1|$ or as $|a_0 b_1|$, or even as $|a_0 a_1|/g$—is the least integer both a_0 and a_1 divides. This number,[25]

$$\operatorname{lcm}(a_0, a_1) = \frac{|a_0 a_1|}{\gcd(a_0, a_1)}, \tag{6.14}$$

is the *least common multiple* (LCM or lcm) of a_0 and a_1. It is an integer because $\gcd(a_0, a_1)$ by definition divides a_1.

Admittedly, one could have skipped § 6.1.5 up through the present subsection, calculating the GCD and LCM of a pair of integers without so much theory simply by factoring the integers and canceling common factors. However, Euclid's algorithm (6.4), Euclid's conclusion (6.7) and this subsection's LCM rule (6.14) together afford an easier, quicker, more comfortable way to do it, especially if the integers are large.

Except in § 6.5, that's all the number theory the book treats; but in applied mathematics, so little will take you pretty far. Now onward we go to other topics.

6.2 The existence and number of polynomial roots

This section shows that an Nth-order polynomial must have exactly N roots.

6.2.1 Polynomial roots

Consider the quotient $B(z)/A(z)$, where

$$\begin{aligned}
A(z) &= z - \alpha, \\
B(z) &= \sum_{k=0}^{N} b_k z^k, \quad N > 0,\ b_N \neq 0, \\
B(\alpha) &= 0.
\end{aligned}$$

[25] In a computer program, it may be preferable to limit the magnitude of the integer temporaries the computer's registers are required to store by expressing (6.14) as

$$\operatorname{lcm}(a_0, a_1) = |a_0| \frac{|a_1|}{\gcd(a_0, a_1)}.$$

Source: [182, "Least common multiple," 17:50, 14 Dec. 2022].

In the long-division symbology of Table 2.3,

$$B(z) = A(z)Q_0(z) + R_0(z),$$

where $Q_0(z)$ is the quotient and $R_0(z)$, a remainder. In this case the divisor $A(z) = z - \alpha$ has first order, and as § 2.6.2 has observed, first-order divisors leave zeroth-order, constant remainders $R_0(z) = \rho$. Thus substituting yields that

$$B(z) = (z - \alpha)Q_0(z) + \rho.$$

When $z = \alpha$, this reduces to

$$B(\alpha) = \rho.$$

But $B(\alpha) = 0$ by assumption, so

$$\rho = 0.$$

Evidently the division leaves no remainder ρ, which is to say that $z - \alpha$ *exactly divides every polynomial $B(z)$ of which $z = \alpha$ is a root.*[26]

Note that if the polynomial $B(z)$ has order N, then the quotient $Q(z) = B(z)/(z - \alpha)$ has exactly order $N - 1$. That is, the leading, z^{N-1} term of the quotient is never null. The reason is that if the leading term were null, if $Q(z)$ had order less than $N - 1$, then $B(z) = (z - \alpha)Q(z)$ could not possibly have order N as we have assumed.

6.2.2 The fundamental theorem of algebra

The *fundamental theorem of algebra* holds that any polynomial $B(z)$ of order N can be factored

$$B(z) = \sum_{k=0}^{N} b_k z^k = b_N \prod_{j=1}^{N} (z - \alpha_j), \quad b_N \neq 0, \tag{6.15}$$

where the α_k are the N roots of the polynomial.[27]

To prove the theorem, it suffices to show that all polynomials of order $N > 0$ have at least one root; for if a polynomial of order N has a root α_N, then according to § 6.2.1 one can divide the polynomial by $z - \alpha_N$ to obtain a new polynomial of order $N - 1$. To the new polynomial the same logic

[26] See also [147, § 5.86], which reaches the same result in nearly the same way.

[27] Professional mathematicians typically state the theorem in a slightly different form. They also usually prove it in rather a different way. [79, chapter 10, Prob. 74][147, § 5.85]

applies: if it has at least one root α_{N-1}, then one can divide *it* by $z-\alpha_{N-1}$ to obtain yet another polynomial of order $N-2$; and so on, one root extracted at each step, factoring the polynomial step by step into the desired form $b_N \prod_{j=1}^{N}(z-\alpha_j)$.

It remains however to show that there exists no polynomial $B(z)$ of order $N > 0$ lacking roots altogether. To show that there is no such polynomial, consider the locus[28] of all $B(\rho e^{i\phi})$ in the Argand range plane (Fig. 2.7), where $z = \rho e^{i\phi}$, ρ is held constant, and ϕ is variable. Because $e^{i(\phi+n2\pi)} = e^{i\phi}$ and no fractional powers of z appear in (6.15), this locus forms a closed loop. At very large ρ, the $b_N z^N$ term dominates $B(z)$, so the locus there evidently has the general character of $b_N \rho^N e^{iN\phi}$. As such, the locus is nearly but not quite a circle at radius $b_N \rho^N$ from the Argand origin $B(z) = 0$, revolving N times at that great distance before exactly repeating. On the other hand, when $\rho = 0$ the entire locus collapses on the single point $B(0) = b_0$.

Now consider the locus at very large ρ again, but this time let ρ slowly shrink. Watch the locus as ρ shrinks. The locus is like a great string or rubber band, joined at the ends and looped in N great loops. As ρ shrinks smoothly, the string's shape changes smoothly. Eventually ρ disappears and the entire string collapses on the point $B(0) = b_0$. Since the string originally has looped N times at great distance about the Argand origin, but at the end has collapsed on a single point, then at some time between it must have swept through the origin and every other point within the original loops. After all, $B(z)$ is everywhere differentiable, so the string can only *sweep* as ρ decreases; it can never skip. The Argand origin lies inside the loops at the start but outside at the end. If so, then the values of ρ and ϕ precisely where the string has swept through the origin by definition constitute a root $B(\rho e^{i\phi}) = 0$. Thus as we were required to show, $B(z)$ does have at least one root, which observation completes the applied demonstration of the fundamental theorem of algebra.

(The purist might object that we have failed to prove that some trick does not exist whereby the N loops smoothly collapsed without passing through *every* point within. The applicationist might reply that, on an applied level, such an objection were less than wholly serious, but anyway, here is at least one formal tactic by which one could rule out the possibility of a trick: as ρ shrinks, observing $\arg\{B[\rho e^{i\phi}] - b_0\}$ as a function of ϕ, keep count of the net number of times the loops wind counterclockwise about the

[28] A *locus* is the geometric collection of points which satisfy a given criterion. For example, the locus of all points in a plane at distance ρ from a point O is a circle; the locus of all points in three-dimensional space equidistant from two points P and Q is a plane; etc.

point $B[0] = b_0$ as, given a particular value of ρ, ϕ is let to sweep the domain $-2\pi/2 < \phi \leq 2\pi/2$. To fill details is left as an exercise to the interested reader.)

The fact that the roots exist is one thing. Actually finding the roots numerically is another matter. For a quadratic (second order) polynomial, (2.2) gives the roots. For cubic (third order) and quartic (fourth order) polynomials, formulas for the roots are known (see chapter 10) though seemingly not so for quintic (fifth order) and higher-order polynomials;[29] but the Newton-Raphson iteration (§ 4.8) can be used to locate a root numerically in any case. The Newton-Raphson is used to extract one root (*any* root) at each step as described above, reducing the polynomial step by step until all the roots are found.

The reverse problem, finding the polynomial given the roots, is much easier: one just multiplies out $\prod_j (z - \alpha_j)$, as in (6.15).

Incidentally, the reverse problem and its attendant multiplication show that an Nth-order polynomial can have no other roots than the N roots $z = \alpha_j$. Reason: the product $\prod_j (z - \alpha_j)$ is nonzero for all other z.

6.3 Addition and averages

This section discusses the two basic ways to add numbers and the three basic ways to calculate averages of them.

6.3.1 Serial and parallel addition

Consider the following problem. There are three masons. The strongest and most experienced of the three, Adam, lays 60 bricks per hour.[30] Next is Brian who lays 45. Charles is new; he lays only 30. Given eight hours, how many bricks can the three men lay? Answer:

$$(8 \text{ hours})(60 + 45 + 30 \text{ bricks per hour}) = 1080 \text{ bricks}.$$

Now suppose that we are told that Adam can lay a brick every 60 seconds; Brian, every 80 seconds; Charles, every 120 seconds. How much time do the

[29] In a celebrated theorem of pure mathematics [177, "Abel's impossibility theorem"], it is said to be shown that no such formula even exists, given that the formula be constructed according to certain rules. Undoubtedly the theorem is interesting to the professional mathematician, but to the applied mathematician it probably suffices to observe merely that no such formula is known.

[30] The figures in the example are in decimal notation.

three men need to lay 1080 bricks? Answer:

$$\begin{aligned}\frac{1080 \text{ bricks}}{\left(\frac{1}{60}+\frac{1}{80}+\frac{1}{120}\right) \text{ bricks per second}} &= 28{,}800 \text{ seconds} \left(\frac{1 \text{ hour}}{3600 \text{ seconds}}\right) \\ &= 8 \text{ hours.}\end{aligned}$$

The two problems are precisely equivalent. Neither is stated in simpler terms than the other. The notation used to solve the second is less elegant, but fortunately there exists a better notation:

$$(1080 \text{ bricks})(60 \parallel 80 \parallel 120 \text{ seconds per brick}) = 8 \text{ hours},$$

where

$$\frac{1}{60 \parallel 80 \parallel 120} = \frac{1}{60} + \frac{1}{80} + \frac{1}{120}.$$

The operator $\parallel$ is called the *parallel addition* operator. It works according to the law

$$\frac{1}{a \parallel b} = \frac{1}{a} + \frac{1}{b}, \tag{6.16}$$

where the familiar operator $+$ is verbally distinguished from the $\parallel$ when necessary by calling the $+$ the *serial addition* or *series addition* operator. With (6.16) and a bit of arithmetic, the several parallel-addition identities of Table 6.1 are soon derived.

The writer knows of no conventional notation for parallel sums of series, but suggests that the notation which appears in the table,

$$\sum_{k=a}^{b} \parallel f(k) \equiv f(a) \parallel f(a+1) \parallel f(a+2) \parallel \cdots \parallel f(b),$$

might serve if needed.

Assuming that none of the values involved is negative, one can readily show that[31]

$$a \parallel x \le b \parallel x \text{ iff } a \le b. \tag{6.17}$$

This is intuitive. Counterintuitive, perhaps, is that

$$a \parallel x \le a. \tag{6.18}$$

Because we have all learned as children to count in the sensible manner $1, 2, 3, 4, 5, \ldots$—rather than as $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots$—serial addition (+) seems

[31] The word *iff* means, "if and only if."

Table 6.1: Parallel and serial addition identities.

$$
\begin{array}{rclrcl}
\dfrac{1}{a \,\|\, b} & = & \dfrac{1}{a} + \dfrac{1}{b} & \dfrac{1}{a + b} & = & \dfrac{1}{a} \,\|\, \dfrac{1}{b} \\[2ex]
a \,\|\, b & = & \dfrac{ab}{a + b} & a + b & = & \dfrac{ab}{a \,\|\, b} \\[2ex]
a \,\|\, \dfrac{1}{b} & = & \dfrac{a}{1 + ab} & a + \dfrac{1}{b} & = & \dfrac{a}{1 \,\|\, ab} \\[2ex]
a \,\|\, b & = & b \,\|\, a & a + b & = & b + a \\[2ex]
a \,\|\, (b \,\|\, c) & = & (a \,\|\, b) \,\|\, c & a + (b + c) & = & (a + b) + c \\[2ex]
a \,\|\, \infty = \infty \,\|\, a & = & a & a + 0 = 0 + a & = & a \\[2ex]
a \,\|\, (-a) & = & \infty & a + (-a) & = & 0 \\[2ex]
(a)(b \,\|\, c) & = & ab \,\|\, ac & (a)(b + c) & = & ab + ac \\[2ex]
\dfrac{1}{\sum_k \,\|\, a_k} & = & \displaystyle\sum_k \dfrac{1}{a_k} & \dfrac{1}{\sum_k a_k} & = & \displaystyle\sum_k \,\|\, \dfrac{1}{a_k}
\end{array}
$$

more natural than parallel addition ($\|$) does. The psychological barrier is hard to breach, yet for many purposes parallel addition is in fact no less fundamental. Its rules are inherently neither more nor less complicated, as Table 6.1 illustrates; yet outside the electrical engineering literature the parallel addition notation is seldom seen.[32] Now that you have seen it, you can use it. There is profit in learning to think both ways. (Exercise: counting from zero serially goes $0, 1, 2, 3, 4, 5, \ldots$; how does the parallel analog go?)[33]

Convention brings no special symbol for parallel subtraction, incidentally. One merely writes

$$a \,\|\, (-b),$$

which means exactly what it appears to mean.

6.3.2 Averages

Let us return to the problem of the preceding section. Among the three masons, what is their average productivity? The answer depends on how you look at it. On the one hand,

$$\frac{(60 + 45 + 30) \text{ bricks per hour}}{3} = 45 \text{ bricks per hour}.$$

On the other hand,

$$\frac{(60 + 80 + 120) \text{ seconds per brick}}{3} = 86\tfrac{2}{3} \text{ seconds per brick}.$$

These two figures are not the same. That is, $1/(86\frac{2}{3}$ seconds per brick$) \neq$ 45 bricks per hour. Yet both figures are valid. Which figure you choose depends on what you want to calculate. Will the masons lay bricks at the same time in different parts of the wall? Then choose the 45 bricks per hour. Will the masons lay bricks at different times in the same part of the wall? Then, especially if the masons have each equal numbers of bricks to lay, choose the $86\frac{2}{3}$ seconds per brick.

When it is unclear which of the two averages is more appropriate, a third average is available, the *geometric mean*

$$[(60)(45)(30)]^{1/3} \text{ bricks per hour}.$$

[32] In electric circuits, loads are connected in parallel as often as, in fact probably more often than, they are connected in series. Parallel addition gives the electrical engineer a neat way of adding the impedances of parallel-connected loads.

[33] [145, eqn. 1.27]

The geometric mean does not have the problem either of the two averages discussed above has. The inverse geometric mean

$$[(60)(80)(120)]^{1/3} \text{ seconds per brick}$$

implies the same average productivity. The mathematically savvy sometimes prefer the geometric mean over either of the others for this reason.

Generally, the *arithmetic, geometric* and *harmonic means* are defined to be

$$\mu \equiv \frac{\sum_k w_k x_k}{\sum_k w_k} = \left(\sum_k \parallel \frac{1}{w_k}\right)\left(\sum_k w_k x_k\right), \tag{6.19}$$

$$\mu_\Pi \equiv \left[\prod_j x_j^{w_j}\right]^{1/\sum_k w_k} = \left[\prod_j x_j^{w_j}\right]^{\sum_k \parallel 1/w_k}, \tag{6.20}$$

$$\mu_\parallel \equiv \frac{\sum_k \parallel x_k/w_k}{\sum_k \parallel 1/w_k} = \left(\sum_k w_k\right)\left(\sum_k \parallel \frac{x_k}{w_k}\right), \tag{6.21}$$

where the x_k are the several samples and the w_k are weights. For two samples weighted equally, these are

$$\mu = \frac{a+b}{2}, \tag{6.22}$$

$$\mu_\Pi = \sqrt{ab}, \tag{6.23}$$

$$\mu_\parallel = 2(a \parallel b). \tag{6.24}$$

If $a \geq 0$ and $b \geq 0$, then, by successive steps,[34]

$$\begin{array}{rcccl}
0 & \leq & (a-b)^2, & & \\
0 & \leq & a^2 - 2ab + b^2, & & \\
4ab & \leq & a^2 + 2ab + b^2, & & \\
2\sqrt{ab} & \leq & a + b, & & \\
\dfrac{2\sqrt{ab}}{a+b} & \leq & 1 & \leq & \dfrac{a+b}{2\sqrt{ab}}, \\
\dfrac{2ab}{a+b} & \leq & \sqrt{ab} & \leq & \dfrac{a+b}{2}, \\
2(a \,\|\, b) & \leq & \sqrt{ab} & \leq & \dfrac{a+b}{2}.
\end{array}$$

That is,

$$\mu_{\|} \leq \mu_{\Pi} \leq \mu. \tag{6.25}$$

The arithmetic mean is greatest and the harmonic mean, least; with the geometric mean falling between.

Does (6.25) hold when there are several nonnegative samples of various nonnegative weights? To show that it does, consider the case of $N = 2^m$ nonnegative samples of equal weight. Nothing prevents one from dividing such a set of samples in half, considering each subset separately, for if (6.25) holds for each subset individually then surely it holds for the whole set (this is so because the average of the whole set is itself the *average of the two subset averages,* where the word "average" signifies the arithmetic, geometric or harmonic mean as appropriate). But each subset can further be divided in half, then each subsubset can be divided in half again, and so on until each smallest group has two members only—in which case we already know that (6.25) obtains. Starting there and recursing back, we have that (6.25)

[34] The steps are logical enough, but the motivation behind them remains inscrutable until the reader realizes that the writer originally worked the steps out backward with his pencil, from the last step to the first. Only then did he reverse the order and write the steps formally here. The writer had no idea that he was supposed to start from $0 \leq (a-b)^2$ until his pencil working backward showed him. "Begin with the end in mind," the saying goes. In this case the saying is right.

The same reading strategy often clarifies inscrutable math. When you can follow the logic but cannot understand what could possibly have inspired the writer to conceive the logic in the first place, try reading backward.

obtains for the entire set. Now consider that a sample of any weight can be approximated arbitrarily closely by several samples of weight $1/2^m$, provided that m is sufficiently large. By this reasoning, (6.25) holds for any nonnegative weights of nonnegative samples, which was to be demonstrated.

6.4 Rounding

The usual rule schoolchildren (in the author's country at least) are taught is to round numbers to the nearest integer or, if positive, upward when two are equally nearest.[35] Graduate-school lore however eventually trains the scientist or engineer that the childhood rule is suboptimal because, for example,[36] the average of 13.5, 14.5, 15.6 and 16.5 is not $(14 + 15 + 16 + 17)/4 \approx 15.5$.

The childhood rule is *biased.*

Whenever data are rounded, some error in averages and other statistics is naturally inevitable. However, *systematic* error can be ameliorated by the following, improved rule:[37]

> Round to the nearest integer, or to the nearest even integer when two are equally nearest.

The improved rule's key word is "even." The improved rule makes the example's average to be $(14+14+16+16)/4 \approx 15.0$, a better approximation.

One can apply the improved rule likewise to hexadecimal and binary numerals in the obvious way. For example, one can round[38] $(\mathrm{0x2.8}, \mathrm{0x3.7}, \mathrm{0x4.8}, \mathrm{0x5.8}, \mathrm{0x6.A}, \mathrm{0x7.8})$ to $(2, 3, 4, 6, 7, 8)$.

6.5 The continued-fraction representation

Besides representing a number like

$$2\pi = \mathrm{0x6.487F}\ldots = 11.01001000\ldots_{\mathrm{binary}}$$

[35] [87]

[36] This example uses decimal notation.

[37] [86, § 4.8.4]

[38] The rule's extension to round, say, 0x6.487F to 0x6.48, preserving one or more (here, two) fractional digits, is easy to state. To state it is left as an exercise.

in the usual way via a sequence of bits,[39] one can represent it in the curious form of a *continued fraction,*[40]

$$2\pi = 6 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{7 + \cfrac{1}{2 + \cfrac{1}{\mathrm{0x92} + \cfrac{1}{3 + \cdots}}}}}}}$$

in which the "···" means that, if truncated, the continued fraction is to be truncated in the manner of

$$2\pi \approx 6 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{7 + \frac{1}{2}}}}} = \frac{\mathrm{0x2C6}}{\mathrm{0x71}}.$$

In terser notation,[41]

$$2\pi = 6 + \frac{1}{3+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{7+}\,\frac{1}{2+}\,\frac{1}{\mathrm{0x92}+}\,\frac{1}{3+}\cdots,$$

which encodes the number by the sequence $[6; 3, 1, 1, 7, 2, \mathrm{0x92}, 3, \ldots]$ of integers, all of which except maybe the leftmost are positive.

In contrast to a power series (2.22), geometric series (§ 2.6.4), Taylor series (chapter 8) or the like, a continued fraction must be computed right to left,[42] for though it resembles a sum it is not a straight sum as the others are. It is noncommutative and nonassociative (Table 2.1) among other distinctions. In the 2π example, to extend the truncation to include the next element, 0x92, would require recalculation of the entire truncation

[39] Or of decimal digits, but this book prefers hexadecimal/binary. See appendix A.

[40] [182, "Continued fraction," 18:08, 7 Sept. 2019]

[41] The notation follows [1, eqn. 6.1.48].

[42] For best floating-point accuracy, it can be preferable to compute the Taylor series and others right to left, as well.

starting from the 0x92 and working leftward thence—for one cannot take the approximation 0x2C6/0x71 that omits the 0x92 and somehow add the 0x92 to it after the fact.[43]

6.5.1 Extraction from a rational

The number 2π, which the section's introduction has put forth, is irrational.[44] Its continued-fraction representation continues forever. A *rational* number's continued-fraction representation however does not continue forever. Conveniently, it ends. For example,

$$\begin{aligned}
\frac{\mathrm{0x46}}{\mathrm{0x13}} &= 3 + \frac{\mathrm{0xD}}{\mathbf{0x13}} \\
&= 3 + \frac{1}{\mathrm{0x13}/\mathrm{0xD}} = 3 + \frac{1}{(1 + 6/\mathrm{0xD})} = 3 + \frac{1}{1+}\frac{6}{\mathbf{0xD}} \\
&= 3 + \frac{1}{1+}\frac{1}{\mathrm{0xD}/6} = 3 + \frac{1}{1+}\frac{1}{(2 + 1/6)} = 3 + \frac{1}{1+}\frac{1}{2+}\frac{1}{\mathbf{6}},
\end{aligned}$$

illustrating an algorithm by which a rational number's continued-fraction sequence—in this case $[3; 1, 2, 6]$—can be extracted. Observe that the chain 0x13, 0xD, 6 of denominators, emphasized in bold print, that arises during the extraction necessarily consists—whether in this or in another rational example—of a procession of finite positive integers, *the procession uniformly decreasing toward 1* (each denominator, after all, being merely the numerator from the preceding line[45]). Because decreasing, this procession and indeed any such procession from a rational must end. When it ends in the

[43] This writer knows no practical way to do it at any rate.

[44] The body of the book does not prove [120][115] that 2π is irrational, since the mere observation that 2π *is not known to be rational* seems adequately to serve physical applications. Inconsistently, the body of the book does prove that $\sqrt{2}$ is irrational—indeed it does so in this very chapter, in § 6.1.4—but the proof for $\sqrt{2}$ is so simple, and features so prominently in the early history of mathematics, that for the body of the book to include it seemed fitting.

The proof that 2π is irrational is interesting, though, even if it lacks a place in applications. Appendix D outlines the proof.

[45] One could object that cancellation of a hypothetical common prime factor might make the denominator smaller than, rather than equal to, the numerator from the preceding line. For instance, if the example's first and second lines respectively ended in 0xF/**0x13** and 6/**0xF** rather than in 0xD/**0x13** and 6/**0xD** as they do, then cancellation of the common prime factor 3 from the 6/**0xF** would leave 2/**5**. However, even if such a cancellation were to occur, the key word would be "smaller". For the denominator **5** either to be smaller than or equal to the numerator 0xF from the preceding line suffices to prove the point.

Whether such a cancellation can occur is unnecessary for the subsection to consider.

example, it leaves the continued-fraction representation

$$\frac{\mathrm{0x46}}{\mathrm{0x13}} = 3 + \frac{1}{1+}\frac{1}{2+}\frac{1}{6},$$

which may in certain circumstances be preferable to the more familiar bit-sequence representation

$$\begin{aligned}\frac{\mathrm{0x46}}{\mathrm{0x13}} &= \mathrm{0x3.AF28}\ldots \\ &= 11.101011110010100001\;101011110010100001\ldots_{\mathrm{binary}},\end{aligned}$$

which though it repeats does not end.

6.5.2 Extraction from an irrational

Application of the last paragraph's algorithm to the irrational 2π yields that

$$\begin{aligned}2\pi &\approx 6 + \mathrm{0x0.487F} \\ &\approx 6 + \frac{1}{(3 + \mathrm{0x0.8800})} \\ &\approx 6 + \frac{1}{3+}\frac{1}{(1 + \mathrm{0x0.E1E0})} \\ &\approx 6 + \frac{1}{3+}\frac{1}{1+}\frac{1}{(1 + \mathrm{0x0.2224})} \\ &\approx 6 + \frac{1}{3+}\frac{1}{1+}\frac{1}{1+}\frac{1}{(7 + \mathrm{0x0.7F90})} \\ &\;\;\vdots\end{aligned}$$

or, more rigorously, bracketing 2π via *interval arithmetic,*

$$\begin{gathered}6 + \mathrm{0x0.487E} < 2\pi < 6 + \mathrm{0x0.487F}, \\ 6 + \frac{1}{(3 + \mathrm{0x0.880B})} < 2\pi < 6 + \frac{1}{(3 + \mathrm{0x0.87FE})}, \\ 6 + \frac{1}{3+}\frac{1}{(1 + \mathrm{0x0.E1BA})} < 2\pi < 6 + \frac{1}{3+}\frac{1}{(1 + \mathrm{0x0.E1E9})}, \\ 6 + \frac{1}{3+}\frac{1}{1+}\frac{1}{(1 + \mathrm{0x0.2256})} < 2\pi < 6 + \frac{1}{3+}\frac{1}{1+}\frac{1}{(1 + \mathrm{0x0.2218})}, \\ 6 + \frac{1}{3+}\frac{1}{1+}\frac{1}{1+}\frac{1}{(7 + \mathrm{0x0.74AB})} < 2\pi < 6 + \frac{1}{3+}\frac{1}{1+}\frac{1}{1+}\frac{1}{(7 + \mathrm{0x0.823B})}, \\ \vdots\end{gathered}$$

Either way, one notices the alternation

$$\begin{aligned}
6 &< 2\pi, \\
6 + \frac{1}{3} &> 2\pi, \\
6 + \frac{1}{3+}\frac{1}{1} &< 2\pi, \\
6 + \frac{1}{3+}\frac{1}{1+}\frac{1}{1} &> 2\pi, \\
6 + \frac{1}{3+}\frac{1}{1+}\frac{1}{1+}\frac{1}{7} &< 2\pi, \\
&\vdots
\end{aligned}$$

by which the continued-fraction expansion is seen to converge from opposite sides.

The rational algorithm of § 6.5.1 thus works on irrationals, too.

Though the continued-fraction representation is curious and though this section has already exampled an algorithm to extract its terms, why one would pursue the representation is admittedly nonobvious—other than for the not-especially-persuasive reason § 6.5.1 has already mentioned regarding repeating bits. Except in the present section, the continued-fraction representation finds no use in this book, after all. However, consider that[46]

$$\sqrt{2} = 1 + \left(\sqrt{2} - 1\right) = 1 + \frac{1}{\left(1 + \sqrt{2}\right)},$$

to which (2.6) has supplied the product $(1 + \sqrt{2})(\sqrt{2} - 1) = 1$. Further developing,

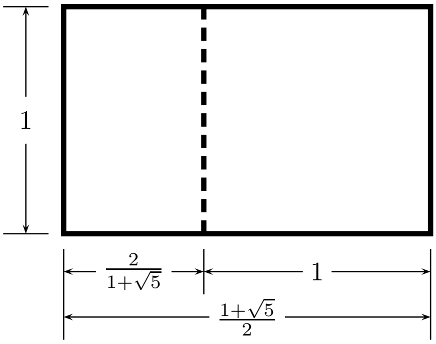
$$\begin{aligned}
\sqrt{2} &= 1 + \frac{1}{2 + \left(\sqrt{2} - 1\right)} = 1 + \frac{1}{2+}\frac{1}{\left(1 + \sqrt{2}\right)} \\
&= 1 + \frac{1}{2+}\frac{1}{2+}\frac{1}{\left(1 + \sqrt{2}\right)} \\
&= 1 + \frac{1}{2+}\frac{1}{2+}\frac{1}{2+}\frac{1}{2+}\frac{1}{2+}\cdots
\end{aligned}$$

a simple, practical result that suggests that continued-fraction representations might be more useful than had been obvious at first glance. The last result might lead one to wonder whether

$$\phi = 1 + \frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\cdots$$

[46] [97, § 1.3]

Figure 6.1: The golden ratio.



were an interesting number. Evidently, one computes this continued fraction right to left by inverting 1 and adding 1, then inverting the sum and adding 1, then inverting the sum and adding 1, and so on iteratively. Assuming that the repetition converges such that, after infinitely many iterations, yet another iteration no longer alters the result—well, the result is ϕ, and the expression $1 + 1/\phi$ symbolizes the aforementioned rule of inverting the sum and adding 1, and "no longer alters" is written by an = equal sign, so

$$\phi = 1 + \frac{1}{\phi}.$$

Subtracting 1 from both sides and multiplying by ϕ, one finds that $(\phi)(\phi - 1) = 1$, which when solved by the method of § 2.2 yields

$$\phi = \frac{1 \pm \sqrt{5}}{2}.$$

Selecting the + sign that ϕ be nonnegative, we discover that

$$\frac{1+\sqrt{5}}{2} = 1 + \frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\cdots$$

This interesting number, called by some *the golden ratio,* has—as is easy to show—the curious property of being the ratio of sides of that rectangle, depicted in Fig. 6.1, from which removal of a square (as by cutting off with scissors) leaves a smaller rectangle in the same proportion as the original.

Mundanely but more importantly, the continued fraction affords a convenient means to approximate a real number accurately with a rational number whose denominator is no larger than it needs to be. See § 6.5.10.

6.5.3 Extraction from a quadratic root

A square root, or more generally a *quadratic root*[47]

$$\begin{gathered} R \equiv \frac{p + s\sqrt{n}}{q}, \\ (n, s, p, q) \in \mathbb{Z}, \quad n \geq 0, \quad q \neq 0, \end{gathered} \tag{6.26}$$

so called because it has the form of a root of a quadratic as in § 2.2, turns out generally to be an easier quantity from which to extract continued-fraction terms than 2π is. However, when n, p, q and s remain unspecified, to extract terms properly wants judicious definitions and some restrictions:

$$\begin{aligned} (j, a_j) \in \mathbb{Z}, \qquad \sqrt{n} &\notin \mathbb{Z}, \quad s \neq 0, \\ u_{-1} \equiv qp, \quad v_{-1} &\equiv q^2, \\ u_j &\equiv v_{j-1}a_j - u_{j-1}, \\ v_j &\equiv \frac{-u_j^2 + (qs)^2 n}{v_{j-1}}, \end{aligned} \tag{6.27}$$

in which the notable restriction that $\sqrt{n} \notin \mathbb{Z}$ requires that $\sqrt{n}$ not be an integer despite that n itself is an integer (for if $\sqrt{n}$ were an integer then the method of § 6.5.7 could be applied instead of this subsection's method).

Admittedly, even if (6.27) is judicious, it comes unmotivated. Nothing in it indicates why a mathematician would have composed it so. Little in it even suggests the purpose it serves. So what *are* u_j and v_j, and why should one want such quantities, anyway?

A precise answer will coalesce in the algebra that follows but, roughly, u_j is a number that serves to shadow the p of (6.26), and v_j to shadow the q. The expressions on the right sides of the $\equiv$ definition signs in (6.27) of u_j and v_j mean nothing to us yet but will soon be found necessary *to iterate* u_j and v_j that these may continue to shadow p and q while the elements of a continued-fraction representation are successively extracted from R. Observe that $u_{-1}/q = p$ and that $v_{-1}/q = q$ so that u_{-1} and v_{-1}, at least, correspond respectively to p and q in a fairly obvious way. (Even more obvious would have been to define $U_{-1} \equiv p$, $V_{-1} \equiv q$, $U_j \equiv V_{j-1}a_j - U_{j-1}$ and $V_j \equiv [-U_j^2 + s^2 n]/V_{j-1}$; but these leave V_j in some instances to be a noninteger, as for example during the extraction—later in this subsection—of the continued-fraction representation of $R = [-5 + 2\sqrt{\mathrm{0xB}}]/3$. If integers are wanted, then the modified definitions of eqn. 6.27 are preferable.)

[47] See footnote 9 regarding (6.26)'s choice of letters.

Stipulating that each a_j be that unique integer which according to (6.27) pulls the corresponding number u_j into the range

$$\begin{aligned} qs\sqrt{n} - v_{j-1} \le u_j < qs\sqrt{n} \qquad & \text{if } v_{j-1} > 0, \\ qs\sqrt{n} < u_j \le qs\sqrt{n} - v_{j-1} \qquad & \text{if } v_{j-1} < 0 \end{aligned} \tag{6.28}$$

(where one verifies satisfaction of an inequality like $qs\sqrt{n} - v_{j-1} \le u_j$ not by evaluating $\sqrt{n}$ but rather by rearranging the inequality to read $qs\sqrt{n} \le u_j + v_{j-1}$, by checking whether the rearranged inequality's two sides have opposite signs, and then, if they don't, by squaring both sides symbolically to obtain a more amenable inequality like $[qs]^2 n \le [u_j + v_{j-1}]^2$; see § 2.1.3) such that

$$0 < \frac{-u_j + qs\sqrt{n}}{v_{j-1}} \le 1 \qquad \text{for all } j \ge 0. \tag{6.29}$$

How so? A compound inequality like $qs\sqrt{n} - v_{j-1} \le u_j < qs\sqrt{n}$ consists of two, separate inequalities: $qs\sqrt{n} - v_{j-1} \le u_j$ and $u_j < qs\sqrt{n}$. Rearrangement of the former yields (6.29)'s "≤ 1" and of latter, its "$0 <$." Since $(u_j + qs\sqrt{n})(-u_j + qs\sqrt{n}) = -u_j^2 + (qs)^2 n$—or, in view of (6.27)'s definition of v_j, since $(u_j + qs\sqrt{n})(-u_j + qs\sqrt{n}) = v_{j-1}v_j$—one can proceed to rewrite (6.29) as

$$0 < \frac{v_j}{u_j + qs\sqrt{n}} \le 1 \qquad \text{for all } j \ge 0, \tag{6.30}$$

eqns. (6.28), (6.29) and (6.30) thus expressing the same stipulation in three different ways. Indeed, when the mathematician first wrote (6.28), it was probably (6.30) or, more precisely, the corresponding line of (6.35) to come, that the mathematician really had in mind.

Observing per (6.27) that $u_{j-1} = v_{j-1}a_j - u_j$ and recalling the algebraic maneuver that has converted (6.29) into (6.30), one finds that

$$\begin{aligned} R \equiv \frac{p + s\sqrt{n}}{q} &= \frac{u_{-1} + qs\sqrt{n}}{v_{-1}} \\ &= \frac{v_{-1}a_0 - u_0 + qs\sqrt{n}}{v_{-1}} \\ &= a_0 + \frac{-u_0 + qs\sqrt{n}}{v_{-1}} \\ &= a_0 + \frac{v_0}{(u_0 + qs\sqrt{n})} \end{aligned}$$

Iterating,

$$\begin{aligned}
R \equiv \frac{p+s\sqrt{n}}{q} &= a_0 + \frac{v_0}{(v_0 a_1 - u_1 + qs\sqrt{n})} \\
&= a_0 + \frac{1}{a_1 +} \frac{(-u_1 + qs\sqrt{n})}{v_0} \\
&= a_0 + \frac{1}{a_1 +} \frac{v_1}{(u_1 + qs\sqrt{n})} \\
&= a_0 + \frac{1}{a_1 +} \frac{1}{a_2 +} \frac{v_2}{(u_2 + qs\sqrt{n})} \\
&= a_0 + \frac{1}{a_1 +} \frac{1}{a_2 +} \frac{1}{a_3 +} \frac{v_3}{(u_3 + qs\sqrt{n})} \\
&\vdots
\end{aligned}$$

After arbitrarily many such iterations,

$$R \equiv \frac{p+s\sqrt{n}}{q} = a_0 + \frac{1}{a_1 +} \frac{1}{a_2 +} \frac{1}{a_3 +} \frac{1}{a_4 +} \frac{1}{a_5 +} \cdots = a_0 + \mathop{\mathrm{K}}_{j=1}^{\infty} \frac{1}{a_j}, \tag{6.31}$$

in which the several a_j are calculated according to (6.27) and (6.28), and in which the K-notation (which resembles the Σ-notation of § 2.3) is introduced to condense the continued fraction's presentation in the manner shown.[48] If the process is halted after the Mth iteration, then

$$\begin{aligned}
R \equiv \frac{p+s\sqrt{n}}{q} &= a_0 + \frac{1}{a_1 +} \frac{1}{a_2 +} \frac{1}{a_3 +} \cdots \frac{1}{a_{M-1} +} \frac{1}{a_M +} \frac{v_M}{(u_M + qs\sqrt{n})} \\
&= a_0 + \mathop{\mathrm{K}}_{j=1}^{M} \frac{1}{a_j} {}_{+} \frac{v_M}{(u_M + qs\sqrt{n})},
\end{aligned} \tag{6.32}$$

[48] A few readers may have read elsewhere that the condensed notation $\mathrm{K}_{j=1}^{M}(1/a_j)$ originated with Gauss. The present author has read the same in two or three scattered places but the most persuasive writer [13] he has read on the topic doubts that Gauss ever used the notation. Lacking further knowledge, the present author offers no opinion in the matter.

The condensed notation is not universally recognized in any event. It does not appear in [97], for example.

The low-printed + sign on (6.32)'s bottom line appears in no other book the author has read but since no more fitting way to represent the notion obviously presents itself, the low-printed + sign shall serve this book. The sign has the meaning (6.32)'s top line implies.

in which the final term, $v_M/(u_M + qs\sqrt{n})$, from which yet further terms could be extracted but have not been, is called the *uncontinued term.*[49]

Though (6.31) and (6.32) give the desired result, the motivational question lingers. The definitions of u_j and v_j in (6.27) work, apparently; but what should have stirred the mathematician to define u_j and v_j so to begin with remains less than clear. However, supposing for a moment that you had not yet read this § 6.5.3, take your pencil in hand and try to extract the first few elements of the continued-fraction representation for example of

$$\begin{aligned}
R = \frac{-5+2\sqrt{\mathrm{0xB}}}{3} &= 0 + \frac{\mathrm{0x13}}{\left(\mathrm{0xF}+6\sqrt{\mathrm{0xB}}\right)} = 0 + \frac{1}{1+}\frac{\mathrm{0xA}}{\left(2+3\sqrt{\mathrm{0xB}}\right)} \\
&= 0 + \frac{1}{1+}\frac{1}{1+}\frac{7}{\left(\mathrm{0x10}+6\sqrt{\mathrm{0xB}}\right)} \\
&= 0 + \frac{1}{1+}\frac{1}{1+}\frac{1}{5+}\frac{5}{\left(\mathrm{0x13}+6\sqrt{\mathrm{0xB}}\right)} \\
&= 0 + \frac{1}{1+}\frac{1}{1+}\frac{1}{5+}\frac{1}{7+}\frac{\mathrm{0xE}}{\left(8+3\sqrt{\mathrm{0xB}}\right)} \\
&= 0 + \frac{1}{1+}\frac{1}{1+}\frac{1}{5+}\frac{1}{7+}\frac{1}{1+}\frac{3}{\left(4+2\sqrt{\mathrm{0xB}}\right)} \\
&= 0 + \frac{1}{1+}\frac{1}{1+}\frac{1}{5+}\frac{1}{7+}\frac{1}{1+}\frac{1}{3+}\frac{\mathrm{0x13}}{\left(\mathrm{0xF}+6\sqrt{\mathrm{0xB}}\right)} \\
&= 0 + \frac{1}{1+}\frac{1}{1+}\frac{1}{5+}\frac{1}{7+}\frac{1}{1+}\frac{1}{3+}\frac{1}{1+}\frac{1}{1+}\frac{1}{5+}\frac{1}{7+}\frac{1}{1+}\frac{1}{3+}\cdots
\end{aligned}$$

When you just did that with your pencil (you *did,* didn't you?), were the steps you took not rather like the steps (6.27) and (6.28) prescribe? The steps you took might not have been precisely the same if you opportunistically canceled factors of 2, 3 and the like as you went, but were the steps not substantially similar? If the steps were, then this is why (6.27) and (6.28) are as they are.

The careful reader may have noticed that, though a desire that v_ℓ be positive has been vaguely expressed, the prospect that $v_\ell = 0$ for some index ℓ, anyway, has not properly been investigated.[50] However, suppose v_ℓ were indeed zero for certain values of the index ℓ and that j were the least

[49] The author has not encountered a standard name for such a final term, though he has not widely read the continued-fraction literature, either. The given name will serve here.

[50] What is this v_ℓ? Should it not be v_j? The answer is that it does not matter which letter one uses for the index as long as the usage remains consistent within the immediate context. The letter ℓ is used here because the paragraph (as you can see) requires the letter j for a different purpose.

such index, such that $v_\ell \neq 0$ for all $-1 \leq \ell < j$. Equation (6.28) has that $u_j \neq qs\sqrt{n}$, which when squared implies that

$$-u_j^2 + (qs)^2 n \neq 0, \tag{6.33}$$

which according to (6.27) makes

$$v_j \neq 0, \tag{6.34}$$

after all.

The especially observant reader might remark that (6.30) ought, as a point of style, to have swapped the $<$ and $\leq$ signs to read $0 \leq \cdots < 1$ instead of $0 < \cdots \leq 1$, and that (6.28) and (6.29) too should have swapped the $<$ and $\leq$ signs accordingly. Your author agrees but see: the u_j and qs are rational while the $\sqrt{n}$ according to § 6.1.4 is not. Thus, $u_j/qs \neq \sqrt{n}$ and therefore

$$\begin{aligned} qs\sqrt{n} - v_{j-1} < u_j < qs\sqrt{n} & \qquad \text{if } v_{j-1} > 0, \\ qs\sqrt{n} < u_j < qs\sqrt{n} - v_{j-1} & \qquad \text{if } v_{j-1} < 0, \\ 0 < \frac{-u_j + qs\sqrt{n}}{v_{j-1}} < 1, & \\ 0 < \frac{v_j}{u_j + qs\sqrt{n}} < 1 & \\ \text{for all } j \geq 0, & \end{aligned} \tag{6.35}$$

anyway. This being so, it matters little which of the two inequality signs should be the $\leq$, whereas the preceding paragraph is simpler when the two signs are as (6.28) through (6.30) have put them.

If one means to use (6.32) to approximate a quadratic root $R \equiv (p + s\sqrt{n})/q$ then one might ask how the uncontinued term $v_M/(u_M + qs\sqrt{n})$ in (6.32) is to be evaluated, since the uncontinued term is itself a quadratic root. However, nothing prevents one from truncating the continued fraction to omit the uncontinued term while approximating R, just as § 6.5.2 has truncated its continued fraction to omit the uncontinued term while approximating 2π. If one does so, then the uncontinued term need not be evaluated in either case. Moreover, just as in the case of 2π, so too in the case of R the continued fraction converges from opposite sides. Section 6.5.8 will develop the point.

6.5.4 Quadratic repetition (integrality)

Table 6.2 lists the continued-fraction representations of several square roots according to the last subsection's procedure. In the table, the terms of the

Table 6.2: The continued-fraction representations of the square roots of the first several primes.

$$\begin{aligned}
\sqrt{2} &= 1+\frac{1}{2+}\frac{1}{2+}\frac{1}{2+}\frac{1}{2+}\frac{1}{2+}\cdots \\
\sqrt{3} &= 1+\frac{1}{1+}\frac{1}{2+}\frac{1}{1+}\frac{1}{2+}\frac{1}{1+}\frac{1}{2+}\cdots \\
\sqrt{5} &= 2+\frac{1}{4+}\frac{1}{4+}\frac{1}{4+}\frac{1}{4+}\frac{1}{4+}\cdots \\
\sqrt{7} &= 2+\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{4+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{4+}\cdots \\
\sqrt{\mathrm{0xB}} &= 3+\frac{1}{3+}\frac{1}{6+}\frac{1}{3+}\frac{1}{6+}\frac{1}{3+}\frac{1}{6+}\cdots \\
\sqrt{\mathrm{0xD}} &= 3+\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{6+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{6+}\cdots \\
\sqrt{\mathrm{0x11}} &= 4+\frac{1}{8+}\frac{1}{8+}\frac{1}{8+}\frac{1}{8+}\frac{1}{8+}\cdots \\
\sqrt{\mathrm{0x13}} &= 4+\frac{1}{2+}\frac{1}{1+}\frac{1}{3+}\frac{1}{1+}\frac{1}{2+}\frac{1}{8+}\frac{1}{2+}\frac{1}{1+}\frac{1}{3+}\frac{1}{1+}\frac{1}{2+}\frac{1}{8+}\cdots \\
\sqrt{\mathrm{0x17}} &= 4+\frac{1}{1+}\frac{1}{3+}\frac{1}{1+}\frac{1}{8+}\frac{1}{1+}\frac{1}{3+}\frac{1}{1+}\frac{1}{8+}\cdots \\
\sqrt{\mathrm{0x1D}} &= 5+\frac{1}{2+}\frac{1}{1+}\frac{1}{1+}\frac{1}{2+}\frac{1}{\mathrm{0xA}+}\frac{1}{2+}\frac{1}{1+}\frac{1}{1+}\frac{1}{2+}\frac{1}{\mathrm{0xA}+}\cdots \\
\sqrt{\mathrm{0x1F}} &= 5+\frac{1}{1+}\frac{1}{1+}\frac{1}{3+}\frac{1}{5+}\frac{1}{3+}\frac{1}{1+}\frac{1}{1+}\frac{1}{\mathrm{0xA}+}\frac{1}{1+}\frac{1}{1+}\frac{1}{3+}\frac{1}{5+}\frac{1}{3+}\frac{1}{1+}\frac{1}{1+}\frac{1}{\mathrm{0xA}+}\cdots \\
\sqrt{\mathrm{0x25}} &= 6+\frac{1}{\mathrm{0xC}+}\frac{1}{\mathrm{0xC}+}\frac{1}{\mathrm{0xC}+}\frac{1}{\mathrm{0xC}+}\frac{1}{\mathrm{0xC}+}\cdots
\end{aligned}$$

continued-fraction representation of each of the square roots seem to repeat, despite that each root is irrational, in contrast to the bits of the bit-sequence representation which do not repeat. The example in § 6.5.3 likewise emits terms that repeat. On the other hand, the terms of the continued-fraction representation of 2π do not seem to repeat, so something subtle is going on here.

This book will not explore the subtlety in great depth, but it will pause to prove at some length—taking two subsections to do it—that *the terms of the continued-fraction representation of each irrational square root and indeed of every irrational quadratic root do in fact repeat.* For this purpose, it suffices to show that, for a given quadratic root and for all $\ell \geq \ell_o$—the $\ell_o \in \mathbb{Z}$ being sufficiently large an integer—all u_ℓ have values drawn from a finite set and all v_ℓ likewise have values drawn from a finite set; for if this is so, then there can exist only a finite collection of distinct pairs (u_ℓ, v_ℓ), whereby—insofar as according to (6.27) and (6.28) mere knowledge of the values of u_{j-1} and v_{j-1} suffices to determine a_j, u_j and v_j—the representation's terms $(1/a_\ell)$ cannot but repeat. The numbers u_{-1} and v_{-1} are integers by definition, as are all a_ℓ. Supposing that u_{j-1} and v_{j-1} were likewise integers, (6.27) would make

$$u_j \in \mathbb{Z}, \tag{6.36}$$

as well. Then (6.27) would have that

$$\begin{aligned} v_j &= \frac{-u_j^2 + (qs)^2 n}{v_{j-1}} = \frac{-(v_{j-1}a_j - u_{j-1})^2 + (qs)^2 n}{v_{j-1}} \\ &= \frac{-v_{j-1}^2 a_j^2 + (v_{j-1})(2u_{j-1}a_j) - u_{j-1}^2 + (qs)^2 n}{v_{j-1}} \\ &= (-v_{j-1}a_j^2 + 2u_{j-1}a_j) + \frac{-u_{j-1}^2 + (qs)^2 n}{v_{j-1}} \end{aligned}$$

The exact value of v_j shall not interest us for the moment, but whether v_j would be an integer shall indeed interest us:

$$v_j \in \mathbb{Z} \quad \text{iff} \quad (-v_{j-1}a_j^2 + 2u_{j-1}a_j) + \frac{-u_{j-1}^2 + (qs)^2 n}{v_{j-1}} \in \mathbb{Z}.$$

The leftward parenthesized expression however is an integer by construction, so

$$v_j \in \mathbb{Z} \quad \text{iff} \quad \frac{-u_{j-1}^2 + (qs)^2 n}{v_{j-1}} \in \mathbb{Z}.$$

Expanding the last line's denominator v_{j-1} according to (6.27),

$$v_j \in \mathbb{Z} \quad \text{iff} \quad \frac{-u_{j-1}^2 + (qs)^2 n}{\left[-u_{j-1}^2 + (qs)^2 n\right]/v_{j-2}} \in \mathbb{Z}.$$

That is, insofar as (6.34) has that $v_{j-1} \neq 0$ and $v_{j-2} \neq 0$ and (6.33) has that $-u_{j-1}^2 + (qs)^2 n \neq 0$,

$$v_j \in \mathbb{Z} \quad \text{iff} \quad v_{j-2} \in \mathbb{Z}.$$

This is a skip-induction, for it connects not v_j to v_{j-1} but v_j to v_{j-2}, so besides that v_{-1} is an integer one must also demonstrate that v_0 is an integer to start the induction. Fortunately, the artifice with which (6.27) has been composed makes it easy to demonstrate that v_0 is an integer:

$$\begin{aligned} v_0 &= \frac{-u_0^2 + (qs)^2 n}{v_{-1}} \\ &= \frac{-(v_{-1}a_0 - u_{-1})^2 + (qs)^2 n}{v_{-1}} \\ &= \frac{-(q^2 a_0 - qp)^2 + (qs)^2 n}{q^2} \\ &= -(qa_0 - p)^2 + s^2 n, \end{aligned}$$

which is plainly an integer, thus starting the induction and thereby implying that

$$v_j \in \mathbb{Z}. \tag{6.37}$$

The foregoing has been the easier half of the proof. The harder half comes next, in § 6.5.5.

6.5.5 Quadratic repetition (boundedness)

The two quantities u_j and v_j being integers, it remains only to show that the two are bounded. This is harder, for its proof *constrains* various quantities including u_j and v_j according to (6.26), (6.27) and (6.28) to be integers—a number-theoretical constraint the typical engineer or physical scientist seldom encounters in problems of such complexity. More specifically, it engages in *modular arithmetic*, constraining the integer u_j to be the greatest that does not exceed, or the least that does not fall short of, the bound (6.28) while still satisfying (6.27). A certain technique is therefore wanted, avoiding explicit mention of a_j (an integer whose value is difficult to determine in the abstract) and, though mention of u_j cannot altogether be avoided, at least avoiding mention of u_{j-1}. The numbers v_{j-1} and v_j are the numbers with which the technique here means to work, abstracting the others away by synthesis of the *parameter*

$$\beta \equiv \frac{-u_j + qs\sqrt{n}}{v_{j-1}}, \tag{6.38}$$

a parameter that conveniently is not constrained to be an integer but which is merely constrained, by (6.35), to lie within the bounds[51]

$$0 < \beta < 1. \tag{6.39}$$

Rearranging (6.38),

$$u_j = -\beta v_{j-1} + qs\sqrt{n}.$$

Substituting the thus-parameterized expression for u_j into (6.27)'s definition of v_j and simplifying, we find that

$$v_j = (\beta)\left(-\beta v_{j-1} + 2qs\sqrt{n}\right), \tag{6.40}$$

in which β again serves as parameter. Dividing by $2qs\sqrt{n}$,

$$\frac{v_j}{2qs\sqrt{n}} = (\beta)\left(-\beta\frac{v_{j-1}}{2qs\sqrt{n}} + 1\right), \tag{6.41}$$

an equation that, since β according to (6.39) is a positive number less than one, makes evident that

$$0 < \frac{v_j}{2qs\sqrt{n}} < 1 \quad \text{if} \quad 0 < \frac{v_{j-1}}{2qs\sqrt{n}} < 1. \tag{6.42}$$

Equation (6.42) says that, if any v_ℓ lies within the specified bounds, then by induction all subsequent v_ℓ must lie also within the same bounds. That all subsequent u_ℓ too are bounded then follows at once from (6.35).

[51] A reader might ask why, if § 6.5.4 was so keen for u_j and v_j to be integers, § 6.5.5 now goes to the trouble of defining a new quantity β specifically not to be an integer. The reason however is twofold. First, that u_j and v_j are integers is convenient when one is trying to show, as this section is, that only a finite collection of distinct pairs (u_ℓ, v_ℓ) exist—a consideration that does not apply to β. Second, when a computer is used to extract continued-fraction representations, integers yield exact results. Anyway, § 6.5.4 having wrung as much proof out of the integers as it can, this § 6.5.5 now turns to the noninteger β.

It will not be necessary to calculate the value of β, so exact results are not at issue. See the narrative.

The sole question remaining is whether a quadratic root exists for which all v_ℓ lie outside the specified bounds. Defining

$$\begin{aligned} \delta_j &\equiv \frac{v_j}{2qs\sqrt{n}} - 1 \neq 0, \\ \epsilon_j &\equiv -\frac{v_j}{2qs\sqrt{n}} \quad \neq 0, \\ \delta_j + \epsilon_j &= -1. \end{aligned} \tag{6.43}$$

respectively to be the amount by which v_j exceeds the upper bound and the amount by which v_j falls short of the lower bound in (6.42)—where δ_j is nonzero because (as the subsection has earlier argued) $\sqrt{n}$ is irrational and ϵ_j is nonzero due to (6.34)—we shall answer the question by proving the following two propositions.

1. If $\delta_{j-1} > 0$ (that is, if v_j lies above the upper bound), then either
   - $\delta_j < 0$ and $\epsilon_j < 0$, or at least
   - $0 < \epsilon_j < \delta_{j-1}$.
2. If $\epsilon_{j-1} > 0$ (that is, if v_j lies below the lower bound), then either
   - $\delta_j < 0$ and $\epsilon_j < 0$, or at least
   - $0 < \delta_j < \epsilon_{j-1}$.

These propositions answer the question because the definitions of (6.43) imply that

$$0 < \frac{v_j}{2qs\sqrt{n}} < 1 \qquad \text{if } \delta_j < 0 \text{ and } \epsilon_j < 0, \tag{6.44}$$

whereupon the induction of (6.42) takes hold; because the propositions, if true, imply in conjunction with (6.43)'s last line that

- if $\delta_{j-1} > 0$ then $\delta_j < 0$, and
- if $\epsilon_{j-1} > 0$ then $\epsilon_j < 0$;

and because the propositions, if true, imply together that

- if $\delta_{j-2} > 0$ and $\epsilon_{j-1} > 0$ then $\delta_j < \epsilon_{j-1} < \delta_{j-2}$, and
- if $\epsilon_{j-2} > 0$ and $\delta_{j-1} > 0$ then $\epsilon_j < \delta_{j-1} < \epsilon_{j-2}$.

Regarding the last two bulleted points, if either $\delta_{j-2} - \delta_j$ or $\epsilon_{j-2} - \epsilon_j$ were a nonzero infinitesimal then further investigation would be required to decide whether a quadratic root existed for which all v_ℓ lay outside the specified bounds. However, v_{j-2} and v_j being integers and q, s and n each being finite, (6.43) permits neither $\delta_{j-2} - \delta_j$ nor $\epsilon_{j-2} - \epsilon_j$ to be a nonzero infinitesimal. Therefore, further investigation is not required. We only need to prove propositions 1 and 2.

If the last paragraph seems too abstract, in plainer language, it proposes that the several v_ℓ spiral in toward the desired bounds. According to the paragraph's propositions, if v_0 for example lies below the lower bound, then v_1 falls either in bounds or above the upper bound—and if above the upper bound, then v_1 at least falls *nearer* to the upper bound than v_0 to the lower bound. Then, insofar as v_1 lies above the upper bound, v_2 falls either in bounds or below the lower bound, again nearer. The sequence v_0, v_1, v_2, v_3, v_4, ..., thus spirals inward, nearer and nearer the bounds at each step, until inevitably a v_ℓ is reached that lies within bounds, after which according to (6.42) all subsequent v_ℓ likewise lie within bounds. The sequence is drawn in, leaping above to below to above to below, closer and closer at each leap, until trapped.

It remains though to prove[52] propositions 1 and 2.

For proposition 1, suppose that

$$\delta_{j-1} > 0.$$

Changing $j \leftarrow j - 1$ in (6.43) and rearranging terms,

$$\frac{v_{j-1}}{2qs\sqrt{n}} = 1 + \delta_{j-1}.$$

[52] Notice the style. Professional mathematicians, at least in the author's generation in the English-speaking world, tend to favor a reverse style of reasoning, in which propositions like 1 and 2 are first proved as "lemmas," after which the lemmas are employed in higher-level demonstrations, thus working through a proof like the one of §§ 6.5.4 and 6.5.5—which you are now reading—backward, more or less from end to beginning. The logic of such a professional approach is unimpeachable but, in the author's university-teaching experience, has a synthetic, unmotivational quality that clashes with the way scientists and engineers—or at least engineering *students*—tend to think.

The engineering style here displayed incidentally brings a subtler benefit: it foreshadows the top-down manner in which an engineer peels away successively more elaborate terms to isolate a term needed to balance or (more likely) approximately balance an equation. This subtler point might not convey much to many readers, but some readers may come to appreciate it if and when, during esoteric work using results from part III, such readers find themselves wrestling an equation (such as in [167], for example) that cannot be exactly solved but wants its balance improved toward an adequate approximate solution.

Substituting this into (6.41),

$$\frac{v_j}{2qs\sqrt{n}} = [\beta][-(\beta)(1+\delta_{j-1})+1].$$

Taking the derivative with respect to the parameter β,

$$\frac{\partial}{\partial\beta}\left(\frac{v_j}{2qs\sqrt{n}}\right) = -(2\beta)(1+\delta_{j-1})+1.$$

Taking the derivative again,

$$\frac{\partial^2}{\partial\beta^2}\left(\frac{v_j}{2qs\sqrt{n}}\right) = -2(1+\delta_{j-1}) < 0.$$

That the second derivative is everywhere negative implies the following.

- As long as the first derivative equals zero at some point within the interval $0 < \beta < 1$, the quotient $v_j/2qs\sqrt{n}$ according to § 4.6 reaches its maximum at the value of β at which first derivative's zero occurs. That is,

$$\begin{aligned}\left[\frac{v_j}{2qs\sqrt{n}}\right]_{\max} &= \{[\beta][-(\beta)(1+\delta_{j-1})+1]\}_{\beta=1/2(1+\delta_{j-1})} \\ &= \frac{1}{4(1+\delta_{j-1})} < \frac{1}{4},\end{aligned}$$

  the specified value of β indeed lying within the interval since $\delta_{j-1} > 0$.

- The quotient $v_j/2qs\sqrt{n}$ reaches its minimum at one or both extremes of β:

$$\begin{aligned}\left[\frac{v_j}{2qs\sqrt{n}}\right]_{\beta=1} &= \{[\beta][-(\beta)(1+\delta_{j-1})+1]\}_{\beta=1} = -\delta_{j-1}; \\ \left[\frac{v_j}{2qs\sqrt{n}}\right]_{\beta=0} &= \{[\beta][-(\beta)(1+\delta_{j-1})+1]\}_{\beta=0} = 0.\end{aligned}$$

  Thus, since $-\delta_{j-1} < 0$,

$$\left[\frac{v_j}{2qs\sqrt{n}}\right]_{\min} = -\delta_{j-1}.$$

Assembling these two points,

$$-\delta_{j-1} < \frac{v_j}{2qs\sqrt{n}} < \frac{1}{4(1+\delta_{j-1})} < \frac{1}{4} \qquad \text{if } \delta_{j-1} > 0,$$

in which $<$ is used instead of $\leq$ because $\beta < 1$ rather than $\beta \leq 1$. Applying (6.43)'s definition of ϵ_j,

$$-\delta_{j-1} < -\epsilon_j < \frac{1}{4(1+\delta_{j-1})} < \frac{1}{4} \qquad \text{if } \delta_{j-1} > 0.$$

Negating all four sides,

$$-\frac{1}{4} < -\frac{1}{4(1+\delta_{j-1})} < \epsilon_j < \delta_{j-1} \qquad \text{if } \delta_{j-1} > 0. \tag{6.45}$$

On the other hand, by the last line of (6.43),

$$\delta_j < -\frac{3}{4} \quad \text{if} \quad -\frac{1}{4} < \epsilon_j. \tag{6.46}$$

Taken together, (6.45) and (6.46), along with (6.43)'s observation that $\epsilon_j \neq 0$, prove proposition 1.

Proposition 2 is proved similarly. Supposing that

$$\epsilon_{j-1} > 0$$

and proceeding by steps resembling the steps the last paragraph has already taken with respect to proposition 1,

$$\begin{aligned}
\frac{v_{j-1}}{2qs\sqrt{n}} &= -\epsilon_{j-1}, \\
\frac{v_j}{2qs\sqrt{n}} &= [\beta][\beta\epsilon_{j-1} + 1], \\
\frac{\partial}{\partial\beta}\left(\frac{v_j}{2qs\sqrt{n}}\right) &= 2\beta\epsilon_{j-1} + 1.
\end{aligned}$$

Unlike in the last paragraph, in this paragraph the first derivative is positive over the entire domain $0 < \beta < 1$. Therefore, the second derivative is unneeded and

$$\begin{aligned}
\left[\frac{v_j}{2qs\sqrt{n}}\right]_{\min} &= \{[\beta][\beta\epsilon_{j-1} + 1]\}_{\beta=0} = 0, \\
\left[\frac{v_j}{2qs\sqrt{n}}\right]_{\max} &= \{[\beta][\beta\epsilon_{j-1} + 1]\}_{\beta=1} = 1 + \epsilon_{j-1}.
\end{aligned}$$

Assembling,

$$0 < \frac{v_j}{2qs\sqrt{n}} < 1 + \epsilon_{j-1}$$

(which, curiously even if not pertinently, is therefore true whether or not $\epsilon_{j-1} > 0$). Applying (6.43)'s definition of δ_j,

$$-1 < \delta_j < \epsilon_{j-1}. \tag{6.47}$$

On the other hand, by the last line of (6.43),

$$\epsilon_j < 0 \quad \text{if} \quad -1 < \delta_j. \tag{6.48}$$

Taken together, (6.47) and (6.48), along with (6.43)'s observation that $\delta_j \neq 0$, prove proposition 2—a proof which, when joined with the last paragraph's proof of proposition 1, completes the admittedly difficult, two-subsection-long demonstration that the terms of the continued-fraction representation of an irrational quadratic root repeat.[53]

6.5.6 Quadratic repetition (reverse hypothesis)

The reverse hypothesis, that a repeating continued fraction represents a quadratic root, is easier to verify; and for the reverse hypothesis we shall dispense with the formality and content ourselves merely to illustrate by example. Let

$$\begin{aligned}\psi &= 1 + \frac{1}{3+}\frac{1}{1+}\frac{1}{1+}\frac{1}{2+}\frac{1}{1+}\frac{1}{1+}\frac{1}{2+}\frac{1}{1+}\frac{1}{1+}\frac{1}{2+}\cdots = 1 + \frac{1}{3+}\frac{1}{\omega},\\ \omega &= 1 + \frac{1}{1+}\frac{1}{2+}\frac{1}{\omega}.\end{aligned}$$

Subtracting 1 from each side of the last equation and inverting,

$$\frac{1}{\omega - 1} = 1 + \frac{1}{2+}\frac{1}{\omega}.$$

Again subtracting 1 from each side and inverting,

$$\frac{\omega - 1}{-\omega + 2} = 2 + \frac{1}{\omega}.$$

[53] The writer gathers that Joseph-Louis Lagrange (1736–1813) was first to demonstrate it. Lagrange's technique was approximately as difficult as this section's but different; [84] reprises it. A shorter, more recent technique—also different from this section's but unfortunately hard to motivate from an applied perspective—is given by [114].

Now subtracting 2 from each side and inverting,

$$\frac{-\omega + 2}{3\omega - 5} = \omega,$$

in which one observes, significantly, that neither numerator nor denominator is of higher order than on the preceding line. Transferring the denominator to the right side,

$$-\omega + 2 = (3\omega - 5)\omega.$$

This is a quadratic equation, soluble by the method of § 2.2:

$$\begin{aligned} \omega &= \frac{2 + \sqrt{\mathrm{0xA}}}{3}, \\ \psi &= \frac{7 - \sqrt{\mathrm{0xA}}}{3}. \end{aligned}$$

Other repeating continued fractions follow a similar pattern and reach results of the same kind.

6.5.7 Extraction from a rational, generalized

Compared to the quadratic extraction of §§ 6.5.3 through 6.5.5, extraction of a rational number's continued-fraction representation is easy:

$$\begin{aligned} R &\equiv \frac{p}{q}, \\ (p, q, j, J, a_j) &\in \mathbb{Z}, \qquad q > 0, \\ u_{-1} &\equiv p, \quad u_0 \equiv q, \\ u_{j+1} &\equiv u_{j-1} - u_j a_j, \\ u_{J+1} &= 0, \\ 0 < u_{j+1} &< u_j \quad \text{for } 0 \le j < J. \end{aligned} \tag{6.49}$$

Iterating,

$$\begin{aligned}R \equiv \frac{p}{q} &= \frac{u_{-1}}{u_0} \\ &= a_0 + \frac{u_1}{u_0} \\ &= a_0 + \frac{1}{a_1 +}\frac{u_2}{u_1} \\ &= a_0 + \frac{1}{a_1 +}\frac{1}{a_2 +}\frac{u_3}{u_2} \\ &= a_0 + \frac{1}{a_1 +}\frac{1}{a_2 +}\frac{1}{a_3 +}\frac{u_4}{u_3} \\ &\;\vdots\end{aligned}$$

Identifying as J the value of j for which $u_{j+1} = 0$ and halting when this value is reached,

$$R \equiv \frac{p}{q} = a_0 + \frac{1}{a_1 +}\frac{1}{a_2 +}\frac{1}{a_3 +}\frac{1}{a_4 +}\frac{1}{a_5 +}\cdots\frac{1}{a_{J-1} +}\frac{1}{a_J} = a_0 + \mathop{\mathrm{K}}_{j=1}^{J}\frac{1}{a_j}. \tag{6.50}$$

See § 6.5.1 for discussion and an example.

6.5.8 Canonical form

If you have worked several examples by now and have thus gained a sense of continued fractions, then it should not surprise that

$$5 < \cdots < 5 + \frac{1}{4} < 5 + \frac{1}{3} < 5 + \frac{1}{2} < 6,$$

that

$$5 + \frac{1}{9} < 5 + \frac{1}{8+}\frac{1}{2} < 5 + \frac{1}{8+}\frac{1}{3} < \cdots < 5 + \frac{1}{8},$$

and so on. Pay particular attention to the nested pattern, though: $5 + (1/8+)(1/2)$ and every other $5 + (1/8+)(1/\cdot)$ lie between $5 + (1/9)$ and $5 + (1/8)$, which themselves lie between 5 and 6. The sense of it resembles the sense of the hexadecimal representation, wherein 0x5.82 along with every other 0x5.8· lies between 0x5.8 and 0x5.9, which themselves lie between 5 and 6. Observe however that $5 + (1/9) < 5 + (1/8)$, whereas $\mathrm{0x5.9} > \mathrm{0x5.8}$: apparently, the sequence of a_1 runs backward, as do the sequences of a_3, of a_5, of a_7 and of other a_{odd}. Otherwise, as you see, the ordering of numbers represented by continued fractions is straightforward as long as one remembers that—unlike the hexadecimal representation, in which no digit larger than 0xF is available—the continued-fraction representation can accommodate any positive integer a_j.

Well, not quite any. If *any* positive a_j were allowed, then a single number could have more than one continued-fraction representation. For example,

$$5 + \frac{1}{9+}\frac{1}{\infty} = 5 + \frac{1}{9} = 5 + \frac{1}{8+}\frac{1}{1}.$$

Therefore, for a given real number R, the unique continued fraction whose elements obey the rule

$$\begin{array}{ll} 2 \le a_J < \infty & \text{if the term } 1/a_J \text{ is the} \\ & \text{representation's final term,} \\ 1 \le a_j < \infty & \text{for all other } j > 0, \end{array} \tag{6.51}$$

is called the *canonical form*.[54]

The continued-fraction representation incidentally is not the only representation that, for sake of uniqueness, restricts its canonical form. The hexadecimal and bit-sequence representations do, too. The canonical hexadecimal representation does not for example admit[55]

$$\text{0x3.97FF FFFF FFFF} \ldots$$

since this number is just 0x3.98. Likewise, the canonical form of the bit-sequence representation does not for example admit

$$11.1001011111111111111111\ldots_{\text{binary}}$$

[54] The uniqueness of the canonical form is too obvious for us to bother with a formal proof. An applicationist must conserve the effort formalism requires for hard proofs that need it!—such as, for example, the quadratic proof of §§ 6.5.3 through 6.5.5.

[55] If this surprises the reader, then here are three ways to think about it. First, consider the difference between 0x3.97FF FFFF FFFF . . . and 0x3.98, which is nothing other than 0x0.0000 0000 0000 . . . Second, observe that 0x0.00FF FFFF FFFF . . . = $(\text{0xF})(\text{0x10}^{-3}) + (\text{0xF})(\text{0x10}^{-4}) + (\text{0xF})(\text{0x10}^{-5}) + \cdots = +(\text{0x0.00F})(\text{0x0.1}^0 + \text{0x0.1}^1 + \text{0x0.1}^2 + \cdots) = \text{0x0.00F}/(1 - \text{0x0.1}) = \text{0x0.01}$ according to (2.36). (Since everybody including the writer is used to reading decimal numerals and especially since the book's main audience consists of scientists, engineers and their students—to whom scientific notation like $k_B \approx 1.381 \times 10^{-23}$ J/K for quantities like, for example, the Boltzmann constant [118] is familiar—there will an unfortunate tendency to misread 0x10^{-3}, the inverse of sixteen cubed, as it were 0×10^{-3}. Don't do that.) Third, given that (2)(5) = 0xA and (3)(5) = 0xF, let the reader ask himself this: if 1/3 = 0x0.5555 . . . and 2/3 = 0x0.AAAA . . . then what does 3/3 equal? For any and all of these reasons, 0x3.97FF FFFF FFFF . . . can be no number other than 0x3.98.

This works for decimal numerals, too, of course: 0.33333 . . . is one third but 0.99999 . . . is just one.

since this number is just $11.10011_{\mathrm{binary}}$. So, restrictions for canonicality are nothing new, even if one has not thought much about them heretofore.

6.5.9 Ratio's construction and relative primes

Given a continued-fraction representation, construction of the fraction, or *ratio,* the continued fraction represents is easy. It's just (6.49) in reverse:[56]

$$\begin{aligned} u_{j-1} &= u_{j+1} + u_j a_j, \\ u_{J+1} &= 0, \quad u_J = 1; \end{aligned} \tag{6.52}$$

and then the ratio is merely $R \equiv p/q = u_{-1}/u_0$ as (6.49) has already said. If the number the continued fraction represents is irrational then the continued fraction will not end, of course, but in that case to compute the number's exact value is impossible, anyway, so one arbitrarily selects a suitably large value for J and thereby truncates the continued fraction at $1/a_J$—as the section's introduction has done for 2π, for example, selecting $J = 5$ and thereby approximating $2\pi = \mathrm{0x2C6}/\mathrm{0x71}$. Treating the example more thoroughly,

$$6 + \frac{1}{3+}\frac{1}{1+}\frac{1}{1+}\frac{1}{7+}\frac{1}{3} < 2\pi < 6 + \frac{1}{3+}\frac{1}{1+}\frac{1}{1+}\frac{1}{7+}\frac{1}{2},$$

whereby

$$\frac{\mathrm{0x413}}{\mathrm{0xA6}} < 2\pi < \frac{\mathrm{0x2C6}}{\mathrm{0x71}}.$$

Generally,

$$R_{\mathrm{min},J} < R < R_{\mathrm{max},J}, \tag{6.53}$$

in which

- if the selected J is even, then $R_{\mathrm{min},J}$ is calculated according to (6.52) using the actual value of a_J and $R_{\mathrm{max},J}$ is similarly calculated except using $a_J \leftarrow a_J + 1$ but,

- if the selected J is odd, then $R_{\mathrm{max},J}$ is calculated according to (6.52) using the actual value of a_J and $R_{\mathrm{min},J}$ is similarly calculated except using $a_J \leftarrow a_J + 1$.

Significantly, *where (6.52) has been applied, u_{j-1} and u_j are* relatively prime *for all* $0 \le j \le J + 1$. That is, no positive integer except 1 divides both u_{j-1} and u_j. How does one know? By induction. The assertion is

[56] If unsure that (6.52) really works, take your pencil and try (6.52) on an example like, say, $5 + (1/8+)(1/2)$.

plainly true for $j = J + 1$, whereupon (6.52) makes the statement true for each $j < J + 1$ in turn, for if u_j and u_{j+1} are relatively prime then, by assertion, neither u_j nor any of the prime factors of which u_j is composed divides u_{j+1}. Therefore, not only are u_j and u_{j+1} relatively prime but u_j and $u_{j+1} + u_j a_j$ are relatively prime for any integer a_j; yet according to (6.52) this is merely to say that u_{j-1} and u_j are relatively prime as was to be shown.

Another way to state the last paragraph's finding is that that the ratio $p/q = u_{-1}/u_0$ iteration of (6.52) yields is *irreducible.*

6.5.10 Estimation and bounding

The principal practical application of the continued-fraction representation is to estimate a real number by a suitable ratio or, better, to bound the number by a suitable pair $R_{\mathrm{min},J}$ and $R_{\mathrm{max},J}$ of ratios as § 6.5.9 has done. The bounding ratios $R_{\mathrm{min},J}$ and $R_{\mathrm{max},J}$ are the best possible in the sense that no ratios with denominators as small afford tighter bounds. *There exist many rational numbers between $R_{min,J}$ and $R_{max,J}$ but all of these,* it is asserted, *have larger denominators than $R_{min,J}$ has and than $R_{max,J}$ has.*

To prove that the estimate the continued fraction yields is in the aforementioned sense best, we begin with the observation that

$$u_{j-1} > u_j \qquad \text{for } j > 0, \tag{6.54}$$

because either

- u_{j+1} and u_j are both positive in which case, since a_j is by canonical construction positive for $j > 0$, eqn. (6.52) implies eqn. (6.54); or

- the term $1/a_j$ is the continued-fraction expansion's final term in which case (6.51) requires that $a_j \geq 2$, by which—despite that $u_{j+1} = 0$—eqn. (6.52) nevertheless again implies eqn. (6.54).

Thus satisfied that (6.54) is correct, we proceed to consider according to the nesting rule of § 6.5.8 the fact that every ratio between $R_{\mathrm{min},J}$ and $R_{\mathrm{max},J}$ has more terms in its canonical continued-fraction expansion than $R_{\mathrm{min},J}$ has and than $R_{\mathrm{max},J}$ has. Supposing for the moment that J were an even number, we let

$$
\begin{aligned}
R_{\min,J} &= a_0 + \mathop{\mathrm{K}}_{j=1}^{J-1} \frac{1}{a_j} + \frac{1}{a_J}, \\
R &= a_0 + \mathop{\mathrm{K}}_{j=1}^{J-1} \frac{1}{a_j} + \frac{1}{a_J +} \frac{1}{a_{J+1} +} \cdots \\
R_{\max,J} &= a_0 + \mathop{\mathrm{K}}_{j=1}^{J-1} \frac{1}{a_j} + \frac{1}{(a_J + 1)}, \\
&J \text{ being even,}
\end{aligned}
$$

and then, for each, we consider the values of u_{J+1}, u_J and u_{J-1} according to (6.52) as follows.

- For $R_{\min,J}$, $u_{J+1} = 0$, $u_J = 1$ and thus $u_{J-1} = a_J$.
- For R, $u_{J+1} \geq 1$, $u_J \geq 2$ (due to eqn. 6.54 evaluated at $j = J + 1$) and thus $u_{J-1} \geq 2a_J + 1$.
- For $R_{\max,J}$, $u_{J+1} = 0$, $u_J = 1$ and thus $u_{J-1} = a_J + 1$.

So we see that R has a larger u_{J+1}, a larger u_J and a larger u_{J-1} than either $R_{\min,J}$ or $R_{\max,J}$ has, from which it follows iteratively by (6.52) that R also has a larger u_{J-2}, a larger u_{J-3}, and so on, down to and including a larger u_0. Since u_0 is the denominator we have been seeking, the last observation completes our proof for even J. For odd J, the proof is similar except that the roles of $R_{\min,J}$ and $R_{\max,J}$ are reversed.

The careful reader might object that the proof has neglected the case $J = 0$ but in that case $R_{\min,J}$ and $R_{\max,J}$ are consecutive integers and thus the theorem's truth is obvious.

One concludes that, when a rational estimate to an irrational number is required, truncation of the irrational's continued-fraction representation is normally, probably the best way to get it.

Chapter 7

The integral

Chapter 4 has observed that the mathematics of calculus concerns a complementary pair of questions:

- Given some function $f(t)$, what is the function's instantaneous rate of change, or *derivative,* $f'(t)$?
- Interpreting some function $f'(t)$ as an instantaneous rate of change, what is the corresponding accretion, or *integral,* $f(t)$?

Chapter 4 has built toward a basic understanding of the first question. This chapter builds toward a basic understanding of the second. The understanding of the second question constitutes the concept of the integral, one of the profoundest ideas in all of mathematics.

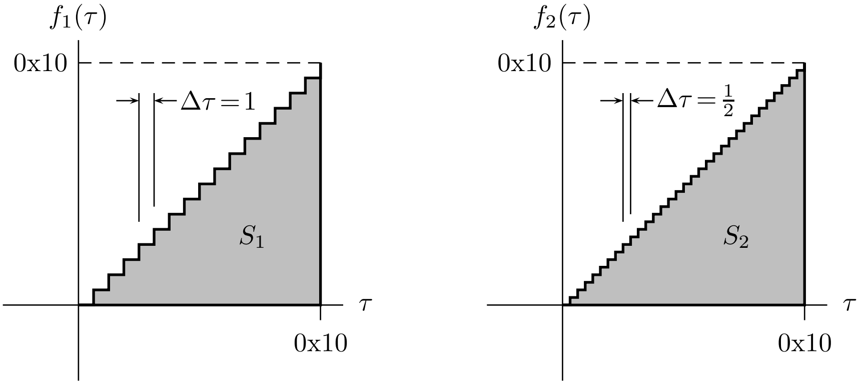
This chapter, which introduces the integral, is undeniably a hard chapter.

Experience knows no reliable way to teach the integral adequately to the uninitiated except through dozens or hundreds of pages of suitable examples and exercises, yet the book you are reading cannot be that kind of book. The sections of the present chapter concisely treat matters which elsewhere rightly command chapters or whole books of their own. Concision can be a virtue—and by design, nothing essential is omitted here—but the bold novice who wishes to learn the integral from these pages alone faces a daunting challenge. It can perhaps be done. Meanwhile, the less intrepid who prefer a gentler initiation might first try a good tutorial like [70].

7.1 The concept of the integral

An *integral* is a finite accretion or sum of an infinite number of infinitesimal elements. This section introduces the concept.

Figure 7.1: Areas representing discrete sums.



7.1.1 An introductory example

Consider the sums

$$\begin{aligned}
S_1 &= \sum_{k=0}^{\text{0x10}-1} k, \\
S_2 &= \frac{1}{2}\sum_{k=0}^{\text{0x20}-1} \frac{k}{2}, \\
S_4 &= \frac{1}{4}\sum_{k=0}^{\text{0x40}-1} \frac{k}{4}, \\
S_8 &= \frac{1}{8}\sum_{k=0}^{\text{0x80}-1} \frac{k}{8}, \\
&\vdots \\
S_n &= \frac{1}{n}\sum_{k=0}^{(\text{0x10})n-1} \frac{k}{n}.
\end{aligned}$$

What do these sums represent? One way to think of them is in terms of the shaded areas of Fig. 7.1. In the figure, S_1 is composed[1] of several tall, thin

[1] If the reader does not fully understand this paragraph's illustration, if the relation of the sum to the area seems unclear, then the reader is urged to pause and consider the

rectangles of width 1 and height k; S_2, of rectangles of width 1/2 and height $k/2$. As n grows, the shaded region in the figure looks more and more like a triangle of base length $b = \mathrm{0x10}$ and height $h = \mathrm{0x10}$. In fact it appears that

$$\lim_{n\to\infty} S_n = \frac{bh}{2} = \mathrm{0x80},$$

or more tersely

$$S_\infty = \mathrm{0x80},$$

is the area the increasingly fine stairsteps approach.

Notice how we have evaluated S_∞, the sum of an infinite number of infinitely narrow rectangles, without actually adding anything up. We have taken a shortcut directly to the total.

In the equation

$$S_n = \frac{1}{n} \sum_{k=0}^{(\mathrm{0x10})n-1} \frac{k}{n},$$

let us now change the variables

$$\begin{aligned} \tau &\leftarrow \frac{k}{n}, \\ \Delta\tau &\leftarrow \frac{1}{n}, \end{aligned}$$

to obtain the representation

$$S_n = \Delta\tau \sum_{k=0}^{(\mathrm{0x10})n-1} \tau;$$

or more properly,

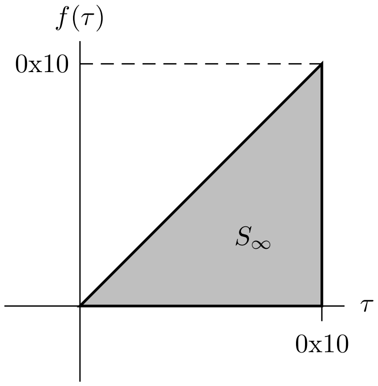
$$S_n = \sum_{k=0}^{(k|_{\tau=\mathrm{0x10}})-1} \tau\,\Delta\tau,$$

where the notation $k|_{\tau=\mathrm{0x10}}$ indicates the value of k when $\tau = \mathrm{0x10}$. Then

$$S_\infty = \lim_{\Delta\tau\to 0^+} \sum_{k=0}^{(k|_{\tau=\mathrm{0x10}})-1} \tau\,\Delta\tau,$$

illustration carefully until he does understand it. If it still seems unclear, then the reader should probably suspend reading here and go to study a good basic calculus text like [70]. The concept is important.

Figure 7.2: An area representing an infinite sum of infinitesimals. (Observe that the infinitesimal $d\tau$ is now too narrow to show on this scale. Compare against $\Delta\tau$ in Fig. 7.1.)



in which it is conventional as $\Delta\tau$ vanishes to change the symbol $d\tau \leftarrow \Delta\tau$, where $d\tau$ is the infinitesimal of chapter 4:

$$S_\infty = \lim_{d\tau\to 0^+} \sum_{k=0}^{(k|_{\tau=\mathrm{0x10}})-1} \tau \, d\tau.$$

The symbol $\lim_{d\tau\to 0^+} \sum_{k=0}^{(k|_{\tau=\mathrm{0x10}})-1}$ is cumbersome, so we replace it with the new symbol[2] $\int_0^{\mathrm{0x10}}$ to obtain the form

$$S_\infty = \int_0^{\mathrm{0x10}} \tau \, d\tau.$$

This means, "stepping in infinitesimal intervals of $d\tau$, the sum of all $\tau\, d\tau$ from $\tau = 0$ to $\tau = \mathrm{0x10}$." Graphically, it is the shaded area of Fig. 7.2.

[2] Like the Greek S, $\sum$, denoting discrete summation, the seventeenth century-styled Roman S, $\int$, stands for Latin "summa," English "sum." See [182, "Long s," 14:54, 7 April 2006].

7.1.2 Generalizing the introductory example

Now consider a generalization of the example of § 7.1.1:

$$S_n = \frac{1}{n} \sum_{k=an}^{bn-1} f\left(\frac{k}{n}\right).$$

(In the example of § 7.1.1, $f[\tau]$ was the simple $f[\tau] = \tau$, but in general it could be any function.) With the change of variables

$$\begin{aligned} \tau &\leftarrow \frac{k}{n}, \\ \Delta\tau &\leftarrow \frac{1}{n}, \end{aligned}$$

whereby

$$\begin{aligned} &k|_{\tau=a} = an, \\ &k|_{\tau=b} = bn, \\ &(k, n) \in \mathbb{Z}, \quad n \neq 0, \end{aligned}$$

(but a and b need not be integers), this is

$$S_n = \sum_{k=(k|_{\tau=a})}^{(k|_{\tau=b})-1} f(\tau)\, \Delta\tau.$$

In the limit,

$$S_\infty = \lim_{d\tau \to 0^+} \sum_{k=(k|_{\tau=a})}^{(k|_{\tau=b})-1} f(\tau)\, d\tau = \int_a^b f(\tau)\, d\tau.$$

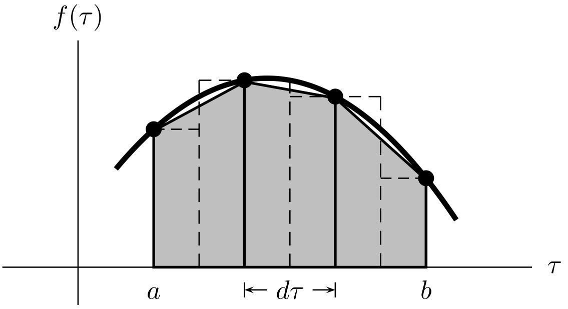
This is the *integral* of $f(\tau)$ in the interval $a < \tau < b$. It represents the area under the curve of $f(\tau)$ in that interval.

7.1.3 The balanced definition and the trapezoid rule

Actually, just as we have defined the derivative in the balanced form (4.8), we do well to define the integral in balanced form, too:

$$\int_a^b f(\tau)\, d\tau \equiv \lim_{d\tau \to 0^+} \left\{ \frac{f(a)\, d\tau}{2} + \sum_{k=(k|_{\tau=a})+1}^{(k|_{\tau=b})-1} f(\tau)\, d\tau + \frac{f(b)\, d\tau}{2} \right\}. \tag{7.1}$$

Figure 7.3: Integration by the trapezoid rule (7.1). Notice that the shaded and dashed areas total the same.



Here, the first and last integration samples are each balanced "on the edge," half within the integration domain and half without.

Equation (7.1) is known as the *trapezoid rule.* Figure 7.3 depicts it. The name "trapezoid" comes of the shapes of the shaded integration elements in the figure. Observe however that it makes no difference whether one regards the shaded trapezoids or the dashed rectangles as the actual integration elements; the total integration area is the same either way.[3] The important point to understand is that the integral is conceptually just a sum. It is a sum of an infinite number of infinitesimal elements as $d\tau$ tends to vanish, but a sum nevertheless; nothing more.

[3] The trapezoid rule (7.1) is perhaps the most straightforward, general, robust way to define the integral, but other schemes are possible, too. For example, taking the trapezoids in adjacent pairs—such that a pair enjoys not only a sample on each end but a third sample in the middle—one can for each pair fit a second-order curve $f(\tau) \approx (c_2)(\tau - \tau_{\text{middle}})^2 + (c_1)(\tau - \tau_{\text{middle}}) + c_0$ to the function, choosing the coefficients c_2, c_1 and c_0 to make the curve match the function exactly at the pair's three sample points; and then substitute the area under the pair's curve (an area which, by the end of § 7.4, we shall know how to calculate exactly) for the areas of the two trapezoids. Changing the symbol $\Delta\tau \leftarrow d\tau$ on one side of the equation to suggest coarse sampling, the result is the unexpectedly simple

$$\int_a^b f(\tau)\,\Delta\tau \approx \left[\frac{1}{3}f(a) + \frac{4}{3}f(a+\Delta\tau) + \frac{2}{3}f(a+2\,\Delta\tau) \right.$$
$$\left. + \frac{4}{3}f(a+3\,\Delta\tau) + \frac{2}{3}f(a+4\,\Delta\tau) + \cdots + \frac{4}{3}f(b-\Delta\tau) + \frac{1}{3}f(b)\right]\Delta\tau,$$

Nothing actually requires the integration element width $d\tau$ to remain constant from element to element, incidentally. Constant widths are usually easiest to handle but variable widths find use in some cases. The only requirement is that $d\tau$ remain infinitesimal. (For further discussion of the point, refer to the treatment of the Leibnitz notation in § 4.4.)

7.2 The antiderivative and the fundamental theorem of calculus

If

$$S(x) \equiv \int_a^x g(\tau)\,d\tau,$$

then what is the derivative dS/dx? After some reflection, one sees that the derivative must be

$$\frac{dS}{dx} = g(x).$$

This is so because the action of the integral is to compile or accrete the area under a curve. The integral accretes area at a rate proportional to the curve's height $f(\tau)$: the higher the curve, the faster the accretion. In this way one sees that the integral and the derivative are inverse operators; the one inverts the other. The integral is the *antiderivative.*

More precisely,

$$\int_a^b \frac{df}{d\tau}\,d\tau = f(\tau)|_a^b, \tag{7.2}$$

where the notation $f(\tau)|_a^b$ or $[f(\tau)]_a^b$ means $f(b) - f(a)$.

as opposed to the trapezoidal

$$\int_a^b f(\tau)\,\Delta\tau \approx \bigg[\frac{1}{2}f(a) + f(a+\Delta\tau) + f(a+2\,\Delta\tau) + f(a+3\,\Delta\tau) + f(a+4\,\Delta\tau) + \cdots + f(b-\Delta\tau) + \frac{1}{2}f(b)\bigg]\Delta\tau$$

implied by (7.1). The curved scheme is called *Simpson's rule.* It is clever and well known.

Simpson's rule had real uses in the slide-rule era when, for practical reasons, one preferred to let $\Delta\tau$ be sloppily large, sampling a curve only a few times to estimate its integral; yet the rule is much less useful when a computer is available to do the arithmetic over an adequate number of samples. At best Simpson's rule does not help much with a computer; at worst it can yield spurious results; and because it is easy to program it tends to encourage thoughtless application. Other than in the footnote you are reading, Simpson's rule is not covered in this book.

The importance of (7.2), fittingly named the *fundamental theorem of calculus*,[4] can hardly be overstated. As the formula which ties together the complementary pair of questions asked at the chapter's start, (7.2) is of utmost importance in the practice of mathematics. The idea behind the formula is indeed simple once grasped, but to grasp the idea firmly in the first place is not entirely trivial.[5] The idea is simple but big. The reader is urged to pause now and ponder the formula thoroughly until he feels reasonably confident that indeed he does grasp it and the important idea it represents. One is unlikely to do much higher mathematics without this formula.

As an example of the formula's use, consider that because $(d/d\tau)(\tau^3/6) = \tau^2/2$, it follows that

$$\int_2^x \frac{\tau^2\,d\tau}{2} = \int_2^x \frac{d}{d\tau}\left(\frac{\tau^3}{6}\right)d\tau = \left.\frac{\tau^3}{6}\right|_2^x = \frac{x^3-8}{6}.$$

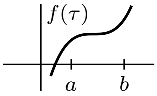
Gathering elements from (4.16) and from Tables 5.2 and 5.3, Table 7.1 lists a handful of the simplest, most useful derivatives for antiderivative use. Section 9.1 speaks further of the antiderivative.

[4] [70, § 11.6][146, § 5-4][182, "Fundamental theorem of calculus," 06:29, 23 May 2006]

[5] Having read from several calculus books and, like millions of others perhaps including the reader, having sat years ago in various renditions of the introductory calculus lectures in school, the author has never yet met a more convincing demonstration of (7.2) than the formula itself. Somehow the underlying idea is too simple, too profound to explain. It's like trying to explain how to drink water, or how to count or to add. Elaborate explanations and their attendant constructs and formalities are indeed possible to contrive, but the idea itself is so simple that somehow such contrivances seem to obscure the idea more than to reveal it.

One ponders the formula (7.2) a while, then the idea dawns on him.

If you want some help pondering, try this: sketch some arbitrary function $f(\tau)$ on a set of axes at the bottom of a piece of paper—some squiggle of a curve like



will do nicely—then on a separate set of axes directly above the first, sketch the corresponding slope function $df/d\tau$. Mark two points a and b on the common horizontal axis; then on the upper, $df/d\tau$ plot, shade the integration area under the curve. Now consider (7.2) in light of your sketch.

There. Does the idea not dawn?

Another way to see the truth of the formula begins by canceling its $(1/d\tau)\,d\tau$ to obtain the form $\int_{\tau=a}^{b} df = f(\tau)|_a^b$. If this way works better for you, fine; but make sure that you understand it the other way, too.

Table 7.1: Basic derivatives for the antiderivative.

$$\int_a^b \frac{df}{d\tau}\,d\tau = f(\tau)|_a^b$$

$$\begin{aligned}
\tau^{a-1} &= \frac{d}{d\tau}\left(\frac{\tau^a}{a}\right), & a &\neq 0\\
\frac{1}{\tau} &= \frac{d}{d\tau}\ln\tau, & \ln 1 &= 0\\
\exp\tau &= \frac{d}{d\tau}\exp\tau, & \exp 0 &= 1\\
\cos\tau &= \frac{d}{d\tau}\sin\tau, & \sin 0 &= 0\\
\sin\tau &= \frac{d}{d\tau}\left(-\cos\tau\right), & \cos 0 &= 1
\end{aligned}$$

7.3 Operators, linearity and multiple integrals

This section presents the operator concept, discusses linearity and its consequences, treats the commutivity of the summational and integrodifferential operators, and introduces the multiple integral.

7.3.1 Operators

An *operator* is a mathematical agent that combines several values of a function.

Such a definition, unfortunately, is extraordinarily unilluminating to those who do not already know what it means. A better way to introduce the operator is by giving examples. Operators include $+$, $-$, multiplication, division, $\sum$, $\prod$, $\int$ and ∂. The essential action of an operator is to take several values of a function and combine them in some way. For example, $\prod$ is an operator in

$$\prod_{j=1}^{5}(2j-1) = (1)(3)(5)(7)(9) = \mathrm{0x3B1}.$$

Notice that the operator has acted to remove the variable j from the expression $2j-1$. The j appears on the equation's left side but not on its right. The operator has used the variable up. Such a variable, used up by an operator, is a *dummy variable,* as encountered earlier in § 2.3.

7.3.2 A formalism

But then how are + and − operators? They don't use any dummy variables up, do they?

Well, that depends on how you look at it. Consider the sum $S = 3 + 5$. One can write this as

$$S = \sum_{k=0}^{1} f(k),$$

where

$$f(k) \equiv \begin{cases} 3 & \text{if } k = 0, \\ 5 & \text{if } k = 1, \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Then,

$$S = \sum_{k=0}^{1} f(k) = f(0) + f(1) = 3 + 5 = 8.$$

By such admittedly excessive formalism, the + operator can indeed be said to use a dummy variable up. The point is that + is in fact an operator just like the others.

Another example of the kind:

$$\begin{aligned} D &= g(z) - h(z) + p(z) + q(z) \\ &= g(z) - h(z) + p(z) - 0 + q(z) \\ &= \Phi(0,z) - \Phi(1,z) + \Phi(2,z) - \Phi(3,z) + \Phi(4,z) \\ &= \sum_{k=0}^{4} (-)^k \Phi(k,z), \end{aligned}$$

where

$$\Phi(k,z) \equiv \begin{cases} g(z) & \text{if } k = 0, \\ h(z) & \text{if } k = 1, \\ p(z) & \text{if } k = 2, \\ 0 & \text{if } k = 3, \\ q(z) & \text{if } k = 4, \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Such unedifying formalism is essentially useless in applications, except as a vehicle for definition. Once you understand why + and − are operators just as $\sum$ and $\int$ are, you can forget the formalism. It doesn't help much.

7.3.3 Linearity

A function $f(z)$ is *linear* iff (if and only if) it has the properties

$$\begin{aligned} f(z_1 + z_2) &= f(z_1) + f(z_2), \\ f(\alpha z) &= \alpha f(z), \\ f(0) &= 0. \end{aligned}$$

The functions $f(z) = 3z$, $f(u, v) = 2u - v$ and $f(z) = 0$ are examples of linear functions. Nonlinear functions include[6] $f(z) = z^2$, $f(u, v) = \sqrt{uv}$, $f(t) = \cos\omega t$, $f(z) = 3z + 1$ and even $f(z) = 1$.

An operator L is linear iff it has the properties

$$\begin{aligned} L(f_1 + f_2) &= Lf_1 + Lf_2, \\ L(\alpha f) &= \alpha Lf, \\ L(0) &= 0. \end{aligned}$$

The operators $\sum$, $\int$, $+$, $-$ and ∂ are examples of linear operators. For instance,[7]

$$\frac{d}{dz}[f_1(z) + f_2(z)] = \frac{df_1}{dz} + \frac{df_2}{dz}.$$

Nonlinear operators include multiplication, division and the various trigonometric functions, among others.

Section 16.1.2 will have more to say about operators and their notation.

7.3.4 Summational and integrodifferential commutivity

Consider the sum

$$S_1 = \sum_{k=a}^{b} \left[\sum_{j=p}^{q} \frac{x^k}{j!} \right].$$

[6] If $3z + 1$ is a *linear expression,* then how is not $f(z) = 3z + 1$ a *linear function?* Answer: the matter is a matter partly of purposeful definition, partly of semantics. The equation $y = 3x + 1$ plots a line, so the expression $3z + 1$ is literally “linear” in this sense; but the definition has more purpose to it than merely this. When you see the linear expression $3z+1$, think $3z+1 = 0$, then $g(z) = 3z = -1$. The $g(z) = 3z$ is linear; the -1 is the constant value it targets. That’s the sense of it.

[7] You don’t see d in the list of linear operators? But d in this context is really just another way of writing ∂, so, yes, d is linear, too. See § 4.4.

This is a sum of the several values of the expression $x^k/j!$, evaluated at every possible pair (j,k) in the indicated domain. Now consider the sum

$$S_2 = \sum_{j=p}^{q}\left[\sum_{k=a}^{b}\frac{x^k}{j!}\right].$$

This is evidently a sum of the same values, only added in a different order. Apparently $S_1 = S_2$. Reflection along these lines must soon lead the reader to the conclusion that, in general,

$$\sum_k\sum_j f(j,k) = \sum_j\sum_k f(j,k).$$

Now consider that an integral is just a sum of many elements, and that a derivative is just a difference of two elements. Integrals and derivatives must then have the same commutative property discrete sums have. For example,

$$\begin{aligned}\int_{v=-\infty}^{\infty}\int_{u=a}^{b} f(u,v)\,du\,dv &= \int_{u=a}^{b}\int_{v=-\infty}^{\infty} f(u,v)\,dv\,du;\\ \int\sum_k f_k(v)\,dv &= \sum_k\int f_k(v)\,dv;\\ \frac{\partial}{\partial v}\int f\,du &= \int\frac{\partial f}{\partial v}\,du.\end{aligned}$$

In general,

$$L_vL_uf(u,v) = L_uL_vf(u,v), \tag{7.3}$$

where L is any of the linear operators $\sum$, $\int$ or ∂.

Some convergent summations, like

$$\sum_{k=0}^{\infty}\sum_{j=0}^{1}\frac{(-)^j}{2k+j+1},$$

diverge once reordered, as

$$\sum_{j=0}^{1}\sum_{k=0}^{\infty}\frac{(-)^j}{2k+j+1}.$$

One cannot blithely swap operators here. This is not because swapping is wrong, but rather because the inner sum after the swap diverges, whence the outer sum after the swap has no concrete summand on which to work. (*Why* does the inner sum after the swap diverge? Answer: $1 + 1/3 + 1/5 + \cdots = [1] + [1/3 + 1/5] + [1/7 + 1/9 + 1/\mathrm{0xB} + 1/\mathrm{0xD}] + \cdots > 1[1/4] + 2[1/8] + 4[1/\mathrm{0x10}] + \cdots = 1/4 + 1/4 + 1/4 + \cdots$. See also § 8.10.5.) For a more twisted example of the same phenomenon, consider[8]

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots = \left(1 - \frac{1}{2} - \frac{1}{4}\right) + \left(\frac{1}{3} - \frac{1}{6} - \frac{1}{8}\right) + \cdots,$$

which associates two negative terms with each positive, but still seems to omit no term. Paradoxically, then,

$$\begin{aligned}1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots &= \left(\frac{1}{2} - \frac{1}{4}\right) + \left(\frac{1}{6} - \frac{1}{8}\right) + \cdots \\ &= \frac{1}{2} - \frac{1}{4} + \frac{1}{6} - \frac{1}{8} + \cdots \\ &= \frac{1}{2}\left(1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots\right),\end{aligned}$$

or so it would seem, but cannot be, for it claims falsely that the sum is half itself. A better way to have handled the example might have been to write the series as

$$\lim_{n\to\infty}\left\{1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots + \frac{1}{2n-1} - \frac{1}{2n}\right\}$$

in the first place, thus explicitly specifying equal numbers of positive and negative terms.[9] So specifying would have forestalled the error. In the earlier example,

$$\lim_{n\to\infty}\sum_{k=0}^{n}\sum_{j=0}^{1}\frac{(-)^j}{2k+j+1}$$

[8] [3, § 1.2.3]

[9] Some students of pure mathematics would assert that the false conclusion had been reached through lack of rigor. Well, maybe. This writer however does not feel sure that *rigor* is quite the right word for what was lacking here. Pure mathematics does bring an elegant notation and a set of formalisms which serve ably to spotlight certain limited kinds of blunders, but these are blunders no less by the applied approach. The stalwart Leonhard Euler—arguably the greatest series-smith in mathematical history—wielded his heavy analytical hammer in thunderous strokes before modern professional mathematics had conceived the notation or the formalisms. If the great Euler did without, then you and I might not always be forbidden to follow his robust example. See also footnote 11.

On the other hand, the professional approach to pure mathematics is worth study if you have the time. Recommended introductions include [100], preceded if necessary by [70] and/or [3, chapter 1].

would likewise have forestalled the error, or at least have made the error explicit.

The *conditional convergence*[10] of the last paragraph, which can occur in integrals as well as in sums, seldom poses much of a dilemma in practice. One can normally swap summational and integrodifferential operators with little worry. The reader however should at least be aware that conditional convergence troubles can arise where a summand or integrand varies in sign or phase.

7.3.5 Multiple integrals

Consider the function

$$f(u, w) = \frac{u^2}{w}.$$

Such a function would not be plotted as a curved line in a plane, but rather as a curved *surface* in a three-dimensional space. Integrating the function seeks not the area under the curve but rather the volume under the surface:

$$V = \int_{u_1}^{u_2} \int_{w_1}^{w_2} \frac{u^2}{w} \, dw \, du.$$

This is a *double integral.* Inasmuch as it can be written in the form

$$\begin{aligned} V &= \int_{u_1}^{u_2} g(u) \, du, \\ g(u) &\equiv \int_{w_1}^{w_2} \frac{u^2}{w} \, dw, \end{aligned}$$

its effect is to cut the area under the surface into flat, upright slices, then the slices crosswise into tall, thin towers. The towers are integrated over w to constitute the slice, then the slices over u to constitute the volume.

In light of § 7.3.4, evidently nothing prevents us from swapping the integrations: u first, then w. Hence

$$V = \int_{w_1}^{w_2} \int_{u_1}^{u_2} \frac{u^2}{w} \, du \, dw.$$

And indeed this makes sense, doesn't it? What difference should it make whether we add the towers by rows first then by columns, or by columns first then by rows? The total volume is the same in any case—albeit the integral

[10] [100, § 16]

over w is potentially ill-behaved[11] near $w = 0$; so that, if for instance w_1 were negative, w_2 were positive, and both were real, one might rather write the double integral as[12]

$$V = \lim_{\epsilon \to 0^+} \left(\int_{w_1}^{-\epsilon} + \int_{+\epsilon}^{w_2} \right) \int_{u_1}^{u_2} \frac{u^2}{w} \, du \, dw.$$

Double integrations arise very frequently in applications. Triple integrations arise about as often. For instance, if $\mu(\mathbf{r}) = \mu(x, y, z)$ represents the position-dependent mass density of some soil,[13] then the total soil mass in some rectangular volume is

$$M = \int_{x_1}^{x_2} \int_{y_1}^{y_2} \int_{z_1}^{z_2} \mu(x, y, z) \, dz \, dy \, dx.$$

As a concise notational convenience, the last is likely to be written,

$$M = \int_V \mu(\mathbf{r}) \, d\mathbf{r},$$

where the V stands for "volume" and is understood to imply a triple integration. Similarly for the double integral,

$$Q = \int_S f(\boldsymbol{\rho}) \, d\boldsymbol{\rho},$$

where the S stands for "surface" and is understood to imply a double integration.

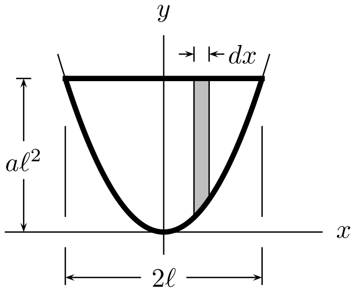
Even more than three nested integrations are possible. If we integrated over time as well as space, the integration would be fourfold. A spatial Fourier transform (§ 19.6) implies a triple integration; and its inverse, another triple: a sixfold integration altogether. Manifold nesting of integrals is thus not just a theoretical mathematical topic; it arises in sophisticated real-world engineering models. The topic concerns us here for this reason.

[11] A great deal of ink is spilled in the applied mathematical literature when summations and/or integrations are interchanged. The author tends to recommend saving the ink, for pure and applied mathematics want different styles. What usually matters in applications is not whether a particular summation or integration satisfies some formal test but rather whether one clearly understands the summand to be summed or the integrand to be integrated. See also footnote 9.

[12] It is interesting to consider the effect of withdrawing the integral's limit at $-\epsilon$ to -2ϵ, as $\lim_{\epsilon \to 0^+} \left(\int_{w_1}^{-2\epsilon} + \int_{+\epsilon}^{w_2} \right) \int_{u_1}^{u_2} \frac{u^2}{w} \, du \, dw$; for, surprisingly—despite that the parameter ϵ is vanishing anyway—the withdrawal does alter the integral unless the limit at $+\epsilon$ also is withdrawn. The reason is that $\lim_{\epsilon \to 0^+} \int_{\epsilon}^{2\epsilon} (1/w) \, dw = \ln 2 \neq 0$.

[13] Conventionally, the Greek letter ρ rather than μ is used for density. However, it happens that we shall need the letter ρ for a different purpose later in the paragraph.

Figure 7.4: The area within a parabola.



7.4 Areas and volumes

By composing and solving appropriate integrals, one can calculate the perimeters, areas and volumes of interesting common shapes and solids.

7.4.1 The area within a parabola

Figure 7.4 depicts an element of the area within the parabola

$$y = ax^2. \tag{7.4}$$

The element—which is shaded in the figure—is nearly rectangular (or, if you prefer, nearly trapezoidal), and indeed is the more nearly rectangular (trapezoidal) the smaller dx is made to be. Evidently, therefore, the element's area is $A_{\mathrm{rectangle}} = (a\ell^2 - ax^2)\,dx$. Integrating the many rectangles, we find the area within the parabola to be

$$\begin{aligned} A_{\mathrm{parabola}} &= \int_{x=-\ell}^{\ell} A_{\mathrm{rectangle}} = \int_{-\ell}^{\ell} (a\ell^2 - ax^2)\,dx \\ &= a\ell^2 \int_{-\ell}^{\ell} dx - a \int_{-\ell}^{\ell} x^2\,dx \\ &= a\ell^2 \Big[x \Big]_{x=-\ell}^{\ell} - a \left[\frac{x^3}{3} \right]_{x=-\ell}^{\ell}. \end{aligned}$$

But $[x]_{x=-\ell}^{\ell} = (\ell) - (-\ell) = 2\ell$ and $[x^3/3]_{x=-\ell}^{\ell} = (\ell)^3/3 - (-\ell)^3/3 = 2\ell^3/3$, so

$$A_{\mathrm{parabola}} = \frac{4a\ell^3}{3}. \tag{7.5}$$

7.4.2 The length of a parabola

Section 7.4.1 has computed the area within the parabola of Fig. 7.4, but what if one wishes instead to compute the parabola's *length?* According to Pythagoras, (1.1),

$$(ds)^2 = (dx)^2 + (dy)^2,$$

the ds being an element of the curve's length. Taking the derivative of (7.4),

$$dy = 2ax\,dx,$$

so

$$(ds)^2 = (dx)^2 + (2ax\,dx)^2,$$

or, solving for ds,

$$ds = dx\,\sqrt{1 + (2ax)^2}.$$

Integrating,

$$s = \int_{x=-\ell}^{\ell} ds = \int_{-\ell}^{\ell} dx\,\sqrt{1 + (2ax)^2} = 2\int_{0}^{\ell} dx\,\sqrt{1 + (2ax)^2},$$

the last step of which observes that, symmetrically, the parabola's left half does not differ in length from its right. Defining

$$u \equiv 2ax,$$

whose derivative is

$$du = 2a\,dx,$$

permits s to be expressed in a slightly simpler form,

$$s = \frac{1}{a}\int_{0}^{2a\ell} du\,\sqrt{1 + u^2}, \tag{7.6}$$

but after this it is not obvious what should be done next.

Several techniques can be tried, most of which however seem to fail against an integrand like $\sqrt{1 + u^2}$. Paradoxically, a modified integrand like $u\sqrt{1 + u^2}$, which looks more complicated, would have been easier to handle, for the technique of[14] § 9.2 would have resolved it neatly; whereas neither the technique of § 9.2 nor any of several others seems to make headway against

[14] This is a forward reference. It is given for information. You need not follow it for now, for the present section's logic does not depend on it.

the simpler-looking $\sqrt{1+u^2}$. Nevertheless, after some trial and error, one may recall Table 5.3 which affords a clue. From the table,[15]

$$\frac{d}{du}\operatorname{arcsinh} u = \frac{1}{\sqrt{1+u^2}}, \tag{7.7}$$

the right side of which resembles our integrand $\sqrt{1+u^2}$. To the above derivative we can append two more,

$$\begin{aligned} \frac{d}{du}\sqrt{1+u^2} &= \frac{u}{\sqrt{1+u^2}}, & (7.8)\\ \frac{d}{du}u\sqrt{1+u^2} &= \sqrt{1+u^2}+\frac{u^2}{\sqrt{1+u^2}}, & (7.9)\end{aligned}$$

computed by the chain and product rules of § 4.5. The last includes the expression $u^2/\sqrt{1+u^2}$ which, after adding and subtracting $1/\sqrt{1+u^2}$, one can alternately write as $u^2/\sqrt{1+u^2} = (1+u^2)/\sqrt{1+u^2} - 1/\sqrt{1+u^2} = \sqrt{1+u^2} - 1/\sqrt{1+u^2}$. Thus,

$$\frac{d}{du}u\sqrt{1+u^2} = 2\sqrt{1+u^2} - \frac{1}{\sqrt{1+u^2}}. \tag{7.10}$$

Not all the derivatives of this paragraph turn out to be useful to the present problem but (7.7) and (7.10) do. The average of those two is

$$\frac{d}{du}\left(\frac{u\sqrt{1+u^2}+\operatorname{arcsinh} u}{2}\right) = \sqrt{1+u^2}, \tag{7.11}$$

whose right side matches our integrand.

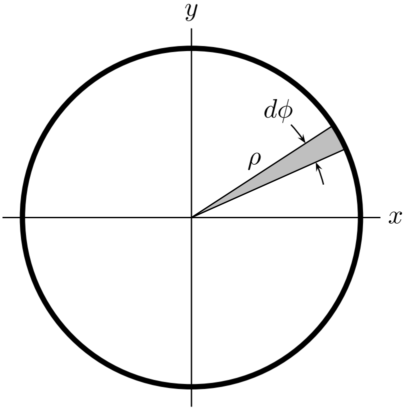
Why should one care that the right side of the average (7.11) matches our integrand? Because the match lets one use the fundamental theorem of calculus, (7.2) of § 7.2. Applying the average, according to the fundamental theorem, to (7.6),

$$s = \frac{1}{a}\left[\frac{u\sqrt{1+u^2}+\operatorname{arcsinh} u}{2}\right]_{u=0}^{2a\ell}.$$

[15] The table's relevant entry includes a ± sign but only the + sign interests us here. Why only? Because we shall have to choose one branch or the other of the hyperbolic arcsine along which to work whether we will or nill. Nothing prevents us from choosing the positive branch.

You can try the negative branch if you wish. After some signs have canceled, you will find that the negative branch arrives at the same result.

Figure 7.5: The area of a circle.



Evaluating,

$$s = \ell\sqrt{1 + (2a\ell)^2} + \frac{1}{2a} \operatorname{arcsinh} 2a\ell, \tag{7.12}$$

or, if you prefer, expanding arcsinh(·) according to Table 5.1,

$$s = \ell\sqrt{1 + (2a\ell)^2} + \frac{1}{2a} \ln\left[2a\ell + \sqrt{1 + (2a\ell)^2}\right] \tag{7.13}$$

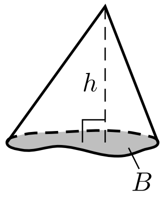
(in which we have chosen the + sign for the table's ± because the − sign would have returned a complex length). This is the parabola's length, measured along its curve.

That wasn't so easy. If you are not sure that you have followed it, that's all right, for you can return to study it again later. Also, you can learn more about the parabola in § 15.7.1 and, in § 9.8, more too about the technique by which the present subsection has evaluated (7.6). Meanwhile, fortunately, the next subsection will be easier and also more interesting. It computes the area of a circle.

7.4.3 The area of a circle

Figure 7.5 depicts an element of a circle's area. The element has wedge shape, but inasmuch as the wedge is infinitesimally narrow, the wedge is indistinguishable from a triangle of base length $\rho\, d\phi$ and height ρ. The area

Figure 7.6: The volume of a cone.



of such a triangle is $A_{\text{triangle}} = \rho^2\,d\phi/2$. Integrating the many triangles, we find the circle's area to be

$$A_{\text{circle}} = \int_{\phi=-\pi}^{\pi} A_{\text{triangle}} = \int_{-\pi}^{\pi} \frac{\rho^2\,d\phi}{2} = \left.\frac{\rho^2\phi}{2}\right|_{\phi=-\pi}^{\pi}.$$

Evaluated,

$$A_{\text{circle}} = \frac{2\pi\rho^2}{2}. \tag{7.14}$$

(The numerical value of 2π—the circumference or perimeter of the unit circle—we have not calculated yet. We will calculate it in § 8.11. The reason to write $2\pi/2$ rather than the deceptively simpler-looking π is that the symbol π alone obscures the sense in which the circle resembles a rolled-up triangle. See appendix A. Sometimes the book uses the symbol π alone, anyway, just to reduce visual clutter; but that an alternate symbol like[16] $\pi\!\!\!\pi = 2\pi$ is not current is unfortunate. If such a symbol were current, then we could have written that $A_{\text{circle}} = \pi\!\!\!\pi\rho^2/2$.)

7.4.4 The volume of a cone

One can calculate the volume of any cone (or pyramid) if one knows its base area B and its altitude h measured normally[17] to the base. Refer to Fig. 7.6. A cross-section of a cone, cut parallel to the cone's base, has the same shape the base has but a different scale. If coordinates are chosen such that the altitude h runs in the $\hat{\mathbf{z}}$ direction with $z = 0$ at the cone's vertex,

[16] The symbol $\pi\!\!\!\pi$ has no name of which the writer is aware. One might provisionally call it "palais" after the mathematician who has suggested it. [124]

[17] *Normally* here means "at right angles."

then the cross-sectional area is evidently[18] $(B)(z/h)^2$. For this reason, the cone's volume is

$$V_{\text{cone}} = \int_0^h (B)\left(\frac{z}{h}\right)^2 dz = \frac{B}{h^2}\int_0^h z^2\,dz = \frac{B}{h^2}\left[\frac{z^3}{3}\right]_{z=0}^h = \frac{B}{h^2}\left(\frac{h^3}{3}\right)$$

Evaluating,

$$V_{\text{cone}} = \frac{Bh}{3}. \tag{7.15}$$

7.4.5 The surface area and volume of a sphere

Of a sphere, Fig. 7.7, one wants to calculate both the surface area and the volume. For the surface area, the sphere's surface is sliced vertically down the z axis into narrow, constant-ϕ, tapered strips (each strip broadest at the sphere's equator, tapering to points at the sphere's $\pm z$ poles) and horizontally across the z axis into narrow, constant-θ rings, as in Fig. 7.8. A surface element so produced (seen as shaded in the latter figure) evidently has the area[19]

$$dS = (r\,d\theta)(\rho\,d\phi) = r^2 \sin\theta\,d\theta\,d\phi.$$

[18] The fact may admittedly not be evident to the reader at first glance. If it is not yet evident to you, then ponder Fig. 7.6 a moment. Consider what it means to cut parallel to a cone's base a cross-section of the cone, and how cross-sections cut nearer a cone's vertex are smaller though the same shape. What if the base were square? Would the cross-sectional area not be $(B)(z/h)^2$ in that case? What if the base were a right triangle with equal legs—in other words, half a square? What if the base were some other strange shape like the base depicted in Fig. 7.6? Could such a strange shape not also be regarded as a definite, well-characterized part of a square? (With a pair of scissors one can cut any shape from a square piece of paper, after all.) Thinking along such lines must soon lead one to the insight that the parallel-cut cross-sectional area of a cone can be nothing other than $(B)(z/h)^2$, regardless of the base's shape.

[19] It can be shown, incidentally—the details are left as an exercise—that $dS = -r\,dz\,d\phi$ also. The subsequent integration arguably goes a little easier if dS is accepted in this mildly clever form. The form is interesting in any event if one visualizes the specific, annular area the expression $\int_{\phi=-\pi}^{\pi} dS = -2\pi r\,dz$ represents: evidently, unexpectedly, an equal portion of the sphere's surface corresponds to each equal step along the z axis, pole to pole; so, should you slice an unpeeled apple into parallel slices of equal thickness, though some slices will be bigger across and thus heavier than others, each slice curiously must take an equal share of the apple's skin. (This is true, anyway, if you judge Fig. 7.8 to represent an apple. The author's children judge it to represent "the Death Star with a square gun" [113], so maybe it depends on your point of view.)

Figure 7.7: A sphere.

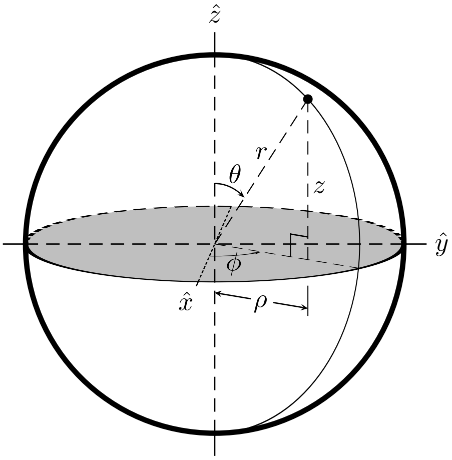
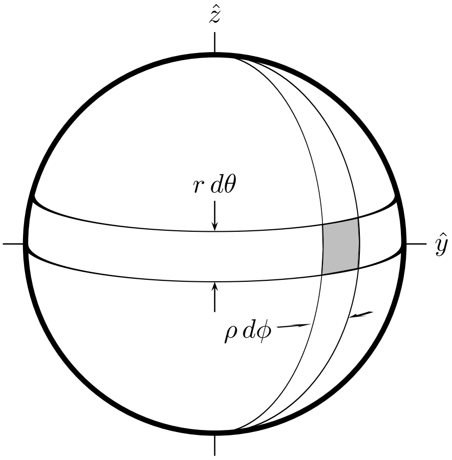


Figure 7.8: An element of the sphere's surface (see Fig. 7.7).



The sphere's total surface area then is the sum of all such elements over the sphere's entire surface:

$$
\begin{aligned}
S_{\text{sphere}} &= \int_{\phi=-\pi}^{\pi} \int_{\theta=0}^{\pi} dS \\
&= \int_{\phi=-\pi}^{\pi} \int_{\theta=0}^{\pi} r^2 \sin\theta \, d\theta \, d\phi \\
&= r^2 \int_{\phi=-\pi}^{\pi} [-\cos\theta]_0^{\pi} \, d\phi \\
&= r^2 \int_{\phi=-\pi}^{\pi} [2] \, d\phi \\
&= 4\pi r^2, \qquad (7.16)
\end{aligned}
$$

where we have used the fact from Table 7.1 that $\sin\tau = (d/d\tau)(-\cos\tau)$.

Having computed the sphere's surface area, one can find its volume just as § 7.4.3 has found a circle's area—except that instead of dividing the circle into many narrow triangles, one divides the sphere into many narrow *cones,* each cone with base area dS and altitude r, with the vertices of all the cones meeting at the sphere's center. Per (7.15), the volume of one such cone is $V_{\text{cone}} = r\,dS/3$. Hence,

$$
V_{\text{sphere}} = \oint_S V_{\text{cone}} = \oint_S \frac{r\,dS}{3} = \frac{r}{3}\oint_S dS = \frac{r}{3} S_{\text{sphere}},
$$

where the useful symbol

$$
\oint_S
$$

indicates *integration over a closed surface.* In light of (7.16), the total volume is

$$
V_{\text{sphere}} = \frac{4\pi r^3}{3}. \qquad (7.17)
$$

(One can compute the same spherical volume more prosaically, without reference to cones, by writing $dV = r^2 \sin\theta \, dr \, d\theta \, d\phi$ then integrating $\int_V dV$. The derivation given above, however, is preferred because it lends the additional insight that a sphere can sometimes be viewed as a great cone rolled up about its own vertex. The circular area derivation of § 7.4.3 lends an analogous insight: that a circle can sometimes be viewed as a great triangle rolled up about *its* own vertex.)

7.5 Checking an integration

Dividing 0x46B/0xD = 0x57 with a pencil, how does one check the result?[20] Answer: by multiplying (0x57)(0xD) = 0x46B. Multiplication inverts division. Easier than division, multiplication provides a quick, reliable check.

Likewise, integrating

$$\int_a^b \frac{\tau^2}{2}\,d\tau = \frac{b^3 - a^3}{6}$$

with a pencil, how does one check the result? Answer: by differentiating

$$\left[\frac{\partial}{\partial b}\left(\frac{b^3 - a^3}{6}\right)\right]_{b=\tau} = \frac{\tau^2}{2}.$$

Differentiation inverts integration. Easier than integration, differentiation like multiplication provides a quick, reliable check.

More formally, according to (7.2),

$$S \equiv \int_a^b \frac{df}{d\tau}\,d\tau = f(b) - f(a). \tag{7.18}$$

Differentiating (7.18) with respect to b and a,

$$\begin{aligned} \left.\frac{\partial S}{\partial b}\right|_{b=\tau} &= \frac{df}{d\tau}, \\ \left.\frac{\partial S}{\partial a}\right|_{a=\tau} &= -\frac{df}{d\tau}. \end{aligned} \tag{7.19}$$

Either line of (7.19) can be used to check an integration. Evaluating (7.18) at $b = a$ yields that

$$S|_{b=a} = 0, \tag{7.20}$$

which can be used to check further.[21]

As useful as (7.19) and (7.20) are, they nevertheless serve only integrals with variable limits. They are of little use to check *definite integrals*

[20] Admittedly, few readers will ever have done much such multidigit *hexadecimal* arithmetic with a pencil, but, hey, go with it. In decimal, it's 1131/13 = 87.

Actually, hexadecimal is just proxy for binary (see appendix A), and long division in straight binary is kind of fun. If you have never tried it, you might. It is simpler than decimal or hexadecimal division, and it's how computers divide. The insight gained is worth the trial.

[21] Using (7.20) to check the example, $(b^3 - a^3)/6|_{b=a} = 0$.

like (9.18) below, which lack variable limits to differentiate. However, many or most integrals one meets in practice have or can be given variable limits. Equations (7.19) and (7.20) do serve such *indefinite integrals.*

It is a rare irony of mathematics that, though numerically differentiation is indeed harder than integration, analytically the opposite is true. Analytically, differentiation is the easier. So far, mostly, the integrals the book has introduced have been easy ones (§ 7.4.2 excepted), but chapter 9 will bring harder ones. Even experienced mathematicians are apt to err in analyzing these. Reversing an integration by taking a relatively easy derivative is thus an excellent way to check a hard-earned integration result.

7.6 Contour integration

To this point we have considered only integrations in which the variable of integration advances in a straight line from one point to another: for instance, $\int_a^b f(\tau)\,d\tau$, in which the function $f(\tau)$ is evaluated at $\tau = a, a + d\tau, a + 2d\tau, \ldots, b$. The integration variable is a real-valued scalar which can do nothing but make a straight line from a to b.

Such is not the case when the integration variable is a vector. Consider the integral

$$S = \int_{\mathbf{r}=\hat{\mathbf{x}}\rho}^{\hat{\mathbf{y}}\rho} (x^2 + y^2)\,d\ell,$$

where $d\ell$ is the infinitesimal length of a step along the path of integration. What does this integral mean? Does it mean to integrate from $\mathbf{r} = \hat{\mathbf{x}}\rho$ to $\mathbf{r} = 0$, then from there to $\mathbf{r} = \hat{\mathbf{y}}\rho$? Or does it mean to integrate along the arc of Fig. 7.9? The two paths of integration begin and end at the same points, but they differ in between, and the integral certainly does not come out the same both ways. Yet many other paths of integration from $\hat{\mathbf{x}}\rho$ to $\hat{\mathbf{y}}\rho$ are possible, not just these two.

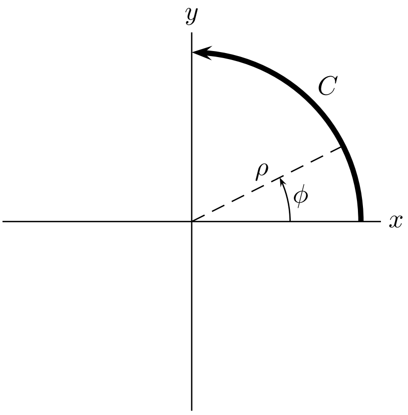
Because multiple paths are possible, we must be more specific:

$$S = \int_C (x^2 + y^2)\,d\ell,$$

where C stands for "contour" and means in this example the specific contour of Fig. 7.9. In the example, $x^2 + y^2 = \rho^2$ (by the Pythagorean theorem) and $d\ell = \rho\,d\phi$, so

$$S = \int_C \rho^2\,d\ell = \int_0^{2\pi/4} \rho^3\,d\phi = \frac{2\pi}{4}\rho^3.$$

Figure 7.9: A contour of integration.



In the example the contour is open, but closed contours which begin and end at the same point are also possible, indeed common. The useful symbol

$$\oint$$

indicates *integration over a closed contour.* It means that the contour ends where it began: the loop is closed. The contour of Fig. 7.9 would be closed, for instance, if it continued to $\mathbf{r} = 0$ and then back to $\mathbf{r} = \hat{\mathbf{x}}\rho$.

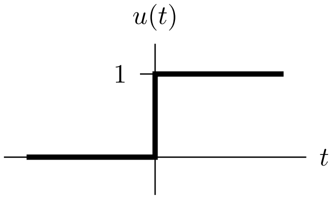
Besides applying where the variable of integration is a vector, contour integration applies equally where the variable of integration is a complex scalar. In the latter case some interesting mathematics emerge, as we shall see in §§ 8.8 and 9.6.

7.7 Discontinuities

The polynomials and trigonometrics studied to this point in the book offer flexible means to model many physical phenomena of interest, but one thing they do not model gracefully is the simple discontinuity. Consider a mechanical valve opened at time $t = t_o$. The flow $x(t)$ past the valve is

$$x(t) = \begin{cases} 0, & t < t_o; \\ x_o, & t > t_o. \end{cases}$$

Figure 7.10: The Heaviside unit step $u(t)$.



One can write this more concisely in the form

$$x(t) = u(t - t_o)x_o,$$

where $u(t)$ is the *Heaviside unit step,*

$$u(t) \equiv \begin{cases} 0, & t < 0; \\ 1/2, & t = 0; \\ 1, & t > 0; \end{cases} \tag{7.21}$$

plotted in Fig. 7.10.

Incidentally, the value $u(0) = 1/2$ of $u(t)$ on the edge is a matter of definition. Equation (7.21) has defined $u(0) = 1/2$ for symmetry's sake but[22]

$$u_1(t) \equiv \begin{cases} 0, & t < 0; \\ 1, & t \geq 0; \end{cases} \tag{7.22}$$

is also possible. The Laplace transform of chapter 19 uses (7.22); most of the rest of the book prefers (7.21).

The derivative of the Heaviside unit step is the curious *Dirac delta*

$$\delta(t) \equiv \frac{d}{dt}u(t), \tag{7.23}$$

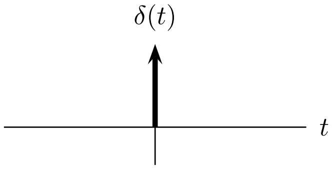
also called[23] the *impulse function,* plotted in Fig. 7.11. This function is zero everywhere except at $t = 0$, where it is infinite, with the property that

$$\int_{-\infty}^{\infty} \delta(t)\,dt = 1, \tag{7.24}$$

[22] [105]

[23] [89, § 19.5]

Figure 7.11: The Dirac delta $\delta(t)$.



and the interesting consequence that

$$\int_{-\infty}^{\infty} \delta(t - t_o) f(t)\, dt = f(t_o) \tag{7.25}$$

for any function $f(t)$, eqn. (7.25) being the Dirac delta's *sifting property.*[24]

[24] It seems inadvisable for the narrative to digress at this point to explore $u(z)$ and $\delta(z)$, the unit step and delta of a complex argument, though by means of Fourier analysis (chapter 18) or by conceiving the Dirac delta as an infinitely narrow Gaussian pulse (§ 18.4) it could perhaps do so. The book has more pressing topics to treat. For the book's present purpose the interesting action of the two functions is with respect to the real argument t.

In the author's country at least, a sort of debate seems to have run for decades between professional and applied mathematicians over the Dirac delta $\delta(t)$. Some professional mathematicians seem to have objected that $\delta(t)$ is not a function inasmuch as it lacks certain properties common to functions as they define them [129, § 2.4][49]. From the applied point of view the objection is admittedly a little hard to understand until one realizes that it is more a dispute over methods and definitions than over facts. What the professionals seem to be saying is that $\delta(t)$ does not fit as neatly as they would like into the abstract mathematical framework they had established for functions in general before Paul Dirac came along in 1930 [182, "Paul Dirac," 05:48, 25 May 2006] and slapped his disruptive $\delta(t)$ down on the table. The objection is not so much that $\delta(t)$ is not allowed as it is that professional mathematics for years after 1930 lacked a fully coherent theory for it.

The objection, if it is an objection, might tell one more about shortcomings of the abstract mathematical framework than it does about the uses of Dirac's $\delta(t)$.

Whether the professional mathematician's definition of the *function* specifically is flawed is not, of course, for this writer to judge. Nevertheless, the fact of the Dirac delta dispute, coupled with the difficulty we applied mathematicians experience in trying to understand the reason the dispute even exists, has unfortunately surrounded the Dirac delta with a kind of mysterious aura, an elusive sense that $\delta(t)$ hides subtle mysteries—when what it really hides is an internal discussion of words and means among the professionals. The professionals who had established the theoretical framework before 1930 justifiably felt reluctant to throw the whole framework away because some scientists and engineers like

Noteworthy is that

$$\delta(\alpha t) = \frac{\delta(t)}{|\alpha|}, \quad \Im(\alpha) = 0, \tag{7.26}$$

a formula that results from changing $\alpha t \leftarrow t$ in (7.24).

The Dirac delta is defined for vectors, too, such that

$$\begin{aligned} \int_S \delta(\boldsymbol{\rho})\, d\boldsymbol{\rho} &= 1, & (7.27)\\ \int_V \delta(\mathbf{r})\, d\mathbf{r} &= 1, & (7.28) \end{aligned}$$

where $d\boldsymbol{\rho} = dx\,dy = \rho\,d\rho\,d\phi$ is a surface infinitesimal and $d\mathbf{r}$ is likewise a volumetric infinitesimal.

7.8 Remarks (and exercises)

The concept of the integral is relatively simple once grasped, but its implications are broad, deep and hard. This chapter is short. One reason introductory calculus texts run so long is that they include many pages of integration examples and exercises. The reader who desires a gentler introduction to the integral might consult among others the textbook the chapter's introduction has recommended.

Even if this book is not an instructional textbook, to let the book include no exercises at all here would seem unmeet. Here are a few. Some of them need material from later chapters, so you should not expect to be able to complete them all now. The harder ones are marked with *asterisks. Work the exercises if you like.

1. Evaluate (a) $\int_0^x \tau\,d\tau$; (b) $\int_0^x \tau^2\,d\tau$. (Answer: $x^2/2$; $x^3/3$.)

2. Evaluate (a) $\int_1^x (1/\tau^2)\,d\tau$; (b) $\int_a^x 3\tau^{-2}\,d\tau$; (c) $\int_a^x C\tau^n\,d\tau$; (d) $\int_0^x (a_2\tau^2 + a_1\tau)\,d\tau$; *(e) $\int_1^x (1/\tau)\,d\tau$.

3. *Evaluate (a) $\int_0^x \sum_{k=0}^{\infty} \tau^k\,d\tau$; (b) $\sum_{k=0}^{\infty} \int_0^x \tau^k\,d\tau$; (c) $\int_0^x \sum_{k=0}^{\infty} (\tau^k/k!)\,d\tau$.

4. Evaluate $\int_0^x \exp \alpha\tau\,d\tau$.

us came along one day with a useful new function which didn't quite fit, but that was the professionals' problem not ours. To us the Dirac delta $\delta(t)$ is just a function. The internal discussion of words and means, we leave to the professionals, who know whereof they speak.

5. Evaluate (a) $\int_{-2}^{5}(3\tau^2 - 2\tau^3)\,d\tau$; (b) $\int_{5}^{-2}(3\tau^2 - 2\tau^3)\,d\tau$. Work the exercise by hand in hexadecimal and give the answer in hexadecimal.

6. Evaluate $\int_{1}^{\infty}(3/\tau^2)\,d\tau$.

7. *Evaluate the integral of the example of § 7.6 along the alternate contour suggested there, from $\hat{\mathbf{x}}\rho$ to 0 to $\hat{\mathbf{y}}\rho$.

8. Evaluate (a) $\int_{0}^{x}\cos\omega\tau\,d\tau$; (b) $\int_{0}^{x}\sin\omega\tau\,d\tau$; *(c)[25] $\int_{0}^{x}\tau\sin\omega\tau\,d\tau$.

9. *Evaluate[26] (a) $\int_{1}^{x}\sqrt{1+2\tau}\,d\tau$; (b) $\int_{x}^{a}[(\cos\sqrt{\tau})/\sqrt{\tau}]\,d\tau$.

10. *Evaluate[27] (a) $\int_{0}^{x}[1/(1+\tau^2)]\,d\tau$ (answer: $\arctan x$); (b) $\int_{0}^{x}[(4+i3)/\sqrt{2-3\tau^2}]\,d\tau$ (hint: the answer involves another inverse trigonometric); (c) $\int_{0}^{x}\sqrt{1-\tau^2}\,d\tau$; (d) $\int_{0}^{x}\tau\sqrt{1-\tau^2}\,d\tau$ (hint: use a different technique than for part c); (e) $\int_{0}^{x}\tau^2\sqrt{1-\tau^2}\,d\tau$ (hint: use a similar technique as for part c).

11. **Evaluate (a) $\int_{-\infty}^{x}\exp[-\tau^2/2]\,d\tau$; (b) $\int_{-\infty}^{\infty}\exp[-\tau^2/2]\,d\tau$.

The last exercise in particular requires some experience to answer. Moreover, it requires a developed sense of applied mathematical style to put the answer in a pleasing form (the right form for part b is very different from that for part a). Some of the easier exercises, of course, you should be able to work right now.

The point of the exercises is to illustrate how hard integrals can be to solve, and in fact how easy it is to come up with an integral which no one really knows how to solve very well. Some solutions to the same integral are better than others (easier to manipulate, faster to numerically calculate, etc.) yet not even the masters can solve them all in practical ways. On the other hand, integrals which arise in practice often can be solved very well with sufficient cleverness—and the more cleverness you develop, the more such integrals you can solve. The ways to solve them are myriad. The mathematical art of solving diverse integrals is well worth cultivating.

Chapter 9 introduces some of the basic, most broadly useful integral-solving techniques. Before addressing techniques of integration, however, as promised earlier we turn our attention in chapter 8 back to the derivative, applied in the form of the Taylor series.

[25] [146, § 8-2]

[26] [146, § 5-6]

[27] Parts (a) and (b) are sourced from [146, back endpaper].

Chapter 8

The Taylor series

The Taylor series is a power series that fits a function in a limited domain neighborhood. Fitting a function in such a way brings at least two advantages:

- it lets us take derivatives and integrals in the same straightforward way (4.15) one can take them given any power series; and

- it implies a simple procedure to calculate values of the function numerically.

This chapter introduces the Taylor series and some of its incidents. It also derives Cauchy's integral formula. The chapter's early sections prepare the ground for the treatment of the Taylor series proper in § 8.3.[1]

(The chapter's early sections, §§ 8.1 and 8.2, are thick with tiny algebraic details. Though the two early sections are interesting enough in their own right the reader who does not wish, for now, to pick through tiny algebraic details may skim the sections and then start reading in § 8.3.)

[1] Because even at the applied level the proper derivation of the Taylor series involves mathematical induction, analytic continuation and the matter of convergence domains, no balance of rigor the chapter might strike seems wholly satisfactory. The chapter errs maybe toward too much rigor; for, with a little less, most of §§ 8.1, 8.2, 8.4 and 8.6 would cease to be necessary. For the impatient, to read only the following sections might not be an unreasonable way to shorten the chapter: §§ 8.3, 8.5, 8.8, 8.9 and 8.11, plus the introduction of § 8.1.

From another point of view, the chapter errs maybe toward too little rigor. Some pretty constructs of pure mathematics serve the Taylor series and Cauchy's integral formula. However, such constructs drive the applied mathematician on too long a detour (a detour appendix C briefly overviews). The chapter as written represents the most nearly satisfactory compromise the writer has been able to strike.

8.1 The power-series expansion of $1/(1-z)^{n+1}$

Before approaching the Taylor series proper in § 8.3, we shall find it both interesting and useful to demonstrate that

$$\frac{1}{(1-z)^{n+1}} = \sum_{k=0}^{\infty} \binom{n+k}{n} z^k, \quad n \geq 0. \tag{8.1}$$

The demonstration comes in three stages. Of the three, it is the second stage (§ 8.1.2) which actually proves (8.1). The first stage (§ 8.1.1) comes up with the formula for the second stage to prove. The third stage (§ 8.1.3) establishes the sum's convergence. In all the section,

$$i, j, k, m, n, K \in \mathbb{Z}.$$

8.1.1 The formula

In § 2.6.4 we found that

$$\frac{1}{1-z} = \sum_{k=0}^{\infty} z^k = 1 + z + z^2 + z^3 + \cdots$$

for $|z| < 1$. What about $1/(1-z)^2$, $1/(1-z)^3$, $1/(1-z)^4$, and so on? By the long-division procedure of Table 2.4, one can calculate the first few terms of $1/(1-z)^2$ to be

$$\frac{1}{(1-z)^2} = \frac{1}{1-2z+z^2} = 1 + 2z + 3z^2 + 4z^3 + \cdots$$

whose coefficients $1, 2, 3, 4, \ldots$ happen to be the numbers down the first diagonal of Pascal's triangle (Fig. 4.2 on page 102; see also Fig. 4.1). Dividing $1/(1-z)^3$ seems to produce the coefficients $1, 3, 6, \mathrm{0xA}, \ldots$ down the second diagonal; dividing $1/(1-z)^4$, the coefficients down the third. A curious pattern seems to emerge, worth investigating more closely. The pattern recommends the conjecture (8.1).

To motivate the conjecture a bit more formally (though without actually proving it yet), suppose that $1/(1-z)^{n+1}$, $n \geq 0$, is expandable in the power series

$$\frac{1}{(1-z)^{n+1}} = \sum_{k=0}^{\infty} a_{nk} z^k, \tag{8.2}$$

where the a_{nk} are coefficients to be determined. Multiplying by $1-z$, we have that

$$\frac{1}{(1-z)^n} = \sum_{k=0}^{\infty}[a_{nk} - a_{n(k-1)}]z^k.$$

This is to say that

$$a_{(n-1)k} = a_{nk} - a_{n(k-1)},$$

or in other words that

$$a_{n(k-1)} + a_{(n-1)k} = a_{nk}. \tag{8.3}$$

Thinking of Pascal's triangle, one is reminded by (8.3) of a formula of Table 4.1, transcribed here in the symbols

$$\binom{m-1}{j-1} + \binom{m-1}{j} = \binom{m}{j}, \tag{8.4}$$

except that (8.3) is not $a_{(m-1)(j-1)} + a_{(m-1)j} = a_{mj}$.

Various changes of variable are possible to make (8.4) better match (8.3). We might try at first a few false ones, but eventually the change

$$\begin{aligned} n+k &\leftarrow m, \\ k &\leftarrow j, \end{aligned}$$

recommends itself. Thus changing in (8.4) gives

$$\binom{n+k-1}{k-1} + \binom{n+k-1}{k} = \binom{n+k}{k}.$$

Transformed according to a rule of Table 4.1, this is

$$\binom{n+[k-1]}{n} + \binom{[n-1]+k}{n-1} = \binom{n+k}{n}, \tag{8.5}$$

which fits (8.3) perfectly if

$$a_{nk} = \binom{n+k}{n}. \tag{8.6}$$

Hence we conjecture that (8.6), applied to (8.2), would make (8.1) true.

Equation (8.1) is thus suggestive. It works at least for the important case of $n = 0$; this much is easy to test. In light of (8.3), it seems to imply a relationship between the $1/(1-z)^{n+1}$ series and the $1/(1-z)^n$ series for any n. But *to seem* is not *to be.* At this point, all we can say is that (8.1) seems right. We will establish that it is right in the next subsection.

8.1.2 The proof by induction

Equation (8.1) is proved by induction as follows. Consider the sum

$$S_n \equiv \sum_{k=0}^{\infty} \binom{n+k}{n} z^k. \tag{8.7}$$

Multiplying by $1 - z$ yields that

$$(1-z)S_n = \sum_{k=0}^{\infty} \left[\binom{n+k}{n} - \binom{n+[k-1]}{n} \right] z^k.$$

Per (8.5), this is

$$(1-z)S_n = \sum_{k=0}^{\infty} \binom{[n-1]+k}{n-1} z^k. \tag{8.8}$$

Now suppose that (8.1) is true for $n = i - 1$ (where i denotes an integer rather than the imaginary unit):

$$\frac{1}{(1-z)^i} = \sum_{k=0}^{\infty} \binom{[i-1]+k}{i-1} z^k. \tag{8.9}$$

In light of (8.8), this means that

$$\frac{1}{(1-z)^i} = (1-z)S_i.$$

Dividing by $1 - z$,

$$\frac{1}{(1-z)^{i+1}} = S_i.$$

Applying (8.7),

$$\frac{1}{(1-z)^{i+1}} = \sum_{k=0}^{\infty} \binom{i+k}{i} z^k. \tag{8.10}$$

Evidently (8.9) implies (8.10). In other words, if (8.1) is true for $n = i - 1$, then it is also true for $n = i$. Thus *by induction,* if it is true for any one n, then it is also true for all greater n.

The "if" in the last sentence is important. Like all inductions, this one needs at least one *start case* to be valid. The $n = 0$ supplies the start case

$$\frac{1}{(1-z)^{0+1}} = \sum_{k=0}^{\infty} \binom{k}{0} z^k = \sum_{k=0}^{\infty} z^k,$$

which per (2.36) we know to be true.

8.1.3 Convergence

The question remains as to the domain over which the sum (8.1) converges.[2] To answer the question, consider that per Table 4.1,

$$\binom{m}{j} = \frac{m}{m-j}\binom{m-1}{j}, \quad m > 0.$$

With the substitution $n + k \leftarrow m$, $n \leftarrow j$, this means that

$$\binom{n+k}{n} = \frac{n+k}{k}\binom{n+[k-1]}{n},$$

or more tersely,

$$a_{nk} = \frac{n+k}{k} a_{n(k-1)},$$

where

$$a_{nk} \equiv \binom{n+k}{n}$$

are per (8.6) the coefficients of the power series (8.1). Rearranging factors,

$$\frac{a_{nk}}{a_{n(k-1)}} = \frac{n+k}{k} = 1 + \frac{n}{k}. \tag{8.11}$$

[2] The meaning of the verb *to converge* may seem clear enough from the context and from earlier references, but if explanation here helps: a series converges if and only if it approaches a specific, finite value after many terms. A more rigorous way of saying the same thing is as follows: the series

$$S = \sum_{k=0}^{\infty} \tau_k$$

converges iff (if and only if), for all possible positive constants ϵ, there exists a finite $K \geq -1$ such that

$$\left| \sum_{k=K+1}^{n} \tau_k \right| < \epsilon,$$

for all $n \geq K$ (of course it is also required that the τ_k be finite, but you knew that already).

The professional mathematical literature calls such convergence "uniform convergence," distinguishing it through a test devised by Weierstrass from the weaker "pointwise convergence" [3, § 1.5]. The applied mathematician can profit by learning the professional view in the matter but the effect of trying to teach the professional view in a book like this would not be pleasing. Here, we avoid error by keeping a clear view of the physical phenomena the mathematics is meant to model.

It is interesting nevertheless to consider an example of an integral for which convergence is not so simple, such as Frullani's integral of § 9.10.

Multiplying (8.11) by z^k/z^{k-1} gives the ratio

$$\frac{a_{nk}z^k}{a_{n(k-1)}z^{k-1}} = \left(1 + \frac{n}{k}\right) z,$$

which is to say that the kth term of (8.1) is $(1 + n/k)z$ times the $(k - 1)$th term. So long as the criterion[3]

$$\left|\left(1 + \frac{n}{k}\right) z\right| \le 1 - \delta$$

is satisfied for all sufficiently large $k > K$—where $0 < \delta \ll 1$ is a small positive constant—then the series evidently converges (see § 2.6.4 and eqn. 3.22). But we can bind $1+n/k$ as close to unity as desired by making K sufficiently large, so to meet the criterion it suffices that

$$|z| < 1. \tag{8.12}$$

The bound (8.12) thus establishes a sure convergence domain for (8.1).

8.1.4 General remarks on mathematical induction

We have proven (8.1) by means of a mathematical induction. The virtue of induction as practiced in § 8.1.2 is that it makes a logically clean, airtight case for a formula. Its vice is that it conceals the subjective process which has led the mathematician to consider the formula in the first place. Once you obtain a formula somehow, maybe you can prove it by induction; but the induction probably does not help you to obtain the formula! A good inductive proof usually begins by motivating the formula proven, as in § 8.1.1.

Richard W. Hamming once said of mathematical induction,

> The theoretical difficulty the student has with mathematical induction arises from the reluctance to ask seriously, "How could I prove a formula for an infinite number of cases when I know that testing a finite number of cases is not enough?" Once you

[3] Although one need not ask the question to understand the proof, the reader may nevertheless wonder why the simpler $|(1 + n/k)z| < 1$ is not given as a criterion. The surprising answer is that not all series $\sum \tau_k$ with $|\tau_k/\tau_{k-1}| < 1$ converge! For example, the unpretentious $\sum 1/k$ does not converge. As we see however, all series $\sum \tau_k$ with $|\tau_k/\tau_{k-1}| < 1 - \delta$ do converge. The distinction is subtle but rather important.

The really curious reader may now ask why $\sum 1/k$ does not converge. Answer: it *majorizes* $\int_1^x (1/\tau)\, d\tau = \ln x$. See (5.8) and § 8.10.

> really face this question, you will understand the ideas behind mathematical induction. It is only when you grasp the problem clearly that the method becomes clear. [70, § 2.3]

Hamming also wrote,

> The function of rigor is mainly critical and is seldom constructive. Rigor is the hygiene of mathematics, which is needed to protect us against careless thinking. [70, § 1.6]

The applied mathematician may tend to avoid rigor for which he finds no immediate use, but he does not disdain mathematical rigor on principle. The style lies in exercising rigor at the right level for the problem at hand. Hamming, a professional mathematician who sympathized with the applied mathematician's needs, wrote further,

> Ideally, when teaching a topic the degree of rigor should follow the student's perceived need for it. . . . It is necessary to require a gradually rising level of rigor so that when faced with a real need for it you are not left helpless. As a result, [one cannot teach] a uniform level of rigor, but rather a gradually rising level. Logically, this is indefensible, but psychologically there is little else that can be done. [70, § 1.6]

Applied mathematics holds that the practice *is* defensible, on the ground that the math serves the model; but Hamming nevertheless makes a pertinent point.

Mathematical induction is a broadly applicable technique for constructing mathematical proofs. We will not always write inductions out as explicitly in this book as we have done in the present section—often we will leave the induction as an implicit exercise for the interested reader—but this section's example at least lays out the general pattern of the technique.

8.2 Shifting a power series' expansion point

One more question we should treat before approaching the Taylor series proper in § 8.3 concerns the shifting of a power series' expansion point. How can the expansion point of the power series

$$\begin{aligned} f(z) &= \sum_{k=K}^{\infty} (a_k)(z - z_o)^k, \qquad (8.13) \\ (k, K) &\in \mathbb{Z}, \quad K \le 0, \end{aligned}$$

which may have terms of negative order, be shifted from $z = z_o$ to $z = z_1$?

The first step in answering the question is straightforward: one rewrites (8.13) in the form

$$f(z) = \sum_{k=K}^{\infty} (a_k)([z - z_1] - [z_o - z_1])^k,$$

and then changes the variables

$$\begin{aligned} w &\equiv \frac{z - z_1}{z_o - z_1}, \\ c_k &\equiv [-(z_o - z_1)]^k a_k, \end{aligned} \tag{8.14}$$

to obtain

$$f(z) = \sum_{k=K}^{\infty} (c_k)(1 - w)^k. \tag{8.15}$$

Splitting the $k < 0$ terms from the $k \geq 0$ terms in (8.15), we have that

$$f(z) = f_-(z) + f_+(z), \tag{8.16}$$

$$\begin{aligned} f_-(z) &\equiv \sum_{k=0}^{-(K+1)} \frac{c_{[-(k+1)]}}{(1 - w)^{k+1}}, \\ f_+(z) &\equiv \sum_{k=0}^{\infty} (c_k)(1 - w)^k. \end{aligned}$$

Of the two subseries, the $f_-(z)$ is expanded term by term using (8.1), after which combining like powers of w yields the form

$$\begin{aligned} f_-(z) &= \sum_{j=0}^{\infty} q_j w^j, \\ q_j &\equiv \sum_{n=0}^{-(K+1)} \left(c_{[-(n+1)]}\right) \binom{n + j}{n}. \end{aligned} \tag{8.17}$$

The $f_+(z)$ is even simpler to expand: one need only multiply the series out term by term per (4.5), combining like powers of w to reach the form

$$\begin{aligned} f_+(z) &= \sum_{j=0}^{\infty} p_j w^j, \\ p_j &\equiv \sum_{n=j}^{\infty} (-)^j (c_n) \binom{n}{j}. \end{aligned} \tag{8.18}$$

Equations (8.13) through (8.18) serve to shift a power series' expansion point, calculating the coefficients of a power series for $f(z)$ about $z = z_1$, given those of a power series about $z = z_o$. Notice that—unlike the original, $z = z_o$ power series—the new, $z = z_1$ power series has terms $(z - z_1)^k$ only for $k \geq 0$; it has no terms of negative order. At the price per (8.12) of restricting the convergence domain to $|w| < 1$, our shift of the expansion point away from the pole at $z = z_o$ has resolved the terms for which $k < 0$. Moreover, though one recognizes the price, we actually pay the price only to the extent to which there are terms for which $k < 0$—which often there aren't.

The method fails if $z = z_1$ happens to be a pole or other nonanalytic point of $f(z)$. The convergence domain vanishes as z_1 approaches such a forbidden point. (Examples of such forbidden points include $z = 0$ in $h[z] = 1/z$ and in $g[z] = \sqrt{z}$. See §§ 8.4 through 8.8.) Furthermore, even if z_1 does represent a fully analytic point of $f(z)$, it also must lie within the convergence domain of the original, $z = z_o$ series for the shift to be trustworthy as derived.

The attentive reader might observe that we have formally established the convergence neither of $f_-(z)$ in (8.17) nor of $f_+(z)$ in (8.18). Regarding the former convergence, that of $f_-(z)$, we have strategically framed the problem so that one needn't worry about it, running the sum in (8.13) from the finite $k = K \leq 0$ rather than from the infinite $k = -\infty$; and since according to (8.12) each term of the original $f_-(z)$ of (8.16) converges for $|w| < 1$, the reconstituted $f_-(z)$ of (8.17) safely converges in the same domain. The latter convergence, that of $f_+(z)$, is harder to establish in the abstract because that subseries has an infinite number of terms. As we will see by pursuing a different line of argument in § 8.3, however, the $f_+(z)$ of (8.18) can be nothing other than the Taylor series about $z = z_1$ of the function $f_+(z)$ in any event, enjoying the same convergence domain any such Taylor series enjoys.[4]

[4] A rigorous argument can be constructed without appeal to § 8.3 if desired, from the ratio $n/(n - k)$ of Table 4.1, which ratio approaches unity with increasing n. A more elegant rigorous argument can be made indirectly by way of a complex contour integral. In applied mathematics, however, one does not normally try to shift the expansion point of an *unspecified* function $f(z)$, anyway. Rather, one shifts the expansion point of some concrete function like $\sin z$ or $\ln(1 - z)$. The imagined difficulty (if any) vanishes in the concrete case. Appealing to § 8.3, the important point is the one made in the narrative: $f_+(z)$ can be nothing other than the Taylor series in any event.

8.3 Expanding functions in Taylor series

Having prepared the ground, we now stand in position to treat the Taylor series proper. The treatment begins with a question: if you had to express some function $f(z)$ by a power series

$$f(z) = \sum_{k=0}^{\infty} (a_k)(z - z_o)^k,$$

with terms of nonnegative order $k \geq 0$ only, then how would you do it? The procedure of § 8.1 worked well enough in the case of $f(z) = 1/(1-z)^{n+1}$, but it is not immediately obvious that the same procedure would work more generally. What if $f(z) = \sin z$, for example?[5]

Fortunately a different way to attack the power-series expansion problem is known. It works by asking the question: what power series, having terms of nonnegative order only, most resembles $f(z)$ in the immediate neighborhood of $z = z_o$? To resemble $f(z)$, the desired power series should have $a_0 = f(z_o)$; otherwise it would not have the right value at $z = z_o$. Then it should have $a_1 = f'(z_o)$ for the right slope. Then, $a_2 = f''(z_o)/2$ for the right second derivative, and so on. With this procedure,

$$f(z) = \sum_{k=0}^{\infty} \left(\left. \frac{d^k f}{dz^k} \right|_{z=z_o} \right) \frac{(z - z_o)^k}{k!}. \tag{8.19}$$

Equation (8.19) is the *Taylor series.* Where it converges, it has all the same derivatives $f(z)$ has, so if $f(z)$ is infinitely differentiable then the Taylor series exactly represents the function.[6] (At the cost of abandoning applied methods, appendix C elaborates for the benefit of readers that would like

[5] The actual Taylor series for $\sin z$ is given in § 8.9.

[6] The professional mathematician demands greater rigor at this juncture [7][58][153][147][148][79][101]. An applicationist ordinarily impatient with professional scruples might nevertheless pause to attend to the professional's objection in this instance. Consider for example the function

$$g(t) \equiv \exp\left(-\frac{1}{t^2}\right), \quad \Im(t) = 0,$$

proposed by [4], a function whose derivatives are all null at $t = 0$ despite that function (along with its derivatives) is nonnull elsewhere over the real domain. Is the Taylor series of $g(t)$ an exact representation?

Extension to the complex domain relieves the trouble but the writer has never encountered, nor been able to devise, a suitable applications-level proof of this fact. Appendix C instead outlines a professional-style proof.

elaboration. However, one might study the present chapter at least as far as the end of § 8.9 before attempting the appendix.)

The Taylor series is not guaranteed to converge outside some neighborhood near $z = z_o$, but where it does converge it is precise.

When $z_o = 0$, the series is also called the *Maclaurin series.* By either name, the series is a construct of great importance and tremendous practical value, as we shall soon see.

8.4 Analytic continuation

As earlier mentioned in § 2.11.3, an *analytic function* is a function which is infinitely differentiable in the domain neighborhood of interest—or, maybe more appropriately for our applied purpose, a function expressible as a Taylor series in that neighborhood. As we have seen, only one Taylor series about z_o is possible for a given function $f(z)$:

$$f(z) = \sum_{k=0}^{\infty} (a_k)(z - z_o)^k.$$

However, nothing prevents one from transposing the series to a different expansion point $z = z_1$ by the method of § 8.2, except that the transposed series may there enjoy a different convergence domain. As it happens, this section's purpose finds it convenient to swap symbols $z_o \leftrightarrow z_1$, transposing rather from expansion about $z = z_1$ to expansion about $z = z_o$. In the swapped notation, so long as the expansion point $z = z_o$ lies fully within (neither outside nor right on the edge of) the $z = z_1$ series' convergence domain, the two series evidently describe the selfsame underlying analytic function.

Since an analytic function $f(z)$ is infinitely differentiable and enjoys a unique Taylor expansion $f_o(z-z_o) = f(z)$ about each point z_o in its domain, it follows that if two Taylor series $f_1(z - z_1)$ and $f_2(z - z_2)$ find even a small neighborhood $|z - z_o| < \epsilon$ which lies in the domain of both, then the two can both be transposed to the common $z = z_o$ expansion point. If the two are found to have the same Taylor series there, then f_1 and f_2 both represent the same function. Moreover, if a series f_3 is found whose domain overlaps that of f_2, then a series f_4 whose domain overlaps that of f_3, and so on, and if each pair in the chain matches at least in a small neighborhood in its region of overlap, then the whole chain of overlapping series necessarily represents the same underlying analytic function $f(z)$. The series f_1 and the series f_n represent the same analytic function even if their domains do not directly overlap at all.

This is a manifestation of the principle of *analytic continuation.* The principle holds that if two analytic functions are the same within some domain neighborhood $|z - z_o| < \epsilon$, then they are the same everywhere.[7] Observe however that the principle fails at poles and other nonanalytic points, because the function is not differentiable there.

The result of § 8.2, which shows general power series to be expressible as Taylor series except at their poles and other nonanalytic points, extends the analytic continuation principle to cover power series in general, including power series with terms of negative order.

Now, observe: though all convergent power series are indeed analytic, one need not actually expand every analytic function in a power series. Sums, products and ratios of analytic functions are hardly less differentiable than the functions themselves—as also, by the derivative chain rule, is an analytic function of analytic functions. For example, where $g(z)$ and $h(z)$ are analytic, there also is $f(z) \equiv g(z)/h(z)$ analytic (except perhaps at isolated points where $h[z] = 0$). Besides, given Taylor series for $g(z)$ and $h(z)$ one can make a power series for $f(z)$ by long division if desired, so that is all right. Section 8.15 speaks further on the point.

The subject of analyticity is rightly a matter of deep concern to the professional mathematician. It is also a long wedge which drives pure and applied mathematics apart. When the professional mathematician speaks generally of a "function," he means *any function at all.* One can construct some pretty unreasonable functions if one wants to, such as

$$\begin{aligned} f([2k+1]2^m) &\equiv (-)^m, \quad (k, m) \in \mathbb{Z}; \\ f(z) &\equiv 0 \text{ otherwise.} \end{aligned}$$

However, neither functions like this $f(z)$ nor more subtly unreasonable functions normally arise in the modeling of physical phenomena. When such functions do arise, one transforms, approximates, reduces, replaces and/or avoids them. The full theory which classifies and encompasses—or explicitly excludes—such functions is thus of limited interest to the applied mathe-

[7] The writer hesitates to mention that he is given to understand [153] that the domain neighborhood can technically be reduced to a domain contour of nonzero length but zero width. Having never met a significant application of this extension of the principle, the writer has neither researched the extension's proof nor asserted its truth. He does not especially recommend that the reader worry over the point. The domain neighborhood $|z - z_o| < \epsilon$ suffices.

matician, and this book does not cover it.[8]

This does not mean that the scientist or engineer never encounters nonanalytic functions. On the contrary, he encounters several, but they are not subtle: $|z|$; $\arg z$; z^*; $\Re(z)$; $\Im(z)$; $u(t)$; $\delta(t)$. Refer to §§ 2.11 and 7.7. Such functions are nonanalytic either because they lack proper derivatives in the Argand plane according to (4.13) or because one has defined them only over a real domain.

8.5 Branch points

The function $g(z) \equiv \sqrt{z}$ is an interesting, troublesome function. Its derivative is $dg/dz = 1/2\sqrt{z}$, so even though the function is finite at $z = 0$, its derivative is not finite there. Evidently $g(z)$ has a nonanalytic point at $z = 0$, yet the point is not a pole. What is it?

We call it a *branch point.* The defining characteristic of the branch point is that, given a function $f(z)$ with such a point at $z = z_o$, if one encircles[9] the point once alone (that is, without also encircling some other branch point) by a closed contour in the Argand domain plane, while simultaneously tracking $f(z)$ in the Argand range plane—and if one demands that z and $f(z)$ move smoothly, that neither of these suddenly skip from one spot to another—then one finds that $f(z)$ ends in a different place than it began, even though z itself has returned precisely to its own starting point. The range contour remains open even though the domain contour is closed.

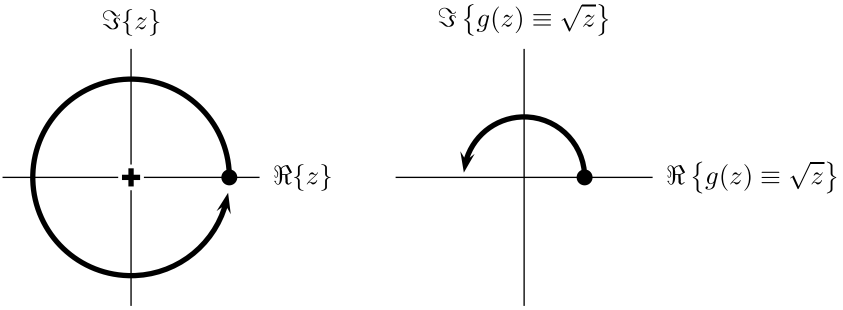
> In complex analysis, a branch point may be thought of informally as a point z_o at which a "multiple-valued function" changes values when one winds once around z_o.[10]

[8] Many books do cover it in varying degrees, including [58][153][79][147] and numerous others. The foundations of the pure theory of a complex variable, though abstract, are beautiful, and though they do not comfortably fit a book like this even an applied mathematician can profit substantially by studying them. The few pages of appendix C trace only the pure theory's main thread. However that may be, the pure theory is probably best appreciated after one already understands its chief conclusions. Though not for the explicit purpose of serving the pure theory, the present chapter does develop just such an understanding.

[9] For readers whose native language is not English, "to encircle" means "to surround" or "to enclose." The verb does not require the boundary to have the shape of an actual, geometrical circle; any closed shape suffices. However, the circle is a typical shape, probably the most fitting shape to imagine when thinking of the concept abstractly.

[10] [182, "Branch point," 18:10, 16 May 2006]

Figure 8.1: A coördinated evolution of z and $g(z) \equiv \sqrt{z}$.


$\Im\{z\}$
$\Re\{z\}$
$\Im\left\{g(z) \equiv \sqrt{z}\right\}$
$\Re\left\{g(z) \equiv \sqrt{z}\right\}$

An analytic function like $g(z) \equiv \sqrt{z}$ having a branch point evidently is not single-valued. It is multiple-valued. For a single z more than one distinct $g(z)$ is possible, as Fig. 8.1 suggests. (Looking at the figure, incidentally, the perceptive reader might ask why the figure does not merely plot $g[z]$ against z on a single pair of axes. Of course, the reader knows why, but the question is worth asking, anyway. The answer is that the figure would indeed like to plot $g[z]$ against z on a single pair of axes but cannot because four dimensions would be required for the visual! Three dimensions are sometimes let to overburden two-dimensional paper—as in Fig. 7.7 for example—but four dimensions are too many; so, instead, Fig. 8.1 coördinates a pair of plots at two visual dimensions per plot. There is irony in this, for the real and imaginary parts of a scalar together constitute only a single actual dimension, but no one seems to know how to display an imaginary part *visually* without using an extra axis.)

An analytic function like $h(z) \equiv 1/z$, by contrast to $g(z)$, is single-valued even though it has a pole. This function does not suffer the syndrome described. When a domain contour encircles a pole of $h(z)$ or of any other function that has a pole, the corresponding range contour is properly closed. Poles do not cause their functions to be multiple-valued and, thus, *poles are not branch points.*

Evidently $f(z) \equiv (z - z_o)^a$ has a branch point at $z = z_o$ if and only if a is not an integer. If $f(z)$ does have a branch point—if a is not an integer—then the mathematician must draw a distinction between $z_1 = z_o + \rho e^{i\phi}$ and $z_2 = z_o + \rho e^{i(\phi+2\pi)}$, *even though the two are exactly the same number.* Indeed $z_1 = z_2$, but paradoxically $f(z_1) \neq f(z_2)$.

This is difficult. It is confusing, too, until one realizes that the fact of a branch point says nothing whatsoever about the argument z. As far as z is concerned, there really is no distinction between $z_1 = z_o + \rho e^{i\phi}$ and $z_2 = z_o + \rho e^{i(\phi+2\pi)}$—none at all. What draws the distinction is the multiple-valued function $f(z)$ which uses the argument.

It is as though I had a mad colleague who called me Thaddeus H. Black, until one day I happened to walk past behind his desk (rather than in front as I usually did), whereupon for some reason he began calling me Gorbag J. Pfufnik. I had not changed at all, but now the colleague calls me by a different name. The change isn't really in me, is it? It's in my colleague, who seems to suffer a branch point. If it is important to me to be sure that my colleague really is addressing me when he cries, "Pfufnik!" then I had better keep a running count of how many times I have turned about his desk, hadn't I, even though the number of turns is personally of no import to me.

The usual analysis strategy when one encounters a branch point is simply to avoid the point. Where an integral follows a closed contour as in § 8.8, the strategy is to compose the contour to exclude the branch point, to shut it out. Such a strategy of avoidance usually prospers.[11]

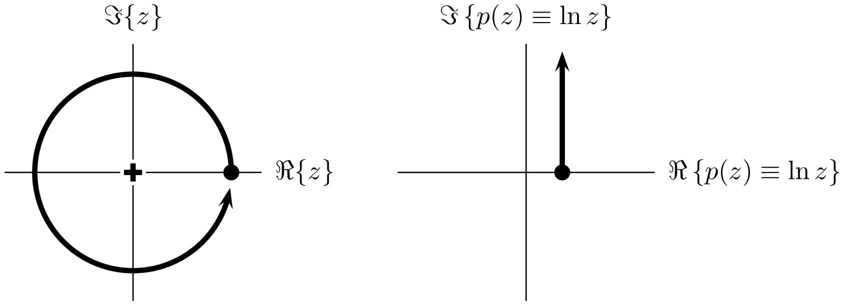
Curiously, the function $p(z) \equiv \ln z$ has a branch point at $z = 0$ despite that the function's order is zero (§ 5.3). By contrast, the order of $g(z) \equiv \sqrt{z}$ was 1/2, a noninteger, so a branch point was to be expected there; whereas a branch point in a zeroth-order function like $\ln(\cdot)$ comes perhaps as a surprise—see Fig. 8.2. Fortunately, as § 8.8 will soon show, the branch point of $\ln(\cdot)$ is not a point one needs to avoid. On the contrary, one often explicitly seeks out such a point. Before addressing that interesting matter, though, let us turn attention briefly to the definitions of entire and meromorphic functions, next, and after that to extrema over a complex domain.

8.6 Entire and meromorphic functions

Though an applied mathematician is unwise to let abstract definitions enthrall his thinking, pure mathematics nevertheless brings some technical definitions the applied mathematician can use. Two such are the definitions

[11] Traditionally associated with branch points in complex variable theory are the notions of *branch cuts* and *Riemann sheets.* These ideas are interesting, but are not central to the analysis as developed in this book and are not covered here. The interested reader might consult a book on complex variables or advanced calculus like [79], among many others.

Figure 8.2: A coördinated evolution of z and $p(z) \equiv \ln z$.



of *entire* and *meromorphic* functions.[12]

A function $f(z)$ which is analytic for all finite z is an *entire function.*[13] Examples include $f(z) = z^2$ and $f(z) = \exp z$, but not $f(z) = 1/z$ which has a pole at $z = 0$.

A function $f(z)$ which is analytic for all finite z except at isolated poles (which can be n-fold poles if n is a finite, positive integer), which has no branch points, of which no circle of finite radius in the Argand domain plane encompasses an infinity of poles, is a *meromorphic function.*[14] Examples include $f(z) = 1/z$, $f(z) = 1/(z+2) + 1/(z-1)^3 + 2z^2$ and $f(z) = \tan z$—the last of which has an infinite number of poles, but of which the poles nowhere cluster in infinite numbers. The function $f(z) = \tan(1/z)$ is not meromorphic since it has an infinite number of poles within (for instance) the Argand unit circle. Even the function $f(z) = \exp(1/z)$ is not meromorphic: it has only the one, isolated nonanalytic point at $z = 0$, and that point is no branch point; but the point is an *essential singularity,* having the character of an infinitifold (∞-fold) pole.[15]

Incidentally, if it seems unclear that the singularities of $\tan z$ are actual

[12] [178]

[13] [153, chapter 6]

[14] The definition follows [107, § 1.1]. At least one competent author [153, chapter 6] however seems (inadvertently?) to exclude functions with an infinite number of poles like the $P(z) \equiv \sum_{k=0}^{\infty} [-]^k/[k!(z+k)]$ of [107]. Nevertheless, *according to the book you are now reading,* a function like $P(z)$ remains meromorphic because, though it has an infinity of poles, it does not crowd this infinity into any finite domain.

[15] [101]

poles, then consider that

$$\tan z = \frac{\sin z}{\cos z} = -\frac{\cos w}{\sin w},$$

wherein we have changed the variable

$$w \leftarrow z - (2n+1)\frac{2\pi}{4}, \quad n \in \mathbb{Z}.$$

Section 8.9 and its Table 8.1, below, give Taylor series for $\cos z$ and $\sin z$, with which

$$\tan z = \frac{-1 + w^2/2 - w^4/\mathrm{0x18} - \cdots}{w - w^3/6 + w^5/\mathrm{0x78} - \cdots}.$$

By long division,

$$\tan z = -\frac{1}{w} + \frac{w/3 - w^3/\mathrm{0x1E} + \cdots}{1 - w^2/6 + w^4/\mathrm{0x78} - \cdots}.$$

(On the other hand, if it is unclear that $z = [2n+1][2\pi/4]$ are the only singularities $\tan z$ has—that it has no singularities of which $\Im[z] \neq 0$—then consider that the singularities of $\tan z$ occur where $\cos z = 0$, which by Euler's formula, eqn. 5.18, occurs where $\exp[+iz] = \exp[-iz]$. This in turn is possible only if $|\exp[+iz]| = |\exp[-iz]|$, which happens only for real z.)

Sections 8.14, 8.15 and 9.7 speak further of the matter.

8.7 Extrema over a complex domain

If a function $f(z)$ is expanded by (8.19) or by other means about an analytic expansion point $z = z_o$ such that

$$f(z) = f(z_o) + \sum_{k=1}^{\infty} (a_k)(z - z_o)^k;$$

and if

$$\begin{aligned} a_k &= 0 \quad \text{for } k < K, \text{ but} \\ a_K &\neq 0, \\ (k, K) &\in \mathbb{Z}, \;\; 0 < K < \infty, \end{aligned}$$

such that a_K is the series' first nonzero coefficient; then, in the immediate neighborhood of the expansion point,[16]

$$f(z) \approx f(z_o) + (a_K)(z - z_o)^K, \quad |z - z_o| \ll 1.$$

Changing $\rho' e^{i\phi'} \leftarrow z - z_o$, this is

$$f(z) \approx f(z_o) + a_K \left(\rho'\right)^K e^{iK\phi'}, \quad 0 \le \rho' \ll 1. \tag{8.20}$$

Evidently one can shift the output of an analytic function $f(z)$ slightly in any desired Argand direction by shifting slightly the function's input z. Specifically according to (8.20), to shift $f(z)$ by $\Delta f \approx \epsilon e^{i\psi}$, one can shift z by $\Delta z \approx (\epsilon/a_K)^{1/K} e^{i(\psi + n2\pi)/K}$, $n \in \mathbb{Z}$. Except at a nonanalytic point of $f(z)$ or in the trivial case that $f(z)$ were everywhere constant, this always works—even where $[df/dz]_{z=z_o} = 0$.

That by varying an analytic function's input one can smoothly shift the function's output in any desired Argand direction has the significant consequence that neither the real nor the imaginary part of the function—nor for that matter any linear combination $\Re[e^{-i\omega} f(z)]$ of the real and imaginary parts—can have an extremum within the interior of a domain over which the function is fully analytic. That is, *a function's extrema over a bounded analytic domain never lie within the domain's interior but always on its boundary*[17,18] (except, as earlier mentioned, in the trivial case of an $f[z]$ that is everywhere constant).

The last consequence is perhaps unexpected. It would not have been so for a real domain bounded by a pair of end points, but for an analytic domain bounded by a complex contour that is the way it is. When looking for a complex function's extrema, one need not search an analytic domain's

[16] The pure logicist might prefer to express this in the δ-ϵ style of § 4.4.9, especially since z and z_o might have physical units of measure attached, in which case the inequality that $|z - z_o| \ll 1$ though suggestive would be strictly meaningless. However, the practical applicationist is probably not so fussy.

What the notation intends to specify, and what the applied mathematician understands it to mean, is that z lie infinitesimally close to z_o, or at any rate that z lie sufficiently close to z_o to emphasize the behavior described. Just *how* close z should lie is not the point. If in doubt, go closer!

In case the reader is still unsatisfied, here it is in δ-ϵ style: for any positive quantity ϵ whose physical units of measure (if any) are compatible as follows, there exists a positive quantity δ such that $|[f(z_o) + (a_K)(z - z_o)^K - f(z)]/[z - z_o]^K| < \epsilon$ for all $|z - z_o| < \delta$. Nevertheless, in applications, the narrative's briefer style, $|z - z_o| \ll 1$, probably suffices.

[17] Professional mathematicians tend to define the domain and its boundary more carefully.

[18] [159][101]

interior, for if the domain contains no poles nor any other nonanalytic points then, apparently, the domain's boundary is the only place an extremum can exist.

8.8 Cauchy's integral formula

In § 7.6 we considered the problem of vector contour integration, in which the sum value of an integration depends not only on the integration's endpoints but also on the path, or *contour,* over which the integration is done, as in Fig. 7.9. Because real scalars are confined to a single line, no alternate choice of path is possible where the variable of integration is a real scalar, so the contour problem does not arise in that case. It does however arise where the variable of integration is a *complex* scalar, because there again different paths are possible. Refer to the Argand plane of Fig. 2.7.

Consider the integral

$$S_n = \int_{z_1}^{z_2} z^{n-1}\,dz, \quad n \in \mathbb{Z}. \tag{8.21}$$

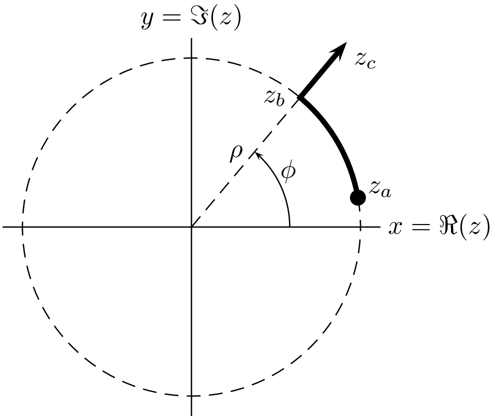
If z were always a real number, then by the antiderivative (§ 7.2) this integral would evaluate to $(z_2^n - z_1^n)/n$; or, in the case of $n = 0$, to $\ln(z_2/z_1)$ [though complications could still arise if $n < 0$ and z_2 differed in sign from z_1]. Inasmuch as z is complex, however, the correct evaluation is less obvious. To evaluate the integral sensibly in the latter case, one must consider some specific path of integration in the Argand plane. One must also consider the meaning of the symbol dz.

8.8.1 The meaning of the symbol dz

The symbol dz represents an infinitesimal step in some direction in the Argand plane:

$$\begin{aligned}
dz &= [z + dz] - [z] \\
&= \left[(\rho + d\rho)e^{i(\phi + d\phi)}\right] - \left[\rho e^{i\phi}\right] \\
&= \left[(\rho + d\rho)e^{i\,d\phi}e^{i\phi}\right] - \left[\rho e^{i\phi}\right] \\
&= \left[(\rho + d\rho)(1 + i\,d\phi)e^{i\phi}\right] - \left[\rho e^{i\phi}\right].
\end{aligned}$$

Figure 8.3: A contour of integration in the Argand plane, in two segments: constant-ρ (z_a to z_b); and constant-ϕ (z_b to z_c).



Since the product of two infinitesimals is negligible even on an infinitesimal scale, we can drop the $d\rho\, d\phi$ term.[19] After canceling finite terms, we are left with the peculiar but fine formula

$$dz = (d\rho + i\rho\, d\phi)e^{i\phi}. \tag{8.22}$$

8.8.2 Integrating along the contour

Now consider the integration (8.21) along the contour of Fig. 8.3. Integrat-

[19] The dropping of second-order infinitesimals like $d\rho\, d\phi$, added to first order infinitesimals like $d\rho$, is a standard calculus technique. One cannot *always* drop them, however. Occasionally one encounters a sum in which not only do the finite terms cancel, but also the first-order infinitesimals. In such a case, the second-order infinitesimals dominate and cannot be dropped. An example of the type is

$$\lim_{\epsilon\to 0} \frac{(1-\epsilon)^3 + 3(1+\epsilon) - 4}{\epsilon^2} = \lim_{\epsilon\to 0} \frac{(1-3\epsilon+3\epsilon^2) + (3+3\epsilon) - 4}{\epsilon^2} = 3.$$

One typically notices that such a case has arisen when the dropping of second-order infinitesimals has left an ambiguous 0/0. To fix the problem, you simply go back to the step during which you dropped the infinitesimal and you restore it, and then you proceed from there. Otherwise there isn't much point in carrying second-order infinitesimals around. In the relatively uncommon event that you need them, you'll know it. The math itself will tell you.

ing along the constant-ϕ segment,

$$\begin{aligned}\int_{z_b}^{z_c} z^{n-1}\,dz &= \int_{\rho_b}^{\rho_c} (\rho e^{i\phi})^{n-1}(d\rho + i\rho\,d\phi)e^{i\phi}\\ &= \int_{\rho_b}^{\rho_c} (\rho e^{i\phi})^{n-1}(d\rho)e^{i\phi}\\ &= e^{in\phi}\int_{\rho_b}^{\rho_c} \rho^{n-1}\,d\rho\\ &= \frac{e^{in\phi}}{n}\,(\rho_c^n - \rho_b^n)\\ &= \frac{z_c^n - z_b^n}{n}.\end{aligned}$$

Integrating along the constant-ρ arc,

$$\begin{aligned}\int_{z_a}^{z_b} z^{n-1}\,dz &= \int_{\phi_a}^{\phi_b} (\rho e^{i\phi})^{n-1}(d\rho + i\rho\,d\phi)e^{i\phi}\\ &= \int_{\phi_a}^{\phi_b} (\rho e^{i\phi})^{n-1}(i\rho\,d\phi)e^{i\phi}\\ &= i\rho^n\int_{\phi_a}^{\phi_b} e^{in\phi}\,d\phi\\ &= \frac{i\rho^n}{in}\left(e^{in\phi_b} - e^{in\phi_a}\right)\\ &= \frac{z_b^n - z_a^n}{n}.\end{aligned}$$

Adding the two, we have that

$$\int_{z_a}^{z_c} z^{n-1}\,dz = \frac{z_c^n - z_a^n}{n},$$

surprisingly the same as for real z. Moreover, contrary to the diagram but nevertheless fairly obviously, nothing prevents one from specifying an alternate path from z_a to z_c that reverses the sequence, traversing a constant-ϕ segment first and then a constant-ρ arc afterward; as long as $n \neq 0$ and, if $n < 0$, the path observes to avoid the point $z = 0$, such a change apparently would not alter the last result. Either way, since any path of integration between any two complex numbers z_1 and z_2 is approximated arbitrarily closely per (8.22) by a succession of short constant-ρ and constant-ϕ elements, it follows generally that

$$\int_{z_1}^{z_2} z^{n-1}\,dz = \frac{z_2^n - z_1^n}{n}, \quad n \in \mathbb{Z},\ n \neq 0. \tag{8.23}$$

The applied mathematician might reasonably ask, "Was (8.23) really worth the trouble? We knew *that* already. It's the same as for real numbers."

Well, we really didn't know it before deriving it, but the point is well taken nevertheless. However, notice the exemption of $n = 0$. Equation (8.23) does not hold in that case. Consider the $n = 0$ integral

$$S_0 = \int_{z_1}^{z_2} \frac{dz}{z}.$$

Following the same steps as before and using (5.8) and (2.42), we find that

$$\int_{\rho_1}^{\rho_2} \frac{dz}{z} = \int_{\rho_1}^{\rho_2} \frac{(d\rho + i\rho\, d\phi)e^{i\phi}}{\rho e^{i\phi}} = \int_{\rho_1}^{\rho_2} \frac{d\rho}{\rho} = \ln \frac{\rho_2}{\rho_1}. \tag{8.24}$$

This is always real-valued, but otherwise it brings no surprise. However,

$$\int_{\phi_1}^{\phi_2} \frac{dz}{z} = \int_{\phi_1}^{\phi_2} \frac{(d\rho + i\rho\, d\phi)e^{i\phi}}{\rho e^{i\phi}} = i \int_{\phi_1}^{\phi_2} d\phi = i(\phi_2 - \phi_1). \tag{8.25}$$

The odd thing about this is in what happens when the contour closes a complete loop in the Argand plane about the $z = 0$ pole. In this case, $\phi_2 = \phi_1 + 2\pi$, so

$$S_0 = i2\pi$$

even though the integration ends where it begins.

Generalizing, we have that

$$\begin{aligned} \oint (z - z_o)^{n-1}\, dz &= 0, \quad n \in \mathbb{Z},\ n \neq 0; \\ \oint \frac{dz}{z - z_o} &= i2\pi; \end{aligned} \tag{8.26}$$

where as in § 7.6 the symbol $\oint$ represents integration about a closed contour that ends where it begins, and where it is implied that the contour loops positively (counterclockwise, in the direction of increasing ϕ) exactly once about the $z = z_o$ pole.

Notice that the formula's $i2\pi$ does not depend on the precise path of integration but only on the fact that the path loops once positively about the pole. Notice also that nothing in the derivation of (8.23) actually requires that n be an integer, so one can write,

$$\int_{z_1}^{z_2} z^{a-1}\, dz = \frac{z_2^a - z_1^a}{a}, \quad a \neq 0. \tag{8.27}$$

However, (8.26) does not hold in the latter case; its integral comes to zero for nonintegral a only if the contour does not enclose the branch point at $z = z_o$.

For a closed contour *which encloses no pole or other nonanalytic point,* (8.27) has that $\oint z^{a-1}\,dz = 0$, or with the change of variable $z - z_o \leftarrow z$,

$$\oint (z - z_o)^{a-1}\,dz = 0.$$

But because any analytic function can be expanded in the form $f(z) = \sum_k (c_k)(z - z_o)^{a_k - 1}$ (which is just a Taylor series if the a_k happen to be positive integers), this means that

$$\oint f(z)\,dz = 0 \tag{8.28}$$

if $f(z)$ is everywhere analytic within the contour.[20]

8.8.3 The formula

The combination of (8.26) and (8.28) is powerful. Consider the closed contour integral

$$\oint \frac{f(z)}{z - z_o}\,dz,$$

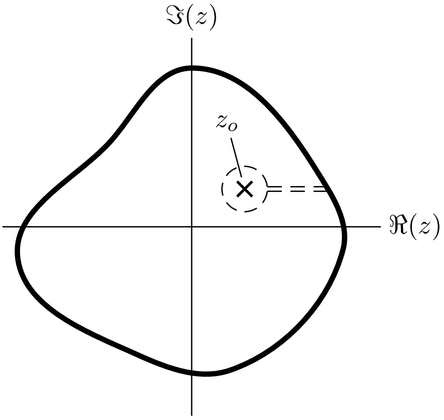
where the contour encloses no nonanalytic point of $f(z)$ itself but does enclose the pole of $f(z)/(z - z_o)$ at $z = z_o$. If the contour were a tiny circle of infinitesimal radius about the pole, then the integrand would reduce to $f(z_o)/(z - z_o)$; and then per (8.26),

$$\oint \frac{f(z)}{z - z_o}\,dz = i2\pi f(z_o). \tag{8.29}$$

But if the contour were not an infinitesimal circle but rather the larger contour of Fig. 8.4? In this case, if the dashed detour which excludes the

[20] The careful reader will observe that (8.28)'s derivation does not explicitly handle an $f(z)$ represented by a Taylor series with an infinite number of terms and a finite convergence domain (for example, $f[z] = \ln[1 - z]$). However, by § 8.2 one can transpose such a series from z_o to an overlapping convergence domain about z_1. Let the contour's interior be divided into several cells, each of which is small enough to enjoy a single convergence domain. Integrate about each cell. Because the cells share boundaries within the contour's interior, each interior boundary is integrated twice, once in each direction, canceling. The original contour—each piece of which is an exterior boundary of some cell—is integrated once piecewise. This is the basis on which a more rigorous proof is constructed.

Figure 8.4: A Cauchy contour integral.



pole is taken, then according to (8.28) the resulting integral totals zero; but the two straight integral segments evidently cancel; and similarly as we have just reasoned, the *reverse-directed* integral about the tiny detour circle is $-i2\pi f(z_o)$; so to bring the total integral to zero the integral about the main contour must be $i2\pi f(z_o)$. Thus, (8.29) holds for any positively-directed contour which once encloses a pole and no other nonanalytic point, whether the contour be small or large. Equation (8.29) is *Cauchy's integral formula.*

If the contour encloses multiple poles (§§ 2.10 and 9.7.2), then by the principle of linear superposition (§ 7.3.3),

$$\oint \left[f_o(z) + \sum_k \frac{f_k(z)}{z - z_k} \right] dz = i2\pi \sum_k f_k(z_k), \tag{8.30}$$

where the $f_o(z)$ is a *regular part*;[21] and again, where neither $f_o(z)$ nor any of the several $f_k(z)$ has a pole or other nonanalytic point within (or on) the contour. The values $f_k(z_k)$, which represent the strengths of the poles, are called *residues*. In words, (8.30) says that an integral about a closed contour in the Argand plane comes to $i2\pi$ times the sum of the residues of the poles (if any) thus enclosed. (Note however that eqn. 8.30 does not handle branch points. If there is a branch point, the contour must exclude it or the formula

[21] [107, § 1.1]

will not work.)

As we shall see in § 9.6, whether in the form of (8.29) or of (8.30) Cauchy's integral formula is an extremely useful result.[22]

8.8.4 Enclosing a multiple pole

When a complex contour of integration encloses a double, triple or other n-fold pole, the integration can be written,

$$S = \oint \frac{f(z)}{(z-z_o)^{m+1}}\,dz, \quad m \in \mathbb{Z},\ m \geq 0,$$

where $m+1 = n$. Expanding $f(z)$ in a Taylor series (8.19) about $z = z_o$,

$$S = \oint \sum_{k=0}^{\infty} \left(\left. \frac{d^k f}{dz^k} \right|_{z=z_o} \right) \frac{dz}{(k!)(z-z_o)^{m-k+1}}.$$

But according to (8.26), only the $k = m$ term contributes, so

$$\begin{aligned} S &= \oint \left(\left. \frac{d^m f}{dz^m} \right|_{z=z_o} \right) \frac{dz}{(m!)(z-z_o)} \\ &= \frac{1}{m!} \left(\left. \frac{d^m f}{dz^m} \right|_{z=z_o} \right) \oint \frac{dz}{(z-z_o)} \\ &= \frac{i2\pi}{m!} \left(\left. \frac{d^m f}{dz^m} \right|_{z=z_o} \right), \end{aligned}$$

where the integral is evaluated in the last step according to (8.29). Altogether,

$$\oint \frac{f(z)}{(z-z_o)^{m+1}}\,dz = \frac{i2\pi}{m!} \left(\left. \frac{d^m f}{dz^m} \right|_{z=z_o} \right), \quad m \in \mathbb{Z},\ m \geq 0. \tag{8.31}$$

Equation (8.31) evaluates a contour integral about an n-fold pole as (8.29) does about a single pole. (When $m = 0$, the two equations are the same.)[23]

[22] [79, § 10.6][153][182, "Cauchy's integral formula," 14:13, 20 April 2006]

[23] [101][153]

8.9 Taylor series for specific functions

With the general Taylor series formula (8.19), the derivatives of Tables 5.2 and 5.3, and the observation from (4.16) that

$$\frac{d(z^a)}{dz} = az^{a-1},$$

one can calculate Taylor series for many functions. For instance, expanding about $z = 1$,

$$\begin{aligned}
\ln z\big|_{z=1} &= & \ln z\big|_{z=1} &= 0, \\
\left.\frac{d}{dz}\ln z\right|_{z=1} &= & \left.\frac{1}{z}\right|_{z=1} &= 1, \\
\left.\frac{d^2}{dz^2}\ln z\right|_{z=1} &= & \left.\frac{-1}{z^2}\right|_{z=1} &= -1, \\
\left.\frac{d^3}{dz^3}\ln z\right|_{z=1} &= & \left.\frac{2}{z^3}\right|_{z=1} &= 2, \\
& & &\ \vdots \\
\left.\frac{d^k}{dz^k}\ln z\right|_{z=1} &= & \left.\frac{-(-)^k(k-1)!}{z^k}\right|_{z=1} &= -(-)^k(k-1)!, \quad k > 0.
\end{aligned}$$

With these derivatives, the Taylor series about $z = 1$ is

$$\ln z = \sum_{k=1}^{\infty}\left[-(-)^k(k-1)!\right]\frac{(z-1)^k}{k!} = -\sum_{k=1}^{\infty}\frac{(1-z)^k}{k},$$

evidently convergent for $|1 - z| < 1$. (And if z lies outside the convergence domain? Several strategies are then possible. One can expand the Taylor series about a different point; but cleverer and easier is to take advantage of some convenient relationship like $\ln w = -\ln[1/w]$. Section 8.10.4 elaborates.) Using such Taylor series, one can relatively efficiently calculate actual numerical values for $\ln z$ and many other functions.

Table 8.1 lists Taylor series for a few functions of interest. All the series converge for $|z| < 1$. The $\exp z$, $\sin z$ and $\cos z$ series converge for all complex z. Among the several series, the series for $\arctan z$ is computed

Table 8.1: Taylor series.

$$
\begin{aligned}
f(z) &= \sum_{k=0}^{\infty} \left(\left. \frac{d^k f}{dz^k} \right|_{z=z_o} \right) \prod_{j=1}^{k} \frac{z - z_o}{j} \\
(1+z)^{a-1} &= \sum_{k=0}^{\infty} \prod_{j=1}^{k} \left(\frac{a}{j} - 1 \right) z \\
\exp z &= \sum_{k=0}^{\infty} \prod_{j=1}^{k} \frac{z}{j} = \sum_{k=0}^{\infty} \frac{z^k}{k!} \\
\sin z &= \sum_{k=0}^{\infty} \left[z \prod_{j=1}^{k} \frac{-z^2}{(2j)(2j+1)} \right] \\
\cos z &= \sum_{k=0}^{\infty} \prod_{j=1}^{k} \frac{-z^2}{(2j-1)(2j)} \\
\sinh z &= \sum_{k=0}^{\infty} \left[z \prod_{j=1}^{k} \frac{z^2}{(2j)(2j+1)} \right] \\
\cosh z &= \sum_{k=0}^{\infty} \prod_{j=1}^{k} \frac{z^2}{(2j-1)(2j)} \\
-\ln(1-z) &= \sum_{k=1}^{\infty} \frac{1}{k} \prod_{j=1}^{k} z = \sum_{k=1}^{\infty} \frac{z^k}{k} \\
\arctan z &= \sum_{k=0}^{\infty} \frac{1}{2k+1} \left[z \prod_{j=1}^{k} (-z^2) \right] = \sum_{k=0}^{\infty} \frac{(-)^k z^{2k+1}}{2k+1}
\end{aligned}
$$

indirectly[24] by way of Table 5.3 and (2.35):

$$\begin{aligned}\arctan z &= \int_0^z \frac{1}{1+w^2}\,dw \\ &= \int_0^z \sum_{k=0}^{\infty} (-)^k w^{2k}\,dw \\ &= \sum_{k=0}^{\infty} \frac{(-)^k z^{2k+1}}{2k+1}.\end{aligned}$$

It is interesting to observe from Table 8.1 the useful first-order approximations that

$$\begin{aligned}\lim_{z\to 0} \exp z &= 1+z, \\ \lim_{z\to 0} \sin z &= z,\end{aligned} \tag{8.32}$$

among others.

Professional mathematicians tend to prefer a different, terser, more abstract development of the Taylor series and its incidents than this chapter's. Appendix C outlines it. We lacked the theoretical preparation to tackle appendix C when the chapter began but we have it now. Some will find appendix C more persuasive. You can read appendix C now if you wish.

8.10 Error bounds

One naturally cannot actually sum a Taylor series to an infinite number of terms. One must add some finite number of terms and then quit—which raises the question: how many terms are enough? How can one know that one has added adequately many terms; that the remaining terms, which constitute the tail of the series, are sufficiently insignificant? How can one set error bounds on the truncated sum?

8.10.1 Examples

Some series alternate sign. For these it is easy if the numbers involved happen to be real. For example, from Table 8.1,

$$\ln\frac{3}{2} = \ln\left(1+\frac{1}{2}\right) = \frac{1}{(1)(2^1)} - \frac{1}{(2)(2^2)} + \frac{1}{(3)(2^3)} - \frac{1}{(4)(2^4)} + \cdots$$

[24] [146, § 11-7]

Each term is smaller in magnitude than the last, so the true value of ln(3/2) necessarily lies between the sum of the series to n terms and the sum to $n+1$ terms. The last and next partial sums bound the result. Up to but not including the fourth-order term, for instance,

$$S_4 - \frac{1}{(4)(2^4)} < \ln\frac{3}{2} < S_4,$$
$$S_4 = \frac{1}{(1)(2^1)} - \frac{1}{(2)(2^2)} + \frac{1}{(3)(2^3)}.$$

Other series however do not alternate sign. For example,

$$\begin{aligned}
\ln 2 &= -\ln\frac{1}{2} = -\ln\left(1-\frac{1}{2}\right) = S_5 + R_5,\\
S_5 &= \frac{1}{(1)(2^1)} + \frac{1}{(2)(2^2)} + \frac{1}{(3)(2^3)} + \frac{1}{(4)(2^4)},\\
R_5 &= \frac{1}{(5)(2^5)} + \frac{1}{(6)(2^6)} + \cdots
\end{aligned}$$

The basic technique in such a case is to find a replacement series (or integral) R'_n which one can collapse analytically, each of whose terms equals or exceeds in magnitude the corresponding term of R_n. For the example, one might choose

$$R'_5 = \frac{1}{5}\sum_{k=5}^{\infty}\frac{1}{2^k} = \frac{2}{(5)(2^5)},$$

wherein (2.36) had been used to collapse the summation. Then,

$$S_5 < \ln 2 < S_5 + R'_5.$$

For real $0 \le x < 1$ generally,

$$\begin{aligned}
S_n &< -\ln(1-x) < S_n + R'_n,\\
S_n &\equiv \sum_{k=1}^{n-1}\frac{x^k}{k},\\
R'_n &\equiv \sum_{k=n}^{\infty}\frac{x^k}{n} = \frac{x^n}{(n)(1-x)}.
\end{aligned}$$

Many variations and refinements are possible, some of which we will meet in the rest of the section, but that is the basic technique: to add several terms of the series to establish a lower bound, then to overestimate the remainder of the series to establish an upper bound. The overestimate R'_n *majorizes* the series' true remainder R_n. Notice that the R'_n in the example is a fairly small number, and that it would have been a lot smaller yet had we included a few more terms in S_n (for instance, $n = \mathrm{0x40}$ would have bound $\ln 2$ tighter than the limit of a computer's typical `double`-type floating-point accuracy). The technique usually works well in practice for this reason.

8.10.2 Majorization

To majorize in mathematics is to be, or to replace by virtue of being, everywhere at least as great as. This is best explained by example. Consider the summation

$$S = \sum_{k=1}^{\infty} \frac{1}{k^2} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots$$

The exact value to which this summation totals is unknown to us at this point, but the summation does rather resemble the integral (refer to Table 7.1)

$$I = \int_1^{\infty} \frac{dx}{x^2} = -\frac{1}{x}\bigg|_1^{\infty} = 1.$$

Figure 8.5 plots S and I together as areas—or more precisely, plots $S - 1$ and I together as areas (the summation's first term is omitted). As the plot shows, the unknown area $S - 1$ cannot possibly be as great as the known area I. In symbols, $S - 1 < I = 1$; or,

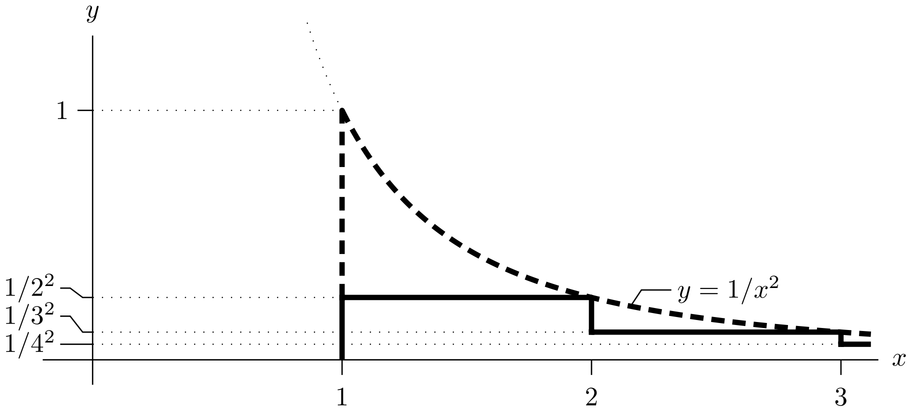
$$S < 2.$$

The integral I majorizes the summation $S - 1$, thus guaranteeing the absolute upper limit on S. (Of course $S < 2$ is a very loose limit, but that isn't the point of the example. In practical calculation, one would let a computer add many terms of the series first numerically, and only then majorize the remainder. Even so, cleverer ways to majorize the remainder of this particular series will occur to the reader, such as in representing the terms graphically—not as flat-topped rectangles—but as slant-topped trapezoids, shifted in the figure a half unit rightward.)

Majorization serves surely to bound an unknown quantity by a larger, known quantity. Reflecting, *minorization*[25] serves surely to bound an un-

[25] The author does not remember ever encountering the word *minorization* heretofore in print, but as a reflection of *majorization* the word seems logical. This book at least will use the word where needed. You can use it too if you like.

Figure 8.5: Majorization. The area I between the dashed curve and the x axis majorizes the area $S - 1$ between the stairstep curve and the x axis, because the height of the dashed curve is everywhere at least as great as that of the stairstep curve.



known quantity by a smaller, known quantity. The quantities in question are often integrals and/or series summations, the two of which are akin as Fig. 8.5 illustrates. The choice of whether to majorize a particular unknown quantity by an integral or by a series summation depends on the convenience of the problem at hand.

The series S of this subsection is interesting, incidentally. It is a *harmonic series* rather than a power series, because though its terms do decrease in magnitude it has no z^k factor (or seen from another point of view, it does have a z^k factor, but $z = 1$), and the ratio of adjacent terms' magnitudes approaches unity as k grows. Harmonic series can be hard to sum accurately, but clever majorization can help.

Incidentally, a much faster method to sum the series S of this particular subsection happens to be known. We are not yet ready to investigate it but shall reach it in § 17.5.3.

8.10.3 Geometric majorization

Harmonic series can be hard to sum as § 8.10.2 has observed, but more common than harmonic series are true power series, easier to sum in that they include a z^k factor in each term. There is no one, ideal bound that works equally well for all power series. However, the point of establishing a bound is not to sum a power series exactly but rather to fence the sum within some sufficiently (rather than optimally) small neighborhood. A simple, general bound which works quite adequately for most power series encountered in practice, including among many others all the Taylor series of Table 8.1, is the *geometric majorization*

$$|\epsilon_n| < \frac{|\tau_n|}{1 - |\rho_n|}. \tag{8.33}$$

Here, τ_n represents the power series' nth-order term (in Table 8.1's series for $\exp z$, for example, $\tau_n = z^n/[n!]$). The $|\rho_n|$ is a positive real number chosen, preferably as small as possible, such that

$$\begin{aligned} \left|\frac{\tau_{k+1}}{\tau_k}\right| &\le |\rho_n| \quad \text{for all } k \ge n, && (8.34)\\ \left|\frac{\tau_{k+1}}{\tau_k}\right| &< |\rho_n| \quad \text{for at least one } k \ge n, \\ 0 &< |\rho_n| < 1; && (8.35) \end{aligned}$$

which is to say, more or less, such that each term in the series' tail is smaller than the last by at least a factor of $|\rho_n|$. Given these definitions, if[26]

$$\begin{aligned} S_n &\equiv \sum_{k=0}^{n-1} \tau_k, \\ \epsilon_n &\equiv S_\infty - S_n, \end{aligned} \tag{8.36}$$

where S_∞ represents the true, exact (but uncalculatable, unknown) infinite series sum, then (2.36) and (3.22) imply the geometric majorization (8.33).

If the last paragraph seems abstract, a pair of concrete examples should serve to clarify. First, if the Taylor series

$$-\ln(1 - z) = \sum_{k=1}^{\infty} \frac{z^k}{k}$$

[26] Some scientists and engineers—as, for example, the authors of [127] and even this writer in earlier years—prefer to define $\epsilon_n \equiv S_n - S_\infty$, oppositely as we define it here. This choice is a matter of taste. Professional mathematicians—as, for example, the author of [171]—seem to tend toward the $\epsilon_n \equiv S_\infty - S_n$ of (8.36).

of Table 8.1 is truncated before the nth-order term, then

$$-\ln(1-z) \approx \sum_{k=1}^{n-1} \frac{z^k}{k},$$
$$|\epsilon_n| < \frac{|z^n|/n}{1-|z|},$$

where ϵ_n is the error in the truncated sum.[27] Here, $|\tau_{k+1}/\tau_k| = [k/(k+1)]\,|z| < |z|$ for all $k \ge n > 0$, so we have chosen $|\rho_n| = |z|$.

Second, if the Taylor series

$$\exp z = \sum_{k=0}^{\infty} \prod_{j=1}^{k} \frac{z}{j} = \sum_{k=0}^{\infty} \frac{z^k}{k!}$$

also of Table 8.1 is truncated before the nth-order term, and if we choose to stipulate that

$$n+1 > |z|,$$

then

$$\exp z \approx \sum_{k=0}^{n-1} \prod_{j=1}^{k} \frac{z}{j} = \sum_{k=0}^{n-1} \frac{z^k}{k!},$$
$$|\epsilon_n| < \frac{|z^n|/n!}{1-|z|/(n+1)}.$$

Here, $|\tau_{k+1}/\tau_k| = |z|/(k+1)$, whose maximum value for all $k \ge n$ occurs when $k = n$, so we have chosen $|\rho_n| = |z|/(n+1)$.

8.10.4 Calculation outside the fast convergence domain

Used directly, the Taylor series of Table 8.1 tend to converge slowly for some values of z and not at all for others. The series for $-\ln(1-z)$ and $(1+z)^{a-1}$ for instance each converge for $|z| < 1$ (though slowly for $|z| \approx 1$); whereas each series diverges when asked to compute a quantity like $-\ln 3$ or 3^{a-1} directly. To shift the series' expansion points per § 8.2 is one way to seek convergence, but for nonentire functions (§ 8.6) like these a more probably

[27] This particular error bound fails for $n = 0$, but that is no flaw. There is no reason to use the error bound for $n = 0$ when, merely by taking one or two more terms into the truncated sum, one can quite conveniently let $n = 1$ or $n = 2$.

profitable strategy is to find and exploit some property of the functions to transform their arguments, such as

$$\begin{aligned} -\ln\gamma &= \ln\frac{1}{\gamma}, \\ \gamma^{a-1} &= \frac{1}{(1/\gamma)^{a-1}}, \end{aligned}$$

which leave the respective Taylor series to compute quantities like $-\ln(1/3)$ and $(1/3)^{a-1}$ they can handle.

Let $f(1+\zeta)$ be a function whose Taylor series about $\zeta = 0$ converges for $|\zeta| < 1$ and which obeys properties of the forms[28]

$$\begin{aligned} f(\gamma) &= g\left[f\left(\frac{1}{\gamma}\right)\right], \\ f(\alpha\gamma) &= h\left[f(\alpha), f(\gamma)\right], \end{aligned} \tag{8.37}$$

where $g[\cdot]$ and $h[\cdot,\cdot]$ are functions we know how to compute like $g[\cdot] = -[\cdot]$ or $g[\cdot] = 1/[\cdot]$; and like $h[\cdot,\cdot] = [\cdot] + [\cdot]$ or $h[\cdot,\cdot] = [\cdot][\cdot]$. Identifying

$$\begin{aligned} \frac{1}{\gamma} &= 1+\zeta, \\ \gamma &= \frac{1}{1+\zeta}, \\ \frac{1-\gamma}{\gamma} &= \zeta, \end{aligned} \tag{8.38}$$

we have that

$$f(\gamma) = g\left[f\left(1+\frac{1-\gamma}{\gamma}\right)\right], \tag{8.39}$$

whose convergence domain $|\zeta| < 1$ is $|1-\gamma|\,/\,|\gamma| < 1$, which is $|\gamma - 1| < |\gamma|$ or in other words

$$\Re(\gamma) > \frac{1}{2}.$$

Although the transformation from ζ to γ has not lifted the convergence limit altogether, we see that it has apparently opened the limit to a broader domain.

[28] This paragraph's notation is necessarily abstract. To make it seem more concrete, consider that the function $f(1+\zeta) = -\ln(1-z)$ has $\zeta = -z$, $f(\gamma) = g[f(1/\gamma)] = -f(1/\gamma)$ and $f(\alpha\gamma) = h[f(\alpha), f(\gamma)] = f(\alpha) + f(\gamma)$; and that the function $f(1+\zeta) = (1+z)^{a-1}$ has $\zeta = z$, $f(\gamma) = g[f(1/\gamma)] = 1/f(1/\gamma)$ and $f(\alpha\gamma) = h[f(\alpha), f(\gamma)] = f(\alpha)f(\gamma)$.

For example, if $f(\gamma) = \ln\gamma$ and $\Re(\gamma) > 1/2$, then[29] $g[\cdot] = -[\cdot]$ and thus, according to (8.39),

$$\begin{aligned} \ln\gamma &= -\ln\left(1 + \frac{1-\gamma}{\gamma}\right) \\ &= -\ln(1-z), \quad z \equiv \frac{\gamma - 1}{\gamma}, \end{aligned}$$

a formulation that lets one apply Table 8.1 to calculate, say, ln 3.

Though this writer knows no way to lift the convergence limit altogether that does not cause more problems than it solves, one can take advantage of the $h[\cdot,\cdot]$ property of (8.37) to sidestep the limit, computing $f(\omega)$ indirectly for any $\omega \neq 0$ by any of several tactics. One nonoptimal but entirely effective tactic is represented by the equations

$$\begin{aligned} \omega &\equiv i^n 2^m \gamma, \\ |\Im(\gamma)| &\le \Re(\gamma), \\ 1 \le \Re(\gamma) &< 2, \\ m, n &\in \mathbb{Z}, \end{aligned} \tag{8.40}$$

whereupon the formula

$$f(\omega) = h[f(i^n 2^m), f(\gamma)] \tag{8.41}$$

calculates $f(\omega)$ fast for any $\omega \neq 0$—provided only that we have other means to compute $f(i^n 2^m)$, which not infrequently we do.[30]

Notice how (8.40) fences γ within a comfortable zone, keeping γ moderately small in magnitude but never too near the $\Re(\gamma) = 1/2$ frontier in the Argand plane. In theory all finite γ rightward of the frontier let the Taylor series converge, but extreme γ of any kind let the series converge only slowly (and due to compound floating-point rounding error perhaps inaccurately) inasmuch as they imply that $|\zeta| \approx 1$. Besides allowing all $\omega \neq 0$, the tactic (8.40) also thus significantly speeds series convergence.

The method and tactic of (8.37) through (8.41) are useful in themselves and also illustrative generally. Of course, most nonentire functions lack

[29] See footnote 28.

[30] Equation (8.41) admittedly leaves open the question of how to compute $f(i^n 2^m)$, but at least for the functions this subsection has used as examples this is not hard. For the logarithm, $-\ln(i^n 2^m) = m \ln(1/2) - in(2\pi/4)$. For the power, $(i^n 2^m)^{a-1} = \operatorname{cis}[(n2\pi/4)(a-1)]/[(1/2)^{a-1}]^m$. The sine and cosine in the cis function are each calculated directly by Taylor series (possibly with the help of Table 3.1), as are the numbers $\ln(1/2)$ and $(1/2)^{a-1}$. The number 2π, we have not calculated yet, but will in § 8.11.

properties of the specific kinds that (8.37) demands, but such functions may have other properties one can analogously exploit.[31]

8.10.5 Divergent series

Variants of this section's techniques can be used to prove that a series does not converge at all. For example,

$$\sum_{k=1}^{\infty} \frac{1}{k}$$

does not converge because

$$\frac{1}{k} > \int_k^{k+1} \frac{d\tau}{\tau};$$

hence,

$$\sum_{k=1}^{\infty} \frac{1}{k} > \sum_{k=1}^{\infty} \int_k^{k+1} \frac{d\tau}{\tau} = \int_1^{\infty} \frac{d\tau}{\tau} = \ln \infty.$$

8.10.6 Remarks

The study of error bounds is not a matter of rules and formulas so much as of ideas, suggestions and tactics. As far as the writer knows, there is no such thing as an optimal error bound—with sufficient cleverness, some tighter bound can usually be discovered—but often easier and more effective than cleverness is simply to add a few extra terms into the series before truncating it (that is, to increase n a little). To eliminate the error entirely usually demands adding an infinite number of terms, which is impossible;

[31] To draw another example from Table 8.1, consider that

$$\begin{aligned} \arctan \omega &= \alpha + \arctan \zeta, \\ \zeta &\equiv \frac{\omega \cos\alpha - \sin\alpha}{\omega \sin\alpha + \cos\alpha}, \end{aligned}$$

where $\arctan \omega$ is interpreted as the geometrical angle the vector $\hat{\mathbf{x}} + \hat{\mathbf{y}}\omega$ makes with $\hat{\mathbf{x}}$. Axes are rotated per (3.7) through some angle α to reduce the tangent from ω to ζ, where $\arctan \zeta$ is interpreted as the geometrical angle the vector $\hat{\mathbf{x}} + \hat{\mathbf{y}}\omega = \hat{\mathbf{x}}'(\omega \sin\alpha + \cos\alpha) + \hat{\mathbf{y}}'(\omega \cos\alpha - \sin\alpha)$ makes with $\hat{\mathbf{x}}'$, thus causing the Taylor series to converge faster or indeed to converge at all.

Any number of further examples and tactics of the kind will occur to the creative reader, shrinking a function's argument by some convenient means before feeding the argument to the function's Taylor series.

but since eliminating the error entirely also requires recording the sum to infinite precision, which is impossible anyway, eliminating the error entirely is not normally a goal one seeks. To eliminate the error to the 0x34-bit (sixteen-decimal place) precision of a computer's `double`-type floating-point representation typically requires something like 0x34 terms—if the series be wisely composed and if care be taken to keep z moderately small and reasonably distant from the edge of the series' convergence domain. Besides, few engineering applications really use much more than 0x10 bits (five decimal places) in any case. Perfect precision is impossible, but adequate precision is usually not hard to achieve.

Occasionally nonetheless a series arises for which even adequate precision is quite hard to achieve. An infamous example is

$$S = -\sum_{k=1}^{\infty} \frac{(-)^k}{\sqrt{k}} = 1 - \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} - \frac{1}{\sqrt{4}} + \cdots ,$$

which obviously converges, but sum it if you can! It is not easy to do.

Before closing the section, we ought to arrest one potential agent of terminological confusion. The "error" in a series summation's error bounds is unrelated to the error of probability theory (chapter 20). The English word "error" is thus overloaded here. A series sum converges to a definite value, and to the same value every time the series is summed; no chance is involved. It is just that we do not necessarily know exactly what that value is. What we can do, by this section's techniques or perhaps by other methods, is to establish a definite neighborhood in which the unknown value is sure to lie; and we can make that neighborhood as tight as we want, merely by including a sufficient number of terms in the sum.

The topic of series error bounds is what G. S. Brown refers to as "trick-based."[32] There is no general answer to the error-bound problem, but there are several techniques which help, some of which this section has introduced. Other techniques, we shall meet later in the book as the need for them arises.

8.11 Calculating 2π

The Taylor series for $\arctan z$ in Table 8.1 implies a neat way of calculating the constant 2π. We already know that $\tan(2\pi/8) = 1$, or in other words that

$$\arctan 1 = \frac{2\pi}{8}.$$

[32] [30]

Applying the Taylor series, we have that

$$2\pi = 8 \sum_{k=0}^{\infty} \frac{(-)^k}{2k+1}. \tag{8.42}$$

The series (8.42) is simple but converges extremely slowly. Much faster convergence is given by angles smaller than $2\pi/8$. For example, from Table 3.2,

$$\arctan\left(2 - \sqrt{3}\right) = \frac{2\pi}{\mathrm{0x18}}.$$

Applying the Taylor series at this angle, we have that[33,34]

$$2\pi = \mathrm{0x18} \sum_{k=0}^{\infty} \frac{(-)^k}{2k+1} \left(2 - \sqrt{3}\right)^{2k+1}$$

or, since $(2 - \sqrt{3})^2 = 7 - 4\sqrt{3}$,

$$\begin{aligned} 2\pi &= \mathrm{0x18} \left(2 - \sqrt{3}\right) \sum_{k=0}^{\infty} \frac{\left(4\sqrt{3} - 7\right)^k}{2k+1} \approx \mathrm{0x6.487F} \qquad (8.43) \\ &= \mathrm{0x6.487E\,D511\,0B46\,11A6\,2633\,145C\,06E0\,E689}\ldots \end{aligned}$$

the series' terms conveniently alternating in sign due to that $(4\sqrt{3})^2 < 7^2$.

8.12 Odd and even functions

An *odd function* is one for which $f(-z) = -f(z)$. Any function whose Taylor series about $z_o = 0$ includes only odd-order terms is an odd function. Examples of odd functions include z^3 and $\sin z$.

An *even function* is one for which $f(-z) = f(z)$. Any function whose Taylor series about $z_o = 0$ includes only even-order terms is an even function. Examples of even functions include z^2 and $\cos z$.

[33] [152, sequence A004601]

[34] The writer is given to understand that clever mathematicians have invented subtle, still much faster-converging iterative schemes toward 2π. However, there is fast and there is fast. The relatively straightforward series this section gives converges to the best accuracy of your computer's floating-point register within a paltry fourteen (0xE) iterations—and, after all, you only need to compute the numerical value of 2π once.

Useful lessons may lurk in the clever mathematics underlying the subtle schemes, but such schemes are not covered here.

For what it's worth, during 2022 the writer's laptop computer took about three minutes to calculate 2^{0x12} bits of 2π—that is, to calculate a little fewer than eighty thousand decimal digits—using (8.43).

Odd and even functions are interesting because of the symmetry they bring—the plot of a real-valued odd function being symmetric about a point, the plot of a real-valued even function being symmetric about a line. Many functions are neither odd nor even, of course, but one can always split an analytic function into two components—one odd, the other even—by the simple expedient of sorting the odd-order terms from the even-order in the function's Taylor series. For example, $\exp z = \sinh z + \cosh z$. Alternately,

$$\begin{aligned} f(z) &= f_{\mathrm{odd}}(z) + f_{\mathrm{even}}(z), \\ f_{\mathrm{odd}}(z) &= \frac{f(z) - f(-z)}{2}, \\ f_{\mathrm{even}}(z) &= \frac{f(z) + f(-z)}{2}, \end{aligned} \tag{8.44}$$

the latter two lines of which are verified by substituting $-z \leftarrow z$ and observing the definitions at the section's head of odd and even, and then the first line of which is verified by adding the latter two.

That the derivative of an odd function is even and that the derivative of an even function is odd should require little explanation if the concepts of oddness, evenness and the derivative are grasped.[35]

Section 18.2.9 will have more to say about odd and even functions.

8.13 Trigonometric poles

The singularities of the trigonometric functions are single poles of residue ± 1 or $\pm i$. For the circular trigonometrics, all the poles lie along the real number line; for the hyperbolic trigonometrics, along the imaginary. Specifically, of the eight trigonometric functions

$$\begin{gathered} \frac{1}{\sin z}, \frac{1}{\cos z}, \frac{1}{\tan z}, \tan z, \\ \frac{1}{\sinh z}, \frac{1}{\cosh z}, \frac{1}{\tanh z}, \tanh z, \end{gathered}$$

[35] Appendix D has little to do with the present chapter but we mention the appendix here in this footnote because a reader that has studied the book up to this point has in principle studied just enough to understand the appendix.

the poles and their respective residues are

$$\begin{aligned}
\left.\frac{z-k\pi}{\sin z}\right|_{z\,=\,k\pi} &= (-)^k, \\
\left.\frac{z-(k-1/2)\pi}{\cos z}\right|_{z\,=\,(k-1/2)\pi} &= (-)^k, \\
\left.\frac{z-k\pi}{\tan z}\right|_{z\,=\,k\pi} &= 1, \\
\left.[z-(k-1/2)\pi]\tan z\right|_{z\,=\,(k-1/2)\pi} &= -1, \\
\left.\frac{z-ik\pi}{\sinh z}\right|_{z\,=\,ik\pi} &= (-)^k, \\
\left.\frac{z-i(k-1/2)\pi}{\cosh z}\right|_{z\,=\,i(k-1/2)\pi} &= (-)^k i, \\
\left.\frac{z-ik\pi}{\tanh z}\right|_{z\,=\,ik\pi} &= 1, \\
\left.[z-i(k-1/2)\pi]\tanh z\right|_{z\,=\,i(k-1/2)\pi} &= 1, \\
k &\in \mathbb{Z}.
\end{aligned} \tag{8.45}$$

To support (8.45)'s claims, we shall marshal the identities of Tables 5.1 and 5.2 plus l'Hôpital's rule (4.29). Before calculating residues and such, however, we should like to verify that the poles (8.45) lists are in fact the only poles that there are; that we have forgotten no poles. Consider for instance the function $1/\sin z = i2/(e^{iz} - e^{-iz})$. This function evidently goes infinite only when $e^{iz} = e^{-iz}$, which is possible only for real z; but for real z, the sine function's very definition establishes the poles $z = k\pi$ (refer to Fig. 3.1). With the observations from Table 5.1 that $i \sinh z = \sin iz$ and $\cosh z = \cos iz$, similar reasoning for each of the eight trigonometrics forbids poles other than those (8.45) lists. Satisfied that we have forgotten no poles, therefore, we finally apply l'Hôpital's rule to each of the ratios

$$\begin{gathered}
\frac{z-k\pi}{\sin z}, \frac{z-(k-1/2)\pi}{\cos z}, \frac{z-k\pi}{\tan z}, \frac{z-(k-1/2)\pi}{1/\tan z}, \\
\frac{z-ik\pi}{\sinh z}, \frac{z-i(k-1/2)\pi}{\cosh z}, \frac{z-ik\pi}{\tanh z}, \frac{z-i(k-1/2)\pi}{1/\tanh z}
\end{gathered}$$

to reach (8.45).

Trigonometric poles evidently are special only in that a trigonometric function has an infinite number of them. The poles are ordinary, single poles, with residues, subject to Cauchy's integral formula and so on. The trigonometrics are meromorphic functions (§ 8.6) for this reason.[36]

The six simpler trigonometrics, $\sin z$, $\cos z$, $\sinh z$, $\cosh z$, $\operatorname{cis} z$ and $\exp z$—conspicuously excluded from this section's gang of eight—have no poles for finite z because $e^{\pm iz}$ and $e^{\pm z}$ are finite. These simpler trigonometrics are not only meromorphic but also entire. Observe however that the *inverse* trigonometrics are multiple-valued and have branch points, and thus are not meromorphic at all.

8.14 The Laurent series

Any analytic function can be expanded in a Taylor series, but never about a pole or branch point of the function. Sometimes one nevertheless wants to expand at least about a pole. Consider for example expanding

$$f(z) = \frac{e^{-z}}{1 - \cos z} \tag{8.46}$$

about the function's pole at $z = 0$. Expanding dividend and divisor separately,

$$\begin{aligned} f(z) &= \frac{1 - z + z^2/2 - z^3/6 + \cdots}{z^2/2 - z^4/\mathrm{0x18} + \cdots} \\ &= \frac{\sum_{j=0}^{\infty} \left[(-)^j z^j / j! \right]}{-\sum_{k=1}^{\infty} (-)^k z^{2k}/(2k)!} \\ &= \frac{\sum_{k=0}^{\infty} \left[-z^{2k}/(2k)! + z^{2k+1}/(2k+1)! \right]}{\sum_{k=1}^{\infty} (-)^k z^{2k}/(2k)!}. \end{aligned}$$

[36] [101]

By long division,

$$\begin{aligned} f(z) &= \frac{2}{z^2} - \frac{2}{z} + \left\{ \left[-\frac{2}{z^2} + \frac{2}{z} \right] \sum_{k=1}^{\infty} \frac{(-)^k z^{2k}}{(2k)!} \right. \\ &\qquad \left. + \sum_{k=0}^{\infty} \left[-\frac{z^{2k}}{(2k)!} + \frac{z^{2k+1}}{(2k+1)!} \right] \right\} \bigg/ \sum_{k=1}^{\infty} \frac{(-)^k z^{2k}}{(2k)!} \\ &= \frac{2}{z^2} - \frac{2}{z} + \left\{ \sum_{k=1}^{\infty} \left[-\frac{(-)^k 2 z^{2k-2}}{(2k)!} + \frac{(-)^k 2 z^{2k-1}}{(2k)!} \right] \right. \\ &\qquad \left. + \sum_{k=0}^{\infty} \left[-\frac{z^{2k}}{(2k)!} + \frac{z^{2k+1}}{(2k+1)!} \right] \right\} \bigg/ \sum_{k=1}^{\infty} \frac{(-)^k z^{2k}}{(2k)!} \\ &= \frac{2}{z^2} - \frac{2}{z} + \left\{ \sum_{k=0}^{\infty} \left[\frac{(-)^k 2 z^{2k}}{(2k+2)!} - \frac{(-)^k 2 z^{2k+1}}{(2k+2)!} \right] \right. \\ &\qquad \left. + \sum_{k=0}^{\infty} \left[-\frac{z^{2k}}{(2k)!} + \frac{z^{2k+1}}{(2k+1)!} \right] \right\} \bigg/ \sum_{k=1}^{\infty} \frac{(-)^k z^{2k}}{(2k)!} \\ &= \frac{2}{z^2} - \frac{2}{z} + \sum_{k=0}^{\infty} \left[\frac{-(2k+1)(2k+2) + (-)^k 2}{(2k+2)!} z^{2k} \right. \\ &\qquad \left. + \frac{(2k+2) - (-)^k 2}{(2k+2)!} z^{2k+1} \right] \bigg/ \sum_{k=1}^{\infty} \frac{(-)^k z^{2k}}{(2k)!}. \end{aligned}$$

The remainder's $k = 0$ terms now disappear as intended; so, factoring z^2/z^2 from the division leaves

$$\begin{aligned} f(z) &= \frac{2}{z^2} - \frac{2}{z} + \sum_{k=0}^{\infty} \left[\frac{(2k+3)(2k+4) + (-)^k 2}{(2k+4)!} z^{2k} \right. \\ &\qquad \left. - \frac{(2k+4) + (-)^k 2}{(2k+4)!} z^{2k+1} \right] \bigg/ \sum_{k=0}^{\infty} \frac{(-)^k z^{2k}}{(2k+2)!}. \end{aligned} \tag{8.47}$$

One can continue dividing to extract further terms if desired, and if all the terms

$$f(z) = \frac{2}{z^2} - \frac{2}{z} + \frac{7}{6} - \frac{z}{2} + \cdots$$

are extracted the result is the *Laurent series* proper,

$$f(z) = \sum_{k=K}^{\infty} (a_k)(z - z_o)^k, \quad (k, K) \in \mathbb{Z}, \ K \le 0. \tag{8.48}$$

However for many purposes (as in eqn. 8.47) the partial Laurent series

$$\begin{aligned} f(z) &= \sum_{k=K}^{-1} (a_k)(z-z_o)^k + \frac{\sum_{k=0}^{\infty}(b_k)(z-z_o)^k}{\sum_{k=0}^{\infty}(c_k)(z-z_o)^k}, \\ (k,K) &\in \mathbb{Z},\ \ K \le 0,\ c_0 \ne 0, \end{aligned} \tag{8.49}$$

suffices and may even be preferable. In either form,

$$f(z) = \sum_{k=K}^{-1} (a_k)(z-z_o)^k + f_o(z-z_o),\ \ (k,K) \in \mathbb{Z},\ K \le 0, \tag{8.50}$$

where, unlike $f(z)$, $f_o(z-z_o)$ is analytic at $z = z_o$. The $f_o(z-z_o)$ of (8.50) is $f(z)$'s *regular part* at $z = z_o$.

The ordinary Taylor series diverges at a function's pole. Handling the pole separately, the Laurent series remedies this defect.[37,38]

Sections 9.6 and 9.7 tell more about poles generally, including multiple poles like the one in the example here.

8.15 Taylor series in $1/z$

A little imagination helps the Taylor series a lot. The Laurent series of § 8.14 represents one way to extend the Taylor series. Several other ways are possible. The typical trouble one has with the Taylor series is that a function's poles and branch points limit the series' convergence domain. Thinking flexibly, however, one can often evade the trouble.

Consider the function

$$f(z) = \frac{\sin(1/z)}{\cos z}.$$

This function has a nonanalytic point of a most peculiar nature at $z = 0$. The point is an essential singularity, and one cannot expand the function directly about it. One could expand the function directly about some other point like $z = 1$, but calculating the Taylor coefficients would take a lot of effort and, even then, the resulting series would suffer a straitly limited

[37] The professional mathematician's treatment of the Laurent series usually seems to begin by defining an annular convergence domain (a convergence domain bounded without by a large circle and within by a small) in the Argand plane. From an applied point of view however what interests us is the basic technique to remove the poles from an otherwise analytic function.

[38] [79, § 10.8][58, § 2.5]

convergence domain. All that however tries too hard. Depending on the application, it may suffice to write,

$$f(z) = \frac{\sin w}{\cos z}, \quad w \equiv \frac{1}{z}.$$

This is

$$f(z) = \frac{z^{-1} - z^{-3}/3! + z^{-5}/5! - \cdots}{1 - z^2/2! + z^4/4! - \cdots},$$

which is all one needs to calculate $f(z)$ numerically—and may be all one needs for analysis, too.

As an example of a different kind, consider

$$g(z) = \frac{1}{(z-2)^2}.$$

Most often, one needs no Taylor series to handle such a function (one simply does the indicated arithmetic). Suppose however that a Taylor series specifically about $z = 0$ were indeed needed for some reason. Then by (8.1) and (4.2),

$$g(z) = \frac{1/4}{(1-z/2)^2} = \frac{1}{4}\sum_{k=0}^{\infty}\binom{1+k}{1}\left[\frac{z}{2}\right]^k = \sum_{k=0}^{\infty}\frac{k+1}{2^{k+2}}z^k,$$

That expansion is good only for $|z| < 2$, but for $|z| > 2$ we also have that

$$g(z) = \frac{1/z^2}{(1-2/z)^2} = \frac{1}{z^2}\sum_{k=0}^{\infty}\binom{1+k}{1}\left[\frac{2}{z}\right]^k = \sum_{k=2}^{\infty}\frac{2^{k-2}(k-1)}{z^k},$$

which expands in negative rather than positive powers of z. Note that we have computed the two series for $g(z)$ without ever actually taking a derivative.

Neither of the section's examples is especially interesting in itself, but their point is that it often pays to think flexibly in extending and applying the Taylor series. One is not required immediately to take the Taylor series of a function as it presents itself; one can first change variables or otherwise rewrite the function in some convenient way, and then take the Taylor series either of the whole function at once or of pieces of it separately. One can expand in negative powers of z equally validly as in positive powers. And, though taking derivatives per (8.19) may be the canonical way to determine Taylor coefficients, any effective means to find the coefficients suffices.

8.16 The multidimensional Taylor series

Equation (8.19) has given the Taylor series for functions of a single variable. The idea of the Taylor series does not differ where there are two or more independent variables, only the details are a little more complicated. For example, consider the function $f(z_1, z_2) = z_1^2 + z_1 z_2 + 2z_2$, which has terms z_1^2 and $2z_2$—these we understand—but also has the cross-term $z_1 z_2$ for which the relevant derivative is the cross-derivative $\partial^2 f/\partial z_1\,\partial z_2$. Where two or more independent variables are involved, one must account for the cross-derivatives, too.

With this idea in mind, the multidimensional Taylor series is

$$f(\mathbf{z}) = \sum_{\mathbf{k}} \left(\left. \frac{\partial^{\mathbf{k}} f}{\partial \mathbf{z}^{\mathbf{k}}} \right|_{\mathbf{z}=\mathbf{z}_o} \right) \frac{(\mathbf{z} - \mathbf{z}_o)^{\mathbf{k}}}{\mathbf{k}!}. \tag{8.51}$$

Well, that's neat. What does it mean?

- The $\mathbf{z}$ is a *vector*[39] incorporating the several independent variables $z_1, z_2, \ldots, z_N$.
- The $\mathbf{k}$ is a nonnegative integer vector of N counters—$k_1, k_2, \ldots, k_N$—one for each of the independent variables. Each of the k_n runs independently from 0 to ∞, and every permutation is possible. For example, if $N = 2$ then

$$\begin{aligned}
\mathbf{k} &= (k_1, k_2) \\
&= (0,0), (0,1), (0,2), (0,3), \ldots; \\
&\ (1,0), (1,1), (1,2), (1,3), \ldots; \\
&\ (2,0), (2,1), (2,2), (2,3), \ldots; \\
&\ (3,0), (3,1), (3,2), (3,3), \ldots; \\
&\ \ldots
\end{aligned}$$

- The $\partial^{\mathbf{k}} f/\partial \mathbf{z}^{\mathbf{k}}$ represents the $\mathbf{k}$th cross-derivative of $f(\mathbf{z})$, meaning that

$$\frac{\partial^{\mathbf{k}} f}{\partial \mathbf{z}^{\mathbf{k}}} \equiv \left(\prod_{n=1}^{N} \frac{\partial^{k_n}}{(\partial z_n)^{k_n}} \right) f.$$

[39] In this generalized sense of the word, a *vector* is an ordered set of N elements. The geometrical vector $\mathbf{v} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y + \hat{\mathbf{z}}z$ of § 3.3, then, is a vector with $N = 3$, $v_1 = x$, $v_2 = y$ and $v_3 = z$. (Generalized vectors of arbitrary N will figure prominently in the book from chapter 11 onward.)

- The $(\mathbf{z}-\mathbf{z}_o)^{\mathbf{k}}$ represents

$$(\mathbf{z}-\mathbf{z}_o)^{\mathbf{k}} \equiv \prod_{n=1}^{N} (z_n - z_{on})^{k_n}.$$

- The $\mathbf{k}!$ represents

$$\mathbf{k}! \equiv \prod_{n=1}^{N} k_n!.$$

With these definitions, the multidimensional Taylor series (8.51) yields all the right derivatives and cross-derivatives at the expansion point $\mathbf{z} = \mathbf{z}_o$. Thus within some convergence domain about $\mathbf{z} = \mathbf{z}_o$, the multidimensional Taylor series (8.51) represents a function $f(\mathbf{z})$ as accurately as the simple Taylor series (8.19) represents a function $f(z)$, and for the same reason.

Chapter 9

Integration techniques

Equation (4.13) implies a general technique for calculating a derivative symbolically. Its counterpart (7.1), unfortunately, implies a general technique only for calculating an integral *numerically*—and even for this purpose it is imperfect; for, when it comes to adding an infinite number of infinitesimal elements, how is one actually to do the sum?

It turns out that there is no one general answer to this question. Some functions are best integrated by one technique, some by another. It is hard to guess in advance which technique might work best.

This chapter surveys several weapons of the intrepid mathematician's arsenal against the integral.

9.1 Integration by antiderivative

The simplest way to solve an integral is just to look at it, recognizing its integrand to be the derivative of something already known:[1]

$$\int_a^z \frac{df}{d\tau}\,d\tau = f(\tau)|_a^z. \tag{9.1}$$

For instance,

$$\int_1^x \frac{1}{\tau}\,d\tau = \ln\tau|_1^x = \ln x.$$

One merely looks at the integrand $1/\tau$, recognizing it to be the derivative of $\ln\tau$, and then directly writes down the solution $\ln\tau|_1^x$. Refer to § 7.2.

[1] The notation $f(\tau)|_a^z$ or $[f(\tau)]_a^z$ means $f(z) - f(a)$.

Again for instance,

$$\int_{1/2}^{x} \frac{1}{\sqrt{1-\tau^2}}\,d\tau = \arcsin \tau|_{1/2}^{x} = \arcsin x - \frac{2\pi}{\mathrm{0xC}},$$

or, if you prefer decimal notation,

$$\int_{1/2}^{x} \frac{1}{\sqrt{1-\tau^2}}\,d\tau = \arcsin \tau|_{1/2}^{x} = \arcsin x - \frac{2\pi}{12}.$$

Refer to Table 3.2 and, more importantly, to Table 5.3.

The technique by itself is pretty limited. However, the frequent object of other integration techniques is to transform an integral into a form to which this basic technique can be applied.

Besides the essential

$$\tau^{a-1} = \frac{d}{d\tau}\left(\frac{\tau^a}{a}\right), \tag{9.2}$$

Tables 7.1, 5.2, 5.3 and 9.1 provide several further good derivatives this antiderivative technique can use.

One particular, nonobvious, useful variation on the antiderivative technique seems worth calling out specially here. If $z = \rho e^{i\phi}$, then (8.24) and (8.25) have that

$$\int_{z_1}^{z_2} \frac{dz}{z} = \ln\frac{\rho_2}{\rho_1} + i(\phi_2 - \phi_1). \tag{9.3}$$

This helps, for example, when z_1 and z_2 are real but negative numbers.

9.2 Integration by substitution

Consider the integral

$$S \equiv \int_{x_1}^{x_2} \frac{x\,dx}{1+x^2}.$$

This integral is not in a form one immediately recognizes. However, with the change of variable

$$u \leftarrow 1 + x^2,$$

whose differential by successive steps is

$$\begin{aligned} d(u) &= d(1+x^2), \\ du &= 2x\,dx, \end{aligned}$$

the integral is

$$\begin{aligned}
S &= \int_{x=x_1}^{x_2} \frac{x\,dx}{u} \\
&= \int_{x=x_1}^{x_2} \frac{2x\,dx}{2u} \\
&= \int_{u=1+x_1^2}^{1+x_2^2} \frac{du}{2u} \\
&= \frac{1}{2} \ln u \bigg|_{u=1+x_1^2}^{1+x_2^2} \\
&= \frac{1}{2} \ln \frac{1+x_2^2}{1+x_1^2}.
\end{aligned}$$

To check the result, we can take the derivative per § 7.5 of the final expression with respect to x_2:

$$\begin{aligned}
\frac{\partial}{\partial x_2} \frac{1}{2} \ln \frac{1+x_2^2}{1+x_1^2}\bigg|_{x_2=x} &= \left[\frac{1}{2} \frac{\partial}{\partial x_2} \left\{ \ln\left(1+x_2^2\right) - \ln\left(1+x_1^2\right) \right\} \right]_{x_2=x} \\
&= \frac{x}{1+x^2},
\end{aligned}$$

which indeed has the form of the integrand with which we started.

The technique is *integration by substitution.* It does not solve all integrals but it does solve many, whether alone or in combination with other techniques.

9.3 Reversal and scaling of the independent variable

Section 9.2 has introduced integration by substitution. Many substitutions are possible but the simple change of variable

$$\begin{aligned}
-u &\leftarrow x, \\
-du &= dx,
\end{aligned} \tag{9.4}$$

is so easy a substitution, and is so often helpful, that it merits a section of its own.

If

$$S \equiv \int_{x=a}^{b} f(x)\,dx, \tag{9.5}$$

where $f(x)$ is a function one wishes to integrate, then changing variables according to (9.4) gives that

$$S = \int_{-u=a}^{b} f(-u)(-du) = -\int_{u=-a}^{-b} f(-u)\,du,$$

which is that

$$S = \int_{u=-b}^{-a} f(-u)\,du. \tag{9.6}$$

Even easier is the case that $a = -\infty$, $b = \infty$, in which

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{\infty} f(-u)\,du. \tag{9.7}$$

The technique of reversal of the independent variable seldom solves an integral on its own but can put an integral in a form to which other techniques can be applied, as for example in § 18.2.6.

Related is the change of variable

$$\begin{aligned} ku &\leftarrow x, \\ k\,du &= dx, \\ \Im(k) &= 0, \end{aligned} \tag{9.8}$$

by which

$$S = k\int_{u=a/k}^{b/k} f(ku)\,du. \tag{9.9}$$

In the case that $a = -\infty$, $b = \infty$,

$$\int_{-\infty}^{\infty} f(x)\,dx = |k|\int_{-\infty}^{\infty} f(ku)\,du. \tag{9.10}$$

9.4 Integration by parts

Integration by parts is a curious but very broadly applicable technique which begins with the derivative product rule (4.22),

$$d(uv) = u\,dv + v\,du,$$

where $u(\tau)$ and $v(\tau)$ are functions of an independent variable τ. Reordering terms,

$$u\,dv = d(uv) - v\,du.$$

Integrating,

$$\int_{\tau=a}^{b} u\,dv = uv|_{\tau=a}^{b} - \int_{\tau=a}^{b} v\,du. \tag{9.11}$$

Equation (9.11) is the rule of *integration by parts.*

For an example of the rule's operation, consider the integral

$$S(x) = \int_0^x \tau \cos\alpha\tau\,d\tau.$$

Unsure how to integrate this, we can begin by integrating *part* of it. We can begin by integrating the $\cos\alpha\tau\,d\tau$ part. Letting

$$\begin{aligned} u &\leftarrow \tau, \\ dv &\leftarrow \cos\alpha\tau\,d\tau, \end{aligned}$$

we find that[2]

$$\begin{aligned} du &= d\tau, \\ v &= \frac{\sin\alpha\tau}{\alpha}. \end{aligned}$$

According to (9.11), then,

$$S(x) = \left.\frac{\tau\sin\alpha\tau}{\alpha}\right|_0^x - \int_0^x \frac{\sin\alpha\tau}{\alpha}\,d\tau = \frac{x}{\alpha}\sin\alpha x + \cos\alpha x - 1.$$

Though integration by parts is a powerful technique, one should understand clearly what it does and does not do. The technique does not just integrate each part of an integral separately. It isn't that simple. What it does is to integrate one part of an integral separately—whichever part one has chosen to identify as dv—while contrarily differentiating the other part u, upon which it rewards the mathematician only with a new integral $\int v\,du$. The new integral may or may not be easier to integrate than was the original $\int u\,dv$. The virtue of the technique lies in that one can often find a part dv which does yield an easier $\int v\,du$. The technique is powerful for this reason.

For another kind of example of the rule's operation, see § 21.3 in the chapter on the gamma function.

[2] The careful reader will observe that $v = (\sin\alpha\tau)/\alpha + C$ matches the chosen dv for any value of C, not just for $C = 0$. This is true. However, nothing in the technique of integration by parts requires us to consider all possible v. Any convenient v suffices. In this case, we choose $v = (\sin\alpha\tau)/\alpha$.

9.5 Integration by unknown coefficients

One of the more powerful integration techniques is relatively inelegant, yet it easily cracks some integrals that give other techniques trouble. The technique is the *method of unknown coefficients,* and it is based on the antiderivative (9.1) plus intelligent guessing. It is best illustrated by example.

Consider the integral (which arises in probability theory)

$$S(x) = \int_0^x e^{-(\rho/\sigma)^2/2}\rho\,d\rho. \tag{9.12}$$

If one does not know how to solve the integral in a more elegant way, one can *guess* a likely-seeming antiderivative form, such as

$$e^{-(\rho/\sigma)^2/2}\rho = \frac{d}{d\rho}ae^{-(\rho/\sigma)^2/2},$$

where the a is an *unknown coefficient.* Having guessed, one has no guarantee that the guess is right, but see: if the guess *were* right, then the antiderivative would have the form

$$\begin{aligned} e^{-(\rho/\sigma)^2/2}\rho &= \frac{d}{d\rho}ae^{-(\rho/\sigma)^2/2} \\ &= -\frac{a\rho}{\sigma^2}e^{-(\rho/\sigma)^2/2}, \end{aligned}$$

implying that

$$a = -\sigma^2$$

(evidently the guess is right, after all). Using this value for a, one can write the specific antiderivative

$$e^{-(\rho/\sigma)^2/2}\rho = \frac{d}{d\rho}\left[-\sigma^2 e^{-(\rho/\sigma)^2/2}\right],$$

with which one can solve the integral, concluding that

$$S(x) = \left[-\sigma^2 e^{-(\rho/\sigma)^2/2}\right]_0^x = \left(\sigma^2\right)\left[1 - e^{-(x/\sigma)^2/2}\right]. \tag{9.13}$$

The same technique solves differential equations, too. Consider for example the differential equation

$$dx = (Ix - P)\,dt, \quad x|_{t=0} = x_o, \; x|_{t=T} = 0, \tag{9.14}$$

which conceptually represents[3] the changing balance x of a bank loan account over time t, where I is the loan's interest rate and P is the borrower's payment rate. If it is desired to find the correct payment rate P which pays the loan off in the time T, then (perhaps after some bad guesses) we guess the form

$$x(t) = Ae^{\alpha t} + B,$$

where α, A and B are unknown coefficients. The guess' derivative is

$$dx = \alpha A e^{\alpha t}\,dt.$$

Substituting the last two equations into (9.14) and dividing by dt yields that

$$\alpha A e^{\alpha t} = IAe^{\alpha t} + IB - P,$$

which at least is satisfied if both of the equations

$$\begin{aligned}\alpha A e^{\alpha t} &= IAe^{\alpha t},\\ 0 &= IB - P,\end{aligned}$$

are satisfied. Evidently good choices for α and B, then, are

$$\begin{aligned}\alpha &= I,\\ B &= \frac{P}{I}.\end{aligned}$$

Substituting these coefficients into the $x(t)$ equation above yields the general solution

$$x(t) = Ae^{It} + \frac{P}{I} \tag{9.15}$$

to (9.14). The constants A and P, we establish by applying the given *boundary conditions* $x|_{t=0} = x_o$ and $x|_{t=T} = 0$. For the former condition, (9.15) is

$$x_o = Ae^{(I)(0)} + \frac{P}{I} = A + \frac{P}{I};$$

and for the latter condition,

$$0 = Ae^{IT} + \frac{P}{I}.$$

[3] Real banks (in the author's country, at least) by law or custom actually use a needlessly more complicated formula—and not only more complicated, but mathematically slightly incorrect, too.

Solving the last two equations simultaneously, we have that

$$\begin{aligned} A &= \frac{-e^{-IT}x_o}{1-e^{-IT}}, \\ P &= \frac{Ix_o}{1-e^{-IT}}. \end{aligned} \tag{9.16}$$

Applying these to the general solution (9.15) yields the specific solution

$$x(t) = \frac{x_o}{1-e^{-IT}}\left[1-e^{(I)(t-T)}\right] \tag{9.17}$$

to (9.14) meeting the boundary conditions, with the payment rate P required of the borrower given by (9.16).

The virtue of the method of unknown coefficients lies in that it permits one to try an entire family of candidate solutions at once, with the family members distinguished by the values of the coefficients. If a solution exists anywhere in the family, the method usually finds it.

The method of unknown coefficients is an elephant. Slightly inelegant the method may be, but it is pretty powerful, too—and it has surprise value (for some reason people seem not to expect it). Such are the kinds of problems the method can solve.

9.6 Integration by closed contour

We pass now from the elephant to the falcon, from the inelegant to the sublime. Consider the definite integral[4]

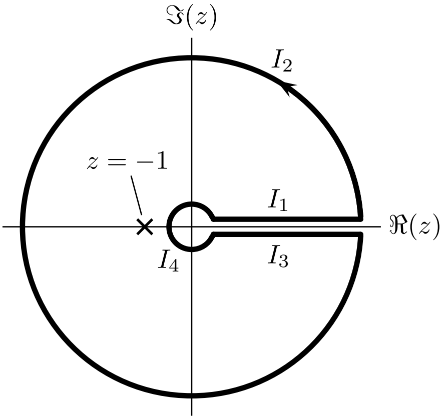
$$S \equiv \int_0^\infty \frac{\tau^a}{\tau+1}\,d\tau, \quad -1 < a < 0.$$

This is a hard integral. No obvious substitution, no evident factoring into parts, seems to solve the integral; but there is a way. The integrand has a pole at $\tau = -1$. Observing that τ is only a dummy integration variable, if one writes the same integral using the complex variable z in place of the real variable τ, then Cauchy's integral formula (8.29) has that integrating once counterclockwise about a closed complex contour, with the contour enclosing the pole at $z = -1$ but shutting out the branch point at $z = 0$, yields that

$$I = \oint \frac{z^a}{z+1}\,dz = i2\pi z^a|_{z=-1} = i2\pi\left(e^{i2\pi/2}\right)^a = i2\pi e^{i2\pi a/2}.$$

[4] [107, § 1.2]

Figure 9.1: Integration by closed contour.



The trouble, of course, is that the integral S does not go about a closed complex contour. One can however construct a closed complex contour I of which S is a part, as in Fig 9.1. If the outer circle in the figure is of infinite radius and the inner, of infinitesimal, then the closed contour I is composed of the four parts

$$\begin{aligned} I &= I_1 + I_2 + I_3 + I_4 \\ &= (I_1 + I_3) + I_2 + I_4. \end{aligned}$$

The figure tempts one to make the mistake of writing that $I_1 = S = -I_3$, but besides being incorrect this defeats the purpose of the closed contour technique. More subtlety is needed. One must take care to interpret the four parts correctly. The integrand $z^a/(z+1)$ is multiple-valued; so, in fact, the two parts $I_1 + I_3 \neq 0$ do not cancel. The integrand has a branch point at $z = 0$, which, in passing from I_3 through I_4 to I_1, the contour has circled. Even though z itself takes on the same values along I_3 as along I_1, the multiple-valued integrand $z^a/(z+1)$ does not. Indeed,

$$\begin{aligned} I_1 &= \int_0^\infty \frac{(\rho e^{i0})^a}{(\rho e^{i0}) + 1}\, d\rho &&= \int_0^\infty \frac{\rho^a}{\rho + 1}\, d\rho &&= S, \\ -I_3 &= \int_0^\infty \frac{(\rho e^{i2\pi})^a}{(\rho e^{i2\pi}) + 1}\, d\rho &&= e^{i2\pi a}\int_0^\infty \frac{\rho^a}{\rho + 1}\, d\rho &&= e^{i2\pi a} S. \end{aligned}$$

Therefore,

$$\begin{aligned}
I &= I_1 + I_2 + I_3 + I_4 \\
&= (I_1 + I_3) + I_2 + I_4 \\
&= (1 - e^{i2\pi a})S + \lim_{\rho\to\infty} \int_{\phi=0}^{2\pi} \frac{z^a}{z+1}\,dz - \lim_{\rho\to 0} \int_{\phi=0}^{2\pi} \frac{z^a}{z+1}\,dz \\
&= (1 - e^{i2\pi a})S + \lim_{\rho\to\infty} \int_{\phi=0}^{2\pi} z^{a-1}\,dz - \lim_{\rho\to 0} \int_{\phi=0}^{2\pi} z^a\,dz \\
&= (1 - e^{i2\pi a})S + \lim_{\rho\to\infty} \left.\frac{z^a}{a}\right|_{\phi=0}^{2\pi} - \lim_{\rho\to 0} \left.\frac{z^{a+1}}{a+1}\right|_{\phi=0}^{2\pi}.
\end{aligned}$$

Since $a < 0$, the first limit vanishes; and because $a > -1$, the second limit vanishes, too, leaving

$$I = (1 - e^{i2\pi a})S.$$

But by Cauchy's integral formula we have already found an expression for I. Substituting this expression into the last equation yields, by successive steps,

$$\begin{aligned}
i2\pi e^{i2\pi a/2} &= (1 - e^{i2\pi a})S, \\
S &= \frac{i2\pi e^{i2\pi a/2}}{1 - e^{i2\pi a}}, \\
S &= \frac{i2\pi}{e^{-i2\pi a/2} - e^{i2\pi a/2}}, \\
S &= -\frac{2\pi/2}{\sin(2\pi a/2)}.
\end{aligned}$$

That is,

$$\int_0^\infty \frac{\tau^a}{\tau+1}\,d\tau = -\frac{2\pi/2}{\sin(2\pi a/2)}, \quad -1 < a < 0, \tag{9.18}$$

an astonishing result.[5] Section 21.6 will use it.

Another example[6] is

$$T \equiv \int_0^{2\pi} \frac{d\theta}{1 + a\cos\theta}, \quad \Im(a) = 0, \ |\Re(a)| < 1.$$

[5] So astonishing is the result, that one is unlikely to believe it at first encounter. However, straightforward (though computationally highly inefficient) numerical integration per (7.1) confirms the result, as the interested reader and his computer can check. Such results vindicate the effort we have spent in deriving Cauchy's integral formula (8.29).

[6] [101]

As in the previous example, here again the contour is not closed. The previous example closed the contour by extending it, excluding the branch point. In this example there is no branch point to exclude, nor need one extend the contour. Rather, one changes the variable

$$z \leftarrow e^{i\theta}$$

and takes advantage of the fact that z, unlike θ, *begins and ends the integration at the same point.* One thus obtains the equivalent integral

$$\begin{aligned} T &= \oint \frac{dz/iz}{1+(a/2)(z+1/z)} = -\frac{i2}{a}\oint \frac{dz}{z^2+2z/a+1} \\ &= -\frac{i2}{a}\oint \frac{dz}{\left[z-\left(-1+\sqrt{1-a^2}\right)/a\right]\left[z-\left(-1-\sqrt{1-a^2}\right)/a\right]}, \end{aligned}$$

whose contour is the unit circle in the Argand plane. The integrand evidently has poles at

$$z = \frac{-1 \pm \sqrt{1-a^2}}{a},$$

whose magnitudes are such that

$$|z|^2 = \frac{2-a^2 \mp 2\sqrt{1-a^2}}{a^2}.$$

One of the two magnitudes is less than unity and one is greater, meaning that one of the two poles lies within the contour and one lies without, as is seen by the successive steps[7]

$$\begin{array}{rclcl}
a^2 & < & 1, && \\
0 & < & 1-a^2, && \\
(-a^2)(0) & > & (-a^2)(1-a^2), && \\
0 & > & -a^2+a^4, && \\
1-a^2 & > & 1-2a^2+a^4, && \\
1-a^2 & > & \left(1-a^2\right)^2, && \\
\sqrt{1-a^2} & > & 1-a^2, && \\
-\sqrt{1-a^2} & < & -(1-a^2) & < & \sqrt{1-a^2}, \\
1-\sqrt{1-a^2} & < & a^2 & < & 1+\sqrt{1-a^2}, \\
2-2\sqrt{1-a^2} & < & 2a^2 & < & 2+2\sqrt{1-a^2}, \\
2-a^2-2\sqrt{1-a^2} & < & a^2 & < & 2-a^2+2\sqrt{1-a^2}, \\
\dfrac{2-a^2-2\sqrt{1-a^2}}{a^2} & < & 1 & < & \dfrac{2-a^2+2\sqrt{1-a^2}}{a^2}.
\end{array}$$

Per Cauchy's integral formula (8.29), integrating about the pole within the contour yields that

$$T = i2\pi \left. \frac{-i2/a}{z-\left(-1-\sqrt{1-a^2}\right)/a} \right|_{z=\left(-1+\sqrt{1-a^2}\right)/a} = \frac{2\pi}{\sqrt{1-a^2}}.$$

Observe that by means of a complex variable of integration, each example has indirectly evaluated an integral whose integrand is purely real. If it seems unreasonable to the reader to expect so flamboyant a technique actually to work, this seems equally unreasonable to the writer—but work it does, nevertheless. It is a great technique.

The technique, *integration by closed contour,* is found in practice to solve many integrals other techniques find almost impossible to crack. The key to making the technique work lies in closing a contour one knows how to treat. The robustness of the technique lies in that any contour of any shape will work, so long as the contour encloses appropriate poles in the Argand domain plane while shutting branch points out.

[7] These steps are perhaps best read from bottom to top. See chapter 6's footnote 34.

The extension

$$\left|\int_{z_1}^{z_2} f(z)\,dz\right| \le \int_{z_1}^{z_2} |f(z)\,dz| \tag{9.19}$$

of the complex triangle sum inequality (3.22) from the discrete to the continuous case sometimes proves useful in evaluating integrals by this section's technique, as in § 17.6.4.

9.7 Integration by partial-fraction expansion

This section treats integration by partial-fraction expansion. It introduces the expansion itself first.[8] Throughout the section,

$$j, j', k, \ell, m, n, p, p_{(\cdot)}, M, N \in \mathbb{Z}.$$

9.7.1 Partial-fraction expansion

Consider the function

$$f(z) = \frac{-4}{z-1} + \frac{5}{z-2}.$$

Combining the two fractions over a common denominator[9] yields that

$$f(z) = \frac{z+3}{(z-1)(z-2)}.$$

Of the two forms, the former is probably the more amenable to analysis. For example, using (9.3),

$$\begin{aligned}\int_{-1}^{0} f(\tau)\,d\tau &= \int_{-1}^{0} \frac{-4}{\tau-1}\,d\tau + \int_{-1}^{0} \frac{5}{\tau-2}\,d\tau \\ &= \left[-4\ln(1-\tau) + 5\ln(2-\tau)\right]_{-1}^{0}.\end{aligned}$$

The trouble is that one is not always given the function in the amenable form.

Given a *rational function*

$$f(z) = \frac{\sum_{k=0}^{N-1} b_k z^k}{\prod_{j=1}^{N} (z-\alpha_j)} \tag{9.20}$$

[8] [129, appendix F][79, §§ 2.7 and 10.12]

[9] Terminology (you probably knew this already): a *fraction* is the ratio of two numbers or expressions B/A. In the fraction, B is the *numerator* and A is the *denominator.* The *quotient* is $Q = B/A$.

in which no two of the several poles α_j are the same, the *partial-fraction expansion* has the form

$$f(z) = \sum_{k=1}^{N} \frac{A_k}{z-\alpha_k}, \tag{9.21}$$

where multiplying each fraction of (9.21) by

$$\frac{\left[\prod_{j=1}^{N}(z-\alpha_j)\right]/(z-\alpha_k)}{\left[\prod_{j=1}^{N}(z-\alpha_j)\right]/(z-\alpha_k)}$$

puts the several fractions over a common denominator, yielding (9.20). Dividing (9.20) by (9.21) gives the ratio

$$1 = \frac{\sum_{k=0}^{N-1} b_k z^k}{\prod_{j=1}^{N}(z-\alpha_j)} \Bigg/ \sum_{k=1}^{N} \frac{A_k}{z-\alpha_k}.$$

In the immediate neighborhood of $z = \alpha_m$, the mth term $A_m/(z-\alpha_m)$ dominates the summation of (9.21). Hence,

$$1 = \lim_{z\to\alpha_m} \frac{\sum_{k=0}^{N-1} b_k z^k}{\prod_{j=1}^{N}(z-\alpha_j)} \Bigg/ \frac{A_m}{z-\alpha_m}.$$

Rearranging factors, we have that

$$A_m = \left.\frac{\sum_{k=0}^{N-1} b_k z^k}{\left[\prod_{j=1}^{N}(z-\alpha_j)\right]/(z-\alpha_m)}\right|_{z=\alpha_m} = \lim_{z\to\alpha_m}\left[(z-\alpha_m)f(z)\right], \tag{9.22}$$

where A_m, the value of $f(z)$ with the pole canceled, is called the *residue* of $f(z)$ at the pole $z = \alpha_m$. Equations (9.21) and (9.22) together give the partial-fraction expansion of (9.20)'s rational function $f(z)$.

9.7.2 Repeated poles

The weakness of the partial-fraction expansion of § 9.7.1 is that it cannot directly handle repeated poles. That is, if $\alpha_n = \alpha_j$, $n \neq j$, then the residue formula (9.22) finds an uncanceled pole remaining in its denominator and thus fails for $A_n = A_j$ (it still works for the other A_m). The conventional way to expand a fraction with repeated poles is presented in § 9.7.6 below; but because at least to this writer that way does not lend much applied insight, the present subsection treats the matter in a different way. Here, we *separate the poles.*

Consider the function

$$g(z) \equiv \sum_{k=0}^{N-1} \frac{Ce^{i2\pi k/N}}{z - \epsilon e^{i2\pi k/N}}, \quad N > 1, \ 0 < \epsilon \ll 1, \tag{9.23}$$

where C is a real-valued constant. This function evidently has a small circle of poles in the Argand plane at $\alpha_k = \epsilon e^{i2\pi k/N}$. Factoring,

$$g(z) = \frac{C}{z} \sum_{k=0}^{N-1} \frac{e^{i2\pi k/N}}{1 - (\epsilon e^{i2\pi k/N})/z}.$$

Using (2.36) to expand the fraction,

$$\begin{aligned}
g(z) &= \frac{C}{z} \sum_{k=0}^{N-1} \left[e^{i2\pi k/N} \sum_{j=0}^{\infty} \left(\frac{\epsilon e^{i2\pi k/N}}{z} \right)^j \right] \\
&= C \sum_{k=0}^{N-1} \sum_{j=1}^{\infty} \frac{\epsilon^{j-1} e^{i2\pi jk/N}}{z^j} \\
&= C \sum_{j=1}^{\infty} \frac{\epsilon^{j-1}}{z^j} \sum_{k=0}^{N-1} \left(e^{i2\pi j/N} \right)^k.
\end{aligned}$$

But[10]

$$\sum_{k=0}^{N-1} \left(e^{i2\pi j/N} \right)^k = \begin{cases} N & \text{if } j = mN, \\ 0 & \text{otherwise,} \end{cases}$$

so

$$g(z) = NC \sum_{m=1}^{\infty} \frac{\epsilon^{mN-1}}{z^{mN}}.$$

For $|z| \gg \epsilon$—that is, except in the immediate neighborhood of the small circle of poles—the first term of the summation dominates. Hence,

$$g(z) \approx NC \frac{\epsilon^{N-1}}{z^N}, \quad |z| \gg \epsilon.$$

[10] If you don't see why, then for $N = 8$ and $j = 3$ plot the several $(e^{i2\pi j/N})^k$ in the Argand plane. Do the same for $j = 2$ then $j = 8$. Only in the $j = 8$ case do the terms add coherently; in the other cases they cancel.

This effect—reinforcing when $j = nN$, canceling otherwise—is a classic manifestation of *Parseval's principle,* which § 17.1 will formally introduce later in the book.

Having achieved this approximation, if we strategically choose

$$C = \frac{1}{N\epsilon^{N-1}},$$

then

$$g(z) \approx \frac{1}{z^N}, \quad |z| \gg \epsilon.$$

But given the chosen value of C, (9.23) is

$$g(z) = \frac{1}{N\epsilon^{N-1}} \sum_{k=0}^{N-1} \frac{e^{i2\pi k/N}}{z - \epsilon e^{i2\pi k/N}}, \quad N > 1, \ 0 < \epsilon \ll 1.$$

Joining the last two equations together, changing $z - z_o \leftarrow z$, and writing more formally, we have that

$$\frac{1}{(z - z_o)^N} = \lim_{\epsilon \to 0} \frac{1}{N\epsilon^{N-1}} \sum_{k=0}^{N-1} \frac{e^{i2\pi k/N}}{z - \left[z_o + \epsilon e^{i2\pi k/N}\right]}, \quad N > 1. \tag{9.24}$$

The significance of (9.24) is that it lets one replace an N-fold pole with a small circle of ordinary poles, which per § 9.7.1 we already know how to handle. Notice incidentally that $1/N\epsilon^{N-1}$ is a large number not a small. The poles are close together but very strong.

9.7.3 An example

An example to illustrate the technique, separating a double pole:

$$
\begin{aligned}
f(z) &= \frac{z^2 - z + 6}{(z-1)^2(z+2)} \\
&= \lim_{\epsilon\to 0} \frac{z^2 - z + 6}{(z - [1 + \epsilon e^{i2\pi(0)/2}])(z - [1 + \epsilon e^{i2\pi(1)/2}])(z+2)} \\
&= \lim_{\epsilon\to 0} \frac{z^2 - z + 6}{(z - [1+\epsilon])(z - [1-\epsilon])(z+2)} \\
&= \lim_{\epsilon\to 0} \left\{ \left(\frac{1}{z - [1+\epsilon]} \right) \left[\frac{z^2 - z + 6}{(z - [1-\epsilon])(z+2)} \right]_{z=1+\epsilon} \right. \\
&\qquad + \left(\frac{1}{z - [1-\epsilon]} \right) \left[\frac{z^2 - z + 6}{(z - [1+\epsilon])(z+2)} \right]_{z=1-\epsilon} \\
&\qquad \left. + \left(\frac{1}{z+2} \right) \left[\frac{z^2 - z + 6}{(z - [1+\epsilon])(z - [1-\epsilon])} \right]_{z=-2} \right\} \\
&= \lim_{\epsilon\to 0} \left\{ \left(\frac{1}{z - [1+\epsilon]} \right) \left[\frac{6 + \epsilon + \epsilon^2}{6\epsilon + 2\epsilon^2} \right] \right. \\
&\qquad + \left(\frac{1}{z - [1-\epsilon]} \right) \left[\frac{6 - \epsilon + \epsilon^2}{-6\epsilon + 2\epsilon^2} \right] \\
&\qquad \left. + \left(\frac{1}{z+2} \right) \left[\frac{\mathrm{0xC}}{9 - \epsilon^2} \right] \right\}
\end{aligned}
$$

As usual when handling infinitesimals like ϵ, we can drop from each sum its insignificant terms in the limit, but which terms are insignificant depends on the ratio in which each sum is:

$$
\begin{aligned}
\lim_{\epsilon\to 0} \frac{6 \pm \epsilon + \epsilon^2}{\pm 6\epsilon + 2\epsilon^2} &= \lim_{\epsilon\to 0} \left\{ \frac{6}{\pm 6\epsilon + 2\epsilon^2} + \frac{\pm\epsilon}{\pm 6\epsilon} \right\} = \lim_{\epsilon\to 0} \left\{ \frac{\pm 1/\epsilon}{1 \pm \epsilon/3} + \frac{1/6}{1} \right\} \\
&= \lim_{\epsilon\to 0} \left\{ \left(\pm \frac{1}{\epsilon} \right) \left(1 \mp \frac{\epsilon}{3} + \cdots \right) + \frac{1}{6} \right\} \\
&= \pm \frac{1}{\epsilon} - \frac{1}{3} + \frac{1}{6} = \pm \frac{1}{\epsilon} - \frac{1}{6}.
\end{aligned}
$$

Thus,

$$
f(z) = \lim_{\epsilon\to 0} \left\{ \frac{1/\epsilon - 1/6}{z - [1+\epsilon]} + \frac{-1/\epsilon - 1/6}{z - [1-\epsilon]} + \frac{4/3}{z+2} \right\}
$$

gives the partial-fraction expansion of $f(z)$.

Incidentally, one can recombine terms to reach the alternate form

$$f(z) = \lim_{\epsilon\to 0}\left\{\frac{1/\epsilon}{z-[1+\epsilon]} + \frac{-1/\epsilon}{z-[1-\epsilon]} + \frac{-1/3}{z-1} + \frac{4/3}{z+2}\right\}.$$

This alternate form is valid and, for many purposes (as in § 9.7.4), is even handy; but one should interpret it cautiously: counting terms in the alternate form, you would think that $f(z)$ had four poles whereas it has in fact only three. The pole of the ratio $(-1/3)/(z-1)$—infinitely weaker than, yet lying infinitely near to, the poles of the two ratios $(1/\epsilon)(z-[1\pm\epsilon])$—is indeed a pole of the ratio $(-1/3)/(z-1)$, but it is not a proper, distinct pole of $f(z)$. It can be understood as a sort of shadow pole, if you like, that lurks near or hides under the twin dominant poles that loom over it. To see the proper, distinct poles of $f(z)$, refer rather to the earlier form.

However one might choose to account for and to describe the shadow pole, one cannot merely omit it. If one did omit it, then recombining the remaining partial fractions over a common denominator (try it!) would fail to recover our original expression for $f(z)$.

9.7.4 Integrating a rational function

If one can find the poles of a rational function of the form (9.20), then one can use (9.21) and (9.22)—and, if needed, (9.24)—to expand the function into a sum of partial fractions, each of which one can integrate individually.

Continuing the example of § 9.7.3, for $0 \le x < 1$,

$$\begin{aligned}
\int_0^x f(\tau)\,d\tau &= \int_0^x \frac{\tau^2-\tau+6}{(\tau-1)^2(\tau+2)}\,d\tau \\
&= \lim_{\epsilon\to 0}\int_0^x \left\{ \frac{1/\epsilon}{\tau-[1+\epsilon]} + \frac{-1/\epsilon}{\tau-[1-\epsilon]} + \frac{-1/3}{\tau-1} + \frac{4/3}{\tau+2} \right\} d\tau \\
&= \lim_{\epsilon\to 0}\left\{ \frac{1}{\epsilon}\ln([1+\epsilon]-\tau) - \frac{1}{\epsilon}\ln([1-\epsilon]-\tau) \right. \\
&\qquad \left. - \frac{1}{3}\ln(1-\tau) + \frac{4}{3}\ln(\tau+2) \right\}_0^x \\
&= \lim_{\epsilon\to 0}\left\{ \frac{1}{\epsilon}\ln\left(\frac{[1+\epsilon]-\tau}{[1-\epsilon]-\tau}\right) - \frac{1}{3}\ln(1-\tau) + \frac{4}{3}\ln(\tau+2) \right\}_0^x \\
&= \lim_{\epsilon\to 0}\left\{ \frac{1}{\epsilon}\ln\left(\frac{[1-\tau]+\epsilon}{[1-\tau]-\epsilon}\right) - \frac{1}{3}\ln(1-\tau) + \frac{4}{3}\ln(\tau+2) \right\}_0^x \\
&= \lim_{\epsilon\to 0}\left\{ \frac{1}{\epsilon}\ln\left(1+\frac{2\epsilon}{1-\tau}\right) - \frac{1}{3}\ln(1-\tau) + \frac{4}{3}\ln(\tau+2) \right\}_0^x \\
&= \lim_{\epsilon\to 0}\left\{ \frac{1}{\epsilon}\left(\frac{2\epsilon}{1-\tau}\right) - \frac{1}{3}\ln(1-\tau) + \frac{4}{3}\ln(\tau+2) \right\}_0^x \\
&= \lim_{\epsilon\to 0}\left\{ \frac{2}{1-\tau} - \frac{1}{3}\ln(1-\tau) + \frac{4}{3}\ln(\tau+2) \right\}_0^x \\
&= \frac{2}{1-x} - 2 - \frac{1}{3}\ln(1-x) + \frac{4}{3}\ln\left(\frac{x+2}{2}\right).
\end{aligned}$$

To check per § 7.5 that the result is correct, we can take the derivative of the final expression:

$$\begin{aligned}
\left[\frac{d}{dx}\left\{\frac{2}{1-x} - 2 - \frac{1}{3}\ln(1-x) + \frac{4}{3}\ln\left(\frac{x+2}{2}\right)\right\}\right]_{x=\tau}& \\
&= \frac{2}{(\tau-1)^2} + \frac{-1/3}{\tau-1} + \frac{4/3}{\tau+2} \\
&= \frac{\tau^2-\tau+6}{(\tau-1)^2(\tau+2)},
\end{aligned}$$

which indeed has the form of the integrand with which we started, confirming the result. (Notice incidentally how much easier it is symbolically to differentiate than to integrate!)

Section 19.4 exercises the technique in a more sophisticated way, applying it in the context of chapter 18's Laplace transform to solve a linear differential equation.

9.7.5 The derivatives of a rational function

Not only the integral of a rational function interests us; its derivatives interest us, too. One needs no special technique to compute such derivatives, of course, but the derivatives do bring some noteworthy properties.

First of interest is the property that a function in the general rational form

$$\Phi(w) = \frac{w^p h_0(w)}{g(w)}, \quad g(0) \neq 0, \tag{9.25}$$

enjoys derivatives in the general rational form

$$\frac{d^k\Phi}{dw^k} = \frac{w^{p-k} h_k(w)}{[g(w)]^{k+1}}, \quad 0 \le k \le p, \tag{9.26}$$

where g and h_k are polynomials in nonnegative powers of w. The property is proved by induction. When $k = 0$, (9.26) is (9.25), so (9.26) is good at least for this case. Then, using the product rule (4.24), if (9.26) holds for $k = n - 1$.

$$\begin{aligned}
\frac{d^n\Phi}{dw^n} &= \frac{d}{dw}\left[\frac{d^{n-1}\Phi}{dw^{n-1}}\right] = \frac{d}{dw}\left[\frac{w^{p-n+1}h_{n-1}(w)}{[g(w)]^n}\right] = \frac{w^{p-n}h_n(w)}{[g(w)]^{n+1}}, \\
h_n(w) &\equiv wg\frac{dh_{n-1}}{dw} - nwh_{n-1}\frac{dg}{dw} + (p-n+1)gh_{n-1}, \quad 0 < n \le p,
\end{aligned}$$

which makes h_n (like h_{n-1}) a polynomial in nonnegative powers of w. By induction on this basis, (9.26) holds for all $0 \le k \le p$, as was to be demonstrated.

A related property is that

$$\left.\frac{d^k\Phi}{dw^k}\right|_{w=0} = 0 \qquad \text{for } 0 \le k < p. \tag{9.27}$$

That is, the function and its first $p-1$ derivatives are all zero at $w = 0$. The reason is that (9.26)'s denominator is $[g(w)]^{k+1} \neq 0$, whereas its numerator has a $w^{p-k} = 0$ factor, when $0 \le k < p$ and $w = 0$.

9.7.6 Repeated poles (the conventional technique)

Though the technique of §§ 9.7.2 and 9.7.4 affords extra insight, it is not the conventional technique to expand in partial fractions a rational function having a repeated pole. The conventional technique is worth learning not only because it is conventional but also because it is usually quicker to apply in practice. This subsection derives it.

A rational function with repeated poles,

$$\begin{aligned} f(z) &= \frac{\sum_{k=0}^{N-1} b_k z^k}{\prod_{j=1}^{M} (z-\alpha_j)^{p_j}}, \\ N &\equiv \sum_{j=1}^{M} p_j, \\ p_j &\geq 0, \\ \alpha_{j'} &\neq \alpha_j \text{ for all } j' \neq j, \end{aligned} \tag{9.28}$$

where j, k, M, N and the several p_j are integers, cannot be expanded solely in the first-order fractions of § 9.7.1, but can indeed be expanded if higher-order fractions are allowed:

$$f(z) = \sum_{j=1}^{M} \sum_{\ell=0}^{p_j-1} \frac{A_{j\ell}}{(z-\alpha_j)^{p_j-\ell}}. \tag{9.29}$$

What the partial-fraction expansion (9.29) lacks are the values of its several coefficients $A_{j\ell}$.

One can determine the coefficients with respect to one (possibly repeated) pole at a time. To determine them with respect to the p_m-fold pole at $z = \alpha_m$, $1 \leq m \leq M$, one multiplies (9.29) by $(z-\alpha_m)^{p_m}$ to obtain the form

$$(z-\alpha_m)^{p_m} f(z) = \sum_{\substack{j=1, \\ j\neq m}}^{M} \sum_{\ell=0}^{p_j-1} \frac{(A_{j\ell})(z-\alpha_m)^{p_m}}{(z-\alpha_j)^{p_j-\ell}} + \sum_{\ell=0}^{p_m-1} (A_{m\ell})(z-\alpha_m)^{\ell}.$$

But (9.27) with $w = z - \alpha_m$ reveals the double summation and its first $p_m - 1$ derivatives all to be null at $z = \alpha_m$; that is,

$$\left. \frac{d^k}{dz^k} \sum_{\substack{j=1, \\ j\neq m}}^{M} \sum_{\ell=0}^{p_j-1} \frac{(A_{j\ell})(z-\alpha_m)^{p_m}}{(z-\alpha_j)^{p_j-\ell}} \right|_{z=\alpha_m} = 0, \qquad 0 \leq k < p_m;$$

so, the $(z-\alpha_m)^{p_m} f(z)$ equation's kth derivative reduces at that point to

$$\begin{aligned} \frac{d^k}{dz^k} \left[(z-\alpha_m)^{p_m} f(z) \right] \bigg|_{z=\alpha_m} &= \sum_{\ell=0}^{p_m-1} \frac{d^k}{dz^k} \left[(A_{m\ell})(z-\alpha_m)^{\ell} \right] \bigg|_{z=\alpha_m} \\ &= k! A_{mk}, \quad 0 \leq k < p_m. \end{aligned}$$

Changing $j \leftarrow m$ and $\ell \leftarrow k$ and solving for $A_{j\ell}$ then produces the coefficients

$$A_{j\ell} = \left(\frac{1}{\ell!}\right)\frac{d^\ell}{dz^\ell}\left[(z-\alpha_j)^{p_j}f(z)\right]\Bigg|_{z=\alpha_j}, \quad 0 \le \ell < p_j, \tag{9.30}$$

to weight the expansion (9.29)'s partial fractions. In case of a repeated pole, these coefficients evidently depend not only on the residual function itself but also on its several derivatives, one derivative per repetition of the pole.

9.7.7 The existence and uniqueness of solutions

Equation (9.30) has solved (9.28) and (9.29). A professional mathematician might object however that it has done so without first proving that a unique solution actually exists.

Comes from us the reply, "Why should we prove that a solution exists, once we have actually found it?"

Ah, but the hypothetical professional's point is that we have found the solution only if in fact it does exist, and uniquely; otherwise what we have *found* is a phantom. A careful review of § 9.7.6's logic discovers no guarantee that all of (9.30)'s coefficients actually come from the same expansion. Maybe there exist two distinct expansions, and some of the coefficients come from the one, some from the other. On the other hand, maybe there exists no expansion at all, in which event it is not even clear what (9.30) means.

"But these are quibbles, cavils and nitpicks!" we are inclined to grumble. "The present book is a book of applied mathematics."

Well, yes, but on this occasion let us nonetheless follow the professional's line of reasoning, if only a short way.

Uniqueness is proved by positing two solutions

$$f(z) = \sum_{j=1}^{M}\sum_{\ell=0}^{p_j-1}\frac{A_{j\ell}}{(z-\alpha_j)^{p_j-\ell}} = \sum_{j=1}^{M}\sum_{\ell=0}^{p_j-1}\frac{B_{j\ell}}{(z-\alpha_j)^{p_j-\ell}}$$

and computing the difference

$$\sum_{j=1}^{M}\sum_{\ell=0}^{p_j-1}\frac{B_{j\ell}-A_{j\ell}}{(z-\alpha_j)^{p_j-\ell}}$$

between them. Logically this difference must be zero for all z if the two solutions are actually to represent the same function $f(z)$. This however is seen to be possible only if $B_{j\ell} = A_{j\ell}$ for each (j, ℓ). Therefore, the two solutions are one and the same. (The professional might request a further demonstration of orthogonality, § 13.8; but we will leave the point in that form.)

Existence comes of combining the several fractions of (9.29) over a common denominator and comparing the resulting numerator against the numerator of (9.28). Each coefficient b_k is seen thereby to be a linear combination of the several $A_{j\ell}$, where the combination's weights depend solely on the locations α_j and multiplicities p_j of $f(z)$'s several poles. From the N coefficients b_k and the N coefficients $A_{j\ell}$, an $N \times N$ system of N linear equations in N unknowns results—which might for example (if, say, $N = 3$) look like

$$\begin{aligned} b_0 &= -2A_{00} + A_{01} + 3A_{10}, \\ b_1 &= A_{00} + A_{01} + A_{10}, \\ b_2 &= 2A_{01} - 5A_{10}. \end{aligned}$$

We will show in chapters 11 through 14 that when such a system has no solution, there always exists an alternate set of b_k for which the same system has multiple solutions. But uniqueness, which we have already established, forbids such multiple solutions in all cases. Therefore it is not possible for the system to have no solution—which is to say, the solution necessarily exists.

We will not often in this book prove existence and uniqueness explicitly, but such proofs when desired tend to fit the pattern outlined here.

9.8 Integration by the manipulation of a Pythagorean expression

In context of the chapter you are reading, a *Pythagorean expression* is an expression of the form of $\pm 1 \pm \tau^2$. This section suggests ways to approach integrands that contain Pythagorean expressions.

9.8.1 Pythagorean radicals

In applications, as for instance in the path-length computation of § 7.4.2, one often meets integrands that contain Pythagorean expressions under a radical sign, like $\sqrt{1-\tau^2}$, $\sqrt{\tau^2-1}$ or $\sqrt{\tau^2+1}$. An example would be

$$S_1(x) \equiv \int_0^x \frac{d\tau}{\sqrt{1-\tau^2}},$$

which contains the *Pythagorean radical* $\sqrt{1-\tau^2}$. Such a Pythagorean radical recalls the inverse trigonometrics of Table 5.3, whereby

$$S_1(x) = \arcsin \tau|_0^x = \arcsin x.$$

Unfortunately, not every integrand that features such a radical appears in the table; so, for example,

$$S_2(x) \equiv \int_0^x d\tau\, \tau\sqrt{1-\tau^2}$$

wants a substitution like $u_2^2 \leftarrow 1-\tau^2$, $u_2\,du_2 = -\tau\,d\tau$, by which the technique of § 9.2 finds that

$$S_2(x) = -\int_1^{\sqrt{1-x^2}} du_2\, u_2^2 = -\left.\frac{u_2^3}{3}\right|_1^{\sqrt{1-x^2}} = \frac{1-(1-x^2)^{3/2}}{3}.$$

That's all relatively straightforward, but now try an integral like

$$S_3(x) \equiv \int_0^x d\tau\, \sqrt{1-\tau^2}.$$

This doesn't *look* harder. Indeed, if anything, it looks slightly easier than $S_1(x)$ or $S_2(x)$. Notwithstanding, the techniques used to solve those somehow don't quite seem to work on $S_3(x)$ [try it!].

As it happens, we have already met, and solved, a similar integral in § 7.4.2. That subsection has illustrated the technique. Applying the same technique here, we assemble by trial a small table of potentially relevant antiderivatives,

$$\begin{aligned}
\frac{d}{d\tau}\arcsin\tau &= \frac{1}{\sqrt{1-\tau^2}},\\
\frac{d}{d\tau}\sqrt{1-\tau^2} &= -\frac{\tau}{\sqrt{1-\tau^2}},\\
\frac{d}{d\tau}\tau\sqrt{1-\tau^2} &= \sqrt{1-\tau^2} - \frac{\tau^2}{\sqrt{1-\tau^2}} = 2\sqrt{1-\tau^2} - \frac{1}{\sqrt{1-\tau^2}}.
\end{aligned}$$

wherein the pivotal step on the last line is to have *manipulated the Pythagorean radical,* observing that

$$\frac{\tau^2}{\sqrt{1-\tau^2}} = -\frac{1-\tau^2}{\sqrt{1-\tau^2}} + \frac{1}{\sqrt{1-\tau^2}} = -\sqrt{1-\tau^2} + \frac{1}{\sqrt{1-\tau^2}}.$$

Using some of the above-computed derivatives, the desired integrand $\sqrt{1-\tau^2}$ can now be built up by stages:

$$\begin{aligned}
\frac{d}{2\,d\tau}\tau\sqrt{1-\tau^2} &= \sqrt{1-\tau^2} - \frac{1}{2\sqrt{1-\tau^2}};\\
\frac{d}{2\,d\tau}\arcsin\tau + \frac{d}{2\,d\tau}\tau\sqrt{1-\tau^2} &= \sqrt{1-\tau^2}.
\end{aligned}$$

Hence,

$$S_3(x) = \left[\frac{1}{2}\arcsin\tau + \frac{1}{2}\tau\sqrt{1-\tau^2}\right]_0^x = \frac{1}{2}\arcsin x + \frac{1}{2}x\sqrt{1-x^2},$$

a result which can (and should) be checked by differentiating in the manner of § 7.5.

Here is another example of integration by the manipulation of a Pythagorean radical:

$$\begin{aligned}
S_4(x) &\equiv \int_1^x d\tau\,\tau^2\sqrt{\tau^2-1};\\
\frac{d}{d\tau}\operatorname{arccosh}\tau &= \frac{1}{\sqrt{\tau^2-1}};\\
\frac{d}{d\tau}\tau\sqrt{\tau^2-1} &= \sqrt{\tau^2-1} + \frac{\tau^2}{\sqrt{\tau^2-1}}\\
&= 2\sqrt{\tau^2-1} + \frac{1}{\sqrt{\tau^2-1}};\\
\frac{d}{d\tau}\tau^3\sqrt{\tau^2-1} &= 3\tau^2\sqrt{\tau^2-1} + \frac{\tau^4}{\sqrt{\tau^2-1}}\\
&= 4\tau^2\sqrt{\tau^2-1} + \frac{\tau^2}{\sqrt{\tau^2-1}}\\
&= (4\tau^2+1)\sqrt{\tau^2-1} + \frac{1}{\sqrt{\tau^2-1}};\\
\frac{d}{4\,d\tau}\tau^3\sqrt{\tau^2-1} &= \left(\tau^2+\frac{1}{4}\right)\sqrt{\tau^2-1} + \frac{1}{4\sqrt{\tau^2-1}}.
\end{aligned}$$

Having assembled the above small table of potentially relevant antiderivatives, we proceed:

$$\begin{aligned}
&-\frac{d}{8\,d\tau}\tau\sqrt{\tau^2-1}+\frac{d}{4\,d\tau}\tau^3\sqrt{\tau^2-1}\\
&\qquad\qquad = \tau^2\sqrt{\tau^2-1}+\frac{1}{8\sqrt{\tau^2-1}};\\
&-\frac{d}{8\,d\tau}\operatorname{arccosh}\tau-\frac{d}{8\,d\tau}\tau\sqrt{\tau^2-1}+\frac{d}{4\,d\tau}\tau^3\sqrt{\tau^2-1}\\
&\qquad\qquad = \tau^2\sqrt{\tau^2-1};\\
&\qquad S_4(x) = \bigg[-\frac{\operatorname{arccosh}\tau}{8}\\
&\qquad\qquad\qquad + \left(\frac{\tau^2}{4}-\frac{1}{8}\right)\tau\sqrt{\tau^2-1}\bigg]_1^x;\\
&\qquad\qquad = -\frac{\operatorname{arccosh}x}{8}+\left(\frac{x^2}{4}-\frac{1}{8}\right)x\sqrt{x^2-1}.
\end{aligned}$$

For yet more examples, consider

$$S_5(x)\equiv\int_1^x\frac{\tau^2\,d\tau}{\sqrt{\tau^2-1}}=\int_1^x d\tau\,\sqrt{\tau^2-1}+\int_1^x\frac{d\tau}{\sqrt{\tau^2-1}},$$

to complete whose evaluation is left as an exercise, and

$$S_6(x)\equiv\int_0^x d\tau\,\sqrt{a^2-\tau^2}=a^2\int_{(\tau/a)=0}^{x/a}d\left(\frac{\tau}{a}\right)\sqrt{1-\left(\frac{\tau}{a}\right)^2}=a^2S_3\left(\frac{x}{a}\right).$$

9.8.2 Pythagorean nonradicals

Besides Pythagorean radicals, Pythagorean nonradicals occur, too. However, these tend to be easier to solve. For example,

$$\begin{aligned}
S_7(x)&\equiv\int_0^x\frac{\tau^2\,d\tau}{1+\tau^2}=\int_0^x\left(1-\frac{1}{1+\tau^2}\right)d\tau=x-\arctan x;\\
S_8(x)&\equiv\int_0^x\frac{\tau^3\,d\tau}{1+\tau^2}=\int_0^x\left(\tau-\frac{\tau}{1+\tau^2}\right)d\tau=\frac{x^2}{2}-\int_0^x\frac{\tau\,d\tau}{1+\tau^2},\\
&\quad u_8^2\leftarrow 1+\tau^2,\ u_8\,du_8=\tau\,d\tau,\\
S_8(x)&=\frac{x^2}{2}-\int_1^{\sqrt{1+x^2}}\frac{du_8}{u_8}=\frac{x^2-\ln(1+x^2)}{2}.
\end{aligned}$$

9.9 Trial derivatives

Besides the technique of the Pythagorean radical, § 9.8.1 has also incidentally demonstrated another, different technique, vaguer but more broadly applicable. It has incidentally demonstrated the technique of *trial derivatives.*[11]

Review the $S_3(x)$ and $S_4(x)$ of § 9.8.1. To solve each has required us to develop a small table of potentially relevant antiderivatives. How did we develop each small table? Well, we began by copying a relevant inverse-trigonometric entry from Table 5.3; but then, to extend the table, we *tried* taking the derivatives of various functions that resembled the integrand or part of the integrand. Not all our trials gave useful antiderivatives but some did.

To decide which derivatives to try during a given integration depends on the mathematician's creativity and experience. However, a typical theme is to multiply the integrand (or part of the integrand) by τ, τ^2 or maybe τ^3, taking the derivative of the product as § 9.8.1 has done. It is not usually necessary, nor helpful, to build up a huge table but—well, when you read § 9.8.1, you saw how it went.

The reason to take trial *derivatives,* incidentally, is that one does not generally know very well how to take trial *antiderivatives!* Analytically, derivatives are the easier to take. To seek an antiderivative by taking derivatives might (or might not) seem counterintuitive, but it's a game of feedback and educated guesses, like nineteenth-century artillery finding the range to its target. It is a game that can prosper, as we have seen.

9.10 Frullani's integral

One occasionally meets an integral of the form

$$S = \int_0^\infty \frac{f(b\tau) - f(a\tau)}{\tau}\, d\tau,$$

where a and b are real, positive coefficients and $f(\tau)$ is an arbitrary complex expression in τ. One wants to split such an integral in two as $\int [f(b\tau)/\tau]\, d\tau - \int [f(a\tau)/\tau]\, d\tau$, except that each half-integral alone may diverge. Nonetheless, splitting the integral in two is the right idea, provided that one first

[11] See [79, § 1.5], which introduces the technique in another guise under a different name.

relaxes the limits of integration as

$$S = \lim_{\epsilon\to 0^+} \left\{ \int_{\epsilon}^{1/\epsilon} \frac{f(b\tau)}{\tau}\,d\tau - \int_{\epsilon}^{1/\epsilon} \frac{f(a\tau)}{\tau}\,d\tau \right\}.$$

Changing $\sigma \leftarrow b\tau$ in the left integral and $\sigma \leftarrow a\tau$ in the right yields that

$$\begin{aligned} S &= \lim_{\epsilon\to 0^+} \left\{ \int_{b\epsilon}^{b/\epsilon} \frac{f(\sigma)}{\sigma}\,d\sigma - \int_{a\epsilon}^{a/\epsilon} \frac{f(\sigma)}{\sigma}\,d\sigma \right\} \\ &= \lim_{\epsilon\to 0^+} \left\{ \int_{a\epsilon}^{b\epsilon} \frac{-f(\sigma)}{\sigma}\,d\sigma + \int_{b\epsilon}^{a/\epsilon} \frac{f(\sigma)-f(\sigma)}{\sigma}\,d\sigma + \int_{a/\epsilon}^{b/\epsilon} \frac{f(\sigma)}{\sigma}\,d\sigma \right\} \\ &= \lim_{\epsilon\to 0^+} \left\{ \int_{a/\epsilon}^{b/\epsilon} \frac{f(\sigma)}{\sigma}\,d\sigma - \int_{a\epsilon}^{b\epsilon} \frac{f(\sigma)}{\sigma}\,d\sigma \right\} \end{aligned}$$

(here on the face of it, we have split the integration as though $a \le b$, but in fact it does not matter which of a and b is the greater, as is easy to verify). So long as each of $f(\epsilon)$ and $f(1/\epsilon)$ approaches a constant value as ϵ vanishes, this is

$$\begin{aligned} S &= \lim_{\epsilon\to 0^+} \left\{ f(+\infty) \int_{a/\epsilon}^{b/\epsilon} \frac{d\sigma}{\sigma} - f(0^+) \int_{a\epsilon}^{b\epsilon} \frac{d\sigma}{\sigma} \right\} \\ &= \lim_{\epsilon\to 0^+} \left\{ f(+\infty) \ln \frac{b/\epsilon}{a/\epsilon} - f(0^+) \ln \frac{b\epsilon}{a\epsilon} \right\} \\ &= [f(\tau)]_0^\infty \ln \frac{b}{a}. \end{aligned}$$

Thus we have *Frullani's integral,*

$$\int_0^\infty \frac{f(b\tau) - f(a\tau)}{\tau}\,d\tau = [f(\tau)]_0^\infty \ln \frac{b}{a}, \tag{9.31}$$

which, if a and b are both real and positive, works for any $f(\tau)$ which has definite $f(0^+)$ and $f(+\infty)$.[12]

9.11 Integrating products of exponentials, powers and logarithms

The products $\exp(\alpha\tau)\tau^n$ (where $n \in \mathbb{Z}$) and $\tau^{a-1} \ln \tau$ tend to arise[13] among other places in integrands related to special functions (as in the book's

[12] [107, § 1.3][3, § 2.5.1][178, "Frullani's integral"]

[13] One could write the latter product more generally as $\tau^{a-1} \ln \beta\tau$. According to Table 2.5, however, $\ln \beta\tau = \ln \beta + \ln \tau$; wherein $\ln \beta$ is just a constant.

part III). The two occur often enough to merit investigation here.

Concerning $\exp(\alpha\tau)\tau^n$, by § 9.5's method of unknown coefficients we guess its antiderivative to fit the form

$$\begin{aligned}\exp(\alpha\tau)\tau^n &= \frac{d}{d\tau}\sum_{k=0}^{n} a_k \exp(\alpha\tau)\tau^k \\ &= \sum_{k=0}^{n} \alpha a_k \exp(\alpha\tau)\tau^k + \sum_{k=1}^{n} k a_k \exp(\alpha\tau)\tau^{k-1} \\ &= \alpha a_n \exp(\alpha\tau)\tau^n + \exp(\alpha\tau)\sum_{k=0}^{n-1}\left[\alpha a_k + (k+1)a_{k+1}\right]\tau^k.\end{aligned}$$

If so, then evidently

$$\begin{aligned}a_n &= \frac{1}{\alpha}; \\ a_k &= -\frac{k+1}{\alpha}a_{k+1}, \quad 0 \le k < n.\end{aligned}$$

That is,

$$a_k = \frac{1}{\alpha}\prod_{j=k+1}^{n}(-j\alpha) = -\frac{n!/k!}{(-\alpha)^{n-k+1}}, \quad 0 \le k \le n.$$

Therefore,[14]

$$\exp(\alpha\tau)\tau^n = \frac{d}{d\tau}\left[-\exp(\alpha\tau)\sum_{k=0}^{n}\frac{n!/k!}{(-\alpha)^{n-k+1}}\tau^k\right], \quad n \in \mathbb{Z},\ n \ge 0,\ \alpha \ne 0. \tag{9.32}$$

The right form to guess for the antiderivative of $\tau^{a-1}\ln\tau$ is less obvious. Remembering however § 5.3's observation that $\ln\tau$ is of zeroth order in τ, after maybe some false tries we eventually do strike the right form

$$\begin{aligned}\tau^{a-1}\ln\tau &= \frac{d}{d\tau}\tau^a[B\ln\tau + C] \\ &= \tau^{a-1}[aB\ln\tau + (B + aC)],\end{aligned}$$

which demands that $B = 1/a$ and that $C = -1/a^2$. Therefore,[15]

$$\tau^{a-1}\ln\tau = \frac{d}{d\tau}\frac{\tau^a}{a}\left(\ln\tau - \frac{1}{a}\right), \quad a \ne 0. \tag{9.33}$$

[14] [154, eqn. 17.25.4][146, appendix 2, eqn. 73]

[15] [154, eqn. 17.26.3][146, appendix 2, eqn. 74]

Table 9.1: Antiderivatives of products of exponentials, powers and logarithms.

$$\begin{aligned}
\exp(\alpha\tau) &= \frac{d}{d\tau}\left[\exp(\alpha\tau)\left(\frac{1}{\alpha}\right)\right] \\
\exp(\alpha\tau)\tau &= \frac{d}{d\tau}\left[\exp(\alpha\tau)\left(\frac{\tau}{\alpha}-\frac{1}{\alpha^2}\right)\right] \\
\exp(\alpha\tau)\tau^2 &= \frac{d}{d\tau}\left[\exp(\alpha\tau)\left(\frac{\tau^2}{\alpha}-\frac{2\tau}{\alpha^2}+\frac{2}{\alpha^3}\right)\right] \\
\exp(\alpha\tau)\tau^n &= \frac{d}{d\tau}\left[-\exp(\alpha\tau)\sum_{k=0}^{n}\frac{n!/k!}{(-\alpha)^{n-k+1}}\tau^k\right] \\
&\qquad n\in\mathbb{Z},\ n\ge 0,\ \alpha\ne 0 \\
\tau^{a-1}\ln\tau &= \frac{d}{d\tau}\frac{\tau^a}{a}\left(\ln\tau-\frac{1}{a}\right),\quad a\ne 0 \\
\frac{\ln\tau}{\tau} &= \frac{d}{d\tau}\frac{(\ln\tau)^2}{2}
\end{aligned}$$

Antiderivatives of terms like $\tau^{a-1}(\ln\tau)^2$, $\exp(\alpha\tau)\tau^n\ln\tau$ and so on can be computed in like manner as the need arises.

Equation (9.33) fails when $a = 0$, but in this case with a little imagination the antiderivative is not hard to guess:[16]

$$\frac{\ln\tau}{\tau} = \frac{d}{d\tau}\frac{(\ln\tau)^2}{2}. \tag{9.34}$$

Table 9.1 summarizes.

9.12 Integration by Taylor series

With sufficient cleverness the techniques of the foregoing sections solve many, many integrals. But not all. When all else fails, as sometimes it does, the Taylor series of chapter 8 and the antiderivative of § 9.1 together offer a concise, practical way to integrate some functions, at the price of losing the

[16] [154, eqn. 17.26.4]

functions' known closed analytic forms. For example,

$$\begin{aligned}\int_0^x \exp\left(-\frac{\tau^2}{2}\right) d\tau &= \int_0^x \sum_{k=0}^{\infty} \frac{(-\tau^2/2)^k}{k!}\, d\tau \\ &= \int_0^x \sum_{k=0}^{\infty} \frac{(-)^k \tau^{2k}}{2^k k!}\, d\tau \\ &= \left[\sum_{k=0}^{\infty} \frac{(-)^k \tau^{2k+1}}{(2k+1)2^k k!}\right]_0^x \\ &= \sum_{k=0}^{\infty} \frac{(-)^k x^{2k+1}}{(2k+1)2^k k!} = (x)\sum_{k=0}^{\infty} \frac{1}{2k+1} \prod_{j=1}^{k} \frac{-x^2}{2j}.\end{aligned}$$

The result is no function one recognizes; it is just a series. This is not necessarily bad, however. After all, when a Taylor series from Table 8.1 is used to calculate $\sin z$, then $\sin z$ is just a series, too. The series above converges just as accurately and just as fast.

Sometimes it helps to give the series a name like[17]

$$\operatorname{myf} z \equiv \sum_{k=0}^{\infty} \frac{(-)^k z^{2k+1}}{(2k+1)2^k k!} = (z)\sum_{k=0}^{\infty} \frac{1}{2k+1} \prod_{j=1}^{k} \frac{-z^2}{2j}.$$

Then,

$$\int_0^x \exp\left(-\frac{\tau^2}{2}\right) d\tau = \operatorname{myf} x.$$

The $\operatorname{myf} z$ is no less a function than $\sin z$ is; it's just a function you hadn't heard of before. You can plot the function, or take its derivative

$$\frac{d}{d\tau} \operatorname{myf} \tau = \exp\left(-\frac{\tau^2}{2}\right),$$

or calculate its value, or do with it whatever else one does with functions. It works just the same.

Beyond the several integration techniques this chapter has introduced, a special-purpose technique of integration by cylindrical transformation will surface in § 18.4.

[17] The myf = "my function," but you can use any name for a function like this.

Chapter 10

Cubics and quartics

Under the heat of noonday, between the hard work of the morning and the heavy lifting of the afternoon, one likes to lay down one's burden and rest a spell in the shade. Chapters 2 through 9 have established the applied mathematical foundations upon which coming chapters will build; and chapter 11, hefting the weighty topic of the matrix, will indeed begin to build on those foundations. But in this short chapter which rests between, we shall refresh ourselves with an interesting but lighter mathematical topic: the topic of cubics and quartics.

The expression

$$z + a_0$$

is a *linear* polynomial, the lone root $z = -a_0$ of which is plain to see. The *quadratic* polynomial

$$z^2 + a_1 z + a_0$$

has of course two roots, which though not plain to see the quadratic formula (2.2) extracts with little effort. So much algebra has been known since antiquity. The roots of higher-order polynomials, the Newton-Raphson iteration (4.30) locates swiftly, but that is an approximate iteration rather than an exact formula like (2.2), and as we have seen in § 4.8 it can occasionally fail to converge. One would prefer an actual formula to extract the roots.

No general formula to extract the roots of the nth-order polynomial seems to be known.[1] However, to extract the roots of the *cubic* and *quartic* polynomials

$$\begin{aligned} z^3 + a_2 z^2 + a_1 z + a_0, \\ z^4 + a_3 z^3 + a_2 z^2 + a_1 z + a_0, \end{aligned}$$

[1] Refer to chapter 6's footnote 29.

though the ancients never discovered how, formulas do exist. The 16th-century algebraists Ferrari, Vieta, Tartaglia and Cardan have given us the clever technique. This chapter explains.[2]

10.1 Vieta's transform

There is a sense to numbers by which 1/2 resembles 2, 1/3 resembles 3, 1/4 resembles 4, and so forth. To capture this sense, one can transform a function $f(z)$ into a function $f(w)$ by the change of variable[3]

$$w + \frac{1}{w} \leftarrow z,$$

or, more generally,

$$w + \frac{w_o^2}{w} \leftarrow z. \tag{10.1}$$

Equation (10.1) is *Vieta's transform.*[4]

For $|w| \gg |w_o|$, we have that $z \approx w$; but as $|w|$ approaches $|w_o|$ this ceases to be true. For $|w| \ll |w_o|$, $z \approx w_o^2/w$. The constant w_o is the *corner value,* in the neighborhood of which w transitions from the one domain to the other. Figure 10.1 plots Vieta's transform for real w in the case that $w_o = 1$.

An interesting alternative to Vieta's transform is

$$w \parallel \frac{w_o^2}{w} \leftarrow z, \tag{10.2}$$

which in light of § 6.3 might be named *Vieta's parallel transform.*

Section 10.2 shows how Vieta's transform can be used.

10.2 Cubics

The general cubic polynomial is too hard to extract the roots of directly, so one begins by changing the variable

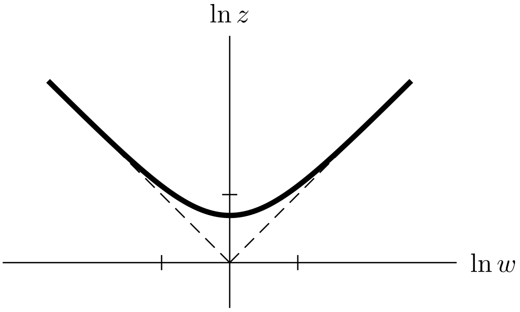
$$x + h \leftarrow z \tag{10.3}$$

[2] [178, "Cubic equation"][178, "Quartic equation"][182, "Quartic equation," 00:26, 9 Nov. 2006][182, "François Viète," 05:17, 1 Nov. 2006][182, "Gerolamo Cardano," 22:35, 31 Oct. 2006][156, § 1.5]

[3] This change of variable broadly recalls the sum-of-exponentials form (5.20) of the cosh(·) function, inasmuch as $\exp[-\phi] = 1/\exp\phi$.

[4] Also called "Vieta's substitution." [178, "Vieta's substitution"]

Figure 10.1: Vieta's transform (10.1) for $w_o = 1$, plotted logarithmically.



to obtain the polynomial

$$x^3 + (a_2 + 3h)x^2 + (a_1 + 2ha_2 + 3h^2)x + (a_0 + ha_1 + h^2a_2 + h^3).$$

The choice

$$h \equiv -\frac{a_2}{3} \tag{10.4}$$

casts the polynomial into the improved form

$$x^3 + \left[a_1 - \frac{a_2^2}{3}\right]x + \left[a_0 - \frac{a_1a_2}{3} + 2\left(\frac{a_2}{3}\right)^3\right],$$

or better yet

$$x^3 - px - q,$$

where

$$\begin{aligned} p &\equiv -a_1 + \frac{a_2^2}{3}, \\ q &\equiv -a_0 + \frac{a_1a_2}{3} - 2\left(\frac{a_2}{3}\right)^3. \end{aligned} \tag{10.5}$$

The solutions to the equation

$$x^3 = px + q, \tag{10.6}$$

then, are the cubic polynomial's three roots.

So we have struck the a_2z^2 term. That was the easy part; what to do next is not so obvious. If one could strike the px term as well, then the roots would follow immediately, but no very simple substitution like (10.3) achieves this—or rather, such a substitution does achieve it, but at the price of reintroducing an unwanted x^2 or z^2 term. That way is no good. Lacking guidance, one might try many, various substitutions, none of which seems to help much; but after weeks or months of such frustration one might eventually discover Vieta's transform (10.1), with the idea of balancing the equation between offsetting w and $1/w$ terms. This works.

Vieta-transforming (10.6) by the change of variable

$$w + \frac{w_o^2}{w} \leftarrow x \tag{10.7}$$

we get the new equation

$$w^3 + (3w_o^2 - p)w + (3w_o^2 - p)\frac{w_o^2}{w} + \frac{w_o^6}{w^3} = q, \tag{10.8}$$

which invites the choice

$$w_o^2 \equiv \frac{p}{3}, \tag{10.9}$$

reducing (10.8) to read

$$w^3 + \frac{(p/3)^3}{w^3} = q.$$

Multiplying by w^3 and rearranging terms, we have the quadratic equation

$$(w^3)^2 = 2\left(\frac{q}{2}\right)w^3 - \left(\frac{p}{3}\right)^3, \tag{10.10}$$

which by (2.2) we know how to solve.

Vieta's transform has reduced the original cubic to a quadratic.

The careful reader will observe that (10.10) seems to imply six roots, double the three the fundamental theorem of algebra (§ 6.2.2) allows a cubic polynomial to have. We shall return to this point in § 10.3. For the moment, however, we should like to improve the notation by defining[5]

$$\begin{aligned} P &\leftarrow -\frac{p}{3}, \\ Q &\leftarrow +\frac{q}{2}, \end{aligned} \tag{10.11}$$

[5] Why did we not define P and Q so to begin with? Well, before unveiling (10.10), we lacked motivation to do so. To define inscrutable coefficients unnecessarily before the need for them is apparent seems poor applied mathematical style.

Table 10.1: A method to extract the three roots of the general cubic polynomial. (In the definition of w^3, one can choose either sign for the ±.)

$$\begin{aligned}
0 &= z^3 + a_2 z^2 + a_1 z + a_0 \\
P &\equiv \frac{a_1}{3} - \left(\frac{a_2}{3}\right)^2 \\
Q &\equiv \frac{1}{2}\left[-a_0 + 3\left(\frac{a_1}{3}\right)\left(\frac{a_2}{3}\right) - 2\left(\frac{a_2}{3}\right)^3\right] \\
w^3 &\equiv \begin{cases} 2Q & \text{if } P = 0, \\ Q \pm \sqrt{Q^2 + P^3} & \text{otherwise.} \end{cases} \\
x &\equiv \begin{cases} 0 & \text{if } P = 0 \text{ and } Q = 0, \\ w - P/w & \text{otherwise.} \end{cases} \\
z &= x - \frac{a_2}{3}
\end{aligned}$$

with which (10.6) and (10.10) are written,

$$x^3 = 2Q - 3Px, \tag{10.12}$$

$$(w^3)^2 = 2Qw^3 + P^3. \tag{10.13}$$

Table 10.1 summarizes the complete cubic-polynomial root-extraction method[6] in the revised notation—including a few fine points regarding superfluous roots and edge cases, treated in §§ 10.3 and 10.4 below.

10.3 Superfluous roots

As § 10.2 has observed, the equations of Table 10.1 seem to imply six roots, double the three the fundamental theorem of algebra (§ 6.2.2) allows a cubic polynomial to have. However, what the equations really imply is not six distinct roots but six distinct w. The definition $x \equiv w - P/w$ maps two w to any one x, so in fact the equations imply only three x and thus three roots z. The question then is: of the six w, which three do we really need and which three can we ignore as superfluous?

The six w naturally come in two groups of three: one group of three from the one w^3 and a second from the other. For this reason, we will guess—and

[6] [154, eqn. 5.3]

logically it is only a guess—that a single w^3 generates three distinct x and thus (because z differs from x only by a constant offset) all three roots z. If the guess is right, then the second w^3 cannot but yield the same three roots, which means that the second w^3 is superfluous and can safely be overlooked. But is the guess right? Does a single w^3 in fact generate three distinct x?

To prove that it does, let us suppose that it did not. Let us suppose that a single w^3 did generate two w which led to the same x. Letting the symbol w_1 represent the third w, then (since all three w come from the same w^3) the two w are $w = e^{\pm i2\pi/3}w_1$. Because $x \equiv w - P/w$, by successive steps,

$$\begin{aligned}
e^{+i2\pi/3}w_1 - \frac{P}{e^{+i2\pi/3}w_1} &= e^{-i2\pi/3}w_1 - \frac{P}{e^{-i2\pi/3}w_1},\\
e^{+i2\pi/3}w_1 + \frac{P}{e^{-i2\pi/3}w_1} &= e^{-i2\pi/3}w_1 + \frac{P}{e^{+i2\pi/3}w_1},\\
e^{+i2\pi/3}\left(w_1 + \frac{P}{w_1}\right) &= e^{-i2\pi/3}\left(w_1 + \frac{P}{w_1}\right),
\end{aligned}$$

which can only be true if

$$w_1^2 = -P.$$

Cubing[7] the last equation,

$$w_1^6 = -P^3;$$

but squaring the table's w^3 definition for $w = w_1$,

$$w_1^6 = 2Q^2 + P^3 \pm 2Q\sqrt{Q^2 + P^3}.$$

Combining the last two on w_1^6,

$$-P^3 = 2Q^2 + P^3 \pm 2Q\sqrt{Q^2 + P^3},$$

or, rearranging terms and halving,

$$Q^2 + P^3 = \mp Q\sqrt{Q^2 + P^3}.$$

Squaring,

$$Q^4 + 2Q^2P^3 + P^6 = Q^4 + Q^2P^3,$$

then canceling offsetting terms and factoring,

$$(P^3)(Q^2 + P^3) = 0.$$

[7] The verb *to cube* in this context means "to raise to the third power," as to change y to y^3, just as the verb *to square* means "to raise to the second power."

The last equation demands rigidly that either $P = 0$ or $P^3 = -Q^2$. Some cubic polynomials do meet the demand—§ 10.4 will treat these and the reader is asked to set them aside for the moment—but most cubic polynomials do not meet it. For most cubic polynomials, then, the contradiction proves false the assumption which gave rise to it. The assumption: that the three x descending from a single w^3 were not distinct. Therefore, provided that $P \neq 0$ and $P^3 \neq -Q^2$, the three x descending from a single w^3 are indeed distinct, as was to be demonstrated.

The conclusion: *either, not both, of the two signs in the table's quadratic solution* $w^3 \equiv Q \pm \sqrt{Q^2 + P^3}$ *demands to be considered.* One can choose either sign; it matters not which.[8] The one sign alone yields all three roots of the general cubic polynomial.

To calculate the three w from w^3, one can apply the Newton-Raphson iteration (4.32), the Taylor series of Table 8.1, or any other convenient root-finding technique to find a single root w_1 such that $w_1^3 = w^3$. The other two roots then come easier. They are $e^{\pm i2\pi/3}w_1$; but $e^{\pm i2\pi/3} = (-1 \pm i\sqrt{3})/2$, so

$$w = w_1, \frac{-1 \pm i\sqrt{3}}{2} w_1. \tag{10.14}$$

10.4 Edge cases

Section 10.3 excepts the edge cases $P = 0$ and $P^3 = -Q^2$. Mostly the book does not worry much about edge cases, but the effects of these cubic edge cases seem sufficiently nonobvious that the book might include here a few words about them, if for no other reason than to offer the reader a model of how to think about edge cases on his own. Table 10.1 gives the quadratic solution

$$w^3 \equiv Q \pm \sqrt{Q^2 + P^3},$$

in which § 10.3 generally finds it sufficient to consider either of the two signs. In the edge case $P = 0$,

$$w^3 = 2Q \text{ or } 0.$$

In the edge case $P^3 = -Q^2$,

$$w^3 = Q.$$

Both edge cases are interesting. In this section, we shall consider first the edge cases themselves, then their effect on the proof of § 10.3.

[8] Numerically, it can matter. As a simple rule, because w appears in the denominator of x's definition, when the two w^3 differ in magnitude one might choose the larger.

The edge case $P = 0$, $Q \neq 0$, like the general non-edge case, gives two distinct quadratic solutions w^3. One of the two however is $w^3 = Q - Q = 0$, which is awkward in light of Table 10.1's definition that $x \equiv w - P/w$. For this reason, in applying the table's method when $P = 0$, one chooses the other quadratic solution, $w^3 = Q + Q = 2Q$. (A reader who wishes to take extra care of the logic might here ask how one can be entirely sure that $w^3 = 0$ is not the w^3 we want to use despite that $x \equiv w - P/w$. More than one answer to this concern could be given. One answer would be that the fundamental theorem of algebra, § 6.2.2, implies three finite roots; so, since $w^3 = 0$ can supply none of the three, it must be that $w^3 = 2Q$ supplies all of them. A different answer is given later in the section.)

The edge case $P^3 = -Q^2 \neq 0$ gives only the one quadratic solution $w^3 = Q$; or, more precisely, it gives two quadratic solutions which happen to have the same value. This is fine. One merely accepts that $w^3 = Q$ and does not worry about choosing one w^3 over the other.

Neither edge case yields more than one, distinct, usable value for w^3, evidently. It would seem that the two edge cases were not troubled by the superfluous roots of § 10.3.

The double edge case, or *corner case,* arises where the two edges meet—where $P = 0$ and $P^3 = -Q^2$, or equivalently where $P = 0$ and $Q = 0$. At the corner, the trouble is that $w^3 = 0$ and that no alternate w^3 is available. However, according to (10.12), $x^3 = 2Q - 3Px$, which in this case means that $x^3 = 0$ and thus that $x = 0$ absolutely, no other x being possible. This implies the triple root $z = -a_2/3$.

And how about merely *double* roots? Section 10.3 has already shown that double roots cannot arise in non-edge cases. One can conclude that all cases of double roots are edge cases. (To identify to which of the two edge cases a double root corresponds is left as an exercise to the interested reader.[9])

Section 10.3 has excluded the edge cases from its proof of the sufficiency of a single w^3. Let us now add the edge cases to the proof. In the edge case $P^3 = -Q^2$, both w^3 are the same, so the one w^3 suffices by default because the other w^3 brings nothing different. The edge case $P = 0$ however does give two distinct w^3, one of which is $w^3 = 0$, which puts an awkward $0/0$ in the table's definition of x. We address this edge in the spirit of l'Hôpital's rule, by sidestepping it, changing P infinitesimally from $P = 0$ to $P = \epsilon$.

[9] The writer has not had cause to investigate the matter.

Then, choosing the $-$ sign in the definition of w^3,

$$\begin{aligned}
w^3 &= Q - \sqrt{Q^2+\epsilon^3} = Q - (Q)\left(1 + \frac{\epsilon^3}{2Q^2}\right) = -\frac{\epsilon^3}{2Q}, \\
w &= -\frac{\epsilon}{(2Q)^{1/3}}, \\
x &= w - \frac{\epsilon}{w} = -\frac{\epsilon}{(2Q)^{1/3}} + (2Q)^{1/3} = (2Q)^{1/3}.
\end{aligned}$$

But choosing the $+$ sign,

$$\begin{aligned}
w^3 &= Q + \sqrt{Q^2+\epsilon^3} = 2Q, \\
w &= (2Q)^{1/3}, \\
x &= w - \frac{\epsilon}{w} = (2Q)^{1/3} - \frac{\epsilon}{(2Q)^{1/3}} = (2Q)^{1/3}.
\end{aligned}$$

Evidently the roots come out the same, either way. This completes the proof.

10.5 Quartics

Having successfully extracted the roots of the general cubic polynomial, we now turn our attention to the general quartic. The kernel of the cubic technique lay in reducing the cubic to a quadratic. The kernel of the quartic technique lies likewise in reducing the quartic to a cubic. The details differ, though; and, strangely enough, in some ways the quartic reduction is actually the simpler.[10]

As with the cubic, one begins solving the quartic by changing the variable

$$x + h \leftarrow z \tag{10.15}$$

to obtain the equation

$$x^4 = sx^2 + px + q, \tag{10.16}$$

[10] Even stranger, historically Ferrari discovered it earlier [178, "Quartic equation"]. Apparently Ferrari discovered the quartic's resolvent cubic (10.22), which he could not solve until Tartaglia applied Vieta's transform to it. What motivated Ferrari to chase the quartic solution while the cubic solution remained still unknown, this writer does not know, but one supposes that it might make an interesting story.

The reason the quartic is simpler to reduce is perhaps related to the fact that $(1)^{1/4} = \pm 1, \pm i$, whereas $(1)^{1/3} = 1, (-1 \pm i\sqrt{3})/2$. The $(1)^{1/4}$ brings a much neater result, the roots lying nicely along the Argand axes. This may also be why the quintic is intractable—but here we trespass the professional mathematician's territory and stray from the scope of this book. See chapter 6's footnote 29.

where

$$\begin{aligned} h &\equiv -\frac{a_3}{4}, \\ s &\equiv -a_2 + 6\left(\frac{a_3}{4}\right)^2, \\ p &\equiv -a_1 + 2a_2\left(\frac{a_3}{4}\right) - 8\left(\frac{a_3}{4}\right)^3, \\ q &\equiv -a_0 + a_1\left(\frac{a_3}{4}\right) - a_2\left(\frac{a_3}{4}\right)^2 + 3\left(\frac{a_3}{4}\right)^4. \end{aligned} \tag{10.17}$$

To reduce (10.16) further, one must be cleverer. Ferrari[11] supplies the cleverness. The clever idea is to transfer some but not all of the sx^2 term to the equation's left side by

$$x^4 + 2ux^2 = (2u + s)x^2 + px + q,$$

where u remains to be chosen; then to complete the square on the equation's left side as in § 2.2, but with respect to x^2 rather than x, as

$$\left(x^2 + u\right)^2 = k^2x^2 + px + j^2, \tag{10.18}$$

where

$$\begin{aligned} k^2 &\equiv 2u + s, \\ j^2 &\equiv u^2 + q. \end{aligned} \tag{10.19}$$

Now, one must regard (10.18) and (10.19) properly. In these equations, s, p and q have definite values fixed by (10.17), but not so u, j or k. The variable u is completely free; we have introduced it ourselves and can assign it any value we like. And though j^2 and k^2 depend on u, still, even after specifying u we remain free at least to choose signs for j and k. As for u, though no choice would truly be wrong, one supposes that a wise choice might at least render (10.18) easier to simplify.

So, what choice for u would be wise? Well, look at (10.18). The left side of that equation is a perfect square. The right side would be, too, if it were that $p = \pm 2jk$; so, arbitrarily choosing the + sign, we propose the constraint that

$$p = 2jk, \tag{10.20}$$

or, better expressed,

$$j = \frac{p}{2k}. \tag{10.21}$$

[11] [178, "Quartic equation"]

Squaring (10.20) and substituting for j^2 and k^2 from (10.19), we have that

$$p^2 = 4(2u+s)(u^2+q);$$

or, after distributing factors, rearranging terms and scaling, that

$$0 = u^3 + \frac{s}{2}u^2 + qu + \frac{4sq-p^2}{8}. \tag{10.22}$$

Equation (10.22) is the *resolvent cubic,* which we know by Table 10.1 how to solve for u, and which we now specify as a second constraint. If the constraints (10.21) and (10.22) are both honored, then we can safely substitute (10.20) into (10.18) to reach the form

$$\left(x^2+u\right)^2 = k^2x^2 + 2jkx + j^2,$$

which is

$$\left(x^2+u\right)^2 = \left(kx+j\right)^2. \tag{10.23}$$

The resolvent cubic (10.22) of course yields three u not one, but the resolvent cubic is a voluntary constraint, so we can just pick one u and ignore the other two. Equation (10.19) then gives k (again, we can just pick one of the two signs), and (10.21) then gives j. With u, j and k established, (10.23) implies the quadratic

$$x^2 = \pm(kx+j) - u, \tag{10.24}$$

which (2.2) solves as

$$x = \pm\frac{k}{2} \pm_o \sqrt{\left(\frac{k}{2}\right)^2 \pm j - u}, \tag{10.25}$$

wherein the two ± signs are tied together but the third, $\pm_o$ sign is independent of the two. Equation (10.25), with the other equations and definitions of this section, reveals the four roots of the general quartic polynomial.

In view of (10.25), the change of variables

$$\begin{aligned} K &\leftarrow \frac{k}{2}, \\ J &\leftarrow j, \end{aligned} \tag{10.26}$$

improves the notation. Using the improved notation, Table 10.2 summarizes the complete quartic-polynomial root-extraction method.

Table 10.2: A method to extract the four roots of the general quartic polynomial. (In the table, the resolvent cubic is solved for u by the method of Table 10.1, where any one of the three resulting u serves. Either of the two K similarly serves. Of the three ± signs in x's definition, the $\pm_o$ is independent but the other two are tied together, the four resulting combinations giving the four roots of the general quartic.)

$$\begin{aligned}
0 &= z^4 + a_3 z^3 + a_2 z^2 + a_1 z + a_0 \\
s &\equiv -a_2 + 6\left(\frac{a_3}{4}\right)^2 \\
p &\equiv -a_1 + 2a_2\left(\frac{a_3}{4}\right) - 8\left(\frac{a_3}{4}\right)^3 \\
q &\equiv -a_0 + a_1\left(\frac{a_3}{4}\right) - a_2\left(\frac{a_3}{4}\right)^2 + 3\left(\frac{a_3}{4}\right)^4 \\
0 &= u^3 + \frac{s}{2}u^2 + qu + \frac{4sq - p^2}{8} \\
K &\equiv \pm\frac{\sqrt{2u+s}}{2} \\
J &\equiv \begin{cases} \pm\sqrt{u^2+q} & \text{if } K = 0, \\ p/4K & \text{otherwise.} \end{cases} \\
x &\equiv \pm K \pm_o \sqrt{K^2 \pm J - u} \\
z &= x - \frac{a_3}{4}
\end{aligned}$$

10.6 Guessing the roots

It is entertaining to put pencil to paper and use Table 10.1's method to extract the roots of the cubic polynomial

$$0 = [z - 1][z - i][z + i] = z^3 - z^2 + z - 1.$$

One finds that

$$\begin{aligned} z &= w + \frac{1}{3} - \frac{2}{3^2 w}, \\ w^3 &\equiv \frac{2\left(5 + \sqrt{3^3}\right)}{3^3}, \end{aligned}$$

which says indeed that $z = 1, \pm i$, but just you try to simplify it! A more baroque, more impenetrable way to write the number 1 is not easy to conceive. One has found the number 1 but cannot recognize it. Figuring the square and cube roots in the expression numerically, the root of the polynomial comes mysteriously to 1.0000, but why? The root's symbolic form gives little clue.

In general no better way is known;[12] we are stuck with the cubic baroquity. However, to the extent to which a cubic, a quartic, a quintic or any other polynomial has real, rational roots, a trick is known to sidestep Tables 10.1 and 10.2 and guess the roots directly. Consider for example the quintic polynomial

$$z^5 - \frac{7}{2}z^4 + 4z^3 + \frac{1}{2}z^2 - 5z + 3.$$

Doubling to make the coefficients all integers produces the polynomial

$$2z^5 - 7z^4 + 8z^3 + 1z^2 - \mathrm{0xA}z + 6,$$

which naturally has the same roots. If the roots are complex or irrational, they are hard to guess; but if any of the roots happens to be real and rational, it must belong to the set

$$\left\{ \pm 1, \pm 2, \pm 3, \pm 6, \pm\frac{1}{2}, \pm\frac{2}{2}, \pm\frac{3}{2}, \pm\frac{6}{2} \right\}.$$

[12] At least, no better way is known to this author. If any reader can straightforwardly simplify the expression without solving a cubic polynomial of some kind, the author would like to hear of it.

No other real, rational root is possible. Trying the several candidates on the polynomial, one finds that 1, −1 and 3/2 are indeed roots. Dividing these out leaves a quadratic which is easy to solve for the remaining roots.

The real, rational candidates are the factors of the polynomial's trailing coefficient (in the example, 6, whose factors are ±1, ±2, ±3 and ±6) divided by the factors of the polynomial's leading coefficient (in the example, 2, whose factors are ±1 and ±2). The reason no other real, rational root is possible is seen[13] by writing $z = p/q$—where $p, q \in \mathbb{Z}$ are integers and the fraction p/q is fully reduced—and then multiplying the nth-order polynomial by q^n to reach the form

$$a_n p^n + a_{n-1} p^{n-1} q + \cdots + a_1 p q^{n-1} + a_0 q^n = 0,$$

where all the coefficients a_k are integers. Moving the q^n term to the equation's right side, we have that

$$\left(a_n p^{n-1} + a_{n-1} p^{n-2} q + \cdots + a_1 q^{n-1}\right) p = -a_0 q^n,$$

which implies that $a_0 q^n$ is a multiple of p. But by demanding that the fraction p/q be fully reduced, we have defined p and q to be *relatively prime* to one another—that is, we have defined them to have no factors but ±1 in common—so, not only $a_0 q^n$ but a_0 itself is a multiple of p. By similar reasoning, a_n is a multiple of q. But if a_0 is a multiple of p, and a_n, a multiple of q, then p and q are factors of a_0 and a_n respectively. We conclude for this reason, as was to be demonstrated, that no real, rational root is possible except a factor of a_0 divided by a factor of a_n.[14]

Such root-guessing is little more than an algebraic trick, of course, but it can be a pretty useful trick if it saves us the embarrassment of inadvertently expressing simple rational numbers in ridiculous ways.

One could write much more about higher-order algebra, but now that the reader has tasted the topic he may feel inclined to agree that, though the general methods this chapter has presented to solve cubics and quartics are interesting, further effort were nevertheless probably better spent elsewhere. The next several chapters turn to the topic of the matrix, harder but much more profitable, toward which we mean to put substantial effort.

[13] The presentation here is quite informal. We do not want to spend many pages on this.

[14] [156, § 3.2]

Part II

Matrices and vectors

Chapter 11

The matrix

Chapters 2 through 9 have laid the foundations of applied mathematics. This chapter begins to build on those foundations, demanding some heavier mathematical lifting.

Taken by themselves, most of the foundational methods of the earlier chapters have handled only one or at most a few numbers (or functions) at a time. However, in practical applications the need to handle large arrays of numbers at once arises often. Some nonobvious effects then emerge, as for example the eigenvalue of chapter 14.

Regarding the eigenvalue: the eigenvalue was always there, but prior to this point in the book it was usually trivial—the eigenvalue of 5 is just 5, for instance—so we didn't bother much to talk about it. It is when numbers are laid out in orderly grids like

$$C = \begin{bmatrix} 6 & 4 & 0 \\ 3 & 0 & 1 \\ 3 & 1 & 0 \end{bmatrix}$$

that nontrivial eigenvalues arise (though you cannot tell just by looking, the eigenvalues of C happen to be -1 and $[7 \pm \sqrt{\mathrm{0x49}}]/2$). But, just what is an *eigenvalue?* Answer: an eigenvalue is the value by which an object like C scales an eigenvector without altering the eigenvector's direction. Of course, we have not yet said what an *eigenvector* is, either, or how C might scale something, but it is to answer precisely such questions that this chapter and the three which follow it are written.

So, we are getting ahead of ourselves. Let's back up.

An object like C is called a *matrix.* It serves as a generalized coefficient or multiplier. Where we have used single numbers as coefficients or multipliers heretofore, one can with sufficient care often use matrices instead. The matrix interests us for this reason among others.

The technical name for the "single number" is the *scalar.* Such a number, as for instance 5 or $-4 + i3$, is called a scalar because its action alone during multiplication is simply to scale the thing it multiplies. Besides acting alone, however, scalars can also act in concert—in orderly formations—thus constituting any of three basic kinds of arithmetical object:

- the *scalar* itself, a single number like $\alpha = 5$ or $\beta = -4 + i3$;
- the *vector,* a column of m scalars like
$$\mathbf{u} = \begin{bmatrix} 5 \\ -4 + i3 \end{bmatrix},$$
which can be written in-line with the notation $\mathbf{u} = [5 \ \ -4 + i3]^T$ (here there are two scalar elements, 5 and $-4+i3$, so in this example $m = 2$);
- the *matrix,* an $m \times n$ grid of scalars, or equivalently a row of n vectors, like
$$A = \begin{bmatrix} 0 & 6 & 2 \\ 1 & 1 & -1 \end{bmatrix},$$
which can be written in-line with the notation $A = [0 \ 6 \ 2; 1 \ 1 \ -1]$ or the notation $A = [0 \ 1; 6 \ 1; 2 \ -1]^T$ (here there are two rows and three columns of scalar elements, so in this example $m = 2$ and $n = 3$).

Several general points are immediately to be observed about these various objects. First, despite the geometrical Argand interpretation of the complex number, a complex number is not a two-element vector but a scalar; therefore any or all of a vector's or matrix's scalar elements can be complex. Second, an m-element vector does not differ for most purposes from an $m \times 1$ matrix; generally the two can be regarded as the same thing. Third, the three-element (that is, three-dimensional) geometrical vector of § 3.3 is just an m-element vector with $m = 3$. Fourth, m and n can be any nonnegative integers, even one, even zero, even infinity.[1]

Where one needs visually to distinguish a symbol like A representing a matrix, one can write it $[A]$, in square brackets.[2] Normally however a simple A suffices.

[1] Fifth, though the progression *scalar, vector, matrix* suggests next a "matrix stack" or stack of p matrices, such objects in fact are seldom used. As we shall see in § 11.1, the chief advantage of the standard matrix is that it neatly represents the linear transformation of one vector into another. "Matrix stacks" bring no such advantage. This book does not treat them.

[2] Alternate notations seen in print include $\overline{\overline{A}}$ and $\mathbf{A}$.

The matrix is a notoriously hard topic to motivate. The idea of the matrix is deceptively simple. The mechanics of matrix arithmetic are deceptively intricate. The most basic body of matrix theory, without which little or no useful matrix work can be done, is deceptively extensive. The matrix neatly encapsulates a substantial knot of arithmetical tedium and clutter, but to understand the matrix one must first understand the tedium and clutter the matrix encapsulates. As far as the author is aware, no one has ever devised a way to introduce the matrix which does not seem shallow, tiresome, irksome, even interminable at first encounter; yet the matrix is too important to ignore. Applied mathematics brings nothing else quite like it.[3]

Chapters 11 through 14 treat the matrix and its algebra. This chapter,

[3] In most of its chapters, the book seeks a balance between terseness the determined beginner cannot penetrate and prolixity the seasoned veteran will not abide. The matrix upsets this balance.

Part of the trouble with the matrix is that its arithmetic is just that, an arithmetic, no more likely to be mastered by mere theoretical study than was the elementary arithmetic of childhood. To master matrix arithmetic, one must drill it; yet the book you hold is fundamentally one of theory not drill.

The reader who has previously drilled matrix arithmetic will meet here the essential applied theory of the matrix. That reader will find this chapter and the next three tedious enough. The reader who has not previously drilled matrix arithmetic, however, is likely to find these chapters positively hostile. Only the doggedly determined beginner will learn the matrix here alone; others will find it more amenable to drill matrix arithmetic first in the early chapters of an introductory linear algebra textbook, dull though such chapters be (see [106] or better yet the fine, surprisingly less dull [75] for instance, though the early chapters of almost any such book give the needed arithmetical drill.) Returning here thereafter, the beginner can expect to find *these* chapters still tedious but no longer impenetrable. The reward is worth the effort. That is the approach the author recommends.

To the mathematical rebel, the young warrior with face painted and sword agleam, still determined to learn the matrix here alone, the author salutes his honorable defiance. Would the rebel consider alternate counsel? If so, then the rebel might compose a dozen matrices of various sizes and shapes, broad, square and tall, decomposing each carefully by pencil per the Gauss-Jordan method of § 12.3, checking results (again by pencil; using a machine defeats the point of the exercise, and using a sword, well, it won't work) by multiplying factors to restore the original matrices. Several hours of such drill should build the young warrior the practical arithmetical foundation to master—with commensurate effort—the theory these chapters bring. The way of the warrior is hard, but conquest is not impossible.

To the matrix veteran, the author presents these four chapters with grim enthusiasm. Substantial, logical, necessary the chapters may be, but exciting they are not. At least, the earlier parts are not very exciting (later parts are better). As a reasonable compromise, the veteran seeking more interesting reading might skip directly to chapters 13 and 14, referring back to chapters 11 and 12 as need arises.

chapter 11, introduces the rudiments of the matrix itself.[4]

11.1 Provenance and basic use

It is in the study of linear transformations that the concept of the matrix first arises. We begin there.

11.1.1 The linear transformation

Section 7.3.3 has introduced the idea of linearity. The *linear transformation*[5] is the operation of an $m \times n$ matrix A, as in

$$A\mathbf{x} = \mathbf{b}, \tag{11.1}$$

to transform an n-element vector $\mathbf{x}$ into an m-element vector $\mathbf{b}$, while respecting the rules of linearity

$$\begin{aligned} A(\mathbf{x}_1 + \mathbf{x}_2) &= A\mathbf{x}_1 + A\mathbf{x}_2 &&= \mathbf{b}_1 + \mathbf{b}_2, \\ A(\alpha\mathbf{x}) &= \alpha A\mathbf{x} &&= \alpha\mathbf{b}, \\ A(0) &= 0. \end{aligned} \tag{11.2}$$

For example,

$$A = \begin{bmatrix} 0 & 6 & 2 \\ 1 & 1 & -1 \end{bmatrix}$$

is the 2×3 matrix which transforms a three-element vector $\mathbf{x}$ into a two-element vector $\mathbf{b}$ such that

$$A\mathbf{x} = \begin{bmatrix} 0x_1 + 6x_2 + 2x_3 \\ 1x_1 + 1x_2 - 1x_3 \end{bmatrix} = \mathbf{b},$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

[4] [14][59][75][106]

[5] Professional mathematicians conventionally are careful to begin by drawing a clear distinction between the ideas of the linear transformation, the basis set and the simultaneous system of linear equations—proving from suitable axioms that the three amount more or less to the same thing, rather than implicitly assuming the fact. The professional approach [14, chapters 1 and 2][106, chapters 1, 2 and 5] has much to recommend it, but it is not the approach we will follow here.

In general, the operation of a matrix A is that[6,7]

$$b_i = \sum_{j=1}^{n} a_{ij} x_j, \tag{11.3}$$

where x_j is the jth element of $\mathbf{x}$, b_i is the ith element of $\mathbf{b}$, and

$$a_{ij} \equiv [A]_{ij}$$

is the element at the ith row and jth column of A, counting from top left (in the example for instance, $a_{12} = 6$).

Besides representing linear transformations as such, matrices can also represent simultaneous systems of linear equations. For example, the system

$$\begin{aligned} 0x_1 + 6x_2 + 2x_3 &= 2, \\ 1x_1 + 1x_2 - 1x_3 &= 4, \end{aligned}$$

is compactly represented as

$$A\mathbf{x} = \mathbf{b},$$

with A as given above and $\mathbf{b} = [2\ 4]^T$. Seen from this point of view, a simultaneous system of linear equations is itself neither more nor less than a linear transformation.

11.1.2 Matrix multiplication (and addition)

Nothing prevents one from lining several vectors $\mathbf{x}_k$ up in a row, industrial mass production-style, transforming them at once into the corresponding

[6] As observed in appendix B, there are unfortunately not enough distinct Roman and Greek letters available to serve the needs of higher mathematics. In matrix work, the Roman letters ijk conventionally serve as indices, but the same letter i also serves as the imaginary unit, which is not an index and has nothing to do with indices. Fortunately, the meaning is usually clear from the context: i in $\sum_i$ or a_{ij} is an index; i in $-4 + i3$ or $e^{i\phi}$ is the imaginary unit. Should a case arise in which the meaning is not clear, one can use ℓjk or some other convenient letters for the indices.

[7] Whether to let the index j run from 0 to $n-1$ or from 1 to n is an awkward question of applied mathematical style. In computers, the index normally runs from 0 to $n-1$, and in many ways this really is the more sensible way to do it. In mathematical theory, however, a 0 index normally implies something special or basic about the object it identifies. The book you are reading tends to let the index run from 1 to n, following mathematical convention in the matter for this reason.

Conceived more generally, an $m \times n$ matrix can be considered an $\infty \times \infty$ matrix with zeros in the unused cells. Here, both indices i and j run from $-\infty$ to $+\infty$ anyway, so the computer's indexing convention poses no dilemma in this case. See § 11.3.

vectors $\mathbf{b}_k$ by the same matrix A. In this case,

$$\begin{aligned} X &\equiv \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \end{bmatrix}, \\ B &\equiv \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_p \end{bmatrix}, \\ AX &= B, \\ b_{ik} &= \sum_{j=1}^{n} a_{ij} x_{jk}. \end{aligned} \tag{11.4}$$

Equation (11.4) implies a definition for matrix multiplication. Such matrix multiplication is associative since

$$\begin{aligned} [(A)(XY)]_{ik} &= \sum_{j=1}^{n} a_{ij} [XY]_{jk} \\ &= \sum_{j=1}^{n} a_{ij} \left[\sum_{\ell=1}^{p} x_{j\ell} y_{\ell k} \right] \\ &= \sum_{\ell=1}^{p} \sum_{j=1}^{n} a_{ij} x_{j\ell} y_{\ell k} \\ &= [(AX)(Y)]_{ik} . \end{aligned} \tag{11.5}$$

Matrix multiplication is not generally commutative, however;

$$AX \neq XA, \tag{11.6}$$

as one can show by a suitable counterexample like $A = [0\ 1; 0\ 0]$, $X = [1\ 0; 0\ 0]$. To multiply a matrix by a scalar, one multiplies each of the matrix's elements individually by the scalar:

$$[\alpha A]_{ij} = \alpha a_{ij}. \tag{11.7}$$

Evidently multiplication by a scalar *is* commutative: $\alpha A\mathbf{x} = A\alpha\mathbf{x}$.

Matrix addition works in the way one would expect, element by element; and as one can see from (11.4), under multiplication, matrix addition is indeed distributive:

$$\begin{aligned} [X + Y]_{ij} &= x_{ij} + y_{ij}; \\ (A)(X + Y) &= AX + AY; \\ (A + C)(X) &= AX + CX. \end{aligned} \tag{11.8}$$

11.1.3 Row and column operators

The matrix equation $A\mathbf{x} = \mathbf{b}$ represents the linear transformation of $\mathbf{x}$ into $\mathbf{b}$, as we have seen. Viewed from another perspective, however, the same matrix equation represents something else; it represents a weighted sum of the columns of A, with the elements of $\mathbf{x}$ as the weights. In this view, one writes (11.3) as

$$\mathbf{b} = \sum_{j=1}^{n} [A]_{*j} x_j, \tag{11.9}$$

where $[A]_{*j}$ is the jth column of A. Here $\mathbf{x}$ is not only a vector; it is also an operator. It operates on A's columns. By virtue of multiplying A from the right, the vector $\mathbf{x}$ is a *column operator* acting on A.

If several vectors $\mathbf{x}_k$ line up in a row to form a matrix X, such that $AX = B$, then the matrix X is likewise a column operator:

$$[B]_{*k} = \sum_{j=1}^{n} [A]_{*j} x_{jk}. \tag{11.10}$$

The kth column of X weights the several columns of A to yield the kth column of B.

If a matrix multiplying from the right is a column operator, is a matrix multiplying from the left a *row operator?* Indeed it is. Another way to write that $AX = B$, besides (11.10), is

$$[B]_{i*} = \sum_{j=1}^{n} a_{ij} [X]_{j*}. \tag{11.11}$$

The ith row of A weights the several rows of X to yield the ith row of B. The matrix A is a row operator. (Observe the notation. The $*$ here means "any" or "all." Hence $[X]_{j*}$ means "jth row, all columns of X"—that is, the jth row of X. Similarly, $[A]_{*j}$ means "all rows, jth column of A"—that is, the jth column of A.)

Column operators attack from the right; row operators, from the left. This rule is worth memorizing; the concept is important. In $AX = B$, the matrix X operates on A's columns; the matrix A operates on X's rows.

Since matrix multiplication produces the same result whether one views it as a linear transformation (11.4), a column operation (11.10) or a row operation (11.11), one might wonder what purpose lies in defining matrix multiplication three separate ways. However, it is not so much for the sake of the mathematics that we define it three ways as it is for the sake of the mathematician. We do it for ourselves. Mathematically, the latter two do indeed expand to yield (11.4), but as written the three represent three different perspectives on the matrix. A tedious, nonintuitive matrix theorem from one perspective can appear suddenly obvious from another (see for example eqn. 11.63). Results hard to visualize one way are easy to visualize another. It is worth developing the mental agility to view and handle matrices all three ways for this reason.

11.1.4 The transpose and the adjoint

One function peculiar to matrix algebra is the *transpose*

$$\begin{aligned} C &= A^T, \\ c_{ij} &= a_{ji}, \end{aligned} \tag{11.12}$$

which mirrors an $m \times n$ matrix into an $n \times m$ matrix. For example,

$$A^T = \begin{bmatrix} 0 & 1 \\ 6 & 1 \\ 2 & -1 \end{bmatrix}.$$

Similar and even more useful is the *conjugate transpose* or *adjoint*[8]

$$\begin{aligned} C &= A^*, \\ c_{ij} &= a_{ji}^*, \end{aligned} \tag{11.13}$$

which again mirrors an $m \times n$ matrix into an $n \times m$ matrix, but conjugates each element as it goes.

The transpose is convenient notationally to write vectors and matrices inline and to express certain matrix-arithmetical mechanics; but algebraically the transpose is artificial. It is the adjoint rather which mirrors a matrix properly. (If the transpose and adjoint functions applied to words as to matrices, then the transpose of "derivations" would be "snoitavired," whereas the adjoint would be "ƨnoitɒviɿɘb." See the difference?) On real-valued matrices like the A in the example, of course, the transpose and the adjoint amount to the same thing.

[8] Alternate notations sometimes seen in print for the adjoint include $A^\dagger$ (a notation which in this book means something unrelated) and A^H (a notation which recalls the name of the mathematician Charles Hermite). However, the book you are reading writes the adjoint only as A^*, a notation which better captures the sense of the thing in the author's view.

If one needed to conjugate the elements of a matrix without transposing the matrix itself, one could contrive notation like A^{*T}. Such a need seldom arises, however.

Observe that

$$\begin{aligned}(A_2 A_1)^T &= A_1^T A_2^T, \\ (A_2 A_1)^* &= A_1^* A_2^*,\end{aligned} \tag{11.14}$$

and more generally that[9]

$$\begin{aligned}\left(\prod_k A_k\right)^T &= \coprod_k A_k^T, \\ \left(\prod_k A_k\right)^* &= \coprod_k A_k^*.\end{aligned} \tag{11.15}$$

11.2 The Kronecker delta

Section 7.7 has introduced the Dirac delta. The discrete analog of the Dirac delta is the *Kronecker delta*[10]

$$\delta_i \equiv \begin{cases} 1 & \text{if } i = 0, \\ 0 & \text{otherwise;} \end{cases} \tag{11.16}$$

or

$$\delta_{ij} \equiv \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{11.17}$$

The Kronecker delta enjoys the Dirac-like properties that

$$\sum_{i=-\infty}^{\infty} \delta_i = \sum_{i=-\infty}^{\infty} \delta_{ij} = \sum_{j=-\infty}^{\infty} \delta_{ij} = 1 \tag{11.18}$$

and that

$$\sum_{j=-\infty}^{\infty} \delta_{ij} a_{jk} = a_{ik}, \tag{11.19}$$

the latter of which is the Kronecker sifting property. The Kronecker equations (11.18) and (11.19) parallel the Dirac equations (7.24) and (7.25).

Chapters 11 and 14 will find frequent use for the Kronecker delta. Later, § 15.4.3 will revisit the Kronecker delta in another light.

[9] Recall from § 2.3 that $\prod_k A_k = \cdots A_3 A_2 A_1$, whereas $\coprod_k A_k = A_1 A_2 A_3 \cdots$.

[10] [182, "Kronecker delta," 15:59, 31 May 2006]

11.3 Dimensionality and matrix forms

An $m \times n$ matrix like

$$X = \begin{bmatrix} -4 & 0 \\ 1 & 2 \\ 2 & -1 \end{bmatrix}$$

can be viewed as the $\infty \times \infty$ matrix

$$X = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & -4 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 1 & 2 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 2 & -1 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

with zeros in the unused cells. As before, $x_{11} = -4$ and $x_{32} = -1$, but now x_{ij} exists for all integral i and j; for instance, $x_{(-1)(-1)} = 0$. For such a matrix, indeed for all matrices, the matrix multiplication rule (11.4) generalizes to

$$\begin{aligned} B &= AX, \\ b_{ik} &= \sum_{j=-\infty}^{\infty} a_{ij} x_{jk}. \end{aligned} \tag{11.20}$$

For square matrices whose purpose is to manipulate other matrices or vectors in place, merely padding with zeros often does not suit. Consider for example the square matrix

$$A_3 = \begin{bmatrix} 1 & 0 & 0 \\ 5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

This A_3 is indeed a matrix, but when it acts A_3X as a row operator on some $3 \times p$ matrix X, its effect is to add to X's second row, 5 times the first. Further consider

$$A_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 5 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

which does the same to a $4 \times p$ matrix X. We can also define $A_5, A_6, A_7, \ldots$, if we want; but, really, all these express the same operation: "to add to the second row, 5 times the first."

The $\infty \times \infty$ matrix

$$A = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 1 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 5 & 1 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

expresses the operation generally. As before, $a_{11} = 1$ and $a_{21} = 5$, but now also $a_{(-1)(-1)} = 1$ and $a_{09} = 0$, among others. By running ones infinitely both ways out the *main diagonal,* we guarantee by (11.20) that when A acts AX on a matrix X of any dimensionality whatsoever, A adds to the second row of X, 5 times the first—and affects no other row. (But what if X is a $1 \times p$ matrix, and *has* no second row? Then the operation AX creates a new second row, 5 times the first—or rather so fills in X's previously null second row.)

In the infinite-dimensional view, the matrices A and X differ essentially.[11] This section explains, developing some nonstandard formalisms the derivations of later sections and chapters can use.[12]

[11] This particular section happens to use the symbols A and X to represent certain specific matrix forms because such usage flows naturally from the usage $A\mathbf{x} = \mathbf{b}$ of § 11.1. Such usage admittedly proves awkward in other contexts. Traditionally in matrix work and elsewhere in the book, the letter A does not necessarily represent an extended operator as it does here, but rather an arbitrary matrix of no particular form.

[12] The idea of infinite dimensionality is sure to discomfit some readers, who have studied matrices before and are used to thinking of a matrix as having some definite size. There is nothing wrong with thinking of a matrix as having some definite size, only that that view does not suit the present book's development. And really, the idea of an $\infty \times 1$ vector or an $\infty \times \infty$ matrix should not seem so strange. After all, consider the vector $\mathbf{u}$ such that

$$u_\ell = \sin \ell\epsilon,$$

where $0 < \epsilon \ll 1$ and ℓ is an integer, which holds all values of the function $\sin\theta$ of a real argument θ. Of course one does not actually write down or store all the elements of an infinite-dimensional vector or matrix, any more than one actually writes down or stores all the bits (or digits) of 2π. Writing them down or storing them is not the point. The point is that infinite dimensionality is all right; that the idea thereof does not threaten to overturn the reader's preëxisting matrix knowledge; that, though the construct seem unfamiliar, no fundamental conceptual barrier rises against it.

Different ways of looking at the same mathematics can be extremely useful to the applied mathematician. The applied mathematical reader who has never heretofore considered

11.3.1 The null and dimension-limited matrices

The *null matrix* is just what its name implies:

$$0 = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix};$$

or more compactly,

$$[0]_{ij} = 0.$$

Special symbols like $\overline{\overline{0}}$, $\mathbf{0}$ or O are possible for the null matrix, but usually a simple 0 suffices. There are no surprises here; the null matrix brings all the expected properties of a zero, like

$$\begin{aligned} 0 + A &= A, \\ [0][X] &= 0. \end{aligned}$$

The same symbol 0 used for the null scalar (zero) and the null matrix is used for the null vector, too. Whether the scalar 0, the vector 0 and the matrix 0 actually represent different things is a matter of semantics, but the three are interchangeable for most practical purposes in any case. Basically, a zero is a zero is a zero; there's not much else to it.[13]

Now a formality: the ordinary $m \times n$ matrix X can be viewed, infinite-dimensionally, as a variation on the null matrix, inasmuch as X differs from the null matrix only in the mn elements x_{ij}, $1 \le i \le m$, $1 \le j \le n$. Though the theoretical dimensionality of X be $\infty \times \infty$, one need record only the mn elements, plus the values of m and n, to retain complete information about such a matrix. So the semantics are these: when we call a matrix X an *$m \times n$ matrix,* or more precisely a *dimension-limited matrix* with an $m \times n$

infinite dimensionality in vectors and matrices would be well served to take the opportunity to do so here. As we shall discover in chapter 12, dimensionality is a poor measure of a matrix's size in any case. What really counts is not a matrix's $m \times n$ dimensionality but rather its *rank.*

[13] Well, of course, there's a lot else to it, when it comes to dividing by zero as in chapter 4, or to summing an infinity of zeros as in chapter 7, but those aren't what we were speaking of here.

active region, we will mean formally that X is an $\infty \times \infty$ matrix whose elements are all zero outside the $m \times n$ rectangle:

$$x_{ij} = 0 \text{ except where } 1 \le i \le m \text{ and } 1 \le j \le n. \tag{11.21}$$

By these semantics, every 3×2 matrix (for example) is also a formally a 4×4 matrix; but a 4×4 matrix is not in general a 3×2 matrix.

11.3.2 The identity and scalar matrices and the extended operator

The *general identity matrix*—or simply, the *identity matrix*—is

$$I = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 1 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 1 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 1 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 1 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

or more compactly,

$$[I]_{ij} = \delta_{ij}, \tag{11.22}$$

where δ_{ij} is the Kronecker delta of § 11.2. The identity matrix I is a matrix 1, as it were,[14] bringing the essential property one expects of a 1:

$$IX = X = XI. \tag{11.23}$$

The *scalar matrix* is

$$\lambda I = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & \lambda & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & \lambda & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & \lambda & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & \lambda & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & \lambda & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

or more compactly,

$$[\lambda I]_{ij} = \lambda\delta_{ij}, \tag{11.24}$$

[14] In fact you can write it as 1 if you like. That is essentially what it is. The I can be regarded as standing for "identity" or as the Roman numeral I.

If the identity matrix I is a matrix 1, then the scalar matrix λI is a matrix λ, such that

$$[\lambda I]X = \lambda X = X[\lambda I]. \tag{11.25}$$

The identity matrix is (to state the obvious) just the scalar matrix with $\lambda = 1$.

The *extended operator* A is a variation on the scalar matrix λI, $\lambda \neq 0$, inasmuch as A differs from λI only in p specific elements, with p a finite number. Symbolically,

$$\begin{aligned} a_{ij} &= \begin{cases} (\lambda)(\delta_{ij} + \alpha_k) & \text{if } (i,j) = (i_k, j_k),\ 1 \le k \le p, \\ \lambda\delta_{ij} & \text{otherwise;} \end{cases} \\ \lambda &\neq 0. \end{aligned} \tag{11.26}$$

The several α_k control how the extended operator A differs from λI. One need record only the several α_k along with their respective addresses (i_k, j_k), plus the scale λ, to retain complete information about such a matrix. For example, for an extended operator fitting the pattern

$$A = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & \lambda & 0 & 0 & 0 & 0 & \cdots \\ \cdots & \lambda\alpha_1 & \lambda & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & \lambda & \lambda\alpha_2 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & \lambda(1+\alpha_3) & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & \lambda & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

one need record only the values of α_1, α_2 and α_3, the respective addresses $(2,1)$, $(3,4)$ and $(4,4)$, and the value of the scale λ; this information alone implies the entire $\infty \times \infty$ matrix A.

When we call a matrix A an *extended $n \times n$ operator,* or an extended operator with an *$n \times n$ active region,* we will mean formally that A is an $\infty \times \infty$ matrix and is further an extended operator for which

$$1 \le i_k \le n \text{ and } 1 \le j_k \le n \text{ for all } 1 \le k \le p. \tag{11.27}$$

That is, an extended $n \times n$ operator is one whose several α_k all lie within the $n \times n$ square. The A in the example is an extended 4×4 operator (and also a 5×5, a 6×6, etc., but not a 3×3).

(Often in practice for smaller operators—especially in the typical case that $\lambda = 1$—one finds it easier just to record all the $n \times n$ elements of the active region. This is fine. Large matrix operators however tend to be *sparse,* meaning that they depart from λI in only a very few of their many elements. It would waste a lot of computer memory explicitly to store all those zeros, so one normally stores just the few elements, instead.)

Implicit in the definition of the extended operator is that the identity matrix I and the scalar matrix λI, $\lambda \neq 0$, are extended operators with 0×0 active regions (and also 1×1, 2×2, etc.). If $\lambda = 0$, however, the scalar matrix λI is just the null matrix, which is no extended operator but rather by definition a 0×0 dimension-limited matrix.

11.3.3 The active region

Though maybe obvious, it bears stating explicitly that a product of dimension-limited and/or extended-operational matrices with $n \times n$ active regions itself has an $n \times n$ active region.[15] (Remember that a matrix with an $m' \times n'$ active region also by definition has an $n \times n$ active region if $m' \leq n$ and $n' \leq n$.) If any of the factors has dimension-limited form then so does the product; otherwise the product is an extended operator.[16]

11.3.4 Other matrix forms

Besides the dimension-limited form of § 11.3.1 and the extended-operational form of § 11.3.2, other infinite-dimensional matrix forms are certainly possible. One could for example advantageously define a "null sparse" form, recording only nonzero elements and their addresses in an otherwise null matrix; or a "tridiagonal extended" form, bearing repeated entries not only along the main diagonal but also along the diagonals just above and just below. Section 11.9 introduces one worthwhile matrix which fits neither the dimension-limited nor the extended-operational form. Still, the dimension-

[15] The section's earlier subsections formally define the term *active region* with respect to each of the two matrix forms.

[16] If symbolic proof of the subsection's claims is wanted, here it is in outline:

$$\begin{aligned}
a_{ij} &= \lambda_a \delta_{ij} \quad \text{unless } 1 \leq (i,j) \leq n, \\
b_{ij} &= \lambda_b \delta_{ij} \quad \text{unless } 1 \leq (i,j) \leq n; \\
[AB]_{ij} &= \sum_k a_{ik} b_{kj} \\
&= \begin{cases} \sum_k (\lambda_a \delta_{ik}) b_{kj} &= \lambda_a b_{ij} \quad \text{unless } 1 \leq i \leq n \\ \sum_k a_{ik} (\lambda_b \delta_{kj}) &= \lambda_b a_{ij} \quad \text{unless } 1 \leq j \leq n \end{cases} \\
&= \lambda_a \lambda_b \delta_{ij} \quad \text{unless } 1 \leq (i,j) \leq n.
\end{aligned}$$

It's probably easier just to sketch the matrices and look at them, though.

limited and extended-operational forms are normally the most useful, and they are the ones we will principally be handling in this book.

One reason to have defined specific infinite-dimensional matrix forms is to show how straightforwardly one can fully represent a practical matrix of an infinity of elements by a modest, finite quantity of information. Further reasons to have defined such forms will soon occur.

11.3.5 The rank-r identity matrix

The *rank-r identity matrix* I_r is the dimension-limited matrix for which

$$[I_r]_{ij} = \begin{cases} \delta_{ij} & \text{if } 1 \le i \le r \text{ and/or } 1 \le j \le r, \\ 0 & \text{otherwise,} \end{cases} \tag{11.28}$$

where either the "and" or the "or" can be regarded (it makes no difference). The effect of I_r is that

$$\begin{aligned} I_m X &= X = X I_n, \\ I_m \mathbf{x} &= \mathbf{x}, \end{aligned} \tag{11.29}$$

where X is an $m \times n$ matrix and $\mathbf{x}$, an $m \times 1$ vector. Examples of I_r include

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

(Remember that in the infinite-dimensional view, I_3, though a 3×3 matrix, is formally an $\infty \times \infty$ matrix with zeros in the unused cells. It has only the three ones and fits the 3×3 dimension-limited form of § 11.3.1. The areas of I_3 not shown are all zero, even along the main diagonal.)

The rank r can be any nonnegative integer, even zero (though the rank-zero identity matrix I_0 is in fact the null matrix, normally just written 0). If alternate indexing limits are needed (for instance for a computer-indexed identity matrix whose indices run from 0 to $r - 1$), the notation I_a^b, where

$$[I_a^b]_{ij} \equiv \begin{cases} \delta_{ij} & \text{if } a \le i \le b \text{ and/or } a \le j \le b, \\ 0 & \text{otherwise,} \end{cases} \tag{11.30}$$

can be used; the rank in this case is $r = b - a + 1$, which is just the count of ones along the matrix's main diagonal.

The name "rank-r" implies that I_r has a "rank" of r, and indeed it does. For the moment, however, we will discern the attribute of rank only in the rank-r identity matrix itself. Section 12.5 defines *rank* for matrices more generally.

11.3.6 The truncation operator

The rank-r identity matrix I_r is also the *truncation operator.* Attacking from the left, as in I_rA, it retains the first through rth rows of A but cancels other rows. Attacking from the right, as in AI_r, it retains the first through rth columns. Such truncation is useful symbolically to reduce an extended operator to dimension-limited form.

Whether a matrix C has dimension-limited or extended-operational form (though not necessarily if it has some other form), if it has an $m \times n$ active region[17] and

$$\begin{aligned} m &\le r, \\ n &\le r, \end{aligned}$$

then

$$I_rC = I_rCI_r = CI_r. \tag{11.31}$$

For such a matrix, (11.31) says at least two things:

- It is superfluous to truncate both rows and columns; it suffices to truncate one or the other.
- The rank-r identity matrix I_r commutes freely past C.

Evidently big identity matrices commute freely where small ones cannot (and the general identity matrix $I = I_{-\infty}^{\infty}$ commutes freely past everything).

11.3.7 The elementary vector and the lone-element matrix

The *lone-element matrix* E_{mn} is the matrix with a one in the mnth cell and zeros elsewhere:

$$[E_{mn}]_{ij} \equiv \delta_{im}\delta_{jn} = \begin{cases} 1 & \text{if } i = m \text{ and } j = n, \\ 0 & \text{otherwise.} \end{cases} \tag{11.32}$$

By this definition, $C = \sum_{i,j} c_{ij}E_{ij}$ for any matrix C. The vector analog of the lone-element matrix is the *elementary vector* $\mathbf{e}_m$, which has a one as the mth element:

$$[\mathbf{e}_m]_i \equiv \delta_{im} = \begin{cases} 1 & \text{if } i = m, \\ 0 & \text{otherwise.} \end{cases} \tag{11.33}$$

By this definition, $[I]_{*j} = \mathbf{e}_j$ and $[I]_{i*} = \mathbf{e}_i^T$.

[17] Refer to the definitions of *active region* in §§ 11.3.1 and 11.3.2. That a matrix has an $m \times n$ active region does not necessarily mean that it is all zero outside the $m \times n$ rectangle. (After all, if it were always all zero outside, then there would be little point in applying a truncation operator. There would be nothing there to truncate.)

11.3.8 Off-diagonal entries

It is interesting to observe and useful to note that if

$$[C_1]_{i*} = [C_2]_{i*} = \mathbf{e}_i^T,$$

then also

$$[C_1C_2]_{i*} = \mathbf{e}_i^T; \tag{11.34}$$

and likewise that if

$$[C_1]_{*j} = [C_2]_{*j} = \mathbf{e}_j,$$

then also

$$[C_1C_2]_{*j} = \mathbf{e}_j. \tag{11.35}$$

The product of matrices has off-diagonal entries in a row or column only if at least one of the factors itself has off-diagonal entries in that row or column. Or, less readably but more precisely, *the ith row or jth column of the product of matrices can depart from $\mathbf{e}_i^T$ or $\mathbf{e}_j$, respectively, only if the corresponding row or column of at least one of the factors so departs.* The reason is that in (11.34), C_1 acts as a row operator on C_2; that if C_1's ith row is $\mathbf{e}_i^T$, then its action is merely to duplicate C_2's ith row, which itself is just $\mathbf{e}_i^T$. Parallel logic naturally applies to (11.35).

11.4 The elementary operator

Section 11.1.3 has introduced the general row or column operator. Denoted T, the *elementary operator* is a simple extended row or column operator from sequences of which more complicated extended operators can be built. The elementary operator T comes in three kinds.[18]

- The first is the *interchange elementary*

  $$T_{[i\leftrightarrow j]} = I - (E_{ii} + E_{jj}) + (E_{ij} + E_{ji}), \tag{11.36}$$

  which by operating $T_{[i\leftrightarrow j]}A$ or $AT_{[i\leftrightarrow j]}$ respectively interchanges A's ith row or column with its jth.[19]

[18] In § 11.3, the symbol A specifically represented an extended operator, but here and generally the symbol represents any matrix.

[19] As a matter of definition, some authors [106] forbid $T_{[i\leftrightarrow i]}$ as an elementary operator, where $j = i$, since after all $T_{[i\leftrightarrow i]} = I$; which is to say that the operator doesn't actually do anything. There exist legitimate tactical reasons to forbid (as in § 11.6), but normally this book permits.

- The second is the *scaling elementary*

  $$T_{\alpha[i]} = I + (\alpha - 1)E_{ii}, \quad \alpha \neq 0, \tag{11.37}$$

  which by operating $T_{\alpha[i]}A$ or $AT_{\alpha[i]}$ scales (multiplies) A's ith row or column, respectively, by the factor α.

- The third and last is the *addition elementary*

  $$T_{\alpha[ij]} = I + \alpha E_{ij}, \quad i \neq j, \tag{11.38}$$

  which by operating $T_{\alpha[ij]}A$ adds to the ith row of A, α times the jth row; or which by operating $AT_{\alpha[ij]}$ adds to the jth column of A, α times the ith column.

Examples of the elementary operators include

$$\begin{aligned}
T_{[1\leftrightarrow 2]} &= \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 0 & 1 & 0 & 0 & 0 & \cdots \\ \cdots & 1 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 1 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 1 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \\
T_{5[4]} &= \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 1 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 1 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 5 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 1 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \\
T_{5[21]} &= \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 1 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 5 & 1 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 1 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 1 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.
\end{aligned}$$

Note that none of these, and in fact no elementary operator of any kind, differs from I in more than four elements.

11.4.1 Properties

Significantly, elementary operators as defined above are always invertible (which is to say, reversible in effect), with

$$\begin{aligned} T^{-1}_{[i\leftrightarrow j]} &= T_{[j\leftrightarrow i]} = T_{[i\leftrightarrow j]}, \\ T^{-1}_{\alpha[i]} &= T_{(1/\alpha)[i]}, \\ T^{-1}_{\alpha[ij]} &= T_{-\alpha[ij]}, \end{aligned} \tag{11.39}$$

being themselves elementary operators such that

$$T^{-1}T = I = TT^{-1} \tag{11.40}$$

in each case.[20] This means that any sequence of elementaries $\prod_k T_k$ can safely be undone by the reverse sequence $\coprod_k T_k^{-1}$:

$$\coprod_k T_k^{-1} \prod_k T_k = I = \prod_k T_k \coprod_k T_k^{-1}. \tag{11.41}$$

The rank-r identity matrix I_r is no elementary operator,[21] nor is the lone-element matrix E_{mn}; but the general identity matrix I is indeed an elementary operator. The last can be considered a distinct, fourth kind of elementary operator if desired; but it is probably easier just to regard it as an elementary of any of the first three kinds, since $I = T_{[i\leftrightarrow i]} = T_{1[i]} = T_{0[ij]}$.

From (11.31), we have that

$$I_rT = I_rTI_r = TI_r \ \ \text{if } 1 \le i \le r \text{ and } 1 \le j \le r \tag{11.42}$$

for any elementary operator T which operates within the given bounds. Equation (11.42) lets an identity matrix with sufficiently high rank pass through a sequence of elementaries as needed.

In general, the transpose of an elementary row operator is the corresponding elementary column operator. Curiously, the interchange elementary is its own transpose and adjoint:

$$T^{*}_{[i\leftrightarrow j]} = T_{[i\leftrightarrow j]} = T^{T}_{[i\leftrightarrow j]}. \tag{11.43}$$

[20] The addition elementary $T_{\alpha[ii]}$ and the scaling elementary $T_{0[i]}$ are forbidden precisely because they are not generally invertible.

[21] If the statement seems to contradict statements of some other books, it is only a matter of definition. This book finds it convenient to define the elementary operator in infinite-dimensional, extended-operational form. The other books are not wrong; their underlying definitions just differ slightly.

11.4.2 Commutation and sorting

Elementary operators often occur in long chains like

$$A = T_{-4[32]}T_{[2\leftrightarrow 3]}T_{(1/5)[3]}T_{(1/2)[31]}T_{5[21]}T_{[1\leftrightarrow 3]},$$

with several elementaries of all kinds intermixed. Some applications demand that the elementaries be sorted and grouped by kind, as

$$A = \left(T_{[2\leftrightarrow 3]}T_{[1\leftrightarrow 3]}\right)\left(T_{-4[21]}T_{(1/0\mathrm{xA})[13]}T_{5[23]}\right)\left(T_{(1/5)[1]}\right)$$

or as

$$A = \left(T_{-4[32]}T_{(1/0\mathrm{xA})[21]}T_{5[31]}\right)\left(T_{(1/5)[2]}\right)\left(T_{[2\leftrightarrow 3]}T_{[1\leftrightarrow 3]}\right),$$

among other possible orderings. Though you probably cannot tell just by looking, the three products above are different orderings of the same elementary chain; they yield the same A and thus represent exactly the same matrix operation. Interesting is that the act of reordering the elementaries has altered some of them into other elementaries of the same kind, but has changed the kind of none of them.

One sorts a chain of elementary operators by repeatedly exchanging adjacent pairs. This of course supposes that one can exchange adjacent pairs, which seems impossible since matrix multiplication is not commutative: $A_1A_2 \neq A_2A_1$. However, at the moment we are dealing in elementary operators only; and for most pairs T_1 and T_2 of elementary operators, though indeed $T_1T_2 \neq T_2T_1$, it so happens that there exists either a T_1' such that $T_1T_2 = T_2T_1'$ or a T_2' such that $T_1T_2 = T_2'T_1$, where T_1' and T_2' are elementaries of the same kinds respectively as T_1 and T_2. The attempt sometimes fails when both T_1 and T_2 are addition elementaries, but all other pairs commute in this way. Significantly, *elementaries of different kinds always commute.* And, though commutation can alter one (never both) of the two elementaries, it changes the kind of neither.

Many qualitatively distinct pairs of elementaries exist; we will list these exhaustively in a moment. First, however, we should like to observe a natural hierarchy among the three kinds of elementary: (i) interchange; (ii) scaling; (iii) addition.

- The interchange elementary is the strongest. Itself subject to alteration only by another interchange elementary, it can alter any elementary by commuting past. When an interchange elementary commutes past another elementary of any kind, what it alters are the other elementary's indices i and/or j (or m and/or n, or whatever symbols happen to represent the indices in question). When two interchange elementaries commute past one another, only one of the two is altered. (Which one? Either. The mathematician chooses.) Refer to Table 11.1.

- Next in strength is the scaling elementary. Only an interchange elementary can alter it, and it in turn can alter only an addition elementary. Scaling elementaries do not alter one another during commutation. When a scaling elementary commutes past an addition elementary, what it alters is the latter's scale α (or β, or whatever symbol happens to represent the scale in question). Refer to Table 11.2.

- The addition elementary, last and weakest, is subject to alteration by either of the other two, itself having no power to alter any elementary during commutation. A pair of addition elementaries are the only pair that can altogether fail to commute—they fail when the row index of one equals the column index of the other—but when they do commute, neither alters the other. Refer to Table 11.3.

Tables 11.1, 11.2 and 11.3 list all possible pairs of elementary operators, as the reader can check. The only pairs that fail to commute are the last three of Table 11.3.

11.5 Inversion and similarity (introduction)

If Tables 11.1, 11.2 and 11.3 exhaustively describe the commutation of one elementary past another elementary, then what can one write of the commutation of an elementary past the general matrix A? With some matrix algebra,

$$\begin{aligned} TA &= (TA)(I) = (TA)(T^{-1}T), \\ AT &= (I)(AT) = (TT^{-1})(AT), \end{aligned}$$

one can write that

$$\begin{aligned} TA &= [TAT^{-1}]T, \\ AT &= T[T^{-1}AT], \end{aligned} \tag{11.44}$$

where T^{-1} is given by (11.39). An elementary commuting rightward changes A to TAT^{-1}; commuting leftward, to $T^{-1}AT$.

Table 11.1: Inverting, commuting, combining and expanding elementary operators: interchange. In the table, $i \neq j \neq m \neq n$; no two indices are the same. Notice that the effect an interchange elementary $T_{[m\leftrightarrow n]}$ has in passing any other elementary, even another interchange elementary, is simply to replace m by n and n by m among the indices of the other elementary.

$$\begin{aligned}
T_{[m\leftrightarrow n]} &= T_{[n\leftrightarrow m]} \\
T_{[m\leftrightarrow m]} &= I \\
IT_{[m\leftrightarrow n]} &= T_{[m\leftrightarrow n]}I \\
T_{[m\leftrightarrow n]}T_{[m\leftrightarrow n]} &= T_{[m\leftrightarrow n]}T_{[n\leftrightarrow m]} = T_{[n\leftrightarrow m]}T_{[m\leftrightarrow n]} = I \\
T_{[m\leftrightarrow n]}T_{[i\leftrightarrow n]} &= T_{[i\leftrightarrow n]}T_{[m\leftrightarrow i]} = T_{[i\leftrightarrow m]}T_{[m\leftrightarrow n]} \\
&= \left(T_{[i\leftrightarrow n]}T_{[m\leftrightarrow n]}\right)^2 \\
T_{[m\leftrightarrow n]}T_{[i\leftrightarrow j]} &= T_{[i\leftrightarrow j]}T_{[m\leftrightarrow n]} \\
T_{[m\leftrightarrow n]}T_{\alpha[m]} &= T_{\alpha[n]}T_{[m\leftrightarrow n]} \\
T_{[m\leftrightarrow n]}T_{\alpha[i]} &= T_{\alpha[i]}T_{[m\leftrightarrow n]} \\
T_{[m\leftrightarrow n]}T_{\alpha[ij]} &= T_{\alpha[ij]}T_{[m\leftrightarrow n]} \\
T_{[m\leftrightarrow n]}T_{\alpha[in]} &= T_{\alpha[im]}T_{[m\leftrightarrow n]} \\
T_{[m\leftrightarrow n]}T_{\alpha[mj]} &= T_{\alpha[nj]}T_{[m\leftrightarrow n]} \\
T_{[m\leftrightarrow n]}T_{\alpha[mn]} &= T_{\alpha[nm]}T_{[m\leftrightarrow n]}
\end{aligned}$$

Table 11.2: Inverting, commuting, combining and expanding elementary operators: scaling. In the table, $i \neq j \neq m \neq n$; no two indices are the same.

$$\begin{aligned}
T_{1[m]} &= I \\
IT_{\beta[m]} &= T_{\beta[m]}I \\
T_{(1/\beta)[m]}T_{\beta[m]} &= I \\
T_{\beta[m]}T_{\alpha[m]} &= T_{\alpha[m]}T_{\beta[m]} = T_{\alpha\beta[m]} \\
T_{\beta[m]}T_{\alpha[i]} &= T_{\alpha[i]}T_{\beta[m]} \\
T_{\beta[m]}T_{\alpha[ij]} &= T_{\alpha[ij]}T_{\beta[m]} \\
T_{\beta[m]}T_{\alpha\beta[im]} &= T_{\alpha[im]}T_{\beta[m]} \\
T_{\beta[m]}T_{\alpha[mj]} &= T_{\alpha\beta[mj]}T_{\beta[m]}
\end{aligned}$$

Table 11.3: Inverting, commuting, combining and expanding elementary operators: addition. In the table, $i \neq j \neq m \neq n$; no two indices are the same. The last three lines give pairs of addition elementaries that do not commute.

$$\begin{aligned}
T_{0[ij]} &= I \\
IT_{\alpha[ij]} &= T_{\alpha[ij]}I \\
T_{-\alpha[ij]}T_{\alpha[ij]} &= I \\
T_{\beta[ij]}T_{\alpha[ij]} &= T_{\alpha[ij]}T_{\beta[ij]} = T_{(\alpha+\beta)[ij]} \\
T_{\beta[mj]}T_{\alpha[ij]} &= T_{\alpha[ij]}T_{\beta[mj]} \\
T_{\beta[in]}T_{\alpha[ij]} &= T_{\alpha[ij]}T_{\beta[in]} \\
T_{\beta[mn]}T_{\alpha[ij]} &= T_{\alpha[ij]}T_{\beta[mn]} \\
T_{\beta[mi]}T_{\alpha[ij]} &= T_{\alpha[ij]}T_{\alpha\beta[mj]}T_{\beta[mi]} \\
T_{\beta[jn]}T_{\alpha[ij]} &= T_{\alpha[ij]}T_{-\alpha\beta[in]}T_{\beta[jn]} \\
T_{\beta[ji]}T_{\alpha[ij]} &\neq T_{\alpha[ij]}T_{\beta[ji]}
\end{aligned}$$

First encountered in § 11.4, the notation T^{-1} means the *inverse* of the elementary operator T, such that

$$T^{-1}T = I = TT^{-1}.$$

Matrix inversion is not for elementary operators only, though. Many matrices C that are more general also have inverses such that

$$C^{-1}C = I = CC^{-1}. \tag{11.45}$$

(Do all matrices have such inverses? No. For example, the null matrix has no such inverse.) The broad question of how to invert a general matrix C, we leave for chapters 12 and 13 to address. For the moment however we should like to observe three simple rules involving matrix inversion.

First, nothing in the logic leading to (11.44) actually requires the matrix T there to be an elementary operator. Any matrix C for which C^{-1} is known can fill the role. Hence,

$$\begin{aligned} CA &= [CAC^{-1}]C, \\ AC &= C[C^{-1}AC]. \end{aligned} \tag{11.46}$$

The transformation CAC^{-1} or $C^{-1}AC$ is called a *similarity transformation.* Sections 12.2 and 14.9 speak further of this.

Second,

$$\begin{aligned} \left(C^T\right)^{-1} &= C^{-T} = \left(C^{-1}\right)^T, \\ \left(C^*\right)^{-1} &= C^{-*} = \left(C^{-1}\right)^*, \end{aligned} \tag{11.47}$$

where C^{-*} is condensed notation for conjugate transposition and inversion in either order and C^{-T} is of like style. Equation (11.47) is a consequence of (11.14), since for conjugate transposition

$$\left(C^{-1}\right)^* C^* = \left[CC^{-1}\right]^* = [I]^* = I = [I]^* = \left[C^{-1}C\right]^* = C^* \left(C^{-1}\right)^*$$

and similarly for nonconjugate transposition.

Third,

$$\left(\prod_k C_k\right)^{-1} = \coprod_k C_k^{-1}. \tag{11.48}$$

This rule emerges upon repeated application of (11.45), which yields that

$$\coprod_k C_k^{-1} \prod_k C_k = I = \prod_k C_k \coprod_k C_k^{-1}.$$

Table 11.4: Matrix inversion properties. (The similarity properties work equally for $C^{-1(r)}$ as for C^{-1} if A honors an $r \times r$ active region. The full notation $C^{-1(r)}$ for the rank-r inverse incidentally is not standard, usually is not needed, and normally is not used.)

$$\begin{array}{rcccl} C^{-1}C &=& I &=& CC^{-1} \\ C^{-1(r)}C &=& I_r &=& CC^{-1(r)} \\ \left(C^T\right)^{-1} &=& C^{-T} &=& \left(C^{-1}\right)^T \\ \left(C^*\right)^{-1} &=& C^{-*} &=& \left(C^{-1}\right)^* \end{array}$$

$$\begin{array}{rcl} CA &=& [CAC^{-1}]C \\ AC &=& C[C^{-1}AC] \\ \left(\prod_k C_k\right)^{-1} &=& \coprod_k C_k^{-1} \end{array}$$

A more limited form of the inverse exists than the infinite-dimensional form of (11.45). This is the rank-r inverse, a matrix $C^{-1(r)}$ such that

$$C^{-1(r)}C = I_r = CC^{-1(r)}. \tag{11.49}$$

The full notation $C^{-1(r)}$ is not standard and usually is not needed, since the context usually implies the rank. When so, one can abbreviate the notation to C^{-1}. In either notation, (11.47) and (11.48) apply equally for the rank-r inverse as for the infinite-dimensional inverse. Because of (11.31), eqn. (11.46) too applies for the rank-r inverse if A's active region is limited to $r \times r$. (Section 13.2 uses the rank-r inverse to solve an exactly determined linear system. This is a famous way to use the inverse, with which many or most readers will already be familiar; but before using it so in chapter 13, we shall first learn how to compute it reliably in chapter 12.)

Table 11.4 summarizes.

11.6 Parity

Consider the sequence of integers or other objects $1, 2, 3, \ldots, n$. By successively interchanging pairs of the objects (any pairs, not just adjacent pairs), one can achieve any desired permutation (§ 4.2.1). For example, beginning with $1, 2, 3, 4, 5$, one can achieve the permutation $3, 5, 1, 4, 2$ by interchanging first the 1 and 3, then the 2 and 5.

Now contemplate all possible pairs:

$$\begin{array}{ccccc}
(1,2) & (1,3) & (1,4) & \cdots & (1,n); \\
 & (2,3) & (2,4) & \cdots & (2,n); \\
 & & (3,4) & \cdots & (3,n); \\
 & & & \ddots & \vdots \\
 & & & & (n-1,n).
\end{array}$$

In a given permutation (like $3, 5, 1, 4, 2$), some pairs will appear in proper sequence with respect to one another, while others will appear in improper sequence. (In $3, 5, 1, 4, 2$, the pair $[1, 2]$ appears in proper sequence in that the larger 2 stands to the right of the smaller 1; but the pair $[1, 3]$ appears in improper sequence in that the larger 3 stands to the *left* of the smaller 1.) If p is the number of pairs which appear in improper sequence (in the example, $p = 6$), and if p is even, then we say that the permutation has *even* or *positive parity;* if odd, then *odd* or *negative parity.*[22]

Now consider: every interchange of adjacent elements must either increment or decrement p by one, reversing parity. Why? Well, think about it. If two elements are adjacent and their order is correct, then interchanging falsifies the order, but only of that pair (no other element interposes, so the interchange affects the ordering of no other pair). Complementarily, if the order is incorrect, then interchanging rectifies the order. Either way, an adjacent interchange alters p by exactly ± 1, thus reversing parity.

What about nonadjacent elements? Does interchanging a pair of these reverse parity, too? To answer the question, let u and v represent the two elements interchanged, with $a_1, a_2, \ldots, a_m$ the elements lying between. Before the interchange:

$$\ldots, u, a_1, a_2, \ldots, a_{m-1}, a_m, v, \ldots$$

After the interchange:

$$\ldots, v, a_1, a_2, \ldots, a_{m-1}, a_m, u, \ldots$$

The interchange reverses with respect to one another just the pairs

$$\begin{array}{llllll}
(u, a_1) & (u, a_2) & \cdots & (u, a_{m-1}) & (u, a_m) \\
(a_1, v) & (a_2, v) & \cdots & (a_{m-1}, v) & (a_m, v) \\
(u, v) & & & &
\end{array}$$

[22] For readers who learned arithmetic in another language than English, the *even* integers are $\ldots, -4, -2, 0, 2, 4, 6, \ldots$; the *odd* integers are $\ldots, -3, -1, 1, 3, 5, 7, \ldots$.

The number of pairs reversed is odd. Since each reversal alters p by ± 1, the net change in p apparently also is odd, reversing parity. It seems that regardless of how distant the pair, *interchanging any pair of elements reverses the permutation's parity.*

The sole exception arises when an element is interchanged with itself. This does not change parity, but it does not change anything else, either, so in parity calculations we ignore it.[23] All other interchanges reverse parity.

We discuss parity in this, a chapter on matrices, because parity concerns the elementary interchange operator of § 11.4. The rows or columns of a matrix can be considered elements in a sequence. If so, then the interchange operator $T_{[i \leftrightarrow j]}$, $i \neq j$, acts precisely in the manner described, interchanging rows or columns and thus reversing parity. It follows that if $i_k \neq j_k$ and q is odd, then $\prod_{k=1}^{q} T_{[i_k \leftrightarrow j_k]} \neq I$. However, it is possible that $\prod_{k=1}^{q} T_{[i_k \leftrightarrow j_k]} = I$ if q is even. In any event, even q implies even p, which means even (positive) parity; odd q implies odd p, which means odd (negative) parity.

We shall have more to say about parity in §§ 11.7.1 and 14.1.

11.7 The quasielementary operator

Multiplying sequences of the elementary operators of § 11.4, one can form much more complicated operators, which per (11.41) are always invertible. Such complicated operators are not trivial to analyze, however, so one finds it convenient to define an intermediate class of operators, called in this book the *quasielementary operators,* more complicated than elementary operators but less so than arbitrary matrices.

A quasielementary operator is composed of elementaries only of a single kind. There are thus three kinds of quasielementary—interchange, scaling and addition—to match the three kinds of elementary. With respect to interchange and scaling, any sequences of elementaries of the respective kinds are allowed. With respect to addition, there are some extra rules, explained in § 11.7.3.

The three subsections which follow respectively introduce the three kinds of quasielementary operator.

[23] This is why some authors forbid self-interchanges, as explained in footnote 19.

11.7.1 The interchange quasielementary or general interchange operator

Any product P of zero or more interchange elementaries,

$$P = \prod_k T_{[i_k \leftrightarrow j_k]}, \tag{11.50}$$

constitutes an *interchange quasielementary, permutation matrix*, *permutor* or *general interchange operator*.[24] An example is

$$P = T_{[2\leftrightarrow5]}T_{[1\leftrightarrow3]} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 1 & \cdots \\ \cdots & 1 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 1 & 0 & \cdots \\ \cdots & 0 & 1 & 0 & 0 & 0 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

This operator resembles I in that it has a single one in each row and in each column, but the ones here do not necessarily run along the main diagonal. The effect of the operator is to shuffle the rows or columns of the matrix it operates on, without altering any of the rows or columns it shuffles.

By (11.41), (11.39), (11.43) and (11.15), the inverse of the general interchange operator is

$$\begin{aligned} P^{-1} &= \left(\prod_k T_{[i_k\leftrightarrow j_k]}\right)^{-1} = \coprod_k T^{-1}_{[i_k\leftrightarrow j_k]} \\ &= \coprod_k T_{[i_k\leftrightarrow j_k]} \\ &= \coprod_k T^{*}_{[i_k\leftrightarrow j_k]} = \left(\prod_k T_{[i_k\leftrightarrow j_k]}\right)^{*} \\ &= P^{*} = P^{T} \end{aligned} \tag{11.51}$$

(where $P^* = P^T$ because P has only real elements). The inverse, transpose and adjoint of the general interchange operator are thus the same:

$$P^TP = P^*P = I = PP^* = PP^T. \tag{11.52}$$

[24] The letter P here recalls the verb "to permute."

A significant attribute of the general interchange operator P is its parity: positive or even parity if the number of interchange elementaries $T_{[i_k \leftrightarrow j_k]}$ which compose it is even; negative or odd parity if the number is odd. This works precisely as described in § 11.6. For the purpose of parity determination, only interchange elementaries $T_{[i_k \leftrightarrow j_k]}$ for which $i_k \neq j_k$ are counted; any $T_{[i \leftrightarrow i]} = I$ noninterchanges are ignored. Thus the example's P above has even parity (two interchanges), as does I itself (zero interchanges), but $T_{[i \leftrightarrow j]}$ alone (one interchange) has odd parity if $i \neq j$. As we shall see in § 14.1, the positive (even) and negative (odd) parities sometimes lend actual positive and negative senses to the matrices they describe. The parity of the general interchange operator P concerns us for this reason.

Parity, incidentally, is a property of the matrix P itself, not just of the operation P represents. No interchange quasielementary P has positive parity as a row operator but negative as a column operator. The reason is that, regardless of whether one ultimately means to use P as a row or column operator, the matrix is nonetheless composable as a definite sequence of interchange elementaries. It is the number of interchanges, not the use, which determines P's parity.

11.7.2 The scaling quasielementary or general scaling operator

Like the interchange quasielementary P of § 11.7.1, the *scaling quasielementary, diagonal matrix* or *general scaling operator* D consists of a product of zero or more elementary operators, in this case elementary scaling operators:[25]

$$D = \prod_{i=-\infty}^{\infty} T_{\alpha_i[i]} = \coprod_{i=-\infty}^{\infty} T_{\alpha_i[i]} = \sum_{i=-\infty}^{\infty} \alpha_i E_{ii} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & * & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & * & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & * & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & * & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & * & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{11.53}$$

(of course it might be that $\alpha_i = 1$, and thus that $T_{\alpha_i[i]} = I$, for some, most or even all i; however, $\alpha_i = 0$ is forbidden by the definition of the scaling

[25] The letter D here recalls the adjective "diagonal."

elementary). An example is

$$D = T_{-5[4]}T_{4[2]}T_{7[1]} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 7 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 4 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & -5 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 1 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

This operator resembles I in that all its entries run down the main diagonal; but these entries, though never zeros, are not necessarily ones, either. They are nonzero scaling factors. The effect of the operator is to scale the rows or columns of the matrix it operates on.

The general scaling operator is a particularly simple matrix. Its inverse is evidently

$$D^{-1} = \coprod_{i=-\infty}^{\infty} T_{(1/\alpha_i)[i]} = \prod_{i=-\infty}^{\infty} T_{(1/\alpha_i)[i]} = \sum_{i=-\infty}^{\infty} \frac{E_{ii}}{\alpha_i}, \tag{11.54}$$

where each element down the main diagonal is individually inverted.

A superset of the general scaling operator is the *diagonal matrix,* defined less restrictively that $[A]_{ij} = 0$ for $i \neq j$, where zeros along the main diagonal are allowed. The conventional notation

$$\begin{aligned} [\mathrm{diag}\{\mathbf{x}\}]_{ij} &\equiv \delta_{ij}x_i = \delta_{ij}x_j, \qquad (11.55) \\ \mathrm{diag}\{\mathbf{x}\} &= \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & x_1 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & x_2 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & x_3 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & x_4 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & x_5 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \end{aligned}$$

converts a vector $\mathbf{x}$ into a diagonal matrix. The diagonal matrix in general is not invertible and is no quasielementary operator, but is sometimes useful nevertheless.

11.7.3 Addition quasielementaries

Any product of interchange elementaries (§ 11.7.1), any product of scaling elementaries (§ 11.7.2), qualifies as a quasielementary operator. Not so, any product of addition elementaries. To qualify as a quasielementary, a product of elementary addition operators must meet some additional restrictions.

Four types of addition quasielementary are defined:[26]

- the *downward multitarget row addition operator,*[27]

$$\begin{aligned} L_{[j]} &= \prod_{i=j+1}^{\infty} T_{\alpha_{ij}[ij]} = \coprod_{i=j+1}^{\infty} T_{\alpha_{ij}[ij]} \qquad (11.56) \\ &= I + \sum_{i=j+1}^{\infty} \alpha_{ij} E_{ij} \\ &= \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 1 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 1 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & * & 1 & 0 & \cdots \\ \cdots & 0 & 0 & * & 0 & 1 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \end{aligned}$$

whose inverse is

$$\begin{aligned} L_{[j]}^{-1} &= \coprod_{i=j+1}^{\infty} T_{-\alpha_{ij}[ij]} = \prod_{i=j+1}^{\infty} T_{-\alpha_{ij}[ij]} \qquad (11.57) \\ &= I - \sum_{i=j+1}^{\infty} \alpha_{ij} E_{ij} = 2I - L_{[j]}; \end{aligned}$$

[26] In this subsection the explanations are briefer than in the last two, but the pattern is similar. The reader can fill in the details.

[27] The letter L here recalls the adjective "lower."

- the *upward multitarget row addition operator,*[28]

$$
\begin{aligned}
U_{[j]} &= \coprod_{i=-\infty}^{j-1} T_{\alpha_{ij}[ij]} = \prod_{i=-\infty}^{j-1} T_{\alpha_{ij}[ij]} \qquad (11.58)\\
&= I + \sum_{i=-\infty}^{j-1} \alpha_{ij} E_{ij}\\
&= \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 1 & 0 & * & 0 & 0 & \cdots \\
\cdots & 0 & 1 & * & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 1 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 1 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix},
\end{aligned}
$$

whose inverse is

$$
\begin{aligned}
U_{[j]}^{-1} &= \prod_{i=-\infty}^{j-1} T_{-\alpha_{ij}[ij]} = \coprod_{i=-\infty}^{j-1} T_{-\alpha_{ij}[ij]} \qquad (11.59)\\
&= I - \sum_{i=-\infty}^{j-1} \alpha_{ij} E_{ij} = 2I - U_{[j]};
\end{aligned}
$$

- the *rightward multitarget column addition operator,* which is the transpose $L_{[j]}^T$ of the downward operator; and

- the *leftward multitarget column addition operator,* which is the transpose $U_{[j]}^T$ of the upward operator.

11.8 The unit triangular matrix

Yet more complicated than the quasielementary of § 11.7 is the *unit triangular matrix,* with which we draw this necessary but tedious chapter toward

[28] The letter U here recalls the adjective "upper."

a long close:

$$
\begin{aligned}
L &= I + \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{i-1} \alpha_{ij} E_{ij} = I + \sum_{j=-\infty}^{\infty} \sum_{i=j+1}^{\infty} \alpha_{ij} E_{ij} && (11.60) \\
&= \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 1 & 0 & 0 & 0 & 0 & \cdots \\
\cdots & * & 1 & 0 & 0 & 0 & \cdots \\
\cdots & * & * & 1 & 0 & 0 & \cdots \\
\cdots & * & * & * & 1 & 0 & \cdots \\
\cdots & * & * & * & * & 1 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}; \\
U &= I + \sum_{i=-\infty}^{\infty} \sum_{j=i+1}^{\infty} \alpha_{ij} E_{ij} = I + \sum_{j=-\infty}^{\infty} \sum_{i=-\infty}^{j-1} \alpha_{ij} E_{ij} && (11.61) \\
&= \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 1 & * & * & * & * & \cdots \\
\cdots & 0 & 1 & * & * & * & \cdots \\
\cdots & 0 & 0 & 1 & * & * & \cdots \\
\cdots & 0 & 0 & 0 & 1 & * & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 1 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}.
\end{aligned}
$$

The former is a *unit lower triangular matrix;* the latter, a *unit upper triangular matrix.* The unit triangular matrix is a generalized addition quasielementary, which adds not only to multiple targets but also from multiple sources—but in one direction only: downward or leftward for L or U^T (or U^*); upward or rightward for U or L^T (or L^*).

The general *triangular matrix* L_S or U_S, which by definition can have any values along its main diagonal, is sometimes of interest, as in the Schur decomposition of § 14.10.[29] The *strictly triangular matrix* $L - I$ or $U - I$ is likewise sometimes of interest, as in Table 11.5.[30] However, such matrices cannot in general be expressed as products of elementary operators and this section does not treat them.

This section presents and derives the basic properties of the unit triangular matrix.

[29] The subscript S here stands for Schur. Other books typically use the symbols L and U for the general triangular matrix of Schur, but this book distinguishes by the subscript.

[30] [182, "Schur decomposition," 00:32, 30 Aug. 2007]

11.8.1 Construction

To make a unit triangular matrix is straightforward:

$$\begin{aligned} L &= \coprod_{j=-\infty}^{\infty} L_{[j]}; \\ U &= \prod_{j=-\infty}^{\infty} U_{[j]}. \end{aligned} \tag{11.62}$$

So long as the multiplication is done in the order indicated,[31] then conveniently,

$$\begin{aligned} \Big[L\Big]_{ij} &= \Big[L_{[j]}\Big]_{ij}, \\ \Big[U\Big]_{ij} &= \Big[U_{[j]}\Big]_{ij}, \end{aligned} \tag{11.63}$$

which is to say that the entries of L and U are respectively nothing more than the relevant entries of the several $L_{[j]}$ and $U_{[j]}$. Equation (11.63) enables one to use (11.62) immediately and directly, without calculation, to build any unit triangular matrix desired.

The correctness of (11.63) is most easily seen if the several $L_{[j]}$ and $U_{[j]}$ are regarded as column operators acting sequentially on I:

$$\begin{aligned} L &= (I)\left(\coprod_{j=-\infty}^{\infty} L_{[j]}\right); \\ U &= (I)\left(\prod_{j=-\infty}^{\infty} U_{[j]}\right). \end{aligned}$$

The reader can construct an inductive proof symbolically on this basis without too much difficulty if desired, but just thinking about how $L_{[j]}$ adds columns leftward and $U_{[j]}$, rightward, then considering the order in which the several $L_{[j]}$ and $U_{[j]}$ act, (11.63) follows at once.

[31] Recall again from § 2.3 that $\prod_k A_k = \cdots A_3 A_2 A_1$, whereas $\coprod_k A_k = A_1 A_2 A_3 \cdots$. This means that $(\prod_k A_k)(C)$ applies first A_1, then A_2, A_3 and so on, as row operators to C; whereas $(C)(\coprod_k A_k)$ applies first A_1, then A_2, A_3 and so on, as column operators to C. The symbols $\prod$ and $\coprod$ as this book uses them thus can be thought of respectively as row and column sequencers.

11.8.2 The product of like unit triangular matrices

The product of like unit triangular matrices,

$$\begin{aligned} L_1 L_2 &= L, \\ U_1 U_2 &= U, \end{aligned} \tag{11.64}$$

is another unit triangular matrix of the same type. The proof for unit lower and unit upper triangular matrices is the same. In the unit lower triangular case, one starts from a form of the definition of a unit lower triangular matrix:

$$[L_1]_{ij} \text{ or } [L_2]_{ij} = \begin{cases} 0 & \text{if } i < j, \\ 1 & \text{if } i = j. \end{cases}$$

Then,

$$[L_1 L_2]_{ij} = \sum_{m=-\infty}^{\infty} [L_1]_{im} [L_2]_{mj}.$$

But as we have just observed, $[L_1]_{im}$ is null when $i < m$, and $[L_2]_{mj}$ is null when $m < j$. Therefore,

$$[L_1 L_2]_{ij} = \begin{cases} 0 & \text{if } i < j, \\ \sum_{m=j}^{i} [L_1]_{im} [L_2]_{mj} & \text{if } i \geq j. \end{cases}$$

Inasmuch as this is true, nothing prevents us from weakening the statement to read

$$[L_1 L_2]_{ij} = \begin{cases} 0 & \text{if } i < j, \\ \sum_{m=j}^{i} [L_1]_{im} [L_2]_{mj} & \text{if } i = j. \end{cases}$$

But this is just

$$[L_1 L_2]_{ij} = \begin{cases} 0 & \text{if } i < j, \\ [L_1]_{ij} [L_2]_{ij} = [L_1]_{ii} [L_2]_{ii} = (1)(1) = 1 & \text{if } i = j, \end{cases}$$

which again is the very definition of a unit lower triangular matrix. Hence (11.64).

11.8.3 Inversion

Inasmuch as any unit triangular matrix can be constructed from addition quasielementaries by (11.62), inasmuch as (11.63) supplies the specific quasielementaries, and inasmuch as (11.57) or (11.59) gives the inverse of each such quasielementary, one can always invert a unit triangular matrix easily by

$$\begin{aligned} L^{-1} &= \prod_{j=-\infty}^{\infty} L_{[j]}^{-1}, \\ U^{-1} &= \coprod_{j=-\infty}^{\infty} U_{[j]}^{-1}. \end{aligned} \tag{11.65}$$

In view of (11.64), therefore, *the inverse of a unit lower triangular matrix is another unit lower triangular matrix; and the inverse of a unit upper triangular matrix, another unit upper triangular matrix.*

It is plain to see but still interesting to note that—unlike the inverse—the adjoint or transpose of a unit lower triangular matrix is a unit upper triangular matrix; and that the adjoint or transpose of a unit upper triangular matrix is a unit lower triangular matrix. The adjoint reverses the sense of the triangle.

11.8.4 The parallel unit triangular matrix

If a unit triangular matrix fits the special, restricted form

$$\begin{aligned} L_{\parallel}^{\{k\}} &= I + \sum_{j=-\infty}^{k} \sum_{i=k+1}^{\infty} \alpha_{ij} E_{ij} \qquad (11.66) \\ &= \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 1 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 1 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\ \cdots & * & * & * & 1 & 0 & \cdots \\ \cdots & * & * & * & 0 & 1 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{aligned}$$

or

$$
\begin{aligned}
U_{\parallel}^{\{k\}} &= I + \sum_{j=k}^{\infty} \sum_{i=-\infty}^{k-1} \alpha_{ij} E_{ij} \qquad (11.67) \\
&= \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 1 & 0 & * & * & * & \cdots \\
\cdots & 0 & 1 & * & * & * & \cdots \\
\cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 1 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 1 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix},
\end{aligned}
$$

confining its nonzero elements to a rectangle within the triangle as shown, then it is a *parallel unit triangular matrix* and has some special properties the general unit triangular matrix lacks.

The general unit lower triangular matrix L acting LA on a matrix A adds the rows of A downward. The parallel unit lower triangular matrix $L_{\parallel}^{\{k\}}$ acting $L_{\parallel}^{\{k\}}A$ also adds rows downward, but with the useful restriction that it makes no row of A both source and target. The addition is *from* A's rows through the kth, *to* A's $(k+1)$th row onward. A horizontal frontier separates source from target, which thus march in A as separate squads.

Similar observations naturally apply with respect to the parallel unit upper triangular matrix $U_{\parallel}^{\{k\}}$, which acting $U_{\parallel}^{\{k\}}A$ adds rows upward, and also with respect to $L_{\parallel}^{\{k\}T}$ and $U_{\parallel}^{\{k\}T}$, which acting $AL_{\parallel}^{\{k\}T}$ and $AU_{\parallel}^{\{k\}T}$ add columns respectively rightward and leftward (remembering that $L_{\parallel}^{\{k\}T}$ is no unit lower but a unit upper triangular matrix; that $U_{\parallel}^{\{k\}T}$ is the lower). Each separates source from target in the matrix A it operates on.

The reason we care about the separation of source from target is that, in matrix arithmetic generally, where source and target are not separate but remain intermixed, the sequence matters in which rows or columns are added. That is, in general,

$$T_{\alpha_1[i_1 j_1]} T_{\alpha_2[i_2 j_2]} \neq I + \alpha_1 E_{i_1 j_1} + \alpha_2 E_{i_2 j_2} \neq T_{\alpha_2[i_2 j_2]} T_{\alpha_1[i_1 j_1]}.$$

It makes a difference whether the one addition comes before, during or after the other—but only because the target of the one addition might be the source of the other. The danger is that $i_1 = j_2$ or $i_2 = j_1$. Remove this danger, and the sequence ceases to matter (refer to Table 11.3).

That is exactly what the parallel unit triangular matrix does: it separates source from target and thus removes the danger. It is for this reason that the parallel unit triangular matrix brings the useful property that

$$\begin{aligned}
L_{\parallel}^{\{k\}} &= I + \sum_{j=-\infty}^{k} \sum_{i=k+1}^{\infty} \alpha_{ij} E_{ij} \\
&= \coprod_{j=-\infty}^{k} \prod_{i=k+1}^{\infty} T_{\alpha_{ij}[ij]} = \coprod_{j=-\infty}^{k} \coprod_{i=k+1}^{\infty} T_{\alpha_{ij}[ij]} \\
&= \prod_{j=-\infty}^{k} \prod_{i=k+1}^{\infty} T_{\alpha_{ij}[ij]} = \prod_{j=-\infty}^{k} \coprod_{i=k+1}^{\infty} T_{\alpha_{ij}[ij]} \\
&= \prod_{i=k+1}^{\infty} \coprod_{j=-\infty}^{k} T_{\alpha_{ij}[ij]} = \coprod_{i=k+1}^{\infty} \coprod_{j=-\infty}^{k} T_{\alpha_{ij}[ij]} \qquad (11.68) \\
&= \prod_{i=k+1}^{\infty} \prod_{j=-\infty}^{k} T_{\alpha_{ij}[ij]} = \coprod_{i=k+1}^{\infty} \prod_{j=-\infty}^{k} T_{\alpha_{ij}[ij]}, \\
U_{\parallel}^{\{k\}} &= I + \sum_{j=k}^{\infty} \sum_{i=-\infty}^{k-1} \alpha_{ij} E_{ij} \\
&= \prod_{j=k}^{\infty} \coprod_{i=-\infty}^{k-1} T_{\alpha_{ij}[ij]} = \cdots ,
\end{aligned}$$

which says that one can build a parallel unit triangular matrix equally well in any sequence—in contrast to the case of the general unit triangular matrix, whose construction per (11.62) one must sequence carefully. (Though eqn. 11.68 does not show them, even more sequences are possible. You can scramble the factors' ordering any random way you like. The multiplication is fully commutative.) Under such conditions, the inverse of the parallel unit triangular matrix is particularly simple:[32]

$$
\begin{aligned}
L_{\parallel}^{\{k\}\,-1} &= I - \sum_{j=-\infty}^{k} \sum_{i=k+1}^{\infty} \alpha_{ij} E_{ij} = 2I - L_{\parallel}^{\{k\}} \\
&= \prod_{j=-\infty}^{k} \coprod_{i=k+1}^{\infty} T_{-\alpha_{ij}[ij]} = \cdots, \\
U_{\parallel}^{\{k\}\,-1} &= I - \sum_{j=k}^{\infty} \sum_{i=-\infty}^{k-1} \alpha_{ij} E_{ij} = 2I - U_{\parallel}^{\{k\}} \\
&= \coprod_{j=k}^{\infty} \prod_{i=-\infty}^{k-1} T_{-\alpha_{ij}[ij]} = \cdots,
\end{aligned}
\tag{11.69}
$$

where again the elementaries can be multiplied in any order. Pictorially,

$$
L_{\parallel}^{\{k\}\,-1} = \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 1 & 0 & 0 & 0 & 0 & \cdots \\
\cdots & 0 & 1 & 0 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\
\cdots & -* & -* & -* & 1 & 0 & \cdots \\
\cdots & -* & -* & -* & 0 & 1 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix},
$$

$$
U_{\parallel}^{\{k\}\,-1} = \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 1 & 0 & -* & -* & -* & \cdots \\
\cdots & 0 & 1 & -* & -* & -* & \cdots \\
\cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 1 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 1 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}.
$$

The inverse of a parallel unit triangular matrix is just the matrix itself, only with each element off the main diagonal negated. Table 11.5 records a few properties that come immediately of the last observation and from the parallel unit triangular matrix's basic layout.

[32] There is some odd parochiality at play in applied mathematics when one calls such collections of symbols as (11.69) "particularly simple." Nevertheless, in the present context the idea (11.69) represents is indeed simple: that one can multiply constituent elementaries in any order and still reach the same parallel unit triangular matrix; that the elementaries in this case do not interfere.

Table 11.5: Properties of the parallel unit triangular matrix. (In the table, the notation I_a^b represents the generalized dimension-limited identity matrix or truncator of eqn. 11.30. Note that the inverses $L_\parallel^{\{k\}\,-1} = L_\parallel^{\{k\}\prime}$ and $U_\parallel^{\{k\}\,-1} = U_\parallel^{\{k\}\prime}$ are parallel unit triangular matrices themselves, such that the table's properties hold for them, too.)

$$\frac{L_\parallel^{\{k\}} + L_\parallel^{\{k\}\,-1}}{2} = I = \frac{U_\parallel^{\{k\}} + U_\parallel^{\{k\}\,-1}}{2}$$

$$\begin{aligned}
I_{k+1}^{\infty} L_\parallel^{\{k\}} I_{-\infty}^{k} &= L_\parallel^{\{k\}} - I &&= I_{k+1}^{\infty}(L_\parallel^{\{k\}} - I)I_{-\infty}^{k} \\
I_{-\infty}^{k-1} U_\parallel^{\{k\}} I_{k}^{\infty} &= U_\parallel^{\{k\}} - I &&= I_{-\infty}^{k-1}(U_\parallel^{\{k\}} - I)I_{k}^{\infty}
\end{aligned}$$

If $L_\parallel^{\{k\}}$ honors an $n \times n$ active region, then

$(I_n - I_k)L_\parallel^{\{k\}} I_k = L_\parallel^{\{k\}} - I = (I_n - I_k)(L_\parallel^{\{k\}} - I)I_k$

and $(I - I_n)(L_\parallel^{\{k\}} - I) = 0 = (L_\parallel^{\{k\}} - I)(I - I_n)$.

If $U_\parallel^{\{k\}}$ honors an $n \times n$ active region, then

$I_{k-1}U_\parallel^{\{k\}}(I_n - I_{k-1}) = U_\parallel^{\{k\}} - I = I_{k-1}(U_\parallel^{\{k\}} - I)(I_n - I_{k-1})$

and $(I - I_n)(U_\parallel^{\{k\}} - I) = 0 = (U_\parallel^{\{k\}} - I)(I - I_n)$.

11.8.5 The partial unit triangular matrix

Besides the notation L and U for the general unit lower and unit upper triangular matrices and the notation $L_{\parallel}^{\{k\}}$ and $U_{\parallel}^{\{k\}}$ for the parallel unit lower and unit upper triangular matrices, we shall find it useful to introduce the additional notation

$$\begin{aligned}
L^{[k]} &= I + \sum_{j=k}^{\infty} \sum_{i=j+1}^{\infty} \alpha_{ij} E_{ij} && (11.70)\\
&= \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 1 & 0 & 0 & 0 & 0 & \cdots \\
\cdots & 0 & 1 & 0 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & * & 1 & 0 & \cdots \\
\cdots & 0 & 0 & * & * & 1 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}, \\
U^{[k]} &= I + \sum_{j=-\infty}^{k} \sum_{i=-\infty}^{j-1} \alpha_{ij} E_{ij} && (11.71)\\
&= \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 1 & * & * & 0 & 0 & \cdots \\
\cdots & 0 & 1 & * & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 1 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 1 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}
\end{aligned}$$

for unit triangular matrices whose off-diagonal content is confined to a narrow wedge and

$$\begin{aligned}
L^{\{k\}} &= I + \sum_{j=-\infty}^{k} \sum_{i=j+1}^{\infty} \alpha_{ij} E_{ij} \qquad (11.72) \\
&= \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 1 & 0 & 0 & 0 & 0 & \cdots \\
\cdots & * & 1 & 0 & 0 & 0 & \cdots \\
\cdots & * & * & 1 & 0 & 0 & \cdots \\
\cdots & * & * & * & 1 & 0 & \cdots \\
\cdots & * & * & * & 0 & 1 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}, \\
U^{\{k\}} &= I + \sum_{j=k}^{\infty} \sum_{i=-\infty}^{j-1} \alpha_{ij} E_{ij} \qquad (11.73) \\
&= \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 1 & 0 & * & * & * & \cdots \\
\cdots & 0 & 1 & * & * & * & \cdots \\
\cdots & 0 & 0 & 1 & * & * & \cdots \\
\cdots & 0 & 0 & 0 & 1 & * & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 1 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}
\end{aligned}$$

for the supplementary forms.[33] Such notation is not standard in the literature, but it serves a purpose in this book and is introduced here for this reason. If names are needed for $L^{[k]}$, $U^{[k]}$, $L^{\{k\}}$ and $U^{\{k\}}$, the former pair can be called *minor partial unit triangular matrices,* and the latter pair, *major partial unit triangular matrices.* Whether minor or major, the partial unit triangular matrix is a matrix which leftward or rightward of the kth column resembles I. Of course partial unit triangular matrices which resemble I above or below the kth *row* are equally possible, and can be denoted $L^{[k]T}$, $U^{[k]T}$, $L^{\{k\}T}$ and $U^{\{k\}T}$.

Observe that the parallel unit triangular matrices $L_{\parallel}^{\{k\}}$ and $U_{\parallel}^{\{k\}}$ of § 11.8.4 are in fact also major partial unit triangular matrices, as the notation suggests.

[33] The notation is arguably imperfect in that $L^{\{k\}} + L^{[k]} - I \neq L$ but rather that $L^{\{k\}} + L^{[k+1]} - I = L$. The conventional notation $\sum_{k=a}^{b} f(k) + \sum_{k=b}^{c} f(k) \neq \sum_{k=a}^{c} f(k)$ suffers the same arguable imperfection.

11.9 The shift operator

Not all useful matrices fit the dimension-limited and extended-operational forms of § 11.3. An exception is the *shift operator* H_k, defined that

$$[H_k]_{ij} = \delta_{i(j+k)}. \tag{11.74}$$

For example,

$$H_2 = \begin{bmatrix} & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 1 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 1 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix}.$$

Operating H_kA, H_k shifts A's rows downward k steps. Operating AH_k, H_k shifts A's columns leftward k steps. Inasmuch as the shift operator shifts all rows or columns of the matrix it operates on, its active region is $\infty \times \infty$ in extent. Obviously, the shift operator's inverse, transpose and adjoint are the same:

$$\begin{aligned} H_k^T H_k = H_k^* H_k = I = H_k H_k^* = H_k H_k^T, \\ H_k^{-1} = H_k^T = H_k^* = H_{-k}. \end{aligned} \tag{11.75}$$

Further obvious but useful identities include that

$$\begin{aligned} (I_\ell - I_k)H_k = H_k I_{\ell-k}, \\ H_{-k}(I_\ell - I_k) = I_{\ell-k} H_{-k}. \end{aligned} \tag{11.76}$$

11.10 The Jacobian derivative

Chapter 4 has introduced the derivative of a function with respect to a scalar variable. One can also take the derivative of a function with respect to a vector variable, and the function itself can be vector-valued. The derivative is

$$\left[\frac{d\mathbf{f}}{d\mathbf{x}}\right]_{ij} = \frac{\partial f_i}{\partial x_j}. \tag{11.77}$$

For instance, if $\mathbf{x}$ has three elements and $\mathbf{f}$ has two, then

$$\frac{d\mathbf{f}}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \frac{\partial f_1}{\partial x_3} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_2}{\partial x_3} \end{bmatrix}.$$

This is called the *Jacobian derivative,* the *Jacobian matrix,* or just the *Jacobian.*[34] Each of its columns is the derivative with respect to one element of $\mathbf{x}$.

The Jacobian derivative of a vector with respect to itself is

$$\frac{d\mathbf{x}}{d\mathbf{x}} = I. \tag{11.78}$$

The derivative is not I_n as one might think, because, even if $\mathbf{x}$ has only n elements, still, one could vary x_{n+1} in principle, and $\partial x_{n+1}/\partial x_{n+1} \neq 0$.

The Jacobian derivative obeys the derivative product rule (4.22) in the form[35]

$$\begin{aligned}
\frac{d}{d\mathbf{x}}(\mathbf{g}^T A\mathbf{f}) &= \left[\mathbf{g}^T A\left(\frac{d\mathbf{f}}{d\mathbf{x}}\right)\right] + \left[\left(\frac{d\mathbf{g}}{d\mathbf{x}}\right)^T A\mathbf{f}\right]^T, \\
\frac{d}{d\mathbf{x}}(\mathbf{g}^* A\mathbf{f}) &= \left[\mathbf{g}^* A\left(\frac{d\mathbf{f}}{d\mathbf{x}}\right)\right] + \left[\left(\frac{d\mathbf{g}}{d\mathbf{x}}\right)^* A\mathbf{f}\right]^T,
\end{aligned} \tag{11.79}$$

valid for any constant matrix A—as is seen by applying the definition (4.13) of the derivative, which here is

$$\frac{\partial\,(\mathbf{g}^* A\mathbf{f})}{\partial x_j} = \lim_{\partial x_j \to 0} \frac{(\mathbf{g} + \partial\mathbf{g}/2)^* A(\mathbf{f} + \partial\mathbf{f}/2) - (\mathbf{g} - \partial\mathbf{g}/2)^* A(\mathbf{f} - \partial\mathbf{f}/2)}{\partial x_j},$$

and simplifying.

The shift operator of § 11.9 and the Jacobian derivative of this section complete the family of matrix rudiments we shall need to begin to do increasingly interesting things with matrices in chapters 13 and 14. Before doing interesting things, however, we must treat two more foundational matrix matters. The two are the Gauss-Jordan decomposition and the matter of matrix rank, which will be the subjects of chapter 12, next.

[34] [182, "Jacobian," 00:50, 15 Sept. 2007]

[35] Notice that the last term on (11.79)'s second line is transposed, not adjointed.

Chapter 12

Matrix rank and the Gauss-Jordan decomposition

Chapter 11 has brought the matrix and its rudiments, the latter including

- lone-element matrix E (§ 11.3.7),
- the null matrix 0 (§ 11.3.1),
- the rank-r identity matrix I_r (§ 11.3.5),
- the general identity matrix I and the scalar matrix λI (§ 11.3.2),
- the elementary operator T (§ 11.4),
- the quasielementary operator P, D, $L_{[k]}$ or $U_{[k]}$ (§ 11.7), and
- the unit triangular matrix L or U (§ 11.8).

Such rudimentary forms have useful properties, as we have seen. The general matrix A does not necessarily have any of these properties, but it turns out that one can factor any matrix whatsoever into a product of rudiments which do have the properties, and that several orderly procedures are known to do so. The simplest of these, and indeed one of the more useful, is the Gauss-Jordan decomposition. This chapter introduces it.

Section 11.3 has deëmphasized the concept of matrix dimensionality $m \times n$, supplying in its place the new concept of matrix rank. However, that section has actually defined rank only for the rank-r identity matrix I_r. In fact all matrices have rank. This chapter explains.

Before treating the Gauss-Jordan decomposition and the matter of matrix rank as such, however, we shall find it helpful to prepare two preliminaries thereto: (i) the matter of the linear independence of vectors; and (ii) the elementary similarity transformation. The chapter begins with these.

Except in § 12.2, the chapter demands more rigor than one likes in such a book as this. However, it is hard to see how to avoid the rigor here, and logically the chapter cannot be omitted. We will drive through the chapter in as few pages as can be managed, and then onward to the more interesting matrix topics of chapters 13 and 14.

12.1 Linear independence

Linear independence is a significant possible property of a set of vectors—whether the set be the several columns of a matrix, the several rows, or some other vectors—the property being defined as follows. A vector is *linearly independent* if its role cannot be served by the other vectors in the set. More formally, the n vectors of the set $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5, \ldots, \mathbf{a}_n\}$ are linearly independent if and only if none of them can be expressed as a linear combination—a weighted sum—of the others. That is, the several $\mathbf{a}_k$ are linearly independent iff

$$\alpha_1 \mathbf{a}_1 + \alpha_2 \mathbf{a}_2 + \alpha_3 \mathbf{a}_3 + \cdots + \alpha_n \mathbf{a}_n \neq 0 \tag{12.1}$$

for all nontrivial α_k, where "nontrivial α_k" means the several α_k, at least one of which is nonzero (*trivial* α_k, by contrast, would be $\alpha_1 = \alpha_2 = \alpha_3 = \cdots = \alpha_n = 0$). Vectors which can combine nontrivially to reach the null vector are by definition *linearly dependent.*

Linear independence is a property of vectors. Technically the property applies to scalars, too, inasmuch as a scalar resembles a one-element vector—so, any nonzero scalar alone is linearly independent—but there is no such thing as a linearly independent pair of scalars, because one of the pair can always be expressed as a complex multiple of the other. Significantly but less obviously, there is also no such thing as a linearly independent set which includes the null vector; (12.1) forbids it. Paradoxically, even the single-member, $n = 1$ set consisting only of $\mathbf{a}_1 = 0$ is, strictly speaking, not linearly independent.

For consistency of definition, we regard the empty, $n = 0$ set as linearly independent, on the technical ground that the only possible linear combination of the empty set is trivial.[1]

[1] This is the kind of thinking which typically governs mathematical edge cases. One

If a linear combination of several independent vectors $\mathbf{a}_k$ forms a vector $\mathbf{b}$, then one might ask: can there exist a different linear combination of the same vectors $\mathbf{a}_k$ which also forms $\mathbf{b}$? That is, if

$$\beta_1\mathbf{a}_1 + \beta_2\mathbf{a}_2 + \beta_3\mathbf{a}_3 + \cdots + \beta_n\mathbf{a}_n = \mathbf{b},$$

where the several $\mathbf{a}_k$ satisfy (12.1), then is

$$\beta_1'\mathbf{a}_1 + \beta_2'\mathbf{a}_2 + \beta_3'\mathbf{a}_3 + \cdots + \beta_n'\mathbf{a}_n = \mathbf{b}$$

possible? To answer the question, suppose that it were possible. The difference of the two equations then would be

$$(\beta_1' - \beta_1)\mathbf{a}_1 + (\beta_2' - \beta_2)\mathbf{a}_2 + (\beta_3' - \beta_3)\mathbf{a}_3 + \cdots + (\beta_n' - \beta_n)\mathbf{a}_n = 0.$$

According to (12.1), this could only be so if the coefficients in the last equation where trivial—that is, only if $\beta_1' - \beta_1 = 0$, $\beta_2' - \beta_2 = 0$, $\beta_3' - \beta_3 = 0$, $\ldots$, $\beta_n' - \beta_n = 0$. But this says no less than that the two linear combinations, which we had supposed to differ, were in fact one and the same. One concludes therefore that, *if a vector* $\mathbf{b}$ *can be expressed as a linear combination of several linearly independent vectors* $\mathbf{a}_k$*, then it cannot be expressed as any other combination of the same vectors.* The combination is unique.

Linear independence can apply in any dimensionality, but it helps to visualize the concept geometrically in three dimensions, using the three-dimensional geometrical vectors of § 3.3. Two such vectors are independent so long as they do not lie along the same line. A third such vector is independent of the first two so long as it does not lie in their common plane. A fourth such vector (unless it points off into some unvisualizable fourth dimension) cannot possibly then be independent of the three.

We discuss the linear independence of vectors in this, a chapter on matrices, because (§ 11.1) a matrix is essentially a sequence of vectors—either of column vectors or of row vectors, depending on one's point of view. As we shall see in § 12.5, the important property of matrix *rank* depends on the number of linearly independent columns or rows a matrix has.

could define the empty set to be linearly dependent if one really wanted to, but what then of the observation that adding a vector to a linearly dependent set never renders the set independent? Surely in this light it is preferable just to define the empty set as independent in the first place. Similar thinking makes $0! = 1$, $\sum_{k=0}^{-1} a_k z^k = 0$, and 2 not 1 the least prime, among other examples.

12.2 The elementary similarity transformation

Section 11.5 and its (11.46) have introduced the *similarity transformation* CAC^{-1} or $C^{-1}AC$, which arises when an operator C commutes respectively rightward or leftward past a matrix A. The similarity transformation has several interesting properties, some of which we are now prepared to discuss, particularly in the case in which the operator happens to be an elementary, $C = T$. In this case, the several rules of Table 12.1 obtain.

Most of the table's rules are fairly obvious if the meaning of the symbols is understood, though to grasp some of the rules it helps to sketch the relevant matrices on a sheet of paper. Of course rigorous symbolic proofs can be constructed after the pattern of § 11.8.2, but they reveal little or nothing sketching the matrices does not. In Table 12.1 as elsewhere, the symbols P, D, L and U represent the quasielementaries and unit triangular matrices of §§ 11.7 and 11.8. The symbols P', D', L' and U' also represent quasielementaries and unit triangular matrices, only not necessarily the same ones P, D, L and U do.

The rules of Table 12.1 permit one to commute some but not all elementaries past a quasielementary operator or unit triangular matrix without fundamentally altering the character of the quasielementary operator or unit triangular matrix, and sometimes without changing it at all. The rules find use among other places in the Gauss-Jordan decomposition of § 12.3.

12.3 The Gauss-Jordan decomposition

The *Gauss-Jordan decomposition* of an arbitrary, dimension-limited, $m \times n$ matrix A is[2]

$$\begin{aligned} A = G_> I_r G_< &= PDLUI_rKS, \\ G_< &\equiv KS, \\ G_> &\equiv PDLU, \end{aligned} \tag{12.2}$$

where

- P and S are general interchange operators (§ 11.7.1);

[2] Most introductory linear algebra texts this writer has met call the Gauss-Jordan decomposition instead the "LU decomposition" and include fewer factors in it, typically merging D into L and omitting K and S. They also omit I_r, since their matrices have pre-defined dimensionality. Perhaps the reader will agree that the decomposition is cleaner as presented here.

Table 12.1: Some elementary similarity transformations.

$$\begin{array}{rcll}
T_{[i\leftrightarrow j]} I T_{[i\leftrightarrow j]} &=& I & \\
T_{[i\leftrightarrow j]} P T_{[i\leftrightarrow j]} &=& P' & \\
T_{[i\leftrightarrow j]} D T_{[i\leftrightarrow j]} &=& \multicolumn{2}{l}{D' = D + ([D]_{jj} - [D]_{ii})\,E_{ii} + ([D]_{ii} - [D]_{jj})\,E_{jj}} \\
T_{[i\leftrightarrow j]} D T_{[i\leftrightarrow j]} &=& D & \text{if } [D]_{ii} = [D]_{jj} \\
T_{[i\leftrightarrow j]} L^{[k]} T_{[i\leftrightarrow j]} &=& L^{[k]} & \text{if } i < k \text{ and } j < k \\
T_{[i\leftrightarrow j]} U^{[k]} T_{[i\leftrightarrow j]} &=& U^{[k]} & \text{if } i > k \text{ and } j > k \\
T_{[i\leftrightarrow j]} L^{\{k\}} T_{[i\leftrightarrow j]} &=& L^{\{k\}\prime} & \text{if } i > k \text{ and } j > k \\
T_{[i\leftrightarrow j]} U^{\{k\}} T_{[i\leftrightarrow j]} &=& U^{\{k\}\prime} & \text{if } i < k \text{ and } j < k \\
T_{[i\leftrightarrow j]} L^{\{k\}}_{\|} T_{[i\leftrightarrow j]} &=& L^{\{k\}\prime}_{\|} & \text{if } i > k \text{ and } j > k \\
T_{[i\leftrightarrow j]} U^{\{k\}}_{\|} T_{[i\leftrightarrow j]} &=& U^{\{k\}\prime}_{\|} & \text{if } i < k \text{ and } j < k \\
T_{\alpha[i]} I T_{(1/\alpha)[i]} &=& I & \\
T_{\alpha[i]} D T_{(1/\alpha)[i]} &=& D & \\
T_{\alpha[i]} A T_{(1/\alpha)[i]} &=& A' & \text{where } A \text{ is any of} \\
 & & & L, U, L_{[k]}, U_{[k]}, L^{[k]}, U^{[k]}, L^{\{k\}}, U^{\{k\}}, L^{\{k\}}_{\|}, U^{\{k\}}_{\|} \\
T_{\alpha[ij]} I T_{-\alpha[ij]} &=& I & \\
T_{\alpha[ij]} D T_{-\alpha[ij]} &=& \multicolumn{2}{l}{D + ([D]_{jj} - [D]_{ii})\,\alpha E_{ij} \neq D'} \\
T_{\alpha[ij]} D T_{-\alpha[ij]} &=& D & \text{if } [D]_{ii} = [D]_{jj} \\
T_{\alpha[ij]} L T_{-\alpha[ij]} &=& L' & \text{if } i > j \\
T_{\alpha[ij]} U T_{-\alpha[ij]} &=& U' & \text{if } i < j
\end{array}$$

- D is a general scaling operator (§ 11.7.2);
- L and U are respectively unit lower and unit upper triangular matrices (§ 11.8);
- $K = L_{\parallel}^{\{r\}T}$ is the transpose of a parallel unit lower triangular matrix, being thus a parallel unit upper triangular matrix (§ 11.8.4);
- $G_>$ and $G_<$ are composites[3] as defined by (12.2); and
- r is an unspecified rank.

The Gauss-Jordan decomposition is also called the *Gauss-Jordan factorization.*

Whether all possible dimension-limited, $m \times n$ matrices A have a Gauss-Jordan decomposition (they do, in fact) is a matter this section addresses. However—at least for matrices which do have one—because $G_>$ and $G_<$ are composed of invertible factors, one can left-multiply the equation $A = G_> I_r G_<$ by $G_>^{-1}$ and right-multiply it by $G_<^{-1}$ to obtain

$$\begin{aligned} U^{-1}L^{-1}D^{-1}P^{-1}AS^{-1}K^{-1} &= G_>^{-1}AG_<^{-1} = I_r, \\ S^{-1}K^{-1} &= G_<^{-1}, \\ U^{-1}L^{-1}D^{-1}P^{-1} &= G_>^{-1}, \end{aligned} \tag{12.3}$$

the Gauss-Jordan's complementary form.

12.3.1 Motive

Equation (12.2) seems inscrutable. The equation itself is easy enough to read, but just as there are many ways to factor a scalar (0xC = $[4][3] = [2]^2[3] = [2][6]$, for example), there are likewise many ways to factor a matrix. Why choose this particular way?

There are indeed many ways. We shall meet some of the others in §§ 13.11, 14.6, 14.10 and 14.12. The Gauss-Jordan decomposition we meet here however has both significant theoretical properties and useful practical applications, and in any case needs less advanced preparation to appreciate than the others, and (at least as developed in this book) precedes the others logically. It emerges naturally when one posits a pair of dimension-limited,

[3] One can pronounce $G_>$ and $G_<$ respectively as "G acting rightward" and "G acting leftward." The letter G itself can be regarded as standing for "Gauss-Jordan," but admittedly it is chosen as much because otherwise we were running out of available Roman capitals!

square, $n \times n$ matrices, A and A^{-1}, for which $A^{-1}A = I_n$, where A is known and A^{-1} is to be determined. [The A^{-1} here is the $A^{-1(n)}$ of eqn. 11.49. However, it is only supposed here that $A^{-1}A = I_n$; it is not *yet* claimed that $AA^{-1} = I_n$. "Square" means that the matrix has an $n \times n$ active region rather than an $m \times n$, $m \neq n$, where "active region" is defined as in § 11.3.1.]

To determine A^{-1} is not an entirely trivial problem. The matrix A^{-1} such that $A^{-1}A = I_n$ may or may not exist (usually it does exist if A is square, but even then it may not, as we shall soon see), and even if it does exist, how to determine it is not immediately obvious. And still, if one can determine A^{-1}, that is only for square A; what if A, having an $m \times n$, $m \neq n$, active region, were not square? In the present subsection however we are not trying to prove anything, only to motivate, so for the moment let us suppose an A for which A^{-1} does exist, let us confine our attention to square A, and let us seek A^{-1} by left-multiplying A by a sequence $\prod T$ of elementary row operators, each of which makes the matrix more nearly resemble I_n. When I_n is finally achieved, then we shall have that

$$\left(\prod T\right)(A) = I_n,$$

or, left-multiplying by I_n and observing that $I_n^2 = I_n$,

$$(I_n)\left(\prod T\right)(A) = I_n,$$

which implies that

$$A^{-1} = (I_n)\left(\prod T\right).$$

The product of elementaries which transforms A to I_n, truncated (§ 11.3.6) to $n \times n$ dimensionality, itself constitutes A^{-1}. This observation is what motivates the Gauss-Jordan decomposition.

By successive steps,[4] a concrete example:

$$\begin{aligned}
A &= \begin{bmatrix} 2 & -4 \\ 3 & -1 \end{bmatrix}, \\
\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} A &= \begin{bmatrix} 1 & -2 \\ 3 & -1 \end{bmatrix}, \\
\begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} A &= \begin{bmatrix} 1 & -2 \\ 0 & 5 \end{bmatrix}, \\
\begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{5} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} A &= \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}, \\
\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{5} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} A &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \\
\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{5} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} A &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.
\end{aligned}$$

Hence,[5]

$$A^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{5} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{\mathrm{A}} & \frac{2}{5} \\ -\frac{3}{\mathrm{A}} & \frac{1}{5} \end{bmatrix}.$$

Using the elementary commutation identity that $T_{\beta[m]} T_{\alpha[mj]} = T_{\alpha\beta[mj]} T_{\beta[m]}$, from Table 11.2, to group like operators, we have that

$$A^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{3}{5} & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{5} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{\mathrm{A}} & \frac{2}{5} \\ -\frac{3}{\mathrm{A}} & \frac{1}{5} \end{bmatrix};$$

or, multiplying the two scaling elementaries to merge them into a single general scaling operator (§ 11.7.2),

$$A^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{3}{5} & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{5} \end{bmatrix} = \begin{bmatrix} -\frac{1}{\mathrm{A}} & \frac{2}{5} \\ -\frac{3}{\mathrm{A}} & \frac{1}{5} \end{bmatrix}.$$

The last equation is written symbolically as

$$A^{-1} = I_2 U^{-1} L^{-1} D^{-1},$$

from which

$$A = DLUI_2 = \begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{3}{5} & 1 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -4 \\ 3 & -1 \end{bmatrix}.$$

[4] Theoretically, all elementary operators including the ones here have extended-operational form (§ 11.3.2), but all those $\cdots$ ellipses clutter the page too much. Only the 2×2 active regions are shown here.

[5] The figures $\frac{1}{\mathrm{A}}$ and $\frac{3}{\mathrm{A}}$ here are respectively one tenth and three tenths. The hexadecimal numeral A is not to be confused with the italic-printed matrix A! The usual hexadecimal prefix 0x is here omitted because it is too bulky to crowd into a cell in a matrix.

Now, admittedly, the equation $A = DLUI_2$ is not (12.2)—or rather, it is (12.2), but only in the special case that $r = 2$ and $P = S = K = I$—which begs the question: why do we need the factors P, S and K in the first place? The answer regarding P and S is that these factors respectively gather row and column interchange elementaries, of which the example given has used none but which other examples sometimes need or want, particularly to avoid dividing by zero when they encounter a zero in an inconvenient cell of the matrix (the reader might try reducing $A = [0\ 1; 1\ 0]$ to I_2, for instance; a row or column interchange is needed here). Regarding K, this factor comes into play when A has broad rectangular ($m < n$) rather than square ($m = n$) shape, and also sometimes when one of the rows of A happens to be a linear combination of the others. The last point, we are not quite ready to detail yet, but at present we are only motivating not proving, so if the reader will accept the other factors and suspend judgment on K until the actual need for it emerges in § 12.3.3, step 12, then we will proceed on this basis.

12.3.2 Method

The Gauss-Jordan decomposition of a matrix A is not discovered at one stroke but rather is gradually built up, elementary by elementary. It begins with the equation

$$A = IIIIAII,$$

where the six I hold the places of the six Gauss-Jordan factors P, D, L, U, K and S of (12.2). By successive elementary operations, the A on the right is gradually transformed into I_r, while the six I are gradually transformed into the six Gauss-Jordan factors. The decomposition thus ends with the equation

$$A = PDLUI_rKS,$$

which is (12.2). In between, while the several matrices are gradually being transformed, the equation is represented as

$$A = \tilde{P}\tilde{D}\tilde{L}\tilde{U}\tilde{I}\tilde{K}\tilde{S}, \tag{12.4}$$

where the initial value of $\tilde{I}$ is A and the initial values of $\tilde{P}$, $\tilde{D}$, etc., are all I.

Each step of the transformation goes as follows. The matrix $\tilde{I}$ is left- or right-multiplied by an elementary operator T. To compensate, one of the six factors is right- or left-multiplied by T^{-1}. Intervening factors are multiplied by both T and T^{-1}, which multiplication constitutes an elementary similarity transformation as described in § 12.2. For example,

$$A = \tilde{P}\left(\tilde{D}T_{(1/\alpha)[i]}\right)\left(T_{\alpha[i]}\tilde{L}T_{(1/\alpha)[i]}\right)\left(T_{\alpha[i]}\tilde{U}T_{(1/\alpha)[i]}\right)\left(T_{\alpha[i]}\tilde{I}\right)\tilde{K}\tilde{S},$$

which is just (12.4), inasmuch as the adjacent elementaries cancel one another; then,

$$\begin{aligned}
\tilde{I} &\leftarrow T_{\alpha[i]}\tilde{I},\\
\tilde{U} &\leftarrow T_{\alpha[i]}\tilde{U}T_{(1/\alpha)[i]},\\
\tilde{L} &\leftarrow T_{\alpha[i]}\tilde{L}T_{(1/\alpha)[i]},\\
\tilde{D} &\leftarrow \tilde{D}T_{(1/\alpha)[i]},
\end{aligned}$$

thus associating the operation with the appropriate factor—in this case, $\tilde{D}$. Such elementary row and column operations are repeated until $\tilde{I} = I_r$, at which point (12.4) has become the Gauss-Jordan decomposition (12.2).

12.3.3 The algorithm

Having motivated the Gauss-Jordan decomposition in § 12.3.1 and having proposed a basic method to pursue it in § 12.3.2, we shall now establish a definite, orderly, failproof algorithm to achieve it. Broadly, the algorithm

- copies A, a dimension-limited, $m \times n$ matrix (not necessarily square), into the variable working matrix $\tilde{I}$ (step 1 below),
- reduces $\tilde{I}$ by suitable row (and maybe column) operations to unit upper triangular form (steps 2 through 7),
- establishes a rank r (step 8), and
- reduces the now unit triangular $\tilde{I}$ further to the rank-r identity matrix I_r (steps 9 through 13).

Specifically, the algorithm decrees the following steps. (The steps as written include many parenthetical remarks—so many that some steps seem to consist more of parenthetical remarks than of actual algorithm. The remarks are unnecessary to execute the algorithm's steps as such. They are however necessary to explain and to justify the algorithm's steps to the reader.)

1. Begin by initializing

$$\begin{gathered}
\tilde{P} \leftarrow I,\ \tilde{D} \leftarrow I,\ \tilde{L} \leftarrow I,\ \tilde{U} \leftarrow I,\ \tilde{K} \leftarrow I,\ \tilde{S} \leftarrow I,\\
\tilde{I} \leftarrow A,\\
i \leftarrow 1,
\end{gathered}$$

where $\tilde{I}$ holds the part of A remaining to be decomposed, where i is a row index, and where the others are the variable working matrices of (12.4). (The eventual goal will be to factor all of $\tilde{I}$ away, leaving $\tilde{I} = I_r$, though the precise value of r will not be known until step 8. Since A is by definition a dimension-limited $m \times n$ matrix, one naturally need not store A beyond the $m \times n$ active region. What is less clear until one has read the whole algorithm, but nevertheless true, is that one also need not store the dimension-limited $\tilde{I}$ beyond the $m \times n$ active region. The other six variable working matrices each have extended-operational form, but they also confine their activity to well-defined regions: $m \times m$ for $\tilde{P}$, $\tilde{D}$, $\tilde{L}$ and $\tilde{U}$; $n \times n$ for $\tilde{K}$ and $\tilde{S}$. One need store none of the matrices beyond these bounds.)

2. (Besides arriving at this point from step 1 above, the algorithm also reënters here from step 7 below. From step 1, $\tilde{I} = A$ and $\tilde{L} = I$, so this step 2 though logical seems unneeded. The need grows clear once one has read through step 7.) Observe that neither the ith row of $\tilde{I}$ nor any row below it has an entry left of the ith column, that $\tilde{I}$ is all-zero below-leftward of and directly leftward of (though not directly below) the *pivot* element $\tilde{\imath}_{ii}$.[6] Observe also that above the ith row, the matrix has proper unit upper triangular form (§ 11.8). Regarding the other factors, notice that $\tilde{L}$ enjoys the major partial unit triangular

[6] The notation $\tilde{\imath}_{ii}$ looks interesting, but this is accidental. The $\tilde{\imath}$ relates not to the doubled, subscribed index ii but to $\tilde{I}$. The notation $\tilde{\imath}_{ii}$ thus means $[\tilde{I}]_{ii}$—in other words, it means the current iith element of the variable working matrix $\tilde{I}$.

form $L^{\{i-1\}}$ (§ 11.8.5) and that $\tilde{d}_{kk} = 1$ for all $k \geq i$. Pictorially,

$$
\begin{aligned}
\tilde{D} &= \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & * & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
\cdots & 0 & * & 0 & 0 & 0 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & * & 0 & 0 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 1 & 0 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}, \\
\tilde{L} = L^{\{i-1\}} &= \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
\cdots & * & 1 & 0 & 0 & 0 & 0 & 0 & \cdots \\
\cdots & * & * & 1 & 0 & 0 & 0 & 0 & \cdots \\
\cdots & * & * & * & 1 & 0 & 0 & 0 & \cdots \\
\cdots & * & * & * & 0 & 1 & 0 & 0 & \cdots \\
\cdots & * & * & * & 0 & 0 & 1 & 0 & \cdots \\
\cdots & * & * & * & 0 & 0 & 0 & 1 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}, \\
\tilde{I} &= \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 1 & * & * & * & * & * & * & \cdots \\
\cdots & 0 & 1 & * & * & * & * & * & \cdots \\
\cdots & 0 & 0 & 1 & * & * & * & * & \cdots \\
\cdots & 0 & 0 & 0 & * & * & * & * & \cdots \\
\cdots & 0 & 0 & 0 & * & * & * & * & \cdots \\
\cdots & 0 & 0 & 0 & * & * & * & * & \cdots \\
\cdots & 0 & 0 & 0 & * & * & * & * & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix},
\end{aligned}
$$

where the ith row and ith column are depicted at center.

3. Choose a nonzero element $\tilde{\imath}_{pq} \neq 0$ on or below the pivot row, where $p \geq i$ and $q \geq i$. (The easiest choice may simply be $\tilde{\imath}_{ii}$, where $p = q = i$, if $\tilde{\imath}_{ii} \neq 0$; but any nonzero element from the ith row downward can in general be chosen. Beginning students of the Gauss-Jordan or LU decomposition are conventionally taught to choose first the least possible q and then the least possible p. When one has no reason to choose otherwise, that is as good a choice as any. There is however no actual need to choose so. In fact alternate choices can sometimes improve

practical numerical accuracy.[7,8] Theoretically nonetheless, when doing exact arithmetic, the choice is quite arbitrary, so long as $\tilde{\imath}_{pq} \neq 0$.) If no nonzero element is available—if all remaining rows $p \geq i$ are now null—then skip directly to step 8.

4. Observing that (12.4) can be expanded to read

$$\begin{aligned} A &= \left(\tilde{P}T_{[p\leftrightarrow i]}\right)\left(T_{[p\leftrightarrow i]}\tilde{D}T_{[p\leftrightarrow i]}\right)\left(T_{[p\leftrightarrow i]}\tilde{L}T_{[p\leftrightarrow i]}\right)\left(T_{[p\leftrightarrow i]}\tilde{U}T_{[p\leftrightarrow i]}\right) \\ &\qquad \times \left(T_{[p\leftrightarrow i]}\tilde{I}T_{[i\leftrightarrow q]}\right)\left(T_{[i\leftrightarrow q]}\tilde{K}T_{[i\leftrightarrow q]}\right)\left(T_{[i\leftrightarrow q]}\tilde{S}\right) \\ &= \left(\tilde{P}T_{[p\leftrightarrow i]}\right)\tilde{D}\left(T_{[p\leftrightarrow i]}\tilde{L}T_{[p\leftrightarrow i]}\right)\tilde{U} \\ &\qquad \times \left(T_{[p\leftrightarrow i]}\tilde{I}T_{[i\leftrightarrow q]}\right)\tilde{K}\left(T_{[i\leftrightarrow q]}\tilde{S}\right), \end{aligned}$$

let

$$\begin{aligned} \tilde{P} &\leftarrow \tilde{P}T_{[p\leftrightarrow i]}, \\ \tilde{L} &\leftarrow T_{[p\leftrightarrow i]}\tilde{L}T_{[p\leftrightarrow i]}, \\ \tilde{I} &\leftarrow T_{[p\leftrightarrow i]}\tilde{I}T_{[i\leftrightarrow q]}, \\ \tilde{S} &\leftarrow T_{[i\leftrightarrow q]}\tilde{S}, \end{aligned}$$

thus interchanging the pth with the ith row and the qth with the ith column, to bring the chosen element to the pivot position. (Re-

[7] A typical Intel or AMD x86-class computer processor represents a C/C++ `double`-type floating-point number, $x = 2^p b$, in 0x40 bits of computer memory. Of the 0x40 bits, 0x34 are for the number's mantissa $2.0 \leq b < 4.0$ (not $1.0 \leq b < 2.0$ as one might expect), 0xB are for the number's exponent $-0\mathrm{x}3\mathrm{FF} \leq p \leq 0\mathrm{x}3\mathrm{FE}$, and one is for the number's ± sign. (The mantissa's high-order bit, which is always 1, is implied not stored, being thus one neither of the 0x34 nor of the 0x40 bits.) The out-of-bounds exponents $p = -0\mathrm{x}400$ and $p = 0\mathrm{x}3\mathrm{FF}$ serve specially respectively to encode 0 and ∞. All this is standard computing practice. Such a floating-point representation is easily accurate enough for most practical purposes, but of course it is not generally exact. [85, § 1-4.2.2]

[8] The Gauss-Jordan's floating-point errors come mainly from dividing by small pivots. Such errors are naturally avoided by avoiding small pivots, at least until as late in the algorithm as possible. Smallness however is relative: a small pivot in a row and a column each populated by even smaller elements is unlikely to cause as much error as is a large pivot in a row and a column each populated by even larger elements.

To choose a pivot, any of several heuristics is reasonable. The following heuristic if programmed intelligently might not be too computationally expensive: define the pivot-smallness metric

$$\tilde{\eta}^2_{pq} \equiv \frac{2\tilde{\imath}^*_{pq}\tilde{\imath}_{pq}}{\sum_{p'=i}^{m}\tilde{\imath}^*_{p'q}\tilde{\imath}_{p'q} + \sum_{q'=i}^{n}\tilde{\imath}^*_{pq'}\tilde{\imath}_{pq'}}.$$

Choose the p and q of least $\tilde{\eta}^2_{pq}$. If two are equally least, then choose first the lesser column index q and then if necessary the lesser row index p.

fer to Table 12.1 for the similarity transformations. The $\tilde{U}$ and $\tilde{K}$ transformations disappear because at this stage of the algorithm, still $\tilde{U} = \tilde{K} = I$. The $\tilde{D}$ transformation disappears because $p \geq i$ and because $\tilde{d}_{kk} = 1$ for all $k \geq i$. Regarding the $\tilde{L}$ transformation, it does not disappear, but $\tilde{L}$ has major partial unit triangular form $L^{\{i-1\}}$, which form according to Table 12.1 it retains since $i - 1 < i \leq p$.)

5. Observing that (12.4) can be expanded to read

$$\begin{aligned} A &= \tilde{P}\left(\tilde{D}T_{\tilde{\imath}_{ii}[i]}\right)\left(T_{(1/\tilde{\imath}_{ii})[i]}\tilde{L}T_{\tilde{\imath}_{ii}[i]}\right)\left(T_{(1/\tilde{\imath}_{ii})[i]}\tilde{U}T_{\tilde{\imath}_{ii}[i]}\right) \\ &\qquad \times \left(T_{(1/\tilde{\imath}_{ii})[i]}\tilde{I}\right)\tilde{K}\tilde{S} \\ &= \tilde{P}\left(\tilde{D}T_{\tilde{\imath}_{ii}[i]}\right)\left(T_{(1/\tilde{\imath}_{ii})[i]}\tilde{L}T_{\tilde{\imath}_{ii}[i]}\right)\tilde{U}\left(T_{(1/\tilde{\imath}_{ii})[i]}\tilde{I}\right)\tilde{K}\tilde{S}, \end{aligned}$$

normalize the new $\tilde{\imath}_{ii}$ pivot by letting

$$\begin{aligned} \tilde{D} &\leftarrow \tilde{D}T_{\tilde{\imath}_{ii}[i]}, \\ \tilde{L} &\leftarrow T_{(1/\tilde{\imath}_{ii})[i]}\tilde{L}T_{\tilde{\imath}_{ii}[i]}, \\ \tilde{I} &\leftarrow T_{(1/\tilde{\imath}_{ii})[i]}\tilde{I}. \end{aligned}$$

This forces $\tilde{\imath}_{ii} = 1$. It also changes the value of $\tilde{d}_{ii}$. Pictorially after this step,

$$\tilde{D} = \left[\begin{array}{ccccccccc} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & * & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & * & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & * & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & * & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right],$$

$$\tilde{I} = \left[\begin{array}{ccccccccc} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 1 & * & * & * & * & * & * & \cdots \\ \cdots & 0 & 1 & * & * & * & * & * & \cdots \\ \cdots & 0 & 0 & 1 & * & * & * & * & \cdots \\ \cdots & 0 & 0 & 0 & 1 & * & * & * & \cdots \\ \cdots & 0 & 0 & 0 & * & * & * & * & \cdots \\ \cdots & 0 & 0 & 0 & * & * & * & * & \cdots \\ \cdots & 0 & 0 & 0 & * & * & * & * & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right].$$

(Though the step changes $\tilde{L}$, too, again it leaves $\tilde{L}$ in the major partial unit triangular form $L^{\{i-1\}}$, because $i-1<i$. Refer to Table 12.1.)

6. Observing that (12.4) can be expanded to read

$$\begin{aligned} A &= \tilde{P}\tilde{D}\left(\tilde{L}T_{\tilde{\imath}_{pi}[pi]}\right)\left(T_{-\tilde{\imath}_{pi}[pi]}\tilde{U}T_{\tilde{\imath}_{pi}[pi]}\right)\left(T_{-\tilde{\imath}_{pi}[pi]}\tilde{I}\right)\tilde{K}\tilde{S} \\ &= \tilde{P}\tilde{D}\left(\tilde{L}T_{\tilde{\imath}_{pi}[pi]}\right)\tilde{U}\left(T_{-\tilde{\imath}_{pi}[pi]}\tilde{I}\right)\tilde{K}\tilde{S}, \end{aligned}$$

clear $\tilde{I}$'s ith column below the pivot by letting

$$\tilde{L} \leftarrow \left(\tilde{L}\right)\left(\coprod_{p=i+1}^{m} T_{\tilde{\imath}_{pi}[pi]}\right),$$

$$\tilde{I} \leftarrow \left(\prod_{p=i+1}^{m} T_{-\tilde{\imath}_{pi}[pi]}\right)\left(\tilde{I}\right).$$

This forces $\tilde{\imath}_{ip}=0$ for all $p>i$. It also fills in $\tilde{L}$'s ith column below the pivot, advancing that matrix from the $L^{\{i-1\}}$ form to the $L^{\{i\}}$ form. Pictorially,

$$\tilde{L} = L^{\{i\}} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & * & 1 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & * & * & 1 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & * & * & * & 1 & 0 & 0 & 0 & \cdots \\ \cdots & * & * & * & * & 1 & 0 & 0 & \cdots \\ \cdots & * & * & * & * & 0 & 1 & 0 & \cdots \\ \cdots & * & * & * & * & 0 & 0 & 1 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

$$\tilde{I} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 1 & * & * & * & * & * & * & \cdots \\ \cdots & 0 & 1 & * & * & * & * & * & \cdots \\ \cdots & 0 & 0 & 1 & * & * & * & * & \cdots \\ \cdots & 0 & 0 & 0 & 1 & * & * & * & \cdots \\ \cdots & 0 & 0 & 0 & 0 & * & * & * & \cdots \\ \cdots & 0 & 0 & 0 & 0 & * & * & * & \cdots \\ \cdots & 0 & 0 & 0 & 0 & * & * & * & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

(Note that it is not necessary actually to apply the addition elementaries here one by one. Together they easily form an addition quasielementary $L_{[i]}$, and thus can be applied all at once. See § 11.7.3.)

7. Increment
$$i \leftarrow i + 1.$$
Go back to step 2.

8. Decrement
$$i \leftarrow i - 1$$
to undo the last instance of step 7 (even if there never was an instance of step 7), thus letting i point to the matrix's last nonzero row. After decrementing, let the rank
$$r \equiv i.$$
Notice that, certainly, $r \le m$ and $r \le n$.

9. (Besides arriving at this point from step 8 above, the algorithm also reënters here from step 11 below.) If $i = 0$, then skip directly to step 12.

10. Observing that (12.4) can be expanded to read
$$A = \tilde{P}\tilde{D}\tilde{L}\left(\tilde{U}T_{\tilde{\imath}_{pi}[pi]}\right)\left(T_{-\tilde{\imath}_{pi}[pi]}\tilde{I}\right)\tilde{K}\tilde{S},$$
clear $\tilde{I}$'s ith column above the pivot by letting
$$\begin{aligned}\tilde{U} &\leftarrow \left(\tilde{U}\right)\left(\prod_{p=1}^{i-1} T_{\tilde{\imath}_{pi}[pi]}\right),\\ \tilde{I} &\leftarrow \left(\coprod_{p=1}^{i-1} T_{-\tilde{\imath}_{pi}[pi]}\right)\left(\tilde{I}\right).\end{aligned}$$
This forces $\tilde{\imath}_{ip} = 0$ for all $p \ne i$. It also fills in $\tilde{U}$'s ith column above the pivot, advancing that matrix from the $U^{\{i+1\}}$ form to the $U^{\{i\}}$ form.

Pictorially,

$$
\begin{aligned}
\tilde{U} = U^{\{i\}} &= \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 1 & 0 & 0 & * & * & * & * & \cdots \\
\cdots & 0 & 1 & 0 & * & * & * & * & \cdots \\
\cdots & 0 & 0 & 1 & * & * & * & * & \cdots \\
\cdots & 0 & 0 & 0 & 1 & * & * & * & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 1 & * & * & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 0 & 1 & * & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}, \\
\tilde{I} &= \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 1 & * & * & 0 & 0 & 0 & 0 & \cdots \\
\cdots & 0 & 1 & * & 0 & 0 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 1 & 0 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}.
\end{aligned}
$$

(As in step 6, here again it is not necessary actually to apply the addition elementaries one by one. Together they easily form an addition quasielementary $U_{[i]}$. See § 11.7.3.)

11. Decrement $i \leftarrow i - 1$. Go back to step 9.

12. Notice that $\tilde{I}$ now has the form of a rank-r identity matrix, except with $n - r$ extra columns dressing its right edge (often $r = n$ however; then there are no extra columns). Pictorially,

$$
\tilde{I} = \begin{bmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 1 & 0 & 0 & 0 & * & * & * & \cdots \\
\cdots & 0 & 1 & 0 & 0 & * & * & * & \cdots \\
\cdots & 0 & 0 & 1 & 0 & * & * & * & \cdots \\
\cdots & 0 & 0 & 0 & 1 & * & * & * & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
\cdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}.
$$

Observing that (12.4) can be expanded to read

$$
A = \tilde{P}\tilde{D}\tilde{L}\tilde{U}\left(\tilde{I}T_{-\tilde{\imath}_{pq}[pq]}\right)\left(T_{\tilde{\imath}_{pq}[pq]}\tilde{K}\right)\tilde{S},
$$

use the now conveniently elementarized columns of $\tilde{I}$'s main body to suppress the extra columns on its right edge by

$$\begin{aligned}
\tilde{I} &\leftarrow \left(\tilde{I}\right)\left(\prod_{q=r+1}^{n}\coprod_{p=1}^{r} T_{-\tilde{\imath}_{pq}[pq]}\right), \\
\tilde{K} &\leftarrow \left(\coprod_{q=r+1}^{n}\prod_{p=1}^{r} T_{\tilde{\imath}_{pq}[pq]}\right)\left(\tilde{K}\right).
\end{aligned}$$

(Actually, entering this step, it was that $\tilde{K} = I$, so in fact $\tilde{K}$ becomes just the product above. As in steps 6 and 10, here again it is not necessary actually to apply the addition elementaries one by one. Together they easily form a parallel unit upper—not lower—triangular matrix $L_{\parallel}^{\{r\}T}$. See § 11.8.4.)

13. Notice now that $\tilde{I} = I_r$. Let

$$P \equiv \tilde{P},\ D \equiv \tilde{D},\ L \equiv \tilde{L},\ U \equiv \tilde{U},\ K \equiv \tilde{K},\ S \equiv \tilde{S}.$$

End.

Never stalling, the algorithm cannot fail to achieve $\tilde{I} = I_r$ and thus a complete Gauss-Jordan decomposition of the form (12.2), though what value the rank r might turn out to have is not normally known to us in advance. (We have not yet proven, but will in § 12.5, that the algorithm always produces the same I_r, the same rank $r \geq 0$, regardless of which pivots $\tilde{\imath}_{pq} \neq 0$ one happens to choose in step 3 along the way. We can safely ignore this unproven fact however for the immediate moment.)

12.3.4 Rank and independent rows

Observe that the Gauss-Jordan algorithm of § 12.3.3 operates always within the bounds of the original $m \times n$ matrix A. Therefore, necessarily,

$$\begin{aligned}
r &\leq m, \\
r &\leq n.
\end{aligned} \tag{12.5}$$

The rank r exceeds the number neither of the matrix's rows nor of its columns. This is unsurprising. Indeed the narrative of the algorithm's step 8 has already noticed the fact.

Observe also however that *the rank always fully reaches $r = m$ if the rows of the original matrix A are linearly independent.* The reason for this observation is that the rank can fall short, $r < m$, only if step 3 finds a null row $i \le m$; but step 3 can find such a null row only if step 6 has created one (or if there were a null row in the original matrix A; but according to § 12.1, such null rows never were linearly independent in the first place). How do we know that step 6 can never create a null row? We know this because the action of step 6 is to add multiples only of *current and earlier* pivot rows to rows in $\tilde{I}$ which have not yet been on pivot.[9] According to (12.1), such action has no power to cancel the independent rows it targets.

12.3.5 Inverting the factors

Inverting the six Gauss-Jordan factors is easy. Sections 11.7 and 11.8 have shown how. One need not however go even to that much trouble. Each of the six factors—P, D, L, U, K and S—is composed of a sequence $\coprod T$ of elementary operators. Each of the six inverse factors—P^{-1}, D^{-1}, L^{-1}, U^{-1}, K^{-1} and S^{-1}—is therefore composed of the *reverse* sequence $\prod T^{-1}$ of *inverse* elementary operators. Refer to (11.41). If one merely records the sequence of elementaries used to build each of the six factors—if one reverses each sequence, inverts each elementary, and multiplies—then the six inverse factors result.

And, in fact, it isn't even that hard. One actually need not record the individual elementaries; one can invert, multiply and forget them in stream. This means starting the algorithm from step 1 with six extra variable working matrices (besides the seven already there):

$$\tilde{P}^{-1} \leftarrow I;\ \tilde{D}^{-1} \leftarrow I;\ \tilde{L}^{-1} \leftarrow I;\ \tilde{U}^{-1} \leftarrow I;\ \tilde{K}^{-1} \leftarrow I;\ \tilde{S}^{-1} \leftarrow I.$$

[9] If the truth of the sentence's assertion regarding the action of step 6 seems nonobvious, one can drown the assertion rigorously in symbols to prove it, but before going to that extreme consider: the action of steps 3 and 4 is to choose a pivot row $p \ge i$ and to shift it upward to the ith position. The action of step 6 then is to add multiples of the chosen pivot row downward only—that is, only to rows which have not yet been on pivot. This being so, steps 3 and 4 in the second iteration find no unmixed rows available to choose as second pivot, but find only rows which already include multiples of the first pivot row. Step 6 in the second iteration therefore adds downward multiples of the second pivot row, *which already includes a multiple of the first pivot row.* Step 6 in the ith iteration adds downward multiples of the ith pivot row, which already includes multiples of the first through $(i-1)$th. So it comes to pass that multiples only of current and earlier pivot rows are added to rows which have not yet been on pivot. To no row is ever added, directly or indirectly, a multiple of itself—until step 10, which does not belong to the algorithm's main loop and has nothing to do with the availability of nonzero rows to step 3.

(There is no $\tilde{I}^{-1}$, not because it would not be useful, but because its initial value would be[10] $A^{-1(r)}$, unknown at algorithm's start.) Then, for each operation on any of $\tilde{P}$, $\tilde{D}$, $\tilde{L}$, $\tilde{U}$, $\tilde{K}$ or $\tilde{S}$, one operates inversely on the corresponding inverse matrix. For example, in step 5,

$$\begin{aligned}
\tilde{D} &\leftarrow \tilde{D}T_{\tilde{\imath}_{ii}[i]}, & \tilde{D}^{-1} &\leftarrow T_{(1/\tilde{\imath}_{ii})[i]}\tilde{D}^{-1}, \\
\tilde{L} &\leftarrow T_{(1/\tilde{\imath}_{ii})[i]}\tilde{L}T_{\tilde{\imath}_{ii}[i]}, & \tilde{L}^{-1} &\leftarrow T_{(1/\tilde{\imath}_{ii})[i]}\tilde{L}^{-1}T_{\tilde{\imath}_{ii}[i]}, \\
\tilde{I} &\leftarrow T_{(1/\tilde{\imath}_{ii})[i]}\tilde{I}.
\end{aligned}$$

With this simple extension, the algorithm yields all the factors not only of the Gauss-Jordan decomposition (12.2) but simultaneously also of the Gauss-Jordan's complementary form (12.3).

12.3.6 Truncating the factors

None of the six factors of (12.2) actually needs to retain its entire extended-operational form (§ 11.3.2). The four factors on the left, row operators, act wholly by their $m \times m$ squares; the two on the right, column operators, by their $n \times n$. Indeed, neither I_r nor A has anything but zeros outside the $m \times n$ rectangle, anyway, so there is nothing for the six operators to act upon beyond those bounds in any event. We can truncate all six operators to dimension-limited forms (§ 11.3.1) for this reason if we want.

To truncate the six operators formally, we left-multiply (12.2) by I_m and right-multiply it by I_n, obtaining

$$I_m A I_n = I_m PDLUI_rKSI_n.$$

According to § 11.3.6, the I_m and I_n respectively truncate rows and columns, actions which have no effect on A since it is already a dimension-limited $m \times n$ matrix. By successive steps, then,

$$\begin{aligned}
A &= I_m PDLUI_rKSI_n \\
&= I_m^7 PDLUI_r^2KSI_n^3;
\end{aligned}$$

and finally, by using (11.31) or (11.42) repeatedly,

$$A = (I_mPI_m)(I_mDI_m)(I_mLI_m)(I_mUI_r)(I_rKI_n)(I_nSI_n), \tag{12.6}$$

where the dimensionalities of the six factors on the equation's right side are respectively $m \times m$, $m \times m$, $m \times m$, $m \times r$, $r \times n$ and $n \times n$. Equation (12.6)

[10] Section 11.5 explains the notation.

expresses any dimension-limited rectangular matrix A as the product of six particularly simple, dimension-limited rectangular factors.

By similar reasoning from (12.2),

$$A = (I_m G_> I_r)(I_r G_< I_n), \tag{12.7}$$

where the dimensionalities of the two factors are $m \times r$ and $r \times n$.

The book will seldom point it out again explicitly, but one can straightforwardly truncate not only the Gauss-Jordan factors but most other factors and operators, too, by the method of this subsection.[11]

12.3.7 Properties of the factors

One would expect such neatly formed operators as the factors of the Gauss-Jordan to enjoy some useful special properties. Indeed they do. Table 12.2 lists a few. The table's properties formally come from (11.52) and Table 11.5; but, if one firmly grasps the matrix forms involved and comprehends the notation (neither of which is trivial to do), if one understands that the operator $(I_n - I_r)$ is a truncator that selects columns $r + 1$ through n of the matrix it operates leftward upon, and if one sketches the relevant factors schematically with a pencil, then the properties are plainly seen without reference to chapter 11 as such.

The table's properties regarding P and S express a general advantage all permutors share. The table's properties regarding K are admittedly less significant, included mostly only because § 13.3 will need them. Still, even the K properties are always true. They might find other uses.

[11] Comes the objection, "Black, why do you make it more complicated than it needs to be? For what reason must all your matrices have infinite dimensionality, anyway? They don't do it that way in my other linear algebra book."

It is a fair question. The answer is that this book is a book of applied mathematical theory; and theoretically in the author's view, infinite-dimensional matrices are significantly neater to handle. To append a null row or a null column to a dimension-limited matrix is to alter the matrix in no essential way, nor is there any real difference between $T_{5[21]}$ when it row-operates on a $3 \times p$ matrix and the same elementary when it row-operates on a $4 \times p$. The relevant theoretical constructs ought to reflect such insights. Hence infinite dimensionality.

Anyway, a matrix displaying an infinite field of zeros resembles a shipment delivering an infinite supply of nothing; one need not be too impressed with either. The two matrix forms of § 11.3 manifest the sense that a matrix can represent a linear transformation, whose rank matters; or a reversible row or column operation, whose rank does not. The extended-operational form, having infinite rank, serves the latter case. In either case, however, the dimensionality $m \times n$ of the matrix is a distraction. It is the rank r, if any, that counts.

Table 12.2: A few properties of the Gauss-Jordan factors.

$$
\begin{aligned}
P^{*} &= P^{-1} = P^{T} \\
S^{*} &= S^{-1} = S^{T} \\
P^{-*} &= \;\; P \;\; = P^{-T} \\
S^{-*} &= \;\; S \;\; = S^{-T}
\end{aligned}
$$

$$\frac{K + K^{-1}}{2} = I$$

$$
\begin{array}{rcccl}
I_r K (I_n - I_r) & = & K - I & = & I_r (K - I)(I_n - I_r) \\
I_r K^{-1} (I_n - I_r) & = & K^{-1} - I & = & I_r (K^{-1} - I)(I_n - I_r) \\
(I - I_n)(K - I) & = & 0 & = & (K - I)(I - I_n) \\
(I - I_n)(K^{-1} - I) & = & 0 & = & (K^{-1} - I)(I - I_n)
\end{array}
$$

Further properties of the several Gauss-Jordan factors can be gleaned from the respectively relevant subsections of §§ 11.7 and 11.8.

12.3.8 Marginalizing the factor I_n

If A happens to be a square, $n \times n$ matrix and if it develops that the rank $r = n$, then one can take advantage of (11.31) to rewrite the Gauss-Jordan decomposition (12.2) in the form

$$PDLUKSI_n = A = I_n PDLUKS, \tag{12.8}$$

thus marginalizing the factor I_n. This is to express the Gauss-Jordan solely in row operations or solely in column operations. It does not change the algorithm and it does not alter the factors; it merely reorders the factors after the algorithm has determined them. It fails however if A is rectangular or $r < n$.

12.3.9 Decomposing an extended operator

Sections 12.5 and 13.1 will demonstrate that, if a matrix A is a extended operator (§ 11.3.2) with an $n \times n$ active region and if the operation the matrix implements is reversible, then the truncated operator $I_n A = I_n A I_n = A I_n$ necessarily enjoys full rank $r = n$. Complementarily, those sections will demonstrate that, if such an extended operator is irreversible, then the truncated operator $I_nA = I_nAI_n = AI_n$ must suffer $r < n$. Full rank is associated with reversibility. This fact has many consequences, among which is the following.

To extend the Gauss-Jordan decomposition of the present section to decompose a reversible, $n \times n$ extended operator A is trivial. One merely writes

$$A = PDLUKS,$$

wherein the I_r has become an I. Or, equivalently, one decomposes the $n \times n$ dimension-limited matrix $I_nA = I_nAI_n = AI_n$ as

$$AI_n = PDLUI_nKS = PDLUKSI_n,$$

from which, inasmuch as all the factors present but I_n are $n \times n$ extended operators, the preceding equation results.

One can decompose only reversible extended operators so. The Gauss-Jordan fails on irreversible extended operators. Fortunately, as we have seen in chapter 11, every extended operator constructible as the product of elementary, quasielementary, unit-triangular and/or shift operators is indeed reversible, so a great variety of extended operators are decomposable. (Note incidentally that shifts, § 11.9, generally prevent the extended operators whose construction includes them from honoring any finite, $n \times n$ active region. Therefore, in this subsection we are generally thinking of extended operators constructible without shifts.)

This subsection's equations remain unnumbered because they say little new. Their only point, really, is that what an operator does outside some appropriately delimited active region is seldom interesting because the vector on which the operator ultimately acts is probably null there in any event. In such a context it may not matter whether one truncates the operator. Indeed, this was also the point of § 12.3.6 and, if you like, of (11.31), too.[12]

[12] If "it may not matter," as the narrative says, then one might just put all matrices in dimension-limited form. Many books do. To put them all in dimension-limited form however brings at least three effects the book you are reading prefers to avoid. First, it leaves shift-and-truncate operations hard to express cleanly (refer to §§ 11.3.6 and 11.9 and, as a typical example of the usage, eqn. 13.7). Second, it confuses the otherwise natural extension of discrete vectors into continuous functions. Third, it leaves one to consider the ranks of reversible operators like $T_{[1\leftrightarrow 2]}$ that naturally should have no rank. The last of the three is arguably most significant: matrix rank is such an important attribute that one prefers to impute it only to those operators about which it actually says something interesting.

Nevertheless, the extended-operational matrix form is hardly more than a formality. All

Regarding the present section as a whole, the Gauss-Jordan decomposition is a significant achievement. It is not the only matrix decomposition—further interesting decompositions include the Gram-Schmidt of § 13.11, the diagonal of § 14.6, the Schur of § 14.10 and the singular-value of § 14.12, among others—but the Gauss-Jordan nonetheless reliably factors an arbitrary $m \times n$ matrix A, which we had not known how to handle very well, into a product of unit triangular matrices and quasielementaries, which we do. We shall put the Gauss-Jordan to good use in chapter 13. However, before closing the present chapter we should like finally, squarely to define and to establish the concept of matrix rank, not only for I_r but for all matrices. To do that, we shall first need one more preliminary: the technique of vector replacement.

12.4 Vector replacement

Consider a set of $m + 1$ (not necessarily independent) vectors

$$\{\mathbf{u}, \mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\}.$$

As a definition, the *space* these vectors *address* consists of all linear combinations of the set's several vectors. That is, the space consists of all vectors $\mathbf{b}$ formable as

$$\beta_o\mathbf{u} + \beta_1\mathbf{a}_1 + \beta_2\mathbf{a}_2 + \cdots + \beta_m\mathbf{a}_m = \mathbf{b}. \tag{12.9}$$

Now consider a specific vector $\mathbf{v}$ in the space,

$$\psi_o\mathbf{u} + \psi_1\mathbf{a}_1 + \psi_2\mathbf{a}_2 + \cdots + \psi_m\mathbf{a}_m = \mathbf{v}, \tag{12.10}$$

for which

$$\psi_o \neq 0.$$

Solving (12.10) for $\mathbf{u}$, we find that

$$\frac{1}{\psi_o}\mathbf{v} - \frac{\psi_1}{\psi_o}\mathbf{a}_1 - \frac{\psi_2}{\psi_o}\mathbf{a}_2 - \cdots - \frac{\psi_m}{\psi_o}\mathbf{a}_m = \mathbf{u}.$$

it says is that the extended operator unobtrusively leaves untouched anything it happens to find outside its operational domain, whereas a dimension-limited operator would have truncated whatever it found there. Since what is found outside the operational domain is typically uninteresting, this may be a distinction without a difference, a distinction one can safely ignore.

With the change of variables

$$\begin{aligned} \phi_o &\leftarrow \frac{1}{\psi_o}, \\ \phi_1 &\leftarrow -\frac{\psi_1}{\psi_o}, \\ \phi_2 &\leftarrow -\frac{\psi_2}{\psi_o}, \\ &\vdots \\ \phi_m &\leftarrow -\frac{\psi_m}{\psi_o}, \end{aligned}$$

for which, quite symmetrically, it happens that

$$\begin{aligned} \psi_o &= \frac{1}{\phi_o}, \\ \psi_1 &= -\frac{\phi_1}{\phi_o}, \\ \psi_2 &= -\frac{\phi_2}{\phi_o}, \\ &\vdots \\ \psi_m &= -\frac{\phi_m}{\phi_o}, \end{aligned}$$

the solution is

$$\phi_o\mathbf{v} + \phi_1\mathbf{a}_1 + \phi_2\mathbf{a}_2 + \cdots + \phi_m\mathbf{a}_m = \mathbf{u}. \tag{12.11}$$

Equation (12.11) has identical form to (12.10), only with the symbols $\mathbf{u} \leftrightarrow \mathbf{v}$ and $\psi \leftrightarrow \phi$ swapped. Since $\phi_o = 1/\psi_o$, assuming that ψ_o is finite it even appears that

$$\phi_o \neq 0;$$

so, the symmetry is complete. Table 12.3 summarizes.

Now further consider an arbitrary vector $\mathbf{b}$ which lies in the space addressed by the vectors

$$\{\mathbf{u}, \mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\}.$$

Does the same $\mathbf{b}$ also lie in the space addressed by the vectors

$$\{\mathbf{v}, \mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\}?$$

Table 12.3: The symmetrical equations of § 12.4.

$$
\begin{array}{rcl@{\qquad\qquad}rcl}
\psi_o\mathbf{u} + \psi_1\mathbf{a}_1 + \psi_2\mathbf{a}_2 & & & \phi_o\mathbf{v} + \phi_1\mathbf{a}_1 + \phi_2\mathbf{a}_2 & & \\
+\cdots+\psi_m\mathbf{a}_m & = & \mathbf{v} & +\cdots+\phi_m\mathbf{a}_m & = & \mathbf{u} \\
0 \neq \dfrac{1}{\psi_o} & = & \phi_o & 0 \neq \dfrac{1}{\phi_o} & = & \psi_o \\
-\dfrac{\psi_1}{\psi_o} & = & \phi_1 & -\dfrac{\phi_1}{\phi_o} & = & \psi_1 \\
-\dfrac{\psi_2}{\psi_o} & = & \phi_2 & -\dfrac{\phi_2}{\phi_o} & = & \psi_2 \\
& \vdots & & & \vdots & \\
-\dfrac{\psi_m}{\psi_o} & = & \phi_m & -\dfrac{\phi_m}{\phi_o} & = & \psi_m
\end{array}
$$

To show that it does, we substitute into (12.9) the expression for $\mathbf{u}$ from (12.11), obtaining the form

$$(\beta_o)(\phi_o\mathbf{v} + \phi_1\mathbf{a}_1 + \phi_2\mathbf{a}_2 + \cdots + \phi_m\mathbf{a}_m) + \beta_1\mathbf{a}_1 + \beta_2\mathbf{a}_2 + \cdots + \beta_m\mathbf{a}_m = \mathbf{b}.$$

Collecting terms, this is

$$\beta_o\phi_o\mathbf{v} + (\beta_o\phi_1 + \beta_1)\mathbf{a}_1 + (\beta_o\phi_2 + \beta_2)\mathbf{a}_2 + \cdots + (\beta_o\phi_m + \beta_m)\mathbf{a}_m = \mathbf{b},$$

in which we see that, yes, $\mathbf{b}$ does indeed also lie in the latter space. Naturally the problem's $\mathbf{u} \leftrightarrow \mathbf{v}$ symmetry then guarantees the converse, that an arbitrary vector $\mathbf{b}$ which lies in the latter space also lies in the former. Therefore, a vector $\mathbf{b}$ must lie in both spaces or neither, never just in one or the other. The two spaces are, in fact, one and the same.

This leads to the following useful conclusion. Given a set of vectors

$$\{\mathbf{u}, \mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\},$$

one can safely replace the $\mathbf{u}$ by a new vector $\mathbf{v}$, obtaining the new set

$$\{\mathbf{v}, \mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\},$$

provided that the replacement vector $\mathbf{v}$ includes at least a little of the replaced vector $\mathbf{u}$ ($\psi_o \neq 0$ in eqn. 12.10) and that $\mathbf{v}$ is otherwise an honest linear combination of the several vectors of the original set, untainted by foreign contribution. *Such vector replacement does not in any way alter the space addressed.* The new space is exactly the same as the old.

As a corollary, *if the vectors of the original set happen to be linearly independent (§ 12.1), then the vectors of the new set are linearly independent, too;* for, if it were that

$$\gamma_o \mathbf{v} + \gamma_1 \mathbf{a}_1 + \gamma_2 \mathbf{a}_2 + \cdots + \gamma_m \mathbf{a}_m = 0$$

for nontrivial γ_o and γ_k, then either $\gamma_o = 0$—impossible since that would make the several $\mathbf{a}_k$ themselves linearly dependent—or $\gamma_o \neq 0$, in which case $\mathbf{v}$ would be a linear combination of the several $\mathbf{a}_k$ alone. But if $\mathbf{v}$ were a linear combination of the several $\mathbf{a}_k$ alone, then (12.10) would still also explicitly make $\mathbf{v}$ a linear combination of the same $\mathbf{a}_k$ plus a nonzero multiple of $\mathbf{u}$. Yet both combinations cannot be, because according to § 12.1, two distinct combinations among a set of independent vectors can never target the same $\mathbf{v}$. The contradiction proves false the assumption which gave rise to it: that the vectors of the new set were linearly dependent. Hence the vectors of the new set are equally as independent as the vectors of the old.

12.5 Rank

Sections 11.3.5 and 11.3.6 have introduced the rank-r identity matrix I_r, where the integer r is the number of ones the matrix has along its main diagonal. Other matrices have rank, too. Commonly, an $n \times n$ matrix has rank $r = n$, but consider the matrix

$$\begin{bmatrix} 5 & 1 & 6 \\ 3 & 6 & 9 \\ 2 & 4 & 6 \end{bmatrix}.$$

The third column of this matrix is the sum of the first and second columns. Also, the third row is just two-thirds the second. Either way, by columns or by rows, the matrix has only two independent vectors. The rank of this 3×3 matrix is not $r = 3$ but $r = 2$.

This section establishes properly the important concept of matrix *rank.* The section demonstrates that every matrix has a definite, unambiguous rank, and shows how this rank can be calculated.

To forestall potential confusion in the matter, we should immediately observe that—like the rest of this chapter but unlike some other parts of the book—this section explicitly trades in exact numbers. If a matrix element here is 5, then it is exactly 5; if 0, then exactly 0. Many real-world matrices, of course—especially matrices populated by measured data—can never truly be exact, but that is not the point here. Here, the numbers are exact.[13]

12.5.1 A logical maneuver

In § 12.5.2 we will execute a pretty logical maneuver, which one might name, "the end justifies the means."[14] When embedded within a larger logical construct as in § 12.5.2, the maneuver if unexpected can confuse, so this subsection is to prepare the reader to expect the maneuver.

The logical maneuver follows this basic pattern.

> If P_1 then Q. If P_2 then Q. If P_3 then Q. Which of P_1, P_2 and P_3 are true is not known, but suppose that it is nevertheless known that *at least one* of the three is true: P_1 or P_2 or P_3. If this is so, then—though one can draw no valid conclusion regarding any one of the three conditions P_1, P_2 or P_3—one can still conclude that their common object Q is true.

One valid way to prove Q, then, would be to suppose P_1 and show that it led to Q; and then alternately to suppose P_2 and show that it separately led to Q; and then again to suppose P_3 and show that it also led to Q. The final step would be to show somehow that P_1, P_2 and P_3 could not possibly all be false at once. Herein, the *means* is to assert several individually suspicious claims, none of which one actually means to prove. The *end* which justifies the means is the conclusion Q, which thereby one can and does prove.

It is a subtle maneuver. Once the reader feels that he grasps its logic, he can proceed to the next subsection where the maneuver is put to use.[15]

[13] It is false to suppose that because applied mathematics *permits* inexact or imprecise quantities, like 3.0 ± 0.1 inches for the length of your thumb, it also requires them. On the contrary, the length of your thumb may indeed be 3.0 ± 0.1 inches, but surely no triangle has 3.0 ± 0.1 sides! A triangle has exactly three sides. The ratio of a circle's circumference to its radius is exactly 2π. The author has exactly one brother. A construction contract might require the builder to finish within exactly 180 days (though the actual construction time might be an inexact $t = 172.6 \pm 0.2$ days), and so on. Exact quantities are every bit as valid in applied mathematics as inexact or imprecise ones are. Where the distinction matters, it is the applied mathematician's responsibility to distinguish between the two kinds of quantity.

[14] The maneuver's name rings a bit sinister, does it not? However, the book is not here setting forth an ethical proposition, but is merely previewing an abstract logical form the mathematics of § 12.5.2 will use.

[15] Pure mathematics admittedly advantages the professional mathematician over the scientific or engineering applicationist when logic like this subsection's arrives. The writer,

12.5.2 The impossibility of identity-matrix promotion

Consider the matrix equation

$$AI_rB = I_s. \tag{12.12}$$

If $r \geq s$, then it is trivial to find matrices A and B for which (12.12) holds: $A = I_s = B$, for instance. If instead

$$r < s,$$

however, it is not so easy to find such matrices A and B. In fact it is impossible. This subsection proves the impossibility. It shows that *one cannot by any row and column operations, reversible or otherwise, ever transform an identity matrix into another identity matrix of greater rank (§ 11.3.5).*

Equation (12.12) can be written in the form

$$(AI_r)B = I_s, \tag{12.13}$$

where, because I_r attacking from the right is the column truncation operator (§ 11.3.6), the product AI_r is a matrix with an unspecified number of rows but only r columns—or, more precisely, with no more than r nonzero columns. Viewed this way, per § 11.1.3, B operates on the r columns of AI_r to produce the s columns of I_s.

The r columns of AI_r are nothing more than the first through rth columns of A. Let the symbols $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5, \ldots, \mathbf{a}_r$ denote these columns. The s columns of I_s, then, are nothing more than the elementary vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5, \ldots, \mathbf{e}_s$ (§ 11.3.7). The claim (12.13) makes is thus that the several vectors $\mathbf{a}_k$ together address each of the several elementary vectors $\mathbf{e}_j$—that is, that a linear combination[16]

$$b_{1j}\mathbf{a}_1 + b_{2j}\mathbf{a}_2 + b_{3j}\mathbf{a}_3 + \cdots + b_{rj}\mathbf{a}_r = \mathbf{e}_j \tag{12.14}$$

exists for each $\mathbf{e}_j$, $1 \leq j \leq s$.

The claim (12.14) will turn out to be false because there are too many $\mathbf{e}_j$, but to prove this, *we shall assume for the moment that the claim were true.* The proof then is by contradiction,[17] and it runs as follows.

an engineer inexpert in symbolic logic except in that restricted form in which the design of digital electronics employs it, will not attempt a symbolic treatment here.

[16] Observe that unlike as in § 12.1, here we have not necessarily assumed that the several $\mathbf{a}_k$ are linearly independent.

[17] As the reader will have observed by this point in the book, the technique—also called *reductio ad absurdum*—is the usual mathematical technique to prove impossibility. One assumes the falsehood to be true, then reasons toward a contradiction which proves the assumption false. Section 6.1.1 among others has already illustrated the technique, but the technique's use here is more sophisticated.

Consider the elementary vector $\mathbf{e}_1$. For $j = 1$, (12.14) is

$$b_{11}\mathbf{a}_1 + b_{21}\mathbf{a}_2 + b_{31}\mathbf{a}_3 + \cdots + b_{r1}\mathbf{a}_r = \mathbf{e}_1,$$

which says that the elementary vector $\mathbf{e}_1$ is a linear combination of the several vectors

$$\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5, \ldots, \mathbf{a}_r\}.$$

Because $\mathbf{e}_1$ is a linear combination, according to § 12.4 one can safely replace any of the vectors in the set by $\mathbf{e}_1$ without altering the space addressed. For example, replacing $\mathbf{a}_1$ by $\mathbf{e}_1$,

$$\{\mathbf{e}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5, \ldots, \mathbf{a}_r\}.$$

The only restriction per § 12.4 is that $\mathbf{e}_1$ contain at least a little of the vector $\mathbf{a}_k$ it replaces—that $b_{k1} \neq 0$. Of course there is no guarantee specifically that $b_{11} \neq 0$, so for $\mathbf{e}_1$ to replace $\mathbf{a}_1$ might not be allowed. However, inasmuch as $\mathbf{e}_1$ is nonzero, then according to (12.14) at least one of the several b_{k1} also is nonzero; and if b_{k1} is nonzero then $\mathbf{e}_1$ can replace $\mathbf{a}_k$. Some of the $\mathbf{a}_k$ might indeed be forbidden, but never all; there is always at least one $\mathbf{a}_k$ which $\mathbf{e}_1$ can replace. (For example, if $\mathbf{a}_1$ were forbidden because $b_{11} = 0$, then $\mathbf{a}_3$ might be available instead because $b_{31} \neq 0$. In this case the new set would be $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{e}_1, \mathbf{a}_4, \mathbf{a}_5, \ldots, \mathbf{a}_r\}$.)

Here comes the hard part. Here is where the logical maneuver of § 12.5.1 comes in. The book to this point has established no general method to tell which of the several $\mathbf{a}_k$ the elementary vector $\mathbf{e}_1$ actually contains (§ 13.2 gives the method, but that section depends logically on this one, so we cannot properly appeal to it here). According to (12.14), the vector $\mathbf{e}_1$ might contain some of the several $\mathbf{a}_k$ or all of them, *but surely it contains at least one of them.* Therefore, even though it is illicit to replace an $\mathbf{a}_k$ by an $\mathbf{e}_1$ which contains none of it, even though we have no idea which of the several $\mathbf{a}_k$ the vector $\mathbf{e}_1$ contains, even though replacing the wrong $\mathbf{a}_k$ logically invalidates any conclusion which flows from the replacement, still we can proceed with the proof—provided only that, in the end, we shall find that the illicit choice of replacement and the licit choice had led alike to the same, identical conclusion. If we do so find then—in the end—the logic will demand of us only an assurance that some licit choice had existed at the time the choice was or might have been made. The logic will never ask, even in retrospect, which specific choice had been the licit one, for only the complete absence of licit choices can threaten the present maneuver.

The claim (12.14) guarantees at least one licit choice. Whether as the maneuver also demands, all the choices, licit and illicit, lead ultimately alike to the same, identical conclusion remains to be determined.

Now consider the elementary vector $\mathbf{e}_2$. According to (12.14), $\mathbf{e}_2$ lies in the space addressed by the original set

$$\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5, \ldots, \mathbf{a}_r\}.$$

Therefore as we have seen, $\mathbf{e}_2$ also lies in the space addressed by the new set

$$\{\mathbf{e}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5, \ldots, \mathbf{a}_r\}$$

(or $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{e}_1, \mathbf{a}_4, \mathbf{a}_5, \ldots, \mathbf{a}_r\}$, or whatever the new set happens to be). That is, not only do coefficients b_{k2} exist such that

$$b_{12}\mathbf{a}_1 + b_{22}\mathbf{a}_2 + b_{32}\mathbf{a}_3 + \cdots + b_{r2}\mathbf{a}_r = \mathbf{e}_2,$$

but also coefficients β_{k2} exist such that

$$\beta_{12}\mathbf{e}_1 + \beta_{22}\mathbf{a}_2 + \beta_{32}\mathbf{a}_3 + \cdots + \beta_{r2}\mathbf{a}_r = \mathbf{e}_2.$$

Again it is impossible for all the coefficients β_{k2} to be zero but, moreover, it is impossible for β_{12} to be the sole nonzero coefficient, for (as should seem plain to the reader who grasps the concept of the elementary vector, § 11.3.7) no elementary vector can ever be a linear combination of other elementary vectors alone! The linear combination which forms $\mathbf{e}_2$ evidently includes a nonzero multiple *of at least one of the remaining* $\mathbf{a}_k$. At least one of the β_{k2} attached to an $\mathbf{a}_k$ (not β_{12}, which is attached to $\mathbf{e}_1$) must be nonzero. Therefore by the same reasoning as before, we now choose an $\mathbf{a}_k$ with a nonzero coefficient $\beta_{k2} \neq 0$ and replace it by $\mathbf{e}_2$, obtaining an even newer set of vectors like

$$\{\mathbf{e}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{e}_2, \mathbf{a}_5, \ldots, \mathbf{a}_r\}.$$

This newer set addresses precisely the same space as the previous set, and thus also as the original set.

And so it goes, replacing one $\mathbf{a}_k$ by an $\mathbf{e}_j$ at a time, until all the $\mathbf{a}_k$ are gone and our set has become

$$\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5, \ldots, \mathbf{e}_r\},$$

which, as we have reasoned, addresses precisely the same space as did the original set

$$\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5, \ldots, \mathbf{a}_r\}.$$

And this is the one, identical conclusion the maneuver of § 12.5.1 has demanded. All intermediate choices, by various paths licit and illicit, ultimately have led alike to the single conclusion of this paragraph, which thereby is properly established.

Admittedly, such subtle logic may not be easy to discern. Here is another, slightly different light by which to illuminate the question. Suppose again that, by making exactly r replacements, we wish to convert the set

$$\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5, \ldots, \mathbf{a}_r\}$$

into

$$\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5, \ldots, \mathbf{e}_r\},$$

assuming again per § 12.4 that the several vectors $\mathbf{a}_k$ of the original set, taken together, address each of the elementary vectors $\mathbf{e}_j$, $1 \le j \le s$, $r < s$. Suppose further again that we wish per § 12.4 to convert the set without altering the space the set addresses. To reach our goal, first we will put the elementary vector $\mathbf{e}_1$ in the place of one of the several vectors $\mathbf{a}_k$, then we will put $\mathbf{e}_2$ in the place of *one of the remaining* $\mathbf{a}_k$; and so on until, at last, we put $\mathbf{e}_r$ in the place of *the last remaining* $\mathbf{a}_k$. We will give the several $\mathbf{e}_j$, $1 \le j \le r$, in their proper order but we might take the several $\mathbf{a}_k$ in any of $r!$ distinct sequences: for instance, in the case that $r = 3$, we might take the several $\mathbf{a}_k$ in any of the $3! = 6$ distinct sequences

$$(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3);\ (\mathbf{a}_1, \mathbf{a}_3, \mathbf{a}_2);\ (\mathbf{a}_2, \mathbf{a}_1, \mathbf{a}_3);\ (\mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_1);\ (\mathbf{a}_3, \mathbf{a}_1, \mathbf{a}_2);\ (\mathbf{a}_3, \mathbf{a}_2, \mathbf{a}_1);$$

except however that we might (or might not) find certain sequences blockaded in the event. Blockaded? Well, consider for example the sequence $(\mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_1)$, and suppose that $\mathbf{e}_1 = 0\mathbf{a}_1 + 4\mathbf{a}_2 - 2\mathbf{a}_3$ and that $\mathbf{e}_2 = 5\mathbf{a}_1 - (1/2)\mathbf{e}_1 + 0\mathbf{a}_3$ (noticing that the latter already has $\mathbf{e}_1$ on the right instead of $\mathbf{a}_2$): in this case the sequence $(\mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_1)$ is blockaded—which is to say, forbidden—because, once $\mathbf{e}_1$ has replaced $\mathbf{a}_2$, since $\mathbf{e}_2$ then contains none of $\mathbf{a}_3$, $\mathbf{e}_2$ cannot according to § 12.4 replace $\mathbf{a}_3$. [Actually, in this example, both sequences beginning $(\mathbf{a}_1, \ldots)$ are blockaded, too, because the turn of $\mathbf{e}_1$ comes first and, at that time, $\mathbf{e}_1$ contains none of $\mathbf{a}_1$.] Clear? No? Too many subscripts? Well, there's nothing for it: if you wish to understand then you will simply have to trace all the subscripts out with your pencil; the example cannot be made any simpler. Now, although some sequences might be blockaded, no unblockaded sequence can run to a dead end, so to speak. After each unblockaded replacement another replacement will always be possible. The reason is as before: that, according to § 12.4, so long as each elementary vector $\mathbf{e}_j$ in turn contains some of the vector it replaces, the replacement cannot alter the space the set addresses; that the space by initial assumption includes all the elementary vectors; that each elementary vector in turn must therefore be found to contain at least one of the vectors then in the set; that no elementary vector can be composed solely of other elementary vectors; and, consequently, that each elementary vector in turn must be found to contain at least one of the set's then remaining $\mathbf{a}_k$. The logic though slightly complicated is nonethemore escapable. The conclusion is that we can indeed convert $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5, \ldots, \mathbf{a}_r\}$ into $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5, \ldots, \mathbf{e}_r\}$, step by step, without altering the space addressed.

The conclusion leaves us with a problem, however. There remain more $\mathbf{e}_j$, $1 \le j \le s$, than there are $\mathbf{a}_k$, $1 \le k \le r$, because, as we have stipulated, $r < s$. Some elementary vectors $\mathbf{e}_j$, $r < j \le s$, are evidently left over. Back at the beginning of the section, the claim (12.14) made was that

$$\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5, \ldots, \mathbf{a}_r\}$$

together addressed each of the several elementary vectors $\mathbf{e}_j$. But as we have seen, this amounts to a claim that

$$\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5, \ldots, \mathbf{e}_r\}$$

together addressed each of the several elementary vectors $\mathbf{e}_j$. Plainly this is impossible with respect to the left-over $\mathbf{e}_j$, $r < j \le s$. The contradiction proves false the claim which gave rise to it. The false claim: that the several $\mathbf{a}_k$, $1 \le k \le r$, addressed all the $\mathbf{e}_j$, $1 \le j \le s$, even when $r < s$.

Equation (12.13), which is just (12.12) written differently, asserts that B is a column operator which does precisely what we have just shown impossible: to combine the r columns of AI_r to yield the s columns of I_s, the latter of which are just the elementary vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5, \ldots, \mathbf{e}_s$. Hence finally we conclude that no matrices A and B exist which satisfy (12.12) when $r < s$. In other words, we conclude that *although row and column operations can demote identity matrices in rank, they can never promote them.* The promotion of identity matrices is impossible.

12.5.3 General matrix rank and its uniqueness

Step 8 of the Gauss-Jordan algorithm (§ 12.3.3) discovers a rank r for any matrix A. One should like to think that this rank r were a definite property of the matrix itself rather than some unreliable artifact of the algorithm, but until now we have lacked the background theory to prove it. Now we have the theory. Here is the proof.

The proof begins with a formal definition of the quantity whose uniqueness we are trying to prove.

- The rank r of an identity matrix I_r is the number of ones along its main diagonal. (This is from § 11.3.5.)
- The rank r of a general matrix A is the rank of an identity matrix I_r to which A can be reduced by *reversible* row and column operations.

Let the symbols $B_>$ and $B_<$ respectively represent the aforementioned reversible row and column operations:

$$\begin{aligned} B_>^{-1}B_> &= I = B_>B_>^{-1}; \\ B_<^{-1}B_< &= I = B_<B_<^{-1}. \end{aligned} \tag{12.15}$$

A matrix A has rank r if and only if row and column operations $B_>$ and $B_<$ exist such that

$$\begin{aligned} B_>AB_< &= I_r, \\ A &= B_>^{-1}I_rB_<^{-1}. \end{aligned} \tag{12.16}$$

The question is whether in (12.16) only a single rank r is possible.

To answer the question, we suppose that another rank were possible, that A had not only rank r but also rank s. Then,

$$\begin{aligned} A &= B_>^{-1}I_rB_<^{-1}, \\ A &= G_>^{-1}I_sG_<^{-1}. \end{aligned}$$

Combining these equations,

$$B_>^{-1}I_rB_<^{-1} = G_>^{-1}I_sG_<^{-1}.$$

Solving first for I_r, then for I_s,

$$\begin{aligned} (B_>G_>^{-1})I_s(G_<^{-1}B_<) &= I_r, \\ (G_>B_>^{-1})I_r(B_<^{-1}G_<) &= I_s. \end{aligned}$$

Were it that $r \neq s$, then one of these two equations would constitute the demotion of an identity matrix and the other, a promotion. But according to § 12.5.2 and its (12.12), promotion is impossible. Therefore $r \neq s$ is also impossible, and

$$r = s$$

is guaranteed. No matrix has two different ranks. *Matrix rank is unique.*

This finding has two immediate implications:

- Reversible row and/or column operations exist to change any matrix of rank r to *any other matrix* of the same rank. The reason is that, according to (12.16), (12.2) and (12.3), reversible operations exist to change both matrices to I_r and back.

- No reversible operation can change a matrix's rank.

The finding further suggests a conjecture:

- The matrices A, A^T and A^* share the same rank r.

The conjecture is proved by using (11.14) or (11.15) to compute the transpose or adjoint of (12.16).

The discovery that every matrix has a single, unambiguous rank and the establishment of a failproof algorithm—the Gauss-Jordan—to ascertain that rank have not been easy to achieve, but they are important achievements nonetheless, worth the effort thereto. The reason these achievements matter is that the mere dimensionality of a matrix is a chimerical measure of the matrix's true size—as for instance for the 3×3 example matrix at the head of the section. Matrix rank by contrast is a solid, dependable measure. We will rely on it often.

Section 12.5.8 comments further.

12.5.4 The full-rank matrix

According to (12.5), the rank r of a dimension-limited matrix (§ 11.3.1)—let us refer to it as a *matrix* (just to reduce excess verbiage)—can exceed the number neither of the matrix's rows nor of its columns. The greatest rank possible for an $m \times n$ matrix is the lesser of m and n. A *full-rank* matrix, then, is defined to be an $m \times n$ matrix with rank $r = m$ or $r = n$—or, if $m = n$, both. A matrix of less than full rank is a *degenerate* matrix.

Consider a tall $m \times n$ matrix C, $m \geq n$, one of whose n columns is a linear combination (§ 12.1) of the others. One could by definition target the dependent column with addition elementaries, using multiples of the other columns to wipe the dependent column out. Having zeroed the dependent column, one could then interchange it over to the matrix's extreme right, effectively throwing the column away, shrinking the matrix to $m \times (n - 1)$ dimensionality. Shrinking the matrix necessarily also shrinks the bound on the matrix's rank to $r \leq n - 1$—which is to say, to $r < n$. But the shrinkage, done by reversible column operations, is itself reversible, by which § 12.5.3 binds the rank of the original, $m \times n$ matrix C likewise to $r < n$. The matrix C, one of whose columns is a linear combination of the others, is necessarily degenerate for this reason.

Now consider a tall matrix A with the same $m \times n$ dimensionality, but with a full n independent columns. The transpose A^T of such a matrix has a full n independent *rows.* One of the conclusions of § 12.3.4 was that a matrix of independent rows always has rank equal to the number of rows. Since A^T is such a matrix, its rank is a full $r = n$. But formally, what this says is that there exist operators $B_<^T$ and $B_>^T$ such that $I_n = B_<^T A^T B_>^T$, the transpose of which equation is $B_> A B_< = I_n$—which in turn says that not only A^T, but also A itself, has full rank $r = n$.

Parallel reasoning rules the rows and columns of broad matrices, $m \leq n$, of course. To square matrices, $m = n$, both lines of reasoning apply.

Gathering findings, we have that

- a tall $m \times n$ matrix, $m \geq n$, has full rank if and only if its columns are linearly independent;
- a broad $m \times n$ matrix, $m \leq n$, has full rank if and only if its rows are linearly independent;
- a square $n \times n$ matrix, $m = n$, has full rank if and only if its columns and/or its rows are linearly independent; and
- a square matrix has both independent columns and independent rows, or neither; never just one or the other.

To say that a matrix has *full column rank* is to say that it is tall or square and has full rank $r = n \leq m$. To say that a matrix has *full row rank* is to say that it is broad or square and has full rank $r = m \leq n$. Only a square matrix can have full column rank and full row rank at the same time, because a tall or broad matrix cannot but include, respectively, more columns or more rows than I_r.

Observe incidentally that extended operators, which per § 11.3.2 define their $m \times n$ active regions differently, have infinite rank.

12.5.5 Underdetermined and overdetermined linear systems (introduction)

The last paragraph of § 12.5.4 provokes yet further terminology. A linear system $A\mathbf{x} = \mathbf{b}$ is *underdetermined* if A lacks full column rank—that is, if $r < n$—because inasmuch as some of A's columns then depend linearly on the others such a system maps multiple n-element vectors $\mathbf{x}$ to the same m-element vector $\mathbf{b}$, meaning that knowledge of $\mathbf{b}$ does not suffice to determine $\mathbf{x}$ uniquely. Complementarily, a linear system $A\mathbf{x} = \mathbf{b}$ is *overdetermined* if A lacks full row rank—that is, if $r < m$. If A lacks both, then the system is paradoxically both underdetermined and overdetermined and is thereby *degenerate.* If A happily has both, then the system is *exactly determined.*

Section 13.2 solves the exactly determined linear system. Section 13.4 solves the nonoverdetermined linear system. Section 13.6 analyzes the unsolvable overdetermined linear system among others. Further generalities await chapter 13; but, regarding the overdetermined system specifically, the present subsection would observe at least the few following facts.

An overdetermined linear system $A\mathbf{x} = \mathbf{b}$ *cannot have a solution for every possible* m*-element driving vector* $\mathbf{b}$*.* The truth of this claim can be seen by decomposing the system's matrix A by Gauss-Jordan and then left-multiplying the decomposed system by $G_>^{-1}$ to reach the form

$$I_r G_< \mathbf{x} = G_>^{-1}\mathbf{b}.$$

If the m-element vector $\mathbf{c} \equiv G_>^{-1}\mathbf{b}$, then $I_r G_< \mathbf{x} = \mathbf{c}$, which is impossible unless the last $m - r$ elements of $\mathbf{c}$ happen to be zero. But since $G_>$ is invertible, each $\mathbf{b}$ corresponds to a unique $\mathbf{c}$ and vice versa; so, if $\mathbf{b}$ is an unrestricted m-element vector then so also is $\mathbf{c}$, which verifies the claim.

Complementarily, *a nonoverdetermined linear system* $A\mathbf{x} = \mathbf{b}$ *does have a solution for every possible* m*-element driving vector* $\mathbf{b}$*.* This is so because in this case the last $m - r$ elements of $\mathbf{c}$ do happen to be zero; or, better stated, because $\mathbf{c}$ in this case has no nonzeros among its last $m-r$ elements, because it *has* no last $m - r$ elements, for the trivial reason that $r = m$.

It is an analytical error, and an easy one innocently to commit, to require that

$$A\mathbf{x} = \mathbf{b}$$

for unrestricted $\mathbf{b}$ when A lacks full row rank. The error is easy to commit because the equation looks right, because such an equation is indeed valid over a broad domain of $\mathbf{b}$ and might very well have been written correctly in that context, only not in the context of unrestricted $\mathbf{b}$. Analysis including such an error can lead to subtly absurd conclusions. It is never such an analytical error however to require that

$$A\mathbf{x} = 0$$

because, whatever other solutions such a system might have, it has at least the solution $\mathbf{x} = 0$.

12.5.6 The full-rank factorization

One sometimes finds dimension-limited matrices of less than full rank inconvenient to handle. However, every dimension-limited, $m \times n$ matrix of rank r can be expressed as the product of two full-rank matrices, one $m \times r$ and the other $r \times n$, both also of rank r:

$$A = BC. \tag{12.17}$$

The truncated Gauss-Jordan (12.7) constitutes one such *full-rank factorization:* $B = I_m G_> I_r$, $C = I_r G_< I_n$, good for any matrix. Other full-rank factorizations are possible, however, including among others the truncated Gram-Schmidt (13.58). The full-rank factorization is not unique.[18]

Of course, if an $m \times n$ matrix already has full rank $r = m$ or $r = n$, then the full-rank factorization is trivial: $A = I_m A$ or $A = AI_n$.

Section 13.6.4 uses the full-rank factorization.

12.5.7 Full column rank and the Gauss-Jordan factors K and S

The Gauss-Jordan decomposition (12.2),

$$A = PDLUI_rKS,$$

of a tall or square $m \times n$ matrix A of full column rank $r = n \le m$ always finds the factor $K = I$, regardless of the pivots one chooses during the Gauss-Jordan algorithm's step 3. If one happens always to choose $q = i$ as pivot column then not only $K = I$ but $S = I$, too.

That $K = I$ is seen by the algorithm's step 12, which creates K. Step 12 nulls the spare columns $q > r$ that dress $\tilde{I}$'s right, but in this case $\tilde{I}$ has only r columns and therefore has no spare columns to null. Hence step 12 does nothing and $K = I$.

That $S = I$ comes immediately of choosing $q = i$ for pivot column during each iterative instance of the algorithm's step 3. But, one must ask, can one choose so? What if column $q = i$ were unusable? That is, what if the only nonzero elements remaining in $\tilde{I}$'s ith column stood above the main diagonal, unavailable for step 4 to bring to pivot? Well, *were* it so, then one would indeed have to choose $q \ne i$ to swap the unusable column away rightward, but see: nothing in the algorithm later fills such a column's zeros with anything else—they remain zeros—so swapping the column away rightward

[18] [14, § 3.3][132, "Moore-Penrose generalized inverse"]

could only delay the crisis. The column would remain unusable. Eventually the column would reappear on pivot when no usable column rightward remained available to swap it with, which contrary to our assumption would mean precisely that $r < n$. Such contradiction can only imply that if $r = n$ then no unusable column can ever appear. One need not swap. We conclude that though one might voluntarily choose $q \neq i$ during the algorithm's step 3, the algorithm cannot force one to do so if $r = n$. Yet if one always does choose $q = i$, as the full-column-rank matrix A evidently leaves one free to do, then indeed $S = I$.

Theoretically, the Gauss-Jordan decomposition (12.2) includes the factors K and S precisely to handle matrices with more columns than rank. Matrices of full column rank $r = n$, common in applications, by definition have no such problem. Therefore, the Gauss-Jordan decomposition theoretically needs no K or S for such matrices, which fact lets us abbreviate the decomposition for such matrices to read

$$A = PDLUI_n. \tag{12.18}$$

Observe however that just because one theoretically can set $S = I$ does not mean that one actually should. The column permutor S exists to be used, after all—especially numerically to avoid small pivots during early invocations of the algorithm's step 5. Equation (12.18) is not mandatory but optional for a matrix A of full column rank (though still $r = n$ and thus $K = I$ for such a matrix, even when the unabbreviated eqn. 12.2 is used). There are however times when it is nice to know that one theoretically could, if doing exact arithmetic, set $S = I$ if one wanted to.

Since $PDLU$ acts as a row operator, (12.18) implies that each row of the square, $n \times n$ matrix A whose rank $r = n$ is full lies in the space the rows of I_n address. This is obvious and boring, but interesting is the converse implication of (12.18)'s complementary form,

$$U^{-1}L^{-1}D^{-1}P^{-1}A = I_n,$$

that each row of I_n lies in the space the rows of A address. The rows of I_n and the rows of A evidently address the same space. One can moreover say the same of A's columns since, according to § 12.5.3, A^T has full rank just as A does. In the whole, *if a matrix A is square and has full rank $r = n$, then A's columns together, A's rows together, I_n's columns together and I_n's rows together each address the same, complete n-dimensional space.*

12.5.8 The significance of rank uniqueness

The result of § 12.5.3, that matrix rank is unique, is an extremely important matrix theorem. It constitutes the chapter's chief result, which we have spent so many pages to attain. Without this theorem, the very concept of matrix rank must remain in doubt, along with all that attends to the concept. The theorem is the rock upon which the general theory of the matrix is built.

The concept underlying the theorem promotes the useful sensibility that a matrix's rank, much more than its mere dimensionality or the extent of its active region, represents the matrix's true size. Dimensionality can deceive, after all. For example, the honest 2×2 matrix

$$\begin{bmatrix} 5 & 1 \\ 3 & 6 \end{bmatrix}$$

has two independent rows or, alternately, two independent columns, and, hence, rank $r = 2$. One can easily construct a phony 3×3 matrix from the honest 2×2, however, simply by applying some 3×3 row and column elementaries:

$$T_{(2/3)[32]} \begin{bmatrix} 5 & 1 \\ 3 & 6 \end{bmatrix} T_{1[13]} T_{1[23]} = \begin{bmatrix} 5 & 1 & 6 \\ 3 & 6 & 9 \\ 2 & 4 & 6 \end{bmatrix}.$$

The 3×3 matrix on the equation's right is the one we met at the head of the section. It *looks* like a rank-three matrix, but really has only two independent columns and two independent rows. Its true rank is $r = 2$. We have here caught a matrix impostor pretending to be bigger than it really is.[19]

Now, admittedly, adjectives like "honest" and "phony," terms like "imposter," are a bit hyperbolic. The last paragraph has used them to convey

[19] An applied mathematician with some matrix experience actually probably recognizes this particular 3×3 matrix as a fraud on sight, but it is a very simple example. No one can just look at some arbitrary matrix and instantly perceive its true rank. Consider for instance the 5×5 matrix (in hexadecimal notation)

$$\begin{bmatrix} 12 & 9 & 3 & 1 & 0 \\ \frac{3}{2} & \frac{\mathrm{F}}{2} & \frac{15}{2} & 2 & 12 \\ \mathrm{D} & 9 & -19 & -\frac{\mathrm{E}}{3} & -6 \\ -2 & 0 & 6 & 1 & 5 \\ 1 & -4 & 4 & 1 & -8 \end{bmatrix}.$$

As the reader can verify by the Gauss-Jordan algorithm, the matrix's rank is not $r = 5$ but $r = 4$.

the subjective sense of the matter, but of course there is nothing mathematically improper or illegal about a matrix of less than full rank so long as the true rank is correctly recognized. When one models a physical phenomenon by a set of equations, one sometimes is dismayed to discover that one of the equations, thought to be independent, is really just a useless combination of the others. This can happen in matrix work, too. The rank of a matrix helps one to recognize how many truly independent vectors, dimensions or equations one actually has available to work with, rather than how many seem available at first glance. Such is the sense of matrix rank.

Chapter 13

Inversion and orthonormalization

The undeniably tedious chapters 11 and 12 have piled the matrix theory deep while affording scant practical reward. Building upon the two tedious chapters, this chapter brings the first rewarding matrix work.

One might be forgiven for forgetting after so many pages of abstract theory that the matrix afforded any reward or had any use at all. Uses however it has. Sections 11.1.1 and 12.5.5 have already broached[1] the matrix's most basic use, the primary subject of this chapter, to represent a system of m linear scalar equations in n unknowns neatly as

$$A\mathbf{x} = \mathbf{b}$$

and to solve the whole system at once by inverting the matrix A that characterizes it.

Now, before we go on, we want to confess that such a use alone, on the surface of it—though interesting—might not have justified the whole uncomfortable bulk of chapters 11 and 12. We already knew how to solve a simultaneous system of linear scalar equations in principle without recourse to the formality of a matrix, after all, as in the last step to derive (3.9) as far back as chapter 3. Why should we have suffered two bulky chapters, if only to prepare to do here something we already knew how to do?

The question is a fair one, but admits at least four answers. First, the matrix neatly solves a linear system not only for a particular driving vector $\mathbf{b}$ but for all possible driving vectors $\mathbf{b}$ at one stroke, as this chapter

[1] The reader who has skipped chapter 12 might at least review § 12.5.5.

explains. Second and yet more impressively, the matrix allows § 13.6 to introduce the *pseudoinverse* to approximate the solution to an unsolvable linear system and, moreover, to do so both optimally and efficiently, whereas such overdetermined systems arise commonly in applications. Third, to solve the linear system neatly is only the primary and most straightforward use of the matrix, not its only use: the even more interesting eigenvalue and its incidents await chapter 14. Fourth, specific applications aside, one should never underestimate the blunt practical benefit of reducing an arbitrarily large grid of scalars to a single symbol A, which one can then manipulate by known algebraic rules. Most students first learning the matrix have probably wondered at this stage whether it were worth all the tedium; so, if the reader now wonders, then he stands in good company. The matrix finally begins to show its worth here.

The chapter opens in § 13.1 by inverting the square matrix to solve the exactly determined, $n \times n$ linear system in § 13.2. It continues in § 13.3 by computing the rectangular matrix's kernel to solve the nonoverdetermined, $m \times n$ linear system in § 13.4. In § 13.6, it brings forth the aforementioned pseudoinverse, which rightly approximates the solution to the unsolvable overdetermined linear system. After briefly revisiting the Newton-Raphson iteration in § 13.7, it concludes by introducing the concept and practice of vector orthonormalization in §§ 13.8 through 13.12.

13.1 Inverting the square matrix

Consider an $n \times n$ square matrix A of full rank $r = n$. Suppose that extended operators $G_>$, $G_<$, $G_>^{-1}$ and $G_<^{-1}$ can be found, each with an $n \times n$ active region (§ 11.3.2), such that[2]

$$\begin{aligned} G_>^{-1}G_> = I = G_>G_>^{-1}, \\ G_<^{-1}G_< = I = G_<G_<^{-1}, \\ A = G_>I_nG_<. \end{aligned} \tag{13.1}$$

Observing from (11.31) that

$$\begin{aligned} I_nA &= \quad A \quad &&= AI_n, \\ I_nG_<^{-1}G_>^{-1} &= G_<^{-1}I_nG_>^{-1} &&= G_<^{-1}G_>^{-1}I_n, \end{aligned}$$

we find by successive steps that

$$\begin{aligned} A &= G_>I_nG_<, \\ I_nA &= G_>G_<I_n, \\ G_<^{-1}G_>^{-1}I_nA &= I_n, \\ (G_<^{-1}I_nG_>^{-1})(A) &= I_n; \end{aligned}$$

[2] The symbology and associated terminology might disorient a reader who had skipped chapters 11 and 12. In this book, the symbol I theoretically represents an $\infty \times \infty$ identity matrix. Outside the $m \times m$ or $n \times n$ square, the operators $G_>$ and $G_<$ each resemble the $\infty \times \infty$ identity matrix I, which means that the operators affect respectively only the first m rows or n columns of the thing they operate on. (In the present section it happens that $m = n$ because the matrix A of interest is square, but this footnote uses both symbols because generally $m \neq n$.)

The symbol I_r contrarily represents an identity matrix of only r ones, though it too can be viewed as an $\infty \times \infty$ matrix with zeros in the unused regions. If interpreted as an $\infty \times \infty$ matrix, the matrix A of the $m \times n$ system $A\mathbf{x} = \mathbf{b}$ has nonzero content only within the $m \times n$ rectangle.

None of this is complicated, really. Its purpose is merely to separate the essential features of a reversible operation like $G_>$ or $G_<$ from the dimensionality of the vector or matrix on which the operation happens to operate. The definitions do however necessarily, slightly diverge from definitions the reader may have been used to seeing in other books. In this book, one can legally multiply any two matrices, because all matrices are theoretically $\infty \times \infty$, anyway (though whether it makes any sense in a given circumstance to multiply mismatched matrices is another question; sometimes it does make sense, as in eqns. 13.25 and 14.50, but more often it does not—which naturally is why the other books tend to forbid such multiplication).

To the extent that the definitions confuse, the reader might briefly review the earlier chapters, especially § 11.3.

or alternately that

$$\begin{aligned} A &= G_> I_n G_<, \\ AI_n &= I_n G_> G_<, \\ AI_n G_<^{-1} G_>^{-1} &= I_n, \\ (A)(G_<^{-1} I_n G_>^{-1}) &= I_n. \end{aligned}$$

Either way, we have that

$$\begin{aligned} A^{-1}A = I_n = AA^{-1}, \\ A^{-1} \equiv G_<^{-1} I_n G_>^{-1}. \end{aligned} \tag{13.2}$$

Of course, for this to work, $G_>$, $G_<$, $G_>^{-1}$ and $G_<^{-1}$ must exist, be known and honor $n \times n$ active regions, which might seem a practical hurdle. However, (12.2), (12.3) and the body of § 12.3 have shown exactly how to find just such a $G_>$, $G_<$, $G_>^{-1}$ and $G_<^{-1}$ for any square matrix A of full rank, without exception; so, there is no trouble here. The factors do exist, and indeed we know how to find them.

Equation (13.2) features the important matrix A^{-1}, the *rank-n inverse* of A.

We have not yet much studied the rank-n inverse, but have at least defined it in (11.49), where we gave it the fuller, nonstandard notation $A^{-1(n)}$. When naming the rank-n inverse in words one usually says simply, "the inverse," because the rank is implied by the size of the square active region of the matrix inverted; but the rank-n inverse from (11.49) is not quite the infinite-dimensional inverse from (11.45), which is what $G_>^{-1}$ and $G_<^{-1}$ are. According to (13.2), the product of A^{-1} and A—or, written more fully, the product of $A^{-1(n)}$ and A—is, not I, but I_n.

Properties that emerge from (13.2) include the following.

- Like A, the rank-n inverse A^{-1} (more fully written as $A^{-1(n)}$) too is an $n \times n$ square matrix of full rank $r = n$.

- Since A is square and has full rank (§ 12.5.4), its rows and, separately, its columns are linearly independent, so it has only the one, unique inverse A^{-1}. No other rank-n inverse of A exists.

- On the other hand, inasmuch as A is square and has full rank, it does per (13.2) indeed have an inverse A^{-1}. The rank-n inverse exists.

- If $B = A^{-1}$ then $B^{-1} = A$. That is, A is itself the rank-n inverse of A^{-1}. The matrices A and A^{-1} thus form an exclusive, reciprocal pair.

- If B is an $n \times n$ square matrix and either $BA = I_n$ or $AB = I_n$, then both equalities in fact hold; thus, $B = A^{-1}$. One can have neither equality without the other.

- Only a square, $n \times n$ matrix of full rank $r = n$ has a rank-n inverse. A matrix A' which is not square, or whose rank falls short of a full $r = n$, is not invertible in the rank-n sense of (13.2).

That A^{-1} is an $n \times n$ square matrix of full rank and that A is itself the inverse of A^{-1} proceed from the definition (13.2) of A^{-1} plus § 12.5.3's finding that reversible operations like $G_>^{-1}$ and $G_<^{-1}$ cannot change I_n's rank. That the inverse exists is plain, inasmuch as the Gauss-Jordan decomposition plus (13.2) reliably calculate it. That the inverse is unique begins from § 12.5.4's observation that the columns (like the rows) of A are linearly independent because A is square and has full rank. From this beginning and the fact that $I_n = AA^{-1}$, it follows that $[A^{-1}]_{*1}$ represents[3] the one and only possible combination of A's columns which achieves $\mathbf{e}_1$, that $[A^{-1}]_{*2}$ represents the one and only possible combination of A's columns which achieves $\mathbf{e}_2$, and so on through $\mathbf{e}_n$. One could observe likewise respecting the independent rows of A. Either way, A^{-1} is unique. Moreover, no other $n \times n$ matrix $B \neq A^{-1}$ satisfies *either* requirement of (13.2)—that $BA = I_n$ or that $AB = I_n$—much less both.

It is not claimed that the matrix factors $G_>$ and $G_<$ themselves are unique, incidentally. On the contrary, many different pairs of matrix factors $G_>$ and $G_<$ can yield $A = G_> I_n G_<$, no less than that many different pairs of scalar factors $\gamma_>$ and $\gamma_<$ can yield $\alpha = \gamma_> 1 \gamma_<$. Though the Gauss-Jordan decomposition is a convenient means to $G_>$ and $G_<$, it is hardly the only means, and any proper $G_>$ and $G_<$ found by any means will serve so long as they satisfy (13.1). What are unique are not the factors but the A and A^{-1} they produce.

What of the degenerate $n \times n$ square matrix A', of rank $r < n$? Rank promotion is impossible as §§ 12.5.2 and 12.5.3 have shown, so in the sense of (13.2) such a matrix has no inverse; for, if it had, then A'^{-1} would by definition represent a row or column operation which impossibly promoted A' to the full rank $r = n$ of I_n. Indeed, in that it has no inverse such a degenerate matrix resembles the scalar 0, which has no reciprocal. Mathematical convention owns a special name for a square matrix which is degenerate and thus has no inverse; it calls it a *singular* matrix.

[3] The notation $[A^{-1}]_{*j}$ means "the jth column of A^{-1}." Refer to § 11.1.3.

And what of a rectangular matrix? Is it degenerate? Well, no, not exactly, not necessarily. The definitions of the present particular section are meant for square matrices; they do not neatly apply to nonsquare ones. Refer to §§ 12.5.3 and 12.5.4. However, appending the right number of null rows or columns to a nonsquare matrix does turn it into a degenerate square, in which case the preceding argument applies. See also §§ 12.5.5, 13.4 and 13.6.

13.2 The exactly determined linear system

Section 11.1.1 has shown how the single matrix equation

$$A\mathbf{x} = \mathbf{b} \tag{13.3}$$

concisely represents an entire simultaneous system of linear scalar equations. If the system has n scalar equations and n scalar unknowns, then the matrix A has square, $n \times n$ dimensionality. Furthermore, if the n scalar equations are independent of one another, then the rows of A are similarly independent, which gives A full rank and makes it invertible. Under these conditions, one can solve (13.3) and the corresponding system of linear scalar equations by left-multiplying (13.3) by the A^{-1} of (13.2) and (13.1) to reach the famous formula

$$\mathbf{x} = A^{-1}\mathbf{b}. \tag{13.4}$$

Inverting the square matrix A of scalar coefficients, (13.4) concisely solves a simultaneous system of n linear scalar equations in n scalar unknowns. It is the classic motivational result of matrix theory.

It has taken the book two long chapters to reach (13.4). If one omits first to prepare the theoretical ground sufficiently to support more advanced matrix work, then one can indeed reach (13.4) with rather less effort than the book has done.[4] As the chapter's introduction has observed, however, we

[4] For motivational reasons, introductory, tutorial linear algebra textbooks like [75] and [106] rightly yet invariably invert the general square matrix of full rank much earlier, reaching (13.4) with less effort. The deferred price the student pays for the simpler-seeming approach of the tutorials is twofold. First, the student fails to develop the Gauss-Jordan decomposition properly, instead learning the less elegant but easier to grasp "row echelon form" of "Gaussian elimination" [75, chapter 1][106, § 1.2]—which makes good matrix-arithmetic drill but leaves the student imperfectly prepared when the time comes to study kernels and eigensolutions or to read and write matrix-handling computer code. Second, in the long run the tutorials save no effort, because the student still must at some point develop the theory underlying matrix rank and supporting each of the several coïncident properties of § 14.2. What the tutorials do is pedagogically necessary—it is how the

shall soon meet additional interesting applications of the matrix which in any case require the theoretical ground to have been prepared. Equation (13.4) is only the first fruit of the effort.

Where the inverse does not exist, where the square matrix A is singular, the rows of the matrix are linearly dependent, meaning that the corresponding system actually contains fewer than n useful scalar equations. Depending on the value of the driving vector $\mathbf{b}$, the superfluous equations either merely reproduce or flatly contradict information the other equations already supply. Either way, no unique solution to a linear system described by a singular square matrix is possible—though a good approximate solution is given by the *pseudoinverse* of § 13.6. In the language of § 12.5.5, the singular square matrix characterizes a system that is both underdetermined and overdetermined, and thus degenerate.

13.3 The kernel

If a matrix A has full column rank (§ 12.5.4), then the columns of A are linearly independent and

$$A\mathbf{x} = 0 \tag{13.5}$$

is impossible if $I_n\mathbf{x} \neq 0$. If the matrix however lacks full column rank then (13.5) is possible even if $I_n\mathbf{x} \neq 0$. In either case, any n-element $\mathbf{x}$ (including $\mathbf{x} = 0$) that satisfies (13.5) belongs to the *kernel* of A.

Let A be an $m \times n$ matrix of rank r. A second matrix,[5] A^K, minimally represents the kernel of A if and only if

- A^K has $n \times (n - r)$ dimensionality (which gives A^K tall rectangular form unless $r = 0$),

writer first learned the matrix and probably how the reader first learned it, too—but it is appropriate to a tutorial, not to a study reference like this book.

In this book, where derivations prevail, the proper place to invert the general square matrix of full rank is here. Indeed, the inversion here goes smoothly, because chapters 11 and 12 have laid under it a firm foundation upon which—and supplied it the right tools with which—to work.

[5] The conventional mathematical notation for the kernel of A is $\ker\{A\}$, $\mathrm{null}\{A\}$ or something nearly resembling one of the two—the notation seems to vary from editor to editor—which technically represent the kernel space itself, as opposed to the notation A^K which represents a matrix whose columns address the kernel space. This book deëmphasizes the distinction and prefers the kernel matrix notation A^K.

If we were really precise, we might write not A^K but $A^{K(n)}$ to match the $A^{-1(r)}$ of (11.49). The abbreviated notation A^K is probably clear enough for most practical purposes, though, and surely more comprehensible to those who do not happen to have read this particular book.

- A^K has full rank $n - r$ (that is, the columns of A^K are linearly independent, which gives A^K full column rank), and
- A^K satisfies the equation
$$AA^K = 0. \tag{13.6}$$

The $n-r$ independent columns of the kernel matrix A^K address the complete space $\mathbf{x} = A^K\mathbf{a}$ of vectors in the kernel, where the $(n - r)$-element vector $\mathbf{a}$ can have any value. In symbols,

$$A\mathbf{x} = A(A^K\mathbf{a}) = (AA^K)\mathbf{a} = 0.$$

The definition does not pretend that the kernel matrix A^K is unique. Except when A has full column rank the kernel matrix is not unique; there are infinitely many kernel matrices A^K to choose from for a given matrix A. What is unique is not the kernel matrix but rather the space its columns address, and it is the latter space rather than A^K as such that is technically the kernel (if you forget and call A^K "a kernel," though, you'll be all right).

The *Gauss-Jordan kernel formula*[6]

$$A^K = S^{-1}K^{-1}H_rI_{n-r} = G_<^{-1}H_rI_{n-r} \tag{13.7}$$

gives a complete kernel A^K of A, where S^{-1}, K^{-1} and $G_<^{-1}$ are the factors their respective symbols indicate of the Gauss-Jordan decomposition's complementary form (12.3) and H_r is the shift operator of § 11.9. Section 13.3.1 derives the formula, next.

13.3.1 The Gauss-Jordan kernel formula

To derive (13.7) is not easy. It begins from the statement of the linear system

$$A\mathbf{x} = \mathbf{b}, \text{ where } \mathbf{b} = 0 \text{ or } r = m, \text{ or both;} \tag{13.8}$$

and where $\mathbf{b}$ and $\mathbf{x}$ are respectively m- and n-element vectors and A is an $m \times n$ matrix of rank r. This statement is broader than (13.5) requires but it serves § 13.4, too; so, for the moment, for generality's sake, we leave $\mathbf{b}$ unspecified but by the given proviso. Gauss-Jordan factoring A, by successive steps,

$$\begin{aligned}
G_>I_rKS\mathbf{x} &= \mathbf{b}, \\
I_rKS\mathbf{x} &= G_>^{-1}\mathbf{b}, \\
I_r(K - I)S\mathbf{x} + I_rS\mathbf{x} &= G_>^{-1}\mathbf{b}.
\end{aligned}$$

[6] The name *Gauss-Jordan kernel formula* is not standard as far as the writer is aware, but we would like a name for (13.7). This name seems as fitting as any.

Applying an identity from Table 12.2 on page 396,

$$I_rK(I_n - I_r)S\mathbf{x} + I_rS\mathbf{x} = G_>^{-1}\mathbf{b}.$$

Rearranging terms,

$$I_rS\mathbf{x} = G_>^{-1}\mathbf{b} - I_rK(I_n - I_r)S\mathbf{x}. \tag{13.9}$$

Equation (13.9) is interesting. It has $S\mathbf{x}$ on both sides, where $S\mathbf{x}$ is the vector $\mathbf{x}$ with elements reordered in some particular way. The equation has however on the left only $I_rS\mathbf{x}$, which is the first r elements of $S\mathbf{x}$; and on the right only $(I_n - I_r)S\mathbf{x}$, which is the remaining $n - r$ elements.[7] No element of $S\mathbf{x}$ appears on both sides. Naturally this is no accident; we have (probably after some trial and error not recorded here) planned the steps leading to (13.9) to achieve precisely this effect. Equation (13.9) implies *that one can choose the last $n - r$ elements of $S\mathbf{x}$ freely, but that the choice then determines the first r elements.*

The implication is significant. To express the implication more clearly we can rewrite (13.9) in the improved form

$$\begin{aligned}
\mathbf{f} &= G_>^{-1}\mathbf{b} - I_rKH_r\mathbf{a},\\
S\mathbf{x} &= \begin{bmatrix}\mathbf{f}\\ \mathbf{a}\end{bmatrix} = \mathbf{f} + H_r\mathbf{a},\\
\mathbf{f} &\equiv I_rS\mathbf{x},\\
\mathbf{a} &\equiv H_{-r}(I_n - I_r)S\mathbf{x},
\end{aligned} \tag{13.10}$$

where $\mathbf{a}$ represents the $n - r$ free elements of $S\mathbf{x}$ and $\mathbf{f}$ represents the r dependent elements. This makes $\mathbf{f}$ and thereby also $\mathbf{x}$ functions of the free parameter $\mathbf{a}$ and the driving vector $\mathbf{b}$:

$$\begin{aligned}
\mathbf{f}(\mathbf{a},\mathbf{b}) &= G_>^{-1}\mathbf{b} - I_rKH_r\mathbf{a},\\
S\mathbf{x}(\mathbf{a},\mathbf{b}) &= \begin{bmatrix}\mathbf{f}(\mathbf{a},\mathbf{b})\\ \mathbf{a}\end{bmatrix} = \mathbf{f}(\mathbf{a},\mathbf{b}) + H_r\mathbf{a}.
\end{aligned} \tag{13.11}$$

If $\mathbf{b} = 0$ as (13.5) requires, then

$$\begin{aligned}
\mathbf{f}(\mathbf{a},0) &= -I_rKH_r\mathbf{a},\\
S\mathbf{x}(\mathbf{a},0) &= \begin{bmatrix}\mathbf{f}(\mathbf{a},0)\\ \mathbf{a}\end{bmatrix} = \mathbf{f}(\mathbf{a},0) + H_r\mathbf{a}.
\end{aligned}$$

[7] Notice how we now associate the factor $(I_n - I_r)$ rightward as a row truncator, though it had first entered acting leftward as a column truncator. The flexibility to reassociate operators in such a way is one of many good reasons chapters 11 and 12 have gone to such considerable trouble to develop the basic theory of the matrix.

Substituting the first line into the second,

$$S\mathbf{x}(\mathbf{a}, 0) = (I - I_r K)H_r\mathbf{a}. \tag{13.12}$$

In the event that $\mathbf{a} = \mathbf{e}_j$, where $1 \le j \le n - r$,

$$S\mathbf{x}(\mathbf{e}_j, 0) = (I - I_r K)H_r\mathbf{e}_j.$$

For all the $\mathbf{e}_j$ at once,

$$S\mathbf{x}(I_{n-r}, 0) = (I - I_r K)H_r I_{n-r}.$$

But if all the $\mathbf{e}_j$ at once—that is, if all the columns of I_{n-r}—exactly address the domain of $\mathbf{a}$, then the columns of $\mathbf{x}(I_{n-r}, 0)$ likewise exactly address the range of $\mathbf{x}(\mathbf{a}, 0)$. Equation (13.6) has already named this range A^K, by which[8]

$$SA^K = (I - I_r K)H_r I_{n-r}. \tag{13.13}$$

Left-multiplying by

$$S^{-1} = S^* = S^T \tag{13.14}$$

produces the alternate kernel formula

$$A^K = S^{-1}(I - I_r K)H_r I_{n-r}. \tag{13.15}$$

[8] These are difficult steps. How does one justify replacing $\mathbf{a}$ by $\mathbf{e}_j$, then $\mathbf{e}_j$ by I_{n-r}, then $\mathbf{x}$ by A^K? One justifies them in that the columns of I_{n-r} are the several $\mathbf{e}_j$, of which any $(n-r)$-element vector $\mathbf{a}$ can be constructed as the linear combination

$$\mathbf{a} = I_{n-r}\mathbf{a} = \left[\begin{array}{ccccc} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 & \cdots & \mathbf{e}_{n-r} \end{array}\right]\mathbf{a} = \sum_{j=1}^{n-r} a_j\mathbf{e}_j$$

weighted by the elements of $\mathbf{a}$. Seen from one perspective, this seems trivial; from another perspective, baffling; until one grasps what is really going on here.

The idea is that if we can solve the problem for each elementary vector $\mathbf{e}_j$—that is, in aggregate, if we can solve the problem for the identity matrix I_{n-r}—then we shall implicitly have solved it for every $\mathbf{a}$ because $\mathbf{a}$ is a weighted combination of the $\mathbf{e}_j$ and the whole problem is linear. The solution

$$\mathbf{x} = A^K\mathbf{a}$$

for a given choice of $\mathbf{a}$ becomes a weighted combination of the solutions for each $\mathbf{e}_j$, with the elements of $\mathbf{a}$ again as the weights. And what are the solutions for each $\mathbf{e}_j$? Answer: the corresponding columns of A^K, which by definition are the independent values of $\mathbf{x}$ that cause $\mathbf{b} = 0$.

The alternate kernel formula (13.15) is correct but not as simple as it could be. By the identity (11.76), eqn. (13.13) is

$$\begin{aligned} SA^K &= (I - I_rK)(I_n - I_r)H_r \\ &= [(I_n - I_r) - I_rK(I_n - I_r)]H_r \\ &= [(I_n - I_r) - (K - I)]H_r, \end{aligned} \tag{13.16}$$

where we have used Table 12.2 again in the last step. How to proceed symbolically from (13.16) is not obvious, but if one sketches the matrices of (13.16) schematically with a pencil, and if one remembers that K^{-1} is just K with elements off the main diagonal negated, then it appears that

$$SA^K = K^{-1}H_rI_{n-r}. \tag{13.17}$$

The appearance is not entirely convincing,[9] but (13.17) though unproven still helps because it posits a hypothesis toward which to target the analysis.

Two variations on the identities of Table 12.2 also help. First, from the identity that

$$\frac{K + K^{-1}}{2} = I,$$

we have that

$$K - I = I - K^{-1}. \tag{13.18}$$

Second, right-multiplying by I_r the identity that

$$I_rK^{-1}(I_n - I_r) = K^{-1} - I$$

and canceling terms, we have that

$$K^{-1}I_r = I_r \tag{13.19}$$

(which actually is pretty obvious if you think about it, since all of K's interesting content lies by construction right of its rth column). Now we have enough to go on with. Substituting (13.18) and (13.19) into (13.16) yields that

$$SA^K = [(I_n - K^{-1}I_r) - (I - K^{-1})]H_r.$$

Adding $0 = K^{-1}I_nH_r - K^{-1}I_nH_r$ and rearranging terms,

$$SA^K = K^{-1}(I_n - I_r)H_r + [K^{-1} - K^{-1}I_n - I + I_n]H_r.$$

[9] Well, no, actually, the appearance pretty much is entirely convincing, but let us finish the proof symbolically nonetheless.

Factoring,

$$SA^K = K^{-1}(I_n - I_r)H_r + [(K^{-1} - I)(I - I_n)]H_r.$$

According to Table 12.2, the quantity in square brackets is zero, so

$$SA^K = K^{-1}(I_n - I_r)H_r,$$

which, considering that the identity (11.76) has that $(I_n - I_r)H_r = H_r I_{n-r}$, proves (13.17). The final step is to left-multiply (13.17) by $S^{-1} = S^* = S^T$, reaching (13.7) that was to be derived.

One would like to feel sure that the columns of (13.7)'s A^K actually addressed the whole kernel space of A rather than only part. One would further like to feel sure that A^K had no redundant columns; that is, that it had full rank. Moreover, the definition of A^K in the section's introduction demands both of these features. In general such features would be hard to establish, but here the factors conveniently are Gauss-Jordan factors. Regarding the whole kernel space, A^K addresses it because A^K comes from all $\mathbf{a}$. Regarding redundancy, A^K lacks it because SA^K lacks it, and SA^K lacks it because according to (13.13) the last rows of SA^K are $H_r I_{n-r}$. So, in fact, (13.7) has both features and does fit the definition.

13.3.2 Converting between kernel matrices

If C is a reversible $(n-r)\times(n-r)$ operator by which we right-multiply (13.6), then the matrix

$$A'^K = A^K C \tag{13.20}$$

like A^K evidently represents the kernel of A:

$$AA'^K = A(A^K C) = (AA^K)C = 0.$$

Indeed this makes sense: because the columns of $A^K C$ address the same space the columns of A^K address, the two matrices necessarily represent the same underlying kernel. Moreover, *some* C exists to convert A^K into every alternate kernel matrix A'^K of A. We know this because § 12.4 lets one replace the columns of A^K with those of A'^K, reversibly, one column at a time, without altering the space addressed. (It might not let one replace the columns in sequence, but if out of sequence then a reversible permutation at the end corrects the order. Refer to §§ 12.5.1 and 12.5.2 for the pattern by which this is done.)

The orthonormalizing column operator R^{-1} of (13.56) below incidentally tends to make a good choice for C.

13.3.3 The degree of freedom

A slightly vague but extraordinarily useful concept has emerged in this section, worth pausing briefly to appreciate. The concept is the concept of the *degree of freedom.*

A degree of freedom is a parameter one remains free to determine within some continuous domain. For example, Napoleon's artillerist[10] might have enjoyed as many as six degrees of freedom in firing a cannonball: two in where he chose to set up his cannon (one degree in north-south position, one in east-west); two in aim (azimuth and elevation); one in muzzle velocity (as governed by the quantity of gunpowder used to propel the ball); and one in time. A seventh potential degree of freedom, the height from which the artillerist fires, is of course restricted by the lay of the land: the artillerist can fire from a high place only if the place he has chosen to fire from happens to be up on a hill, for Napoleon had no flying cannon. Yet even among the six remaining degrees of freedom, the artillerist might find some impractical to exercise. The artillerist probably preloads the cannon always with a standard charge of gunpowder because, when he finds his target in the field, he cannot spare the time to unload the cannon and alter the charge: this costs one degree of freedom. Likewise, the artillerist must limber up the cannon and hitch it to a horse to shift it to better ground; for this too he cannot spare time in the heat of battle: this costs two degrees. And Napoleon might yell, "Fire!" canceling the time degree as well. Two degrees of freedom remain to the artillerist; but, since exactly two degrees are needed to hit some particular target on the battlefield, the two are enough.

Now consider what happens if the artillerist loses one of his last two remaining degrees of freedom. Maybe the cannon's carriage wheel is broken and the artillerist can no longer turn the cannon; that is, he can still choose firing elevation but no longer azimuth. In such a strait to hit some particular target on the battlefield, the artillerist needs somehow to recover another degree of freedom, for he needs two but has only one. If he disregards Napoleon's order, "Fire!" (maybe not a wise thing to do, but, anyway, ...) and waits for the target to traverse the cannon's fixed line of fire, then he can still hope to hit even with the broken carriage wheel; for could he choose neither azimuth nor the moment to fire, then he would almost surely miss.

Some apparent degrees of freedom are not real. For example, muzzle velocity gives the artillerist little control firing elevation does not also give.

[10] The author, who has never fired an artillery piece (unless an arrow from a Boy Scout bow counts), invites any real artillerist among the readership to write in to improve the example.

Other degrees of freedom are nonlinear in effect: a certain firing elevation gives maximum range; nearer targets can be hit by firing either higher or lower at the artillerist's discretion. On the other hand, too much gunpowder might break the cannon.

All of this is hard to generalize in unambiguous mathematical terms, but the count of the degrees of freedom in a system is of high conceptual importance to the engineer nonetheless. Basically, the count captures the idea that to control n output variables of some system takes at least n independent input variables. The n may possibly for various reasons still not suffice—it might be wise in some cases to allow $n + 1$ or $n + 2$—but in no event will fewer than n do. Engineers of all kinds think in this way: an aeronautical engineer knows in advance that an airplane needs at least n ailerons, rudders and other control surfaces for the pilot adequately to control the airplane; an electrical engineer knows in advance that a circuit needs at least n potentiometers for the technician adequately to tune the circuit; and so on.

In geometry, a line brings a single degree of freedom. A plane brings two. A point brings none. If the line bends and turns like a mountain road, it still brings a single degree of freedom. And if the road reaches an intersection? Answer: still one degree. A degree of freedom has some continuous nature, not merely a discrete choice to turn left or right. On the other hand, a swimmer in a swimming pool enjoys three degrees of freedom (up-down, north-south, east-west) even though his domain in any of the three is limited to the small volume of the pool. The driver on the mountain road cannot claim a second degree of freedom at the mountain intersection (he can indeed claim a choice, but the choice being discrete lacks the proper character of a degree of freedom), but he might plausibly claim a second degree of freedom upon reaching the city, where the web or grid of streets is dense enough to approximate access to any point on the city's surface. Just how many streets it takes to turn the driver's "line" experience into a "plane" experience is a matter for the mathematician's discretion.

Reviewing (13.11), we find $n-r$ degrees of freedom in the general underdetermined linear system, represented by the $n-r$ free elements of $\mathbf{a}$. If the underdetermined system is not also overdetermined, if it is nondegenerate such that $r = m$, then it is guaranteed to have a family of solutions $\mathbf{x}$. This family is the topic of the next section.

13.4 The nonoverdetermined linear system

The exactly determined linear system of § 13.2 is common, but also common is the more general, nonoverdetermined linear system

$$A\mathbf{x} = \mathbf{b}, \tag{13.21}$$

in which $\mathbf{b}$ is a known, m-element vector; $\mathbf{x}$ is an unknown, n-element vector; and A is a square or broad, $m \times n$ matrix of full row rank (§ 12.5.4)

$$r = m \leq n. \tag{13.22}$$

Except in the exactly determined edge case $r = m = n$ of § 13.2, the nonoverdetermined linear system has no unique solution but rather a family of solutions. This section delineates the family.

13.4.1 Particular and homogeneous solutions

The nonoverdetermined linear system (13.21) by definition admits more than one solution $\mathbf{x}$ for a given driving vector $\mathbf{b}$. Such a system is hard to solve all at once, though, so we prefer to split the system as

$$\begin{aligned} A\mathbf{x}_1 &= \mathbf{b}, \\ A(A^K\mathbf{a}) &= 0, \\ \mathbf{x} &= \mathbf{x}_1 + A^K\mathbf{a}, \end{aligned} \tag{13.23}$$

which, when the second line is added to the first and the third is substituted, makes the whole form (13.21). Splitting the system does not change it but does let us treat the system's first and second lines in (13.23) separately. In the split form, the symbol $\mathbf{x}_1$ represents any one n-element vector that happens to satisfy the form's first line—many are possible; the mathematician just picks one—and is called *a particular solution* of (13.21). The $(n - r)$-element vector $\mathbf{a}$ remains unspecified, whereupon $A^K\mathbf{a}$ represents the complete family of n-element vectors that satisfy the form's second line. The family of vectors expressible as $A^K\mathbf{a}$ is called *the homogeneous solution* of (13.21).

Notice the italicized articles *a* and *the.*

The Gauss-Jordan kernel formula (13.7) has given us A^K and thereby the homogeneous solution, which renders the analysis of (13.21) already half done. To complete the analysis, it remains in § 13.4.2 to find a particular solution.

13.4.2 A particular solution

Any particular solution will do. Equation (13.11) has that

$$\begin{aligned}\mathbf{f}(\mathbf{a},\mathbf{b}) &= G_>^{-1}\mathbf{b} - I_r K H_r \mathbf{a},\\ (S)\left[\mathbf{x}_1(\mathbf{a},\mathbf{b}) + A^K\mathbf{a}\right] &= \begin{bmatrix} \mathbf{f}(\mathbf{a},\mathbf{b}) \\ \mathbf{a} \end{bmatrix} = \mathbf{f}(\mathbf{a},\mathbf{b}) + H_r\mathbf{a},\end{aligned}$$

where we have substituted the last line of (13.23) for $\mathbf{x}$. This holds for any $\mathbf{a}$ and $\mathbf{b}$. We are not free to choose the driving vector $\mathbf{b}$, but since we need only one particular solution, $\mathbf{a}$ can be anything we want. Why not

$$\mathbf{a} = 0?$$

Then

$$\begin{aligned}\mathbf{f}(0,\mathbf{b}) &= G_>^{-1}\mathbf{b},\\ S\mathbf{x}_1(0,\mathbf{b}) &= \begin{bmatrix} \mathbf{f}(0,\mathbf{b}) \\ 0 \end{bmatrix} = \mathbf{f}(0,\mathbf{b}).\end{aligned}$$

That is,

$$\mathbf{x}_1 = S^{-1}G_>^{-1}\mathbf{b}. \tag{13.24}$$

13.4.3 The general solution

Assembling (13.7), (13.23) and (13.24) in light of (12.3) yields the general solution

$$\mathbf{x} = S^{-1}(G_>^{-1}\mathbf{b} + K^{-1}H_r I_{n-r}\mathbf{a}) \tag{13.25}$$

to the nonoverdetermined linear system (13.21).

In exact arithmetic (13.25) solves the nonoverdetermined linear system in theory exactly. Therefore (13.25) properly concludes the section. Nevertheless, one should like to add a significant practical observation regarding *inexact arithmetic* as follows.

Practical calculations are usually done in inexact arithmetic insofar as they are done in the limited precision of a computer's floating-point registers. Exceptions are possible—exact-arithmetic libraries are available for a programmer to call—but exact-arithmetic libraries are slow and memory-intensive and, for this reason among others, are only occasionally used in practice. When they are not used, compounded rounding error in a floating-point register's last bit of mantissa[11] eventually disrupts (13.25) for matrices larger than some moderately large size. Avoiding unduly small pivots early in the Gauss-Jordan extends (13.25)'s reach to larger matrices, and for yet larger matrices a bewildering variety of more sophisticated techniques exists to mitigate the problem, which can be vexing because the problem arises even when the matrix A is exactly known.

Equation (13.25) is thus useful and correct, but one should at least be aware that it can in practice lose floating-point accuracy when the matrix it attacks grows too large. (It can also lose accuracy when the matrix's rows are almost dependent, but that is more the fault of the matrix than of the formula. See § 14.8, which addresses a related problem.)

13.5 The residual

Equations (13.2) and (13.4) solve the exactly determined linear system $A\mathbf{x} = \mathbf{b}$. Equation (13.25) broadens the solution to include the nonoverdetermined linear system. None of those equations however can handle the overdetermined linear system, because for general $\mathbf{b}$ the overdetermined linear system

$$A\mathbf{x} \approx \mathbf{b} \tag{13.26}$$

has no exact solution. (See § 12.5.5 for the definitions of *underdetermined, overdetermined,* etc.)

One is tempted to declare the overdetermined system uninteresting because it has no solution and to leave the matter there, but this would be a serious mistake. In fact the overdetermined system is especially interesting, and the more so because it arises so frequently in applications. One seldom

[11] What is a *mantissa?* Illustration: in the number 1.65×10^6, the mantissa is 1.65. However, computers do it in binary rather than in decimal, typically with fifty-two (0x34) stored bits of mantissa not counting the leading bit which, in binary, is not stored because it is always 1. (There exist implementational details like floating-point "denormals" which might seem pedantically to contradict the always-1 rule, but that is a computer-engineering technicality uninteresting in the context of the present discussion. What might be interesting in the present context is this: a standard double-precision floating-point representation has—besides fifty-two bits of mantissa—also eleven, 0xB, bits for an exponent and one bit for a sign. The smallest positive number representable without denormalization is $2^{-0\text{x}3\text{FE}}$; the largest is $0\text{x}1.\text{FFFF FFFF FFFF F} \times 2^{0\text{x}3\text{FF}}$, just less than $2^{0\text{x}400}$. If the code for a full $2^{0\text{x}400}$ is entered, then that is held to represent infinity. Similarly, the code for $2^{-0\text{x}3\text{FF}}$ represents zero. If you think that the code for $2^{-0\text{x}400}$ should instead represent zero, no such code can actually be entered, for the exponent's representation is offset by one; and there are many other details beyond the book's scope.)

trusts a minimal set of data for important measurements, yet extra data imply an overdetermined system. We need to develop the mathematics to handle the overdetermined system properly.

The quantity[12,13]

$$\mathbf{r}(\mathbf{x}) \equiv \mathbf{b} - A\mathbf{x} \tag{13.27}$$

measures how nearly some candidate solution $\mathbf{x}$ solves the system (13.26). We call this quantity the *residual,* and the smaller, the better. More precisely, the smaller the nonnegative real scalar

$$[\mathbf{r}(\mathbf{x})]^*[\mathbf{r}(\mathbf{x})] = \sum_i |r_i(\mathbf{x})|^2 \tag{13.28}$$

is, called the *squared residual norm,* the more favorably we regard the candidate solution $\mathbf{x}$.

13.6 The Moore-Penrose pseudoinverse and the least-squares problem

A typical problem is to fit a straight line to some data. For example, suppose that we are building-construction contractors with a unionized work force, whose labor union can supply additional, fully trained labor on demand. Suppose further that we are contracted to build a long freeway and have been adding workers to the job in recent weeks to speed construction. On Saturday morning at the end of the second week, we gather and plot the production data on the left of Fig. 13.1. If u_i and b_i respectively represent the number of workers and the length of freeway completed during week i, then we can fit a straight line $b = \sigma u + \gamma$ to the measured production data such that

$$\begin{bmatrix} u_1 & 1 \\ u_2 & 1 \end{bmatrix} \begin{bmatrix} \sigma \\ \gamma \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

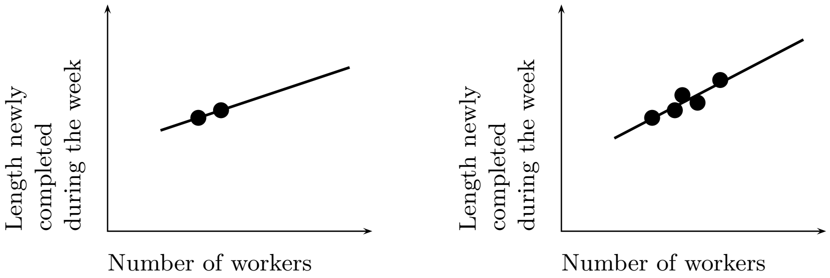
inverting the matrix per §§ 13.1 and 13.2 to solve for $\mathbf{x} \equiv [\sigma\ \gamma]^T$, in the hope that the resulting line will predict future production accurately.

The foregoing is all mathematically irreproachable. By the fifth Saturday however we shall have gathered more production data, plotted on the figure's right, to which we should like to fit a better line to predict production more accurately. The added data present a problem. Statistically, the added

[12] Alas, the alphabet has only so many letters (see appendix B). The $\mathbf{r}$ here is unrelated to matrix rank r.

[13] This is as [171] defines it. Some authors [127] however prefer to define $\mathbf{r}(\mathbf{x}) \equiv A\mathbf{x} - \mathbf{b}$, instead.

Figure 13.1: Fitting a line to measured data.



data are welcome, but geometrically we need only two points to specify a line; what are we to do with the other three? The five points together overdetermine the linear system

$$\begin{bmatrix} u_1 & 1 \\ u_2 & 1 \\ u_3 & 1 \\ u_4 & 1 \\ u_5 & 1 \end{bmatrix} \begin{bmatrix} \sigma \\ \gamma \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix}.$$

There is no way to draw a single straight line $b = \sigma u + \gamma$ exactly through all five, for in placing the line we enjoy only two degrees of freedom.[14]

The proper approach is to draw among the data points a single straight line that misses the points as narrowly as possible. More precisely, the proper approach chooses parameters σ and γ to minimize the squared residual norm $[\mathbf{r}(\mathbf{x})]^*[\mathbf{r}(\mathbf{x})]$ of § 13.5, given that

$$A = \begin{bmatrix} u_1 & 1 \\ u_2 & 1 \\ u_3 & 1 \\ u_4 & 1 \\ u_5 & 1 \\ & \vdots \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} \sigma \\ \gamma \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ \vdots \end{bmatrix}.$$

[14] Section 13.3.3 characterized a line as enjoying only one degree of freedom. Why now two? The answer is that § 13.3.3 discussed travel along a line rather than placement of a line as here. Though both involve lines, they differ as driving an automobile differs from washing one. Do not let this confuse you.

Such parameters constitute a *least-squares* solution.

The matrix A in the example has two columns, data marching on the left, all ones on the right. This is a typical structure for A, but in general any matrix A with any number of columns of any content might arise (because there were more than two relevant variables or because some data merited heavier weight than others, among many further reasons). Whatever matrix A might arise from whatever source, this section attacks the difficult but important problem of approximating optimally a solution to the general, possibly unsolvable linear system (13.26), $A\mathbf{x} \approx \mathbf{b}$.

13.6.1 Least squares in the real domain

The least-squares problem is simplest when the matrix A enjoys full column rank and no complex numbers are involved. In this case, we seek to minimize the squared residual norm

$$\begin{aligned}
[\mathbf{r}(\mathbf{x})]^T[\mathbf{r}(\mathbf{x})] &= (\mathbf{b} - A\mathbf{x})^T(\mathbf{b} - A\mathbf{x}) \\
&= \mathbf{x}^T A^T A\mathbf{x} + \mathbf{b}^T\mathbf{b} - \left(\mathbf{x}^T A^T\mathbf{b} + \mathbf{b}^T A\mathbf{x}\right) \\
&= \mathbf{x}^T A^T A\mathbf{x} + \mathbf{b}^T\mathbf{b} - 2\mathbf{x}^T A^T\mathbf{b} \\
&= \mathbf{x}^T A^T (A\mathbf{x} - 2\mathbf{b}) + \mathbf{b}^T\mathbf{b},
\end{aligned}$$

in which the transpose is used interchangeably for the adjoint because all the numbers involved happen to be real. The norm is minimized where

$$\frac{d}{d\mathbf{x}}\left(\mathbf{r}^T\mathbf{r}\right) = 0$$

(in which $d/d\mathbf{x}$ is the Jacobian operator of § 11.10). A requirement that

$$\frac{d}{d\mathbf{x}}\left[\mathbf{x}^T A^T (A\mathbf{x} - 2\mathbf{b}) + \mathbf{b}^T\mathbf{b}\right] = 0$$

comes of combining the last two equations. Differentiating by the Jacobian product rule (11.79) yields the equation

$$\mathbf{x}^T A^T A + \left[A^T (A\mathbf{x} - 2\mathbf{b})\right]^T = 0;$$

or, after transposing the equation, rearranging terms and dividing by 2, the simplified equation

$$A^T A\mathbf{x} = A^T\mathbf{b}.$$

Assuming (as warranted by § 13.6.2, next) that the $n \times n$ square matrix $A^T A$ is invertible, the simplified equation implies the approximate but optimal least-squares solution

$$\mathbf{x} = \left(A^T A\right)^{-1} A^T \mathbf{b} \tag{13.29}$$

to the unsolvable linear system (13.26) in the restricted but quite typical case that A and $\mathbf{b}$ are real and A has full column rank.

Equation (13.29) plots the line on Fig. 13.1's right. As the reader can see, the line does not pass through all the points, for no line can; but it does pass pretty convincingly nearly among them. In fact it passes optimally nearly among them. No line can pass more nearly, in the squared-residual norm sense of (13.28).[15]

[15] Here is a nice example of the use of the mathematical adjective *optimal* in its adverbial form. "Optimal" means "best." Many problems in applied mathematics involve discovering the best of something. What constitutes the best however can be a matter of judgment, even of dispute. We will leave to the philosopher and the theologian the important question of what constitutes objective good, for applied mathematics is a poor guide to such mysteries. The role of applied mathematics is to construct suitable models to calculate quantities needed to achieve some definite good; its role is not, usually, to identify the good as good in the first place.

One generally establishes mathematical optimality by some suitable, nonnegative, real *cost function* or *metric,* and the less, the better. Strictly speaking, the mathematics cannot tell us which metric to use, but where no other consideration prevails the applied mathematician tends to choose the metric that best simplifies the mathematics at hand—and, really, that is about as good a way to choose a metric as any. The metric (13.28) is so chosen.

"But," comes the objection, "what if some more complicated metric is better?"

Well, if the other metric really, objectively is better, then one should probably use it. In general however the mathematical question is: what does one mean by "better?" Better by which metric? Each metric is better according to itself. This is where the mathematician's experience, taste and judgment come in.

In the present section's example, too much labor on the freeway job might actually slow construction rather than speed it. One could therefore seek to fit not a line but some downward-turning curve to the data. Mathematics offers many downward-turning curves. A circle, maybe? Not likely. An experienced mathematician would probably reject the circle on the aesthetic yet practical ground that the parabola $b = \alpha u^2 + \sigma u + \gamma$ lends itself to easier analysis. Yet even fitting a mere straight line offers choices. One might fit the line to the points (b_i, u_i) or $(\ln u_i, \ln b_i)$ rather than to the points (u_i, b_i). The three resulting lines differ subtly. They predict production differently. The adjective "optimal" alone evidently does not always tell us all we need to know.

Section 6.3 offers a choice between averages that resembles in spirit this footnote's choice between metrics.

13.6.2 The invertibility of A^*A

Section 13.6.1 has assumed correctly but unwarrantedly that the product A^TA were invertible for real A of full column rank. For real A, it happens that $A^T = A^*$, so it only broadens the same assumption to suppose that the product A^*A were invertible for complex A of full column rank.[16] This subsection warrants the latter assumption, thereby incidentally also warranting the former.

Let A be a complex, $m \times n$ matrix of full column rank $r = n \le m$. Suppose falsely that A^*A were not invertible but singular. Since the product A^*A is a square, $n \times n$ matrix, this is to suppose (§ 13.1) that the product's rank $r' < n$ were less than full, implying (§ 12.5.4) that its columns (as its rows) depended on one another. This would mean that there existed a nonzero, n-element vector $\mathbf{u}$ for which

$$A^*A\mathbf{u} = 0, \quad I_n\mathbf{u} \neq 0.$$

Left-multiplying by $\mathbf{u}^*$ would give that

$$\mathbf{u}^*A^*A\mathbf{u} = 0, \quad I_n\mathbf{u} \neq 0,$$

or in other words that

$$(A\mathbf{u})^*(A\mathbf{u}) = \sum_{i=1}^{n} |[A\mathbf{u}]_i|^2 = 0, \quad I_n\mathbf{u} \neq 0.$$

But this could only be so if

$$A\mathbf{u} = 0, \quad I_n\mathbf{u} \neq 0,$$

impossible when the columns of A are independent. The contradiction proves false the assumption which gave rise to it. The false assumption: that A^*A were singular.

Thus, *the $n \times n$ product A^*A is invertible for any tall or square, $m \times n$ matrix A of full column rank $r = n \le m$.*

13.6.3 Positive definiteness

An $n \times n$ matrix C is *positive definite* if and only if

$$\Im(\mathbf{u}^*C\mathbf{u}) = 0 \text{ and } \Re(\mathbf{u}^*C\mathbf{u}) > 0 \text{ for all } I_n\mathbf{u} \neq 0. \tag{13.30}$$

[16] Notice that if A is tall, then A^*A is a compact, $n \times n$ square, whereas AA^* is a big, $m \times m$ square. It is the compact square that concerns this section. The big square is not very interesting and in any case is not invertible.

As in § 13.6.2, here also when a matrix A has full column rank $r = n \le m$ the product $\mathbf{u}^*A^*A\mathbf{u} = (A\mathbf{u})^*(A\mathbf{u})$ is real and positive for all nonzero, n-element vectors $\mathbf{u}$. Thus per (13.30) *the product A^*A is positive definite for any matrix A of full column rank.*

An $n \times n$ matrix C is *nonnegative definite* if and only if

$$\Im(\mathbf{u}^*C\mathbf{u}) = 0 \text{ and } \Re(\mathbf{u}^*C\mathbf{u}) \ge 0 \text{ for all } \mathbf{u}. \tag{13.31}$$

By reasoning like the last paragraph's, *the product A^*A is nonnegative definite for any matrix A whatsoever.*

Such definitions might seem opaque, but their sense is that a positive definite operator never reverses the thing it operates on, that the product $A\mathbf{u}$ points more in the direction of $\mathbf{u}$ than of $-\mathbf{u}$. Section 13.8 explains further. A positive definite operator resembles a positive scalar in this sense.

13.6.4 The Moore-Penrose pseudoinverse

Not every $m \times n$ matrix A enjoys full rank. According to (12.17), however, every $m \times n$ matrix A of rank r can be factored into a product[17]

$$A = BC$$

of an $m \times r$ tall or square matrix B and an $r \times n$ broad or square matrix C, both of which factors themselves enjoy full rank r. (If A happens to have full row or column rank, then one can just choose $B = I_m$ or $C = I_n$; but even if A lacks full rank, the Gauss-Jordan decomposition of eqn. 12.2 finds at least the full-rank factorization $B = G_>I_r$, $C = I_rG_<$.) This being so, a conjecture seems warranted. Suppose that, inspired by (13.29), we manipulated (13.26) by the successive steps

$$\begin{aligned}
A\mathbf{x} &\approx \mathbf{b}, \\
BC\mathbf{x} &\approx \mathbf{b}, \\
(B^*B)^{-1}B^*BC\mathbf{x} &\approx (B^*B)^{-1}B^*\mathbf{b}, \\
C\mathbf{x} &\approx (B^*B)^{-1}B^*\mathbf{b}.
\end{aligned}$$

Then suppose that we changed

$$C^*\mathbf{u} \leftarrow \mathbf{x},$$

[17] This subsection uses the symbols B and $\mathbf{b}$ for unrelated purposes, which is unfortunate but conventional. See footnote 12.

thus restricting $\mathbf{x}$ to the space addressed by the independent columns of C^*. Continuing,

$$\begin{aligned} CC^*\mathbf{u} &\approx (B^*B)^{-1}B^*\mathbf{b}, \\ \mathbf{u} &\approx (CC^*)^{-1}(B^*B)^{-1}B^*\mathbf{b}. \end{aligned}$$

Changing the variable back and (because we are conjecturing and can do as we like), altering the "$\approx$" sign to "=,"

$$\mathbf{x} = C^*(CC^*)^{-1}(B^*B)^{-1}B^*\mathbf{b}. \tag{13.32}$$

Equation (13.32) has a pleasingly symmetrical form, and we know from § 13.6.2 at least that the two matrices it tries to invert are invertible. So here is our conjecture:

- no $\mathbf{x}$ enjoys a smaller squared residual norm $\mathbf{r}^*\mathbf{r}$ than the $\mathbf{x}$ of (13.32) does; and
- among all $\mathbf{x}$ that enjoy the same, minimal squared residual norm, the $\mathbf{x}$ of (13.32) is strictly least in magnitude.

The conjecture is bold, but if you think about it in the right way it is not unwarranted under the circumstance. After all, (13.32) does resemble (13.29), the latter of which admittedly requires real A of full column rank but does minimize the residual when its requirements are met; and, even if there were more than one $\mathbf{x}$ which minimized the residual, one of them might be smaller than the others: why not the $\mathbf{x}$ of (13.32)? One can but investigate.

The first point of the conjecture is symbolized

$$\mathbf{r}^*(\mathbf{x})\mathbf{r}(\mathbf{x}) \le \mathbf{r}^*(\mathbf{x}+\Delta\mathbf{x})\mathbf{r}(\mathbf{x}+\Delta\mathbf{x}),$$

where $\Delta\mathbf{x}$ represents the deviation, whether small, moderate or large, of some alternate $\mathbf{x}$ from the $\mathbf{x}$ of (13.32). According to (13.27), this is

$$[\mathbf{b} - A\mathbf{x}]^*[\mathbf{b} - A\mathbf{x}] \le [\mathbf{b} - (A)(\mathbf{x}+\Delta\mathbf{x})]^*[\mathbf{b} - (A)(\mathbf{x}+\Delta\mathbf{x})].$$

Reorganizing,

$$[\mathbf{b} - A\mathbf{x}]^*[\mathbf{b} - A\mathbf{x}] \le [(\mathbf{b} - A\mathbf{x}) - A\,\Delta\mathbf{x}]^*[(\mathbf{b} - A\mathbf{x}) - A\,\Delta\mathbf{x}].$$

Distributing factors and canceling like terms,

$$0 \le -\Delta\mathbf{x}^*\,A^*(\mathbf{b} - A\mathbf{x}) - (\mathbf{b} - A\mathbf{x})^*A\,\Delta\mathbf{x} + \Delta\mathbf{x}^*\,A^*A\,\Delta\mathbf{x}.$$

But according to (13.32) and the full-rank factorization $A = BC$,

$$\begin{aligned} A^*(\mathbf{b} - A\mathbf{x}) &= A^*\mathbf{b} - A^*A\mathbf{x} \\ &= [C^*B^*][\mathbf{b}] - [C^*B^*][BC][C^*(CC^*)^{-1}(B^*B)^{-1}B^*\mathbf{b}] \\ &= C^*B^*\mathbf{b} - C^*(B^*B)(CC^*)(CC^*)^{-1}(B^*B)^{-1}B^*\mathbf{b} \\ &= C^*B^*\mathbf{b} - C^*B^*\mathbf{b} = 0, \end{aligned}$$

which reveals two of the inequality's remaining three terms to be zero, leaving an assertion that

$$0 \le \Delta\mathbf{x}^*\, A^*A\,\Delta\mathbf{x}.$$

Each step in the present paragraph is reversible,[18] so the assertion in the last form is logically equivalent to the conjecture's first point, with which the paragraph began. Moreover, the assertion in the last form is correct because the product of any matrix and its adjoint according to § 13.6.3 is a nonnegative definite operator, thus establishing the conjecture's first point.

The conjecture's first point, now established, has it that no $\mathbf{x}+\Delta\mathbf{x}$ enjoys a smaller squared residual norm than the $\mathbf{x}$ of (13.32) does. It does not claim that no $\mathbf{x}+\Delta\mathbf{x}$ enjoys the same, minimal squared residual norm. The latter case is symbolized

$$\mathbf{r}^*(\mathbf{x})\mathbf{r}(\mathbf{x}) = \mathbf{r}^*(\mathbf{x}+\Delta\mathbf{x})\mathbf{r}(\mathbf{x}+\Delta\mathbf{x}),$$

or equivalently by the last paragraph's logic,

$$0 = \Delta\mathbf{x}^*\, A^*A\,\Delta\mathbf{x};$$

or in other words,

$$A\,\Delta\mathbf{x} = 0.$$

But $A = BC$, so this is to claim that

$$B(C\,\Delta\mathbf{x}) = 0,$$

which since B has full column rank is possible only if

$$C\,\Delta\mathbf{x} = 0.$$

Considering the product $\Delta\mathbf{x}^*\,\mathbf{x}$ in light of (13.32) and the last equation, we observe that

$$\begin{aligned} \Delta\mathbf{x}^*\,\mathbf{x} &= \Delta\mathbf{x}^*\,[C^*(CC^*)^{-1}(B^*B)^{-1}B^*\mathbf{b}] \\ &= [C\,\Delta\mathbf{x}]^*[(CC^*)^{-1}(B^*B)^{-1}B^*\mathbf{b}], \end{aligned}$$

[18] The paragraph might inscrutably but logically instead have ordered the steps in reverse as in §§ 6.3.2 and 9.6. See chapter 6's footnote 34.

which is to observe that

$$\Delta\mathbf{x}^*\,\mathbf{x} = 0$$

for any $\Delta\mathbf{x}$ for which $\mathbf{x} + \Delta\mathbf{x}$ achieves minimal squared residual norm.

Returning attention to the conjecture, its second point is symbolized

$$\mathbf{x}^*\mathbf{x} < (\mathbf{x} + \Delta\mathbf{x})^*(\mathbf{x} + \Delta\mathbf{x})$$

for any

$$\Delta\mathbf{x} \neq 0$$

for which $\mathbf{x}+\Delta\mathbf{x}$ achieves minimal squared residual norm (note that it's "$<$" this time, not "$\leq$" as in the conjecture's first point). Distributing factors and canceling like terms,

$$0 < \mathbf{x}^*\,\Delta\mathbf{x} + \Delta\mathbf{x}^*\,\mathbf{x} + \Delta\mathbf{x}^*\,\Delta\mathbf{x}.$$

But the last paragraph has found that $\Delta\mathbf{x}^*\,\mathbf{x} = 0$ for precisely such $\Delta\mathbf{x}$ as we are considering here, so the last inequality reduces to read

$$0 < \Delta\mathbf{x}^*\,\Delta\mathbf{x},$$

which naturally for $\Delta\mathbf{x} \neq 0$ is true. Since each step in the paragraph is reversible, reverse logic establishes the conjecture's second point.

With both its points established, the conjecture is true.

If $A = BC$ is a full-rank factorization, then the matrix[19]

$$A^\dagger \equiv C^*(CC^*)^{-1}(B^*B)^{-1}B^* \tag{13.33}$$

of (13.32) is called the *Moore-Penrose pseudoinverse* of A, more briefly the *pseudoinverse* of A. Whether underdetermined, exactly determined, overdetermined or even degenerate, every matrix has a Moore-Penrose pseudoinverse. Yielding the optimal approximation

$$\mathbf{x} = A^\dagger\mathbf{b}, \tag{13.34}$$

the Moore-Penrose solves the linear system (13.26) as well as the system can be solved—exactly if possible, with minimal squared residual norm if impossible. If A is square and invertible, then the Moore-Penrose $A^\dagger = A^{-1}$ is just the inverse, and then of course (13.34) solves the system uniquely and exactly. Nothing can solve the system uniquely if A has broad shape but the Moore-Penrose still solves the system exactly in that case as long as A has

[19] Some books print $A^\dagger$ as A^+.

full row rank, moreover minimizing the solution's squared magnitude $\mathbf{x}^*\mathbf{x}$ (which the solution of eqn. 13.24 fails to do). If A lacks full row rank, then the Moore-Penrose solves the system as nearly as the system can be solved (as in Fig. 13.1) and as a side-benefit also minimizes $\mathbf{x}^*\mathbf{x}$. The Moore-Penrose is thus a general-purpose solver and approximator for linear systems. It is a significant discovery.[20]

13.7 The multivariate Newton-Raphson iteration

When we first met the Newton-Raphson iteration in § 4.8 we lacked the matrix notation and algebra to express and handle vector-valued functions adeptly. Now that we have the notation and algebra we can write down the multivariate Newton-Raphson iteration almost at once.

The iteration approximates the nonlinear vector function $\mathbf{f}(\mathbf{x})$ by its tangent

$$\tilde{\mathbf{f}}_k(\mathbf{x}) = \mathbf{f}(\mathbf{x}_k) + \left[\frac{d}{d\mathbf{x}}\mathbf{f}(\mathbf{x})\right]_{\mathbf{x}=\mathbf{x}_k}(\mathbf{x}-\mathbf{x}_k),$$

where $d\mathbf{f}/d\mathbf{x}$ is the Jacobian derivative of § 11.10. It then approximates the root $\mathbf{x}_{k+1}$ as the point at which $\tilde{\mathbf{f}}_k(\mathbf{x}_{k+1}) = 0$:

$$\tilde{\mathbf{f}}_k(\mathbf{x}_{k+1}) = 0 = \mathbf{f}(\mathbf{x}_k) + \left[\frac{d}{d\mathbf{x}}\mathbf{f}(\mathbf{x})\right]_{\mathbf{x}=\mathbf{x}_k}(\mathbf{x}_{k+1}-\mathbf{x}_k).$$

Solving for $\mathbf{x}_{k+1}$ (approximately if necessary), we have that

$$\mathbf{x}_{k+1} = \left\{\mathbf{x} - \left[\frac{d}{d\mathbf{x}}\mathbf{f}(\mathbf{x})\right]^{\dagger}\mathbf{f}(\mathbf{x})\right\}_{\mathbf{x}=\mathbf{x}_k}, \tag{13.35}$$

where $[\cdot]^\dagger$ is the Moore-Penrose pseudoinverse of § 13.6—which is just the ordinary inverse $[\cdot]^{-1}$ of § 13.1 if $\mathbf{f}$ and $\mathbf{x}$ happen each to have the same number of elements. Refer to § 4.8 and Fig. 4.6.[21]

Despite the Moore-Penrose notation of (13.35), the Newton-Raphson iteration is not normally meant to be applied at a value of $\mathbf{x}$ for which the

[20] [14, § 3.3][132, "Moore-Penrose generalized inverse"]. For further background and context, see also [69, chapter 3], which does not mention the Moore-Penrose pseudoinverse by name but offers a gentler introduction to the topic and affords examples, drawn from the field of economics, to which one can apply Moore-Penrose.

[21] [142]

Jacobian is degenerate. The iteration intends rather in light of (13.33) that

$$\left[\frac{d}{d\mathbf{x}}\mathbf{f}(\mathbf{x})\right]^\dagger = \begin{cases} [d\mathbf{f}/d\mathbf{x}]^* \left([d\mathbf{f}/d\mathbf{x}]\,[d\mathbf{f}/d\mathbf{x}]^*\right)^{-1} & \text{if } r = m \le n, \\ [d\mathbf{f}/d\mathbf{x}]^{-1} & \text{if } r = m = n, \\ \left([d\mathbf{f}/d\mathbf{x}]^*\,[d\mathbf{f}/d\mathbf{x}]\right)^{-1} [d\mathbf{f}/d\mathbf{x}]^* & \text{if } r = n \le m, \end{cases} \tag{13.36}$$

where $B = I_m$ in the first case and $C = I_n$ in the last. It does not intend to use the full (13.33). If both $r < m$ and $r < n$—which is to say, if the Jacobian is degenerate—then (13.36) fails, as though the curve of Fig. 4.6 ran horizontally at the test point—when one quits, restarting the iteration from another point.

13.8 The dot product

The *dot product* of two vectors, also called the *inner product,*[22] is the product of the two vectors to the extent to which they run in the same direction. It is written as

$$\mathbf{a} \cdot \mathbf{b}.$$

In general,

$$\mathbf{a} \cdot \mathbf{b} = (a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + \cdots + a_n\mathbf{e}_n) \cdot (b_1\mathbf{e}_1 + b_2\mathbf{e}_2 + \cdots + b_n\mathbf{e}_n).$$

But if the dot product is to mean anything, it must be that

$$\mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij}, \tag{13.37}$$

where the Kronecker delta δ_{ij} is as defined in § 11.2. Therefore,

$$\mathbf{a} \cdot \mathbf{b} = a_1b_1 + a_2b_2 + \cdots + a_nb_n;$$

or, more concisely,

$$\mathbf{a} \cdot \mathbf{b} \equiv \mathbf{a}^T\mathbf{b} = \sum_{j=-\infty}^{\infty} a_jb_j. \tag{13.38}$$

[22] The term *inner product* is often used to indicate a broader class of products than the one defined here, especially in some of the older literature. Where used, the notation usually resembles $\langle \mathbf{a}, \mathbf{b} \rangle$ or $(\mathbf{b}, \mathbf{a})$, both of which mean $\mathbf{a}^* \cdot \mathbf{b}$ (or, more broadly, some similar product), except that which of $\mathbf{a}$ and $\mathbf{b}$ is conjugated depends on the author. Most recently, at least in the author's country, the usage $\langle \mathbf{a}, \mathbf{b} \rangle \equiv \mathbf{a}^* \cdot \mathbf{b}$ seems to be emerging as standard where the dot is not used, as in [14, § 3.1][59, chapter 4] (but slightly contrary to [43, § 2.1], for example). At any rate, this book prefers the dot.

The dot notation does not worry whether its arguments are column or row vectors, incidentally:

$$\mathbf{a}\cdot\mathbf{b} = \mathbf{a}\cdot\mathbf{b}^T = \mathbf{a}^T\cdot\mathbf{b} = \mathbf{a}^T\cdot\mathbf{b}^T = \mathbf{a}^T\mathbf{b}.$$

That is, if either vector is wrongly oriented, the notation implicitly reorients it before using it. (The more orderly notation $\mathbf{a}^T\mathbf{b}$ by contrast assumes that both are proper column vectors.)

Where vectors may have complex elements, usually one is not interested in $\mathbf{a}\cdot\mathbf{b}$ so much as in

$$\mathbf{a}^*\cdot\mathbf{b} \equiv \mathbf{a}^*\mathbf{b} = \sum_{j=-\infty}^{\infty} a_j^* b_j. \tag{13.39}$$

The reason is that

$$\Re(\mathbf{a}^*\cdot\mathbf{b}) = \Re(\mathbf{a})\cdot\Re(\mathbf{b}) + \Im(\mathbf{a})\cdot\Im(\mathbf{b}),$$

with the product of the imaginary parts added not subtracted, thus honoring the right Argand sense of "the product of the two vectors to the extent to which they run in the same direction."

By the Pythagorean theorem, the dot product

$$|\mathbf{a}|^2 = \mathbf{a}^*\cdot\mathbf{a} \tag{13.40}$$

gives the square of a vector's magnitude, always real, never negative. The unit vector in $\mathbf{a}$'s direction then is

$$\hat{\mathbf{a}} \equiv \frac{\mathbf{a}}{|\mathbf{a}|} = \frac{\mathbf{a}}{\sqrt{\mathbf{a}^*\cdot\mathbf{a}}}, \tag{13.41}$$

from which

$$|\hat{\mathbf{a}}|^2 = \hat{\mathbf{a}}^*\cdot\hat{\mathbf{a}} = 1. \tag{13.42}$$

When two vectors do not run in the same direction at all, such that

$$\mathbf{a}^*\cdot\mathbf{b} = 0, \tag{13.43}$$

the two vectors are said to lie *orthogonal* to one another. Geometrically this puts them at right angles. For other angles θ between two vectors,

$$\hat{\mathbf{a}}^*\cdot\hat{\mathbf{b}} = \cos\theta, \tag{13.44}$$

which formally defines the angle θ even when $\mathbf{a}$ and $\mathbf{b}$ have more than three elements each.

13.9 Complex vector inequalities

The triangle inequalities (2.47) and (3.21) lead one to hypothesize generally that

$$|\mathbf{a}| - |\mathbf{b}| \le |\mathbf{a} + \mathbf{b}| \le |\mathbf{a}| + |\mathbf{b}| \tag{13.45}$$

for any complex, n-dimensional vectors $\mathbf{a}$ and $\mathbf{b}$. Section 13.9.2 will prove (13.45); but first, § 13.9.1 develops a related inequality by Schwarz.

13.9.1 The Schwarz inequality

The *Schwarz inequality,* alternately the *Cauchy-Schwarz inequality,*[23] has that

$$|\mathbf{a}^* \cdot \mathbf{b}| \le |\mathbf{a}||\mathbf{b}|. \tag{13.46}$$

Roughly in words: the dot product does not exceed the product of lengths.

If the three-dimensional geometrical vectors with their dot products of chapters 3 and 15 are already familiar to you then (13.46) might seem too obvious to bother proving. The present chapter however brings an arbitrary number n of dimensions. Furthermore, elements in any or every dimension can be complex. Therefore, the geometry is not so easy to visualize in the general case. One would prefer an algebraic proof.[24]

The proof is by contradiction. We suppose falsely that

$$|\mathbf{a}^* \cdot \mathbf{b}| > |\mathbf{a}||\mathbf{b}|.$$

Squaring and using (2.71) and (13.40),

$$(\mathbf{a}^* \cdot \mathbf{b})(\mathbf{b}^* \cdot \mathbf{a}) > (\mathbf{a}^* \cdot \mathbf{a})(\mathbf{b}^* \cdot \mathbf{b}),$$

or in other words,

$$\sum_{i,j} a_i^* b_i b_j^* a_j > \sum_{i,j} a_i^* a_i b_j^* b_j,$$

wherein each side of the inequality is real-valued by construction (that is, each side is real-valued because we had started with a real inequality and—despite that elements on either side may be complex—no step since the start

[23] Pronounced as "Schwartz," almost as "Schwortz." You can sound out the German w like an English v if you wish. The other name being French is pronounced as "Co-shee," preferably with little stress but—to the extent necessary while speaking English—with stress laid on the first syllable.

[24] See [187] and [43, § 2.1] for various other proofs, one of which partly resembles the proof given here. See also [182, "Cauchy-Schwarz inequality," 17:56, 22 May 2017].

has made either side of the inequality complex as a whole). One would like to segregate conjugated elements for separate handling; it is not easy to see how to segregate them all at once but to reorder factors as

$$\sum_{i,j} [(a_i b_j)^* (b_i a_j)] > \sum_{i,j} [(a_i b_j)^* (a_i b_j)]$$

at least makes a step in the right direction. The last inequality is unhelpfully asymmetric, though, so we swap indices $i \leftrightarrow j$ to write the same inequality as that

$$\sum_{i,j} [(b_i a_j)^* (a_i b_j)] > \sum_{i,j} [(b_i a_j)^* (b_i a_j)] .$$

The swapped inequality is asymmetric too but one can add it to the earlier, unswapped inequality to achieve the symmetric form

$$\sum_{i,j} [(a_i b_j)^* (b_i a_j) + (b_i a_j)^* (a_i b_j)] > \sum_{i,j} [(a_i b_j)^* (a_i b_j) + (b_i a_j)^* (b_i a_j)] .$$

Does this help? Indeed it does. Transferring all terms to the inequality's right side,

$$0 > \sum_{i,j} [(a_i b_j)^* (a_i b_j) + (b_i a_j)^* (b_i a_j) - (a_i b_j)^* (b_i a_j) - (b_i a_j)^* (a_i b_j)] ,$$

Factoring,

$$0 > \sum_{i,j} [(a_i b_j - b_i a_j)^* (a_i b_j - b_i a_j)] = \sum_{i,j} |a_i b_j - b_i a_j|^2 ,$$

which inequality is impossible because $0 \le |\cdot|^2$ regardless of what the $|\cdot|$ might be. The contradiction proves false the assumption that gave rise to it, thus establishing the Schwarz inequality of (13.46).

13.9.2 Triangle inequalities

The proof of the sum hypothesis (13.45) that $|\mathbf{a} + \mathbf{b}| \le |\mathbf{a}| + |\mathbf{b}|$ is again by contradiction. We suppose falsely that

$$|\mathbf{a} + \mathbf{b}| > |\mathbf{a}| + |\mathbf{b}| .$$

Squaring and using (13.40),

$$(\mathbf{a} + \mathbf{b})^* \cdot (\mathbf{a} + \mathbf{b}) > \mathbf{a}^* \cdot \mathbf{a} + 2\,|\mathbf{a}|\,|\mathbf{b}| + \mathbf{b}^* \cdot \mathbf{b}.$$

Distributing factors and canceling like terms,

$$\mathbf{a}^* \cdot \mathbf{b} + \mathbf{b}^* \cdot \mathbf{a} > 2 \left|\mathbf{a}\right| \left|\mathbf{b}\right|,$$

where both sides of the inequality remain real for the same reason as in the last subsection. On the left, the imaginary parts offset because $(\mathbf{b}^* \cdot \mathbf{a})^* = \mathbf{a}^* \cdot \mathbf{b}$, leaving

$$2\Re\left(\mathbf{a}^* \cdot \mathbf{b}\right) > 2 \left|\mathbf{a}\right| \left|\mathbf{b}\right|.$$

However, the real part of the Schwarz inequality (13.46) has that

$$\Re\left(\mathbf{a}^* \cdot \mathbf{b}\right) \le \left|\mathbf{a}^* \cdot \mathbf{b}\right| \le \left|\mathbf{a}\right| \left|\mathbf{b}\right|,$$

which, when doubled, contradicts the last finding. The contradiction proves false the assumption that gave rise to it, thus establishing the sum hypothesis of (13.45).

The difference hypothesis that $|\mathbf{a}| - |\mathbf{b}| \le |\mathbf{a} + \mathbf{b}|$ is established by defining a vector $\mathbf{c}$ such that

$$\mathbf{a} + \mathbf{b} + \mathbf{c} = 0,$$

whereupon according to the sum hypothesis

$$\begin{aligned} \left|\mathbf{a} + \mathbf{c}\right| &\le \left|\mathbf{a}\right| + \left|\mathbf{c}\right|, \\ \left|\mathbf{b} + \mathbf{c}\right| &\le \left|\mathbf{b}\right| + \left|\mathbf{c}\right|. \end{aligned}$$

That is,

$$\begin{aligned} \left|-\mathbf{b}\right| &\le \left|\mathbf{a}\right| + \left|-\mathbf{a} - \mathbf{b}\right|, \\ \left|-\mathbf{a}\right| &\le \left|\mathbf{b}\right| + \left|-\mathbf{a} - \mathbf{b}\right|, \end{aligned}$$

which is the difference hypothesis in disguise. This completes the proof of the triangle inequalities (13.45)

The triangle sum inequality is alternately called the *Minkowski inequality*.[25]

As in § 3.10, here too we can extend the sum inequality to the even more general form

$$\left|\sum_k \mathbf{a}_k\right| \le \sum_k \left|\mathbf{a}_k\right|. \tag{13.47}$$

[25] [43, § 2.1]

13.10 The orthogonal complement

The $m \times (m - r)$ kernel (§ 13.3)[26]

$$A^{\perp} \equiv A^{*K} \tag{13.48}$$

is an interesting matrix. By definition of the kernel, the columns of A^{*K} are the independent vectors $\mathbf{u}_j$ for which $A^*\mathbf{u}_j = 0$, which—inasmuch as the *rows* of A^* are the adjoints of the columns of A—is possible only when each $\mathbf{u}_j$ lies orthogonal to every column of A. This says that the columns of $A^{\perp} \equiv A^{*K}$ address the complete space of vectors that lie orthogonal to A's columns, such that

$$A^{\perp *}A = 0 = A^*A^{\perp}. \tag{13.49}$$

The matrix $A^{\perp}$ is called the *orthogonal complement*[27] or *perpendicular matrix* to A.

Among other uses, the orthogonal complement $A^{\perp}$ supplies the columns A lacks to reach full row rank. Properties include that

$$\begin{aligned} A^{*K} &= A^{\perp}, \\ A^{*\perp} &= A^{K}. \end{aligned} \tag{13.50}$$

13.11 Gram-Schmidt orthonormalization

If a vector $\mathbf{x} = A^K\mathbf{a}$ belongs to a kernel space A^K (§ 13.3), then so equally does any $\alpha\mathbf{x}$. If the vectors $\mathbf{x}_1 = A^K\mathbf{a}_1$ and $\mathbf{x}_2 = A^K\mathbf{a}_2$ both belong, then so does $\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2$. If I claim $A^K = [3\ 4\ 5; -1\ 1\ 0]^T$ to represent a kernel, then you are not mistaken arbitrarily to rescale each column of my A^K by a separate nonzero factor, instead for instance representing the same kernel as $A^K = [6\ 8\ 0\mathrm{xA}; \frac{1}{7}\ {-\frac{1}{7}}\ 0]^T$. Kernel vectors have no inherent scale. Style generally asks one to remove the false appearance of scale by using (13.41) *to normalize* the columns of a kernel matrix to unit magnitude before reporting them. The same goes for the eigenvectors of chapter 14 to come.

Where a kernel matrix A^K has two or more columns (or a repeated eigenvalue has two or more eigenvectors), style generally asks one not only to normalize but also *to orthogonalize* the columns before reporting them.

[26] The symbol $A^{\perp}$ [75][14][106] can be pronounced "A perp," short for "A perpendicular," since by (13.49) $A^{\perp}$ is in some sense perpendicular to A.

If we were really precise, we might write not $A^{\perp}$ but $A^{\perp(m)}$. Refer to footnote 5.

[27] [75, § 3.VI.3]

One orthogonalizes a vector $\mathbf{b}$ with respect to a vector $\mathbf{a}$ by subtracting from $\mathbf{b}$ a multiple of $\mathbf{a}$ such that

$$\begin{aligned}\mathbf{a}^* \cdot \mathbf{b}_\perp &= 0,\\ \mathbf{b}_\perp &\equiv \mathbf{b} - \beta\mathbf{a},\end{aligned}$$

where the symbol $\mathbf{b}_\perp$ represents the orthogonalized vector. Substituting the second of these equations into the first and solving for β yields that

$$\beta = \frac{\mathbf{a}^* \cdot \mathbf{b}}{\mathbf{a}^* \cdot \mathbf{a}}.$$

Hence,

$$\begin{aligned}\mathbf{a}^* \cdot \mathbf{b}_\perp &= 0,\\ \mathbf{b}_\perp &\equiv \mathbf{b} - \frac{\mathbf{a}^* \cdot \mathbf{b}}{\mathbf{a}^* \cdot \mathbf{a}}\mathbf{a}.\end{aligned} \tag{13.51}$$

But according to (13.41), $\mathbf{a} = \hat{\mathbf{a}}\sqrt{\mathbf{a}^* \cdot \mathbf{a}}$; and according to (13.42), $\hat{\mathbf{a}}^* \cdot \hat{\mathbf{a}} = 1$; so,

$$\mathbf{b}_\perp = \mathbf{b} - \hat{\mathbf{a}}(\hat{\mathbf{a}}^* \cdot \mathbf{b}); \tag{13.52}$$

or, in matrix notation,

$$\mathbf{b}_\perp = \mathbf{b} - \hat{\mathbf{a}}(\hat{\mathbf{a}}^*)(\mathbf{b}).$$

This is arguably better written,

$$\mathbf{b}_\perp = [I - (\hat{\mathbf{a}})(\hat{\mathbf{a}}^*)]\,\mathbf{b} \tag{13.53}$$

(observe that it's $[\hat{\mathbf{a}}][\hat{\mathbf{a}}^*]$, a matrix, rather than the scalar $[\hat{\mathbf{a}}^*][\hat{\mathbf{a}}]$).

One *orthonormalizes* a set of vectors by orthogonalizing them with respect to one another and then normalizing each of them to unit magnitude. The procedure to orthonormalize several vectors

$$\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n\}$$

therefore is as follows. First, normalize $\mathbf{x}_1$ by (13.41); call the result $\hat{\mathbf{x}}_{1\perp}$. Second, orthogonalize $\mathbf{x}_2$ with respect to $\hat{\mathbf{x}}_{1\perp}$ by (13.52) or (13.53), then normalize it; call the result $\hat{\mathbf{x}}_{2\perp}$. Third, orthogonalize $\mathbf{x}_3$ with respect to $\hat{\mathbf{x}}_{1\perp}$ then to $\hat{\mathbf{x}}_{2\perp}$, then normalize it; call the result $\hat{\mathbf{x}}_{3\perp}$. Proceed in this manner through the several $\mathbf{x}_j$. Symbolically,

$$\begin{aligned}\hat{\mathbf{x}}_{j\perp} &= \frac{\mathbf{x}_{j\perp}}{\sqrt{\mathbf{x}_{j\perp}^* \mathbf{x}_{j\perp}}},\\ \mathbf{x}_{j\perp} &\equiv \left[\prod_{i=1}^{j-1}(I - \hat{\mathbf{x}}_{i\perp}\hat{\mathbf{x}}_{i\perp}^*)\right]\mathbf{x}_j.\end{aligned} \tag{13.54}$$

By the vector replacement principle of § 12.4 in light of (13.51), the resulting orthonormal set of vectors

$$\{\hat{\mathbf{x}}_{1\perp}, \hat{\mathbf{x}}_{2\perp}, \hat{\mathbf{x}}_{3\perp}, \ldots, \hat{\mathbf{x}}_{n\perp}\}$$

addresses the same space as did the original set.

Orthonormalization naturally works equally for any linearly independent set of vectors, not only for kernel vectors or eigenvectors. By the technique, one can conveniently replace a set of independent vectors by an equivalent, neater, orthonormal set which addresses precisely the same space.

13.11.1 Efficient implementation

To turn an equation like the latter line of (13.54) into an efficient numerical algorithm sometimes demands some extra thought, in perspective of whatever it happens to be that one is trying to accomplish. If all one wants is some vectors orthonormalized, then the equation as written is neat but is overkill because the product $\hat{\mathbf{x}}_{i\perp}\hat{\mathbf{x}}_{i\perp}^*$ is a matrix, whereas the product $\hat{\mathbf{x}}_{i\perp}^*\mathbf{x}_j$ implied by (13.52) is just a scalar. Fortunately, one need not apply the latter line of (13.54) exactly as written. One can instead introduce intermediate vectors $\mathbf{x}_{ji}$, representing the $\prod$ multiplication in the admittedly messier form

$$\begin{aligned}\mathbf{x}_{j1} &\equiv \mathbf{x}_j,\\ \mathbf{x}_{j(i+1)} &\equiv \mathbf{x}_{ji} - (\hat{\mathbf{x}}_{i\perp}^* \cdot \mathbf{x}_{ji})\,\hat{\mathbf{x}}_{i\perp},\\ \mathbf{x}_{j\perp} &= \mathbf{x}_{jj}.\end{aligned} \tag{13.55}$$

Besides obviating the matrix $I - \hat{\mathbf{x}}_{i\perp}\hat{\mathbf{x}}_{i\perp}^*$ and the associated matrix multiplication, the messier form (13.55) has the significant additional practical virtue that it lets one forget each intermediate vector $\mathbf{x}_{ji}$ immediately after using it. (A well-written orthonormalizing computer program reserves memory for one intermediate vector only, which memory it repeatedly overwrites—and, actually, probably does not even reserve that much, working rather in the memory space it has already reserved for $\hat{\mathbf{x}}_{j\perp}$.)[28]

Other equations one algorithmizes can likewise benefit from thoughtful rendering.

13.11.2 The Gram-Schmidt decomposition

The orthonormalization technique this section has developed is named the *Gram-Schmidt process.* One can turn it into the *Gram-Schmidt decomposi-*

[28] [182, "Gram-Schmidt process," 04:48, 11 Aug. 2007]

tion

$$
\begin{aligned}
A &= QR = QUDS, \\
R &\equiv UDS,
\end{aligned}
\tag{13.56}
$$

also called the *orthonormalizing* or *QR decomposition,* by an algorithm that somewhat resembles the Gauss-Jordan algorithm of § 12.3.3; except that (12.4) here becomes

$$A = \tilde{Q}\tilde{U}\tilde{D}\tilde{S} \tag{13.57}$$

and initially $\tilde{Q} \leftarrow A$. By elementary column operations based on (13.54) and (13.55), the algorithm gradually transforms $\tilde{Q}$ into a dimension-limited, $m \times r$ matrix Q of orthonormal columns, distributing the inverse elementaries to $\tilde{U}$, $\tilde{D}$ and $\tilde{S}$ according to Table 12.1—where the latter three working matrices ultimately become the extended-operational factors U, D and S of (13.56).

Borrowing the language of computer science we observe that the indices i and j of (13.54) and (13.55) imply a two-level nested loop, one level looping over j and the other over i. The equations suggest *j-major nesting,* with the loop over j at the outer level and the loop over i at the inner, such that the several (i, j) index pairs occur in the sequence (reading left to right then top to bottom)

$$
\begin{array}{llll}
(1,2) & & & \\
(1,3) & (2,3) & & \\
(1,4) & (2,4) & (3,4) & \\
\cdots & \cdots & \cdots & \ddots
\end{array}
$$

In reality, however, (13.55)'s middle line requires only that no $\hat{\mathbf{x}}_{i\perp}$ be used before it is fully calculated; otherwise that line does not care which (i, j) pair follows which. The i-major nesting

$$
\begin{array}{llll}
(1,2) & (1,3) & (1,4) & \cdots \\
 & (2,3) & (2,4) & \cdots \\
 & & (3,4) & \cdots \\
 & & & \ddots
\end{array}
$$

bringing the very same index pairs in a different sequence, is just as valid. We choose i-major nesting on the subtle ground that it affords better information to the choice of column index p during the algorithm's step 3.

The algorithm, in detail:

1. Begin by initializing

$$\begin{gathered}\tilde{U} \leftarrow I,\ \tilde{D} \leftarrow I,\ \tilde{S} \leftarrow I,\\ \tilde{Q} \leftarrow A,\\ i \leftarrow 1.\end{gathered}$$

2. (Besides arriving at this point from step 1 above, the algorithm also reënters here from step 9 below.) Observe that $\tilde{U}$ enjoys the major partial unit triangular form $L^{\{i-1\}T}$ (§ 11.8.5), that $\tilde{D}$ is a general scaling operator (§ 11.7.2) with $\tilde{d}_{jj} = 1$ for all $j \geq i$, that $\tilde{S}$ is permutor (§ 11.7.1), and that the first through $(i-1)$th columns of $\tilde{Q}$ consist of mutually orthonormal unit vectors.

3. Choose a column $p \geq i$ of $\tilde{Q}$ containing at least one nonzero element. (The simplest choice is perhaps $p = i$ as long as the ith column does not happen to be null, but one might instead prefer to choose the column of greatest magnitude, or to choose randomly, among other heuristics.) If $\tilde{Q}$ is null in and rightward of its ith column such that no column $p \geq i$ remains available to choose, then skip directly to step 10.

4. Observing that (13.57) can be expanded to read

$$\begin{aligned} A &= \left(\tilde{Q}T_{[i\leftrightarrow p]}\right)\left(T_{[i\leftrightarrow p]}\tilde{U}T_{[i\leftrightarrow p]}\right)\left(T_{[i\leftrightarrow p]}\tilde{D}T_{[i\leftrightarrow p]}\right)\left(T_{[i\leftrightarrow p]}\tilde{S}\right)\\ &= \left(\tilde{Q}T_{[i\leftrightarrow p]}\right)\left(T_{[i\leftrightarrow p]}\tilde{U}T_{[i\leftrightarrow p]}\right)\tilde{D}\left(T_{[i\leftrightarrow p]}\tilde{S}\right), \end{aligned}$$

   where the latter line has applied a rule from Table 12.1, interchange the chosen pth column to the ith position by

$$\begin{aligned} \tilde{Q} &\leftarrow \tilde{Q}T_{[i\leftrightarrow p]},\\ \tilde{U} &\leftarrow T_{[i\leftrightarrow p]}\tilde{U}T_{[i\leftrightarrow p]},\\ \tilde{S} &\leftarrow T_{[i\leftrightarrow p]}\tilde{S}. \end{aligned}$$

5. Observing that (13.57) can be expanded to read

$$A = \left(\tilde{Q}T_{(1/\alpha)[i]}\right)\left(T_{\alpha[i]}\tilde{U}T_{(1/\alpha)[i]}\right)\left(T_{\alpha[i]}\tilde{D}\right)\tilde{S},$$

   normalize the ith column of $\tilde{Q}$ by

$$\begin{aligned} \tilde{Q} &\leftarrow \tilde{Q}T_{(1/\alpha)[i]},\\ \tilde{U} &\leftarrow T_{\alpha[i]}\tilde{U}T_{(1/\alpha)[i]},\\ \tilde{D} &\leftarrow T_{\alpha[i]}\tilde{D}, \end{aligned}$$

where

$$\alpha = \sqrt{\left[\tilde{Q}\right]_{*i}^{*} \cdot \left[\tilde{Q}\right]_{*i}}.$$

6. Initialize

$$j \leftarrow i + 1.$$

7. (Besides arriving at this point from step 6 above, the algorithm also reënters here from step 8 below.) If $j > n$ then skip directly to step 9. Otherwise, observing that (13.57) can be expanded to read

$$A = \left(\tilde{Q}T_{-\beta[ij]}\right)\left(T_{\beta[ij]}\tilde{U}\right)\tilde{D}\tilde{S},$$

orthogonalize the jth column of $\tilde{Q}$ per (13.55) with respect to the ith column by

$$\begin{aligned} \tilde{Q} &\leftarrow \tilde{Q}T_{-\beta[ij]}, \\ \tilde{U} &\leftarrow T_{\beta[ij]}\tilde{U}, \end{aligned}$$

where

$$\beta = \left[\tilde{Q}\right]_{*i}^{*} \cdot \left[\tilde{Q}\right]_{*j}.$$

8. Increment

$$j \leftarrow j + 1$$

and return to step 7.

9. Increment

$$i \leftarrow i + 1$$

and return to step 2.

10. Let

$$Q \equiv \tilde{Q},\ U \equiv \tilde{U},\ D \equiv \tilde{D},\ S \equiv \tilde{S},$$
$$r = i - 1.$$

End.

Though the Gram-Schmidt algorithm broadly resembles the Gauss-Jordan, at least two significant differences stand out: (i) the Gram-Schmidt is one-sided because it operates only on the columns of $\tilde{Q}$, never on the rows; (ii) since Q is itself dimension-limited, the Gram-Schmidt decomposition (13.56) needs and has no explicit factor I_r.

As in § 12.5.7, here also one sometimes prefers that $S = I$. The algorithm optionally supports this preference if the $m \times n$ matrix A has full column rank $r = n$, when null columns cannot arise, if one always chooses $p = i$ during the algorithm's step 3. Such optional discipline maintains $S = I$ when desired.

Whether $S = I$ or not, the matrix $Q = QI_r$ has only r columns, so one can write (13.56) as

$$A = (QI_r)(R).$$

Reassociating factors, this is

$$A = (Q)(I_rR), \tag{13.58}$$

which per (12.17) is a proper full-rank factorization with which one can compute the pseudoinverse $A^\dagger$ of A (see eqn. 13.33, above; but see also eqn. 13.67, below).

If the Gram-Schmidt decomposition (13.56) looks useful, it is even more useful than it looks. The most interesting of its several factors is the $m \times r$ orthonormalized matrix Q, whose orthonormal columns address the same space the columns of A themselves address. If Q reaches the maximum possible rank $r = m$, achieving square, $m \times m$ shape, then it becomes a *unitary matrix*—the subject of § 13.12.

Before treating the unitary matrix, however, let us pause to develop the orthogonal complement by Gram-Schmidt in § 13.11.3, next.

13.11.3 The orthogonal complement by Gram-Schmidt

Having decomposed an $m \times n$ matrix as

$$A = QR = (QI_r)R, \tag{13.59}$$

observing that the r independent columns of the $m \times r$ matrix $Q = QI_r$ address the same space the columns of A address, Gram-Schmidt computes an orthogonal complement (13.48) by constructing the $m \times (r + m)$ matrix

$$A' \equiv QI_r + I_mH_{-r} = \begin{bmatrix} QI_r & I_m \end{bmatrix}. \tag{13.60}$$

This constructed matrix A' is then itself decomposed,

$$A' = Q'R', \tag{13.61}$$

again by Gram-Schmidt—with the differences that, this time, one chooses $p = 1, 2, 3, \ldots, r$ during the first r instances of the algorithm's step 3 and that one skips the unnecessary step 7 for all $j \leq r$, on the ground that the earlier Gram-Schmidt application of (13.59) has already orthonormalized first r columns of A', which columns, after all, are just $Q = QI_r$. The resulting $m \times m$, full-rank square matrix

$$Q' = QI_r + A^{\perp}H_{-r} = \left[\begin{array}{cc} QI_r & A^{\perp} \end{array} \right] \tag{13.62}$$

consists of

- r columns on the left that address the same space the columns of A address and
- $m-r$ columns on the right that give a complete orthogonal complement (§ 13.10) $A^{\perp}$ of A.

Each column has unit magnitude and conveniently lies orthogonal—indeed, orthonormal—to every other column, left and right.

Equation (13.62) is probably the more useful form, but the *Gram-Schmidt orthogonal-complement formula* as such is that

$$A^{*K} = A^{\perp} = Q'H_rI_{m-r}. \tag{13.63}$$

The writer has not encountered a Gram-Schmidt kernel in the style of (13.7) to accompany the Gram-Schmidt orthogonal complement of (13.63)—though if necessary one could maybe combine (13.63) with (13.48) for the purpose. Instead, normally, as far as the writer knows, the Gauss-Jordan (13.7) is used. Meanwhile however, the matrix Q' of this subsection is interesting. Being square and orthonormal, the $m \times m$ matrix Q' is a unitary matrix. Unitary matrices will be the subject of § 13.12, next.

13.12 The unitary matrix

When the orthonormalized matrix Q of the Gram-Schmidt decomposition (13.56) is square, having the maximum possible rank $r = m$, it brings one property so interesting that the property merits a section of its own. The property is that

$$Q^*Q = I_m = QQ^*. \tag{13.64}$$

The reason that $Q^*Q = I_m$ is that Q's columns are orthonormal, and that the very definition of orthonormality demands that the dot product $[Q]^*_{*i} \cdot [Q]_{*j}$ of orthonormal columns be zero unless $i = j$, when the dot product of a unit vector with itself is unity. That $I_m = QQ^*$ is unexpected, however, until one realizes[29] that the equation $Q^*Q = I_m$ characterizes Q^* to be the rank-m inverse of Q, and that § 13.1 lets any rank-m inverse (orthonormal or otherwise) attack just as well from the right as from the left. Thus,

$$Q^{-1} = Q^*, \tag{13.65}$$

a very useful property. A matrix Q that satisfies (13.64), whether derived from the Gram-Schmidt or from elsewhere, is called a *unitary matrix.* (Note that the permutor of § 11.7.1 enjoys the property of eqn. 13.65 precisely because it is unitary.)

One immediate consequence of (13.64) is that *a square matrix with either orthonormal columns or orthonormal rows is unitary and has both.*

The product of two or more unitary matrices is itself unitary if the matrices are of the same dimensionality. To prove it, consider the product

$$Q = Q_a Q_b \tag{13.66}$$

of $m \times m$ unitary matrices Q_a and Q_b. Let the symbols $\mathbf{q}_j$, $\mathbf{q}_{aj}$ and $\mathbf{q}_{bj}$ respectively represent the jth columns of Q, Q_a and Q_b and let the symbol q_{bij} represent the ith element of $\mathbf{q}_{bj}$. By the columnwise interpretation (§ 11.1.3) of matrix multiplication,

$$\mathbf{q}_j = \sum_i q_{bij}\mathbf{q}_{ai}.$$

The adjoint dot product of any two of Q's columns then is

$$\mathbf{q}_{j'}^* \cdot \mathbf{q}_j = \sum_{i,i'} q_{bi'j'}^* q_{bij} \mathbf{q}_{ai'}^* \cdot \mathbf{q}_{ai}.$$

But $\mathbf{q}_{ai'}^* \cdot \mathbf{q}_{ai} = \delta_{i'i}$ because Q_a is unitary,[30] so

$$\mathbf{q}_{j'}^* \cdot \mathbf{q}_j = \sum_i q_{bij'}^* q_{bij} = \mathbf{q}_{bj'}^* \cdot \mathbf{q}_{bj} = \delta_{j'j},$$

which says neither more nor less than that the columns of Q are orthonormal, which is to say that Q is unitary, as was to be demonstrated.

Unitary operations preserve length. That is, operating on an m-element vector by an $m \times m$ unitary matrix does not alter the vector's magnitude. To prove it, consider the system

$$Q\mathbf{x} = \mathbf{b}.$$

[29] [59, § 4.4]

[30] This is true only for $1 \le i \le m$, but you knew that already.

Multiplying the system by its own adjoint yields that

$$\mathbf{x}^*Q^*Q\mathbf{x} = \mathbf{b}^*\mathbf{b}.$$

But according to (13.64), $Q^*Q = I_m$; so,

$$\mathbf{x}^*\mathbf{x} = \mathbf{b}^*\mathbf{b},$$

as was to be demonstrated.

Equation (13.65) lets one use the Gram-Schmidt decomposition (13.56) to invert a square matrix as

$$A^{-1} = R^{-1}Q^* = S^*D^{-1}U^{-1}Q^*. \tag{13.67}$$

Unitary extended operators are certainly possible, for if Q is an $m \times m$ dimension-limited matrix, then the extended operator

$$Q_\infty = Q + (I - I_m),$$

which is just Q with ones running out the main diagonal from its active region, itself meets the unitary criterion (13.64) for $m = \infty$.

Unitary matrices are so easy to handle that they can sometimes justify significant effort to convert a model to work in terms of them if possible. We shall meet the unitary matrix again in §§ 14.10 and 14.12.

The chapter as a whole has demonstrated at least in theory (and usually in practice) techniques to solve any linear system characterized by a matrix of finite dimensionality, whatever the matrix's rank or shape. It has explained how to orthonormalize a set of vectors and has derived from the explanation the useful Gram-Schmidt decomposition. As the chapter's introduction had promised, the matrix has shown its worth here; for without the matrix's notation, arithmetic and algebra most of the chapter's findings would have lain beyond practical reach. And even so, the single most interesting agent of matrix arithmetic remains yet to be treated. This last is the eigenvalue, and it is the subject of chapter 14, next.

Chapter 14

The eigenvalue

The *eigenvalue* is a scalar by which a square matrix scales a vector without otherwise changing it, such that

$$A\mathbf{v} = \lambda\mathbf{v}.$$

This chapter analyzes the eigenvalue and the associated *eigenvector* it scales.

Before treating the eigenvalue proper, the chapter gathers from across chapters 11 through 14 several properties all invertible square matrices share, assembling them in § 14.2 for reference. One of these regards the *determinant,* which opens the chapter.

14.1 The determinant

Through chapters 11, 12 and 13 the theory of the matrix has developed slowly but pretty straightforwardly. Here comes the first unexpected turn.

It begins with an arbitrary-seeming definition. The *determinant* of an $n \times n$ square matrix A is the sum of $n!$ terms, each term the product of n elements, no two elements from the same row or column, terms of positive parity adding to and terms of negative parity subtracting from the sum—a term's parity (§ 11.6) being the parity of the permutor (§ 11.7.1) marking the positions of the term's elements.

Unless you already know about determinants, the definition alone might seem hard to parse, so try this. The inverse of the general 2×2 square matrix

$$A_2 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

by the Gauss-Jordan method or any other convenient technique, is found to be

$$A_2^{-1} = \frac{\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}}{a_{11}a_{22} - a_{12}a_{21}}.$$

The quantity[1]

$$\det A_2 = a_{11}a_{22} - a_{12}a_{21}$$

in the denominator is defined to be the *determinant* of A_2. Each of the determinant's terms includes one element from each column of the matrix and one from each row, with parity giving the term its ± sign. The determinant of the general 3×3 square matrix by the same rule is

$$\begin{aligned} \det A_3 \;\; = \;\; & (a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}) \\ & - (a_{13}a_{22}a_{31} + a_{12}a_{21}a_{33} + a_{11}a_{23}a_{32}); \end{aligned}$$

and indeed if we tediously invert such a matrix symbolically, we do find that quantity in the denominator there.

The parity rule merits a more careful description. The parity of a term like $a_{12}a_{23}a_{31}$ is positive because the parity of the permutor, or interchange quasielementary (§ 11.7.1),

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

marking the positions of the term's elements is positive. The parity of a term like $a_{13}a_{22}a_{31}$ is negative for the same reason. The determinant comprehends all possible such terms, $n!$ in number, half of positive parity and half of negative. (How do we know that exactly half are of positive and half, negative? Answer: by pairing the terms. For every term like $a_{12}a_{23}a_{31}$ whose marking permutor is P, there is a corresponding $a_{13}a_{22}a_{31}$ whose marking permutor is $T_{[1\leftrightarrow 2]}P$, necessarily of opposite parity. The sole exception to the rule is the 1×1 square matrix, which has no second term to pair.)

[1] The determinant $\det A$ used to be written $|A|$, an appropriately terse notation for which the author confesses some nostalgia. The older notation $|A|$ however unluckily suggests "the magnitude of A," which though not quite the wrong idea is not quite the right idea, either. The magnitude $|z|$ of a scalar or $|\mathbf{u}|$ of a vector is a real-valued, nonnegative, nonanalytic function of the elements of the quantity in question, whereas the determinant $\det A$ is a complex-valued, analytic function. The book follows convention by denoting the determinant as $\det A$ for this reason among others.

Normally the context implies a determinant's rank n, but the nonstandard notation

$$\det^{(n)} A$$

is available especially to call the rank out, stating explicitly that the determinant has exactly $n!$ terms. (See also §§ 11.3.5 and 11.5 and eqn. 11.49.[2])

It is admitted[3] that we have not, as yet, actually shown the determinant to be a generally useful quantity; we have merely motivated and defined it. The true history of the determinant is unknown to this writer, but one might suppose that the determinant had originally emerged not from abstract considerations but for the mundane reason that the quantity it represents occurs frequently in practice (as in the A_2^{-1} of the example above). Nothing however logically prevents one from simply defining some quantity which, at first, one merely suspects will later prove useful. So we do here.[4]

14.1.1 Basic properties

The determinant $\det A$ enjoys several useful basic properties.

- If

$$c_{i*} = \begin{cases} a_{i''*} & \text{when } i = i', \\ a_{i'*} & \text{when } i = i'', \\ a_{i*} & \text{otherwise,} \end{cases}$$

or if

$$c_{*j} = \begin{cases} a_{*j''} & \text{when } j = j', \\ a_{*j'} & \text{when } j = j'', \\ a_{*j} & \text{otherwise,} \end{cases}$$

where $i'' \neq i'$ and $j'' \neq j'$, then

$$\det C = -\det A. \tag{14.1}$$

Interchanging rows or columns negates the determinant.

- If

$$c_{i*} = \begin{cases} \alpha a_{i*} & \text{when } i = i', \\ a_{i*} & \text{otherwise,} \end{cases}$$

[2] And see further chapter 13's footnotes 5 and 26.

[3] [59, § 1.2]

[4] [59, chapter 1]

or if

$$c_{*j} = \begin{cases} \alpha a_{*j} & \text{when } j = j', \\ a_{*j} & \text{otherwise,} \end{cases}$$

then

$$\det C = \alpha \det A. \tag{14.2}$$

Scaling a single row or column of a matrix scales the matrix's determinant by the same factor. (Equation 14.2 tracks the linear scaling property of § 7.3.3 and of eqn. 11.2.)

- If

$$c_{i*} = \begin{cases} a_{i*} + b_{i*} & \text{when } i = i', \\ a_{i*} = b_{i*} & \text{otherwise,} \end{cases}$$

or if

$$c_{*j} = \begin{cases} a_{*j} + b_{*j} & \text{when } j = j', \\ a_{*j} = b_{*j} & \text{otherwise,} \end{cases}$$

then

$$\det C = \det A + \det B. \tag{14.3}$$

If one row or column of a matrix C is the sum of the corresponding rows or columns of two other matrices A and B, while the three matrices remain otherwise identical, then the determinant of the one matrix is the sum of the determinants of the other two. (Equation 14.3 tracks the linear superposition property of § 7.3.3 and of eqn. 11.2.)

- If

$$c_{i'*} = 0,$$

or if

$$c_{*j'} = 0,$$

then

$$\det C = 0. \tag{14.4}$$

A matrix with a null row or column also has a null determinant.

- If

$$c_{i''*} = \gamma c_{i'*},$$

or if

$$c_{*j''} = \gamma c_{*j'},$$

where $i'' \neq i'$ and $j'' \neq j'$, then

$$\det C = 0. \tag{14.5}$$

The determinant is zero if one row or column of the matrix is a multiple of another.

- The determinant of the adjoint is just the determinant's conjugate, and the determinant of the transpose is just the determinant itself:

$$\begin{aligned} \det C^* &= (\det C)^* \,; \\ \det C^T &= \det C. \end{aligned} \tag{14.6}$$

These basic properties are all fairly easy to see if the definition of the determinant is clearly understood. Equations (14.2), (14.3) and (14.4) come because each of the $n!$ terms in the determinant's expansion has exactly one element from row i' or column j'. Equation (14.1) comes because a row or column interchange reverses parity. Equation (14.6) comes because according to § 11.7.1, the permutors P and P^* always have the same parity, and because the adjoint operation individually conjugates each element of C. Finally, (14.5) comes because, in this case, every term in the determinant's expansion finds an equal term of opposite parity to offset it. Or, more formally, (14.5) comes because the following procedure does not alter the matrix: (i) scale row i'' or column j'' by $1/\gamma$; (ii) scale row i' or column j' by γ; (iii) interchange rows $i' \leftrightarrow i''$ or columns $j' \leftrightarrow j''$. Not altering the matrix, the procedure does not alter the determinant either; and indeed according to (14.2), step (ii)'s effect on the determinant cancels that of step (i). However, according to (14.1), step (iii) negates the determinant. Hence the net effect of the procedure is to negate the determinant—to negate the very determinant the procedure is not permitted to alter. The apparent contradiction can be reconciled only if the determinant is zero to begin with.

From the foregoing properties the following further property can be deduced.

- If

$$c_{i*} = \begin{cases} a_{i*} + \alpha a_{i'*} & \text{when } i = i'', \\ a_{i*} & \text{otherwise,} \end{cases}$$

or if

$$c_{*j} = \begin{cases} a_{*j} + \alpha a_{*j'} & \text{when } j = j'', \\ a_{*j} & \text{otherwise,} \end{cases}$$

where $i'' \neq i'$ and $j'' \neq j'$, then

$$\det C = \det A. \tag{14.7}$$

Adding to a row or column of a matrix a multiple of another row or column does not change the matrix's determinant.

To derive (14.7) for rows (the column proof is similar), one defines a matrix B such that

$$b_{i*} \equiv \begin{cases} \alpha a_{i'*} & \text{when } i = i'', \\ a_{i*} & \text{otherwise.} \end{cases}$$

From this definition, $b_{i''*} = \alpha a_{i'*}$ whereas $b_{i'*} = a_{i'*}$, so

$$b_{i''*} = \alpha b_{i'*},$$

which by (14.5) guarantees that

$$\det B = 0.$$

On the other hand, the three matrices A, B and C differ only in the (i'')th row, where $[C]_{i''*} = [A]_{i''*} + [B]_{i''*}$; so, according to (14.3),

$$\det C = \det A + \det B.$$

Equation (14.7) results from combining the last two equations.

14.1.2 The determinant and the elementary operator

Section 14.1.1 has it that interchanging, scaling or adding rows or columns of a matrix respectively negates, scales or does not alter the matrix's determinant. But the three operations named are precisely the operations of the three elementaries of § 11.4. Therefore,

$$\begin{array}{rcccl} \det T_{[i\leftrightarrow j]}A &=& -\det A &=& \det AT_{[i\leftrightarrow j]}, \\ \det T_{\alpha[i]}A &=& \alpha \det A &=& \det AT_{\alpha[j]}, \\ \det T_{\alpha[ij]}A &=& \det A &=& \det AT_{\alpha[ij]}, \\ && 1 \le (i,j) \le n,\ i \ne j, && \end{array} \tag{14.8}$$

for any $n \times n$ square matrix A. Obviously also,

$$\begin{array}{rcccl} \det IA &=& \det A &=& \det AI, \\ \det I_nA &=& \det A &=& \det AI_n, \\ \det I &=& 1 &=& \det I_n. \end{array} \tag{14.9}$$

If A is taken to represent an arbitrary product of identity matrices (I_n and/or I) and elementary operators, then a significant consequence of (14.8) and (14.9), applied recursively, is that the determinant of a product is the product of the determinants, at least where identity matrices and elementary operators are concerned. In symbols,[5]

$$\begin{aligned} \det\left(\prod_k M_k\right) &= \prod_k \det M_k, \\ M_k &\in \left\{I_n, I, T_{[i\leftrightarrow j]}, T_{\alpha[i]}, T_{\alpha[ij]}\right\}, \\ 1 \le (i,j) &\le n. \end{aligned} \tag{14.10}$$

This matters because, as the Gauss-Jordan decomposition of § 12.3 has shown, one can build up any square matrix of full rank by applying elementary operators to I_n. Section 14.1.4 will put the rule (14.10) to good use.

14.1.3 The determinant of a singular matrix

Equation (14.8) gives elementary operators the power to alter a matrix's determinant almost arbitrarily—almost arbitrarily, but not quite. What an $n \times n$ elementary operator[6] cannot do is to change an $n \times n$ matrix's determinant to or from zero. Once zero, a determinant remains zero under the action of elementary operators. Once nonzero, always nonzero. Elementary operators being reversible have no power to breach this barrier.

Another thing $n \times n$ elementaries cannot do according to § 12.5.3 is to change an $n \times n$ matrix's rank. Nevertheless, such elementaries can reduce any $n \times n$ matrix reversibly to I_r, where $r \le n$ is the matrix's rank, by the Gauss-Jordan algorithm of § 12.3. Equation (14.4) has that the $n \times n$ determinant of I_r is zero if $r < n$, so it follows that the $n \times n$ determinant of every rank-r matrix is similarly zero if $r < n$; and complementarily that the $n \times n$ determinant of a rank-n matrix is never zero. Singular matrices always have zero determinants; full-rank square matrices never do. One can evidently tell the singularity or invertibility of a square matrix from its determinant alone.

[5] Notation like "$\in$", first met in § 2.3, can be too fancy for applied mathematics, but it does help here. The notation $M_k \in \{\ldots\}$ restricts M_k to be any of the things between the braces. As it happens though, in this case, (14.11) below is going to erase the restriction.

[6] That is, an elementary operator which honors an $n \times n$ active region. See § 11.3.2.

14.1.4 The determinant of a matrix product

Sections 14.1.2 and 14.1.3 suggest the useful rule that

$$\det AB = \det A \det B. \tag{14.11}$$

To prove the rule, we consider three distinct cases.

The first case is that A is singular. In this case, B acts as a column operator on A, whereas according to § 12.5.2 no operator has the power to promote A in rank. Hence the product AB is no higher in rank than A, which says that AB is no less singular than A, which implies that AB like A has a null determinant. Evidently (14.11) holds in the first case.

The second case is that B is singular. The proof here resembles that of the first case.

The third case is that neither matrix is singular. Here, we use Gauss-Jordan to decompose both matrices into sequences of elementary operators and rank-n identity matrices, for which

$$\begin{aligned}
\det AB &= \det\{[A]\,[B]\} \\
&= \det\left\{\left[\left(\prod T\right) I_n \left(\coprod T\right)\right]\left[\left(\prod T\right) I_n \left(\coprod T\right)\right]\right\} \\
&= \left(\prod \det T\right) \det I_n \left(\coprod \det T\right) \left(\prod \det T\right) \det I_n \left(\coprod \det T\right) \\
&= \det\left[\left(\prod T\right) I_n \left(\coprod T\right)\right] \det\left[\left(\prod T\right) I_n \left(\coprod T\right)\right] \\
&= \det A \det B,
\end{aligned}$$

which is a schematic way of pointing out in light of (14.10) merely that since A and B are products of identity matrices and elementaries, the determinant of the product is the product of the determinants.

So it is that (14.11) holds in all three cases, as was to be demonstrated. *The determinant of a matrix product is the product of the matrix determinants.*

14.1.5 Determinants of inverse and unitary matrices

From (14.11) it follows that

$$\det A^{-1} = \frac{1}{\det A} \tag{14.12}$$

because $A^{-1}A = I_n$ and $\det I_n = 1$.

From (14.6) it follows that if Q is a unitary matrix (§ 13.12), then

$$\begin{aligned} \det Q^* \det Q &= 1, \\ |\det Q| &= 1. \end{aligned} \tag{14.13}$$

This reason is that $|\det Q|^2 = (\det Q)^*(\det Q) = \det Q^* \det Q = \det Q^*Q = \det Q^{-1}Q = \det I_n = 1$.

14.1.6 Inverting the square matrix by determinant

The Gauss-Jordan algorithm comfortably inverts concrete matrices of moderate size, but swamps one in nearly interminable algebra when *symbolically* inverting general matrices larger than the A_2 at the section's head. Slogging through the algebra to invert A_3 symbolically nevertheless (the reader need not actually do this unless he desires a long exercise), one quite incidentally discovers a clever way to factor the determinant:

$$\begin{aligned} C^TA &= (\det A)I_n = AC^T; \\ c_{ij} &\equiv \det R_{ij}; \\ [R_{ij}]_{i'j'} &\equiv \begin{cases} 1 & \text{if } i' = i \text{ and } j' = j, \\ 0 & \text{if } i' = i \text{ or } j' = j \text{ but not both,} \\ a_{i'j'} & \text{otherwise.} \end{cases} \end{aligned} \tag{14.14}$$

Pictorially,

$$R_{ij} = \begin{bmatrix} & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & * & * & 0 & * & * & \cdots \\ \cdots & * & * & 0 & * & * & \cdots \\ \cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\ \cdots & * & * & 0 & * & * & \cdots \\ \cdots & * & * & 0 & * & * & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix},$$

same as A except in the ith row and jth column. The matrix C, called the *cofactor* of A, then consists of the determinants of the various R_{ij}.

Another way to write (14.14) is

$$[C^TA]_{ij} = (\det A)\delta_{ij} = [AC^T]_{ij}, \tag{14.15}$$

which comprises two cases. In the case that $i = j$,

$$[AC^T]_{ij} = [AC^T]_{ii} = \sum_\ell a_{i\ell}c_{i\ell} = \sum_\ell a_{i\ell} \det R_{i\ell} = \det A = (\det A)\delta_{ij},$$

wherein the equation $\sum_\ell a_{i\ell} \det R_{i\ell} = \det A$ states that $\det A$, being a determinant, consists of several terms, each term including one factor from each row of A, where $a_{i\ell}$ provides the ith row and $R_{i\ell}$ provides the other rows.[7] In the case that $i \neq j$,

$$[AC^T]_{ij} = \sum_\ell a_{i\ell} c_{j\ell} = \sum_\ell a_{i\ell} \det R_{j\ell} = 0 = (\det A)(0) = (\det A)\delta_{ij},$$

wherein $\sum_\ell a_{i\ell} \det R_{j\ell}$ is the determinant, not of A itself, but rather of A with the jth row replaced by a copy of the ith, which according to (14.5) evaluates to zero. Similar equations can be written for $[C^T A]_{ij}$ in both cases. The two cases together prove (14.15), hence also (14.14).

Dividing (14.14) by $\det A$, we have that[8]

$$\begin{aligned} A^{-1}A &= I_n = AA^{-1}, \\ A^{-1} &= \frac{C^T}{\det A}. \end{aligned} \tag{14.16}$$

Equation (14.16) inverts a matrix by determinant. In practice, it inverts small matrices nicely, through about 4×4 dimensionality (the A_2^{-1} equation at the head of the section is just eqn. 14.16 for $n = 2$). It inverts 5×5 and even 6×6 matrices reasonably, too—especially with the help of a computer to do the arithmetic. Though (14.16) still holds in theory for yet larger matrices, and though symbolically it expresses the inverse of an abstract, $n \times n$ matrix concisely whose entries remain unspecified, for concrete matrices much bigger than 4×4 to 6×6 or so its several determinants begin to grow too great and too many for practical calculation. The Gauss-Jordan technique (or even the Gram-Schmidt technique) is preferred to invert concrete matrices above a certain size for this reason.[9]

14.2 Coïncident properties

Chapters 11, 12 and 13, plus this chapter up to the present point, have discovered several coïncident properties of the invertible $n \times n$ square matrix.

[7] This is a bit subtle, but if you actually write out A_3 and its cofactor C_3 symbolically, trying (14.15) on them, then you will soon see what is meant.

[8] Cramer's rule [59, § 1.6], of which the reader may have heard, results from applying (14.16) to (13.4). However, Cramer's rule is really nothing more than (14.16) in a less pleasing form, so this book does not treat Cramer's rule as such.

[9] For very large matrices, even the Gauss-Jordan grows impractical, due to compound floating-point rounding error and the maybe large but nonetheless limited quantity of available computer memory. Iterative techniques, regrettably beyond this edition's scope, serve to invert such matrices approximately.

One does not feel the full impact of the coïncidence when these properties are left scattered across the long chapters; so, let us gather and summarize the properties here. A square, $n \times n$ matrix evidently has either all of the following properties or none of them, never some but not others.

- The matrix is invertible (§ 13.1).
- Its rows are linearly independent (§§ 12.1 and 12.3.4).
- Its columns are linearly independent (§ 12.5.4).
- Its columns address the same space the columns of I_n address, and its rows address the same space the rows of I_n address (§ 12.5.7).
- The Gauss-Jordan algorithm reduces it to I_n (§ 12.3.3). (In this, per § 12.5.3, the choice of pivots does not matter.)
- Decomposing it, the Gram-Schmidt algorithm achieves a fully square, unitary, $n \times n$ factor Q (§ 13.11.2).
- It has full rank $r = n$ (§ 12.5.4).
- The linear system $A\mathbf{x} = \mathbf{b}$ it represents has a unique n-element solution $\mathbf{x}$, given any specific n-element driving vector $\mathbf{b}$ (§ 13.2).
- The determinant $\det A \neq 0$ (§ 14.1.3).
- None of its eigenvalues is zero (§ 14.3, below).

The square matrix which has one of these properties, has all of them. The square matrix which lacks one, lacks all. Assuming exact arithmetic, a square matrix is either invertible, with all that that implies, or singular; never both. The distinction between invertible and singular matrices is theoretically as absolute as (and is indeed analogous to) the distinction between nonzero and zero scalars.

Whether the distinction is always useful is another matter. Usually the distinction is indeed useful, but a matrix can be *almost* singular just as a scalar can be almost zero. Such a matrix is known, among other ways, by its unexpectedly small determinant. Now it is true: in exact arithmetic, a nonzero determinant, no matter how small, implies a theoretically invertible matrix. Practical matrices however often have entries whose values are imprecisely known; and even when they don't, the computers that invert them tend to do arithmetic imprecisely in floating-point. Matrices which live on the hazy frontier between invertibility and singularity resemble the infinitesimals of § 4.1.1. They are called *ill-conditioned* matrices. Section 14.8 develops the topic.

14.3 The eigenvalue itself

We stand ready at last to approach the final major agent of matrix arithmetic, the *eigenvalue.* Suppose a square, $n \times n$ matrix A, a nonzero n-element vector

$$\mathbf{v} = I_n \mathbf{v} \neq 0, \tag{14.17}$$

and a scalar λ, together such that

$$A\mathbf{v} = \lambda \mathbf{v}, \tag{14.18}$$

or in other words such that $A\mathbf{v} = \lambda I_n \mathbf{v}$. If so, then

$$[A - \lambda I_n]\mathbf{v} = 0. \tag{14.19}$$

Since $I_n\mathbf{v}$ is nonzero, the last equation is true if and only if the matrix $[A - \lambda I_n]$ is singular—which in light of § 14.1.3 is to demand that

$$\det(A - \lambda I_n) = 0. \tag{14.20}$$

The left side of (14.20) is an nth-order polynomial in λ, the *characteristic polynomial,* whose n roots are the *eigenvalues*[10] of the matrix A.

What is an eigenvalue, really? An eigenvalue is a scalar a matrix resembles under certain conditions. When a matrix happens to operate on the right *eigenvector* $\mathbf{v}$, it is all the same whether one applies the entire matrix or just the eigenvalue to the vector. The matrix scales the eigenvector by the eigenvalue without otherwise altering the vector, changing the vector's

[10] An example:

$$\begin{aligned}
A &= \begin{bmatrix} 2 & 0 \\ 3 & -1 \end{bmatrix}, \\
\det(A - \lambda I_n) &= \det \begin{bmatrix} 2-\lambda & 0 \\ 3 & -1-\lambda \end{bmatrix} \\
&= (2-\lambda)(-1-\lambda) - (0)(3) \\
&= \lambda^2 - \lambda - 2 = 0, \\
\lambda &= -1 \text{ or } 2.
\end{aligned}$$

magnitude but not its direction. The eigenvalue alone takes the place of the whole, hulking matrix. This is what (14.18) means. Of course it works only when $\mathbf{v}$ happens to be the right *eigenvector,* which § 14.4 discusses.

Observe incidentally that the characteristic polynomial of an $n \times n$ matrix always enjoys full order n regardless of the matrix's rank. The reason lies in the determinant $\det(A - \lambda I_n)$, which comprises exactly $n!$ determinant-terms (we say "determinant-terms" rather than "terms" here only to avoid confusing the determinant's terms with the characteristic polynomial's), only one of which, $(a_{11} - \lambda)(a_{22} - \lambda) \cdots (a_{nn} - \lambda)$, gathers elements straight down the main diagonal of the matrix $[A - \lambda I_n]$. When multiplied out, this main-diagonal determinant-term evidently contributes a $(-\lambda)^n$ to the characteristic polynomial, whereas none of the other determinant-terms finds enough factors of λ to reach order n. (If unsure, take your pencil and just calculate the characteristic polynomials of the 3×3 matrices I_3 and 0. You will soon see what is meant.)

On the other hand, nothing prevents $\lambda = 0$. When $\lambda = 0$, (14.20) makes $\det A = 0$, which as we have said is the sign of a singular matrix. Zero eigenvalues and singular matrices always travel together. *Singular matrices each have at least one zero eigenvalue; nonsingular matrices never do.*

The eigenvalues of a matrix's inverse are the inverses of the matrix's eigenvalues. That is,

$$\lambda_j' \lambda_j = 1 \text{ for all } 1 \le j \le n \text{ if } A'A = I_n = AA'. \tag{14.21}$$

The reason behind (14.21) comes from answering the question: if $A\mathbf{v}_j$ scales $\mathbf{v}_j$ by the factor λ_j, then what does $A'A\mathbf{v}_j = I\mathbf{v}_j$ do to $\mathbf{v}_j$?

Naturally one must solve (14.20)'s nth-order polynomial to locate the actual eigenvalues. One solves it by the same techniques by which one solves any polynomial: the quadratic formula (2.2); the cubic and quartic methods of chapter 10; the Newton-Raphson iteration (4.30). On the other hand, the determinant (14.20) can be impractical to expand for a large matrix; here iterative techniques[11] help.[12]

14.4 The eigenvector

It is an odd fact that (14.19) and (14.20) reveal the eigenvalues λ of a square matrix A while obscuring the associated *eigenvectors* $\mathbf{v}$. Once one has calculated an eigenvalue, though, one can feed it back to calculate the

[11] Such iterative techniques are regrettably not treated by this edition.

[12] The inexpensive [59] also covers the topic competently and readably.

associated eigenvector. According to (14.19), the eigenvectors are the n-element vectors for which

$$[A - \lambda I_n]\mathbf{v} = 0,$$

which is to say that the eigenvectors are the vectors of the kernel space of the degenerate matrix $[A-\lambda I_n]$—which one can calculate (among other ways) by the Gauss-Jordan kernel formula (13.7) or by a method exploiting (13.48).

An eigenvalue and its associated eigenvector, taken together, are sometimes called an *eigensolution.*

14.5 Eigensolution facts

Many useful or interesting mathematical facts concern the eigensolution, among them the following.

- *If the eigensolutions of A are $(\lambda_j, \mathbf{v}_j)$, then the eigensolutions of $A + \alpha I_n$ are $(\lambda_j + \alpha, \mathbf{v}_j)$.* The eigenvalues move over by αI_n while the eigenvectors remain fixed. This is seen by adding $\alpha\mathbf{v}_j$ to both sides of the definition $A\mathbf{v}_j = \lambda\mathbf{v}_j$.

- *A matrix and its inverse share the same eigenvectors with inverted eigenvalues.* Refer to (14.21) and its explanation in § 14.3.

- *Eigenvectors corresponding to distinct eigenvalues are always linearly independent of one another.* To prove this fact, consider several independent eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k-1}$ respectively with distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{k-1}$, and further consider another eigenvector $\mathbf{v}_k$ which might or might not be independent but which too has a distinct eigenvalue λ_k. Were $\mathbf{v}_k$ dependent, which is to say, did nontrivial coefficients c_j exist such that

$$\mathbf{v}_k = \sum_{j=1}^{k-1} c_j \mathbf{v}_j,$$

then left-multiplying the equation by $A - \lambda_k I_n$ would yield

$$0 = \sum_{j=1}^{k-1} (\lambda_j - \lambda_k) c_j \mathbf{v}_j,$$

impossible since the $k-1$ eigenvectors are independent. Thus $\mathbf{v}_k$ too is independent, whereupon by induction from a start case of $k = 1$

we conclude that there exists no dependent eigenvector with a distinct eigenvalue.

- *If an $n \times n$ square matrix A has n independent eigenvectors* (which is always so if the matrix has n distinct eigenvalues and often so even otherwise), *then any n-element vector can be expressed as a unique linear combination of the eigenvectors.* This is a simple consequence of the fact that the $n \times n$ matrix V whose columns are the several eigenvectors $\mathbf{v}_j$ has full rank $r = n$. Unfortunately, some matrices with repeated eigenvalues also have repeated eigenvectors—as for example, curiously,[13] $[1\ 0; 1\ 1]^T$, whose double eigenvalue $\lambda = 1$ has the single eigenvector $[1\ 0]^T$. Section 14.10.2 speaks of matrices of the last kind.

- *An $n \times n$ square matrix whose eigenvectors are linearly independent of one another cannot share all eigensolutions with any other $n \times n$ square matrix.* This fact proceeds from the last point, that every n-element vector $\mathbf{x}$ is a unique linear combination of independent eigenvectors. Neither of the two proposed matrices A_1 and A_2 could scale any of the eigenvector components of $\mathbf{x}$ differently than the other matrix did, so $A_1\mathbf{x} - A_2\mathbf{x} = (A_1 - A_2)\mathbf{x} = 0$ for all $\mathbf{x}$, which in turn is possible only if $A_1 = A_2$.

- *A positive definite matrix has only real, positive eigenvalues. A nonnegative definite matrix has only real, nonnegative eigenvalues.* Were it not so, then $\mathbf{v}^*A\mathbf{v} = \lambda\mathbf{v}^*\mathbf{v}$ (in which $\mathbf{v}^*\mathbf{v}$ naturally is a positive real scalar) would violate the criterion for positive or nonnegative definiteness. See § 13.6.3.

- *Every $n \times n$ square matrix has at least one eigensolution* if $n > 0$, because according to the fundamental theorem of algebra (6.15) the matrix's characteristic polynomial (14.20) has at least one root, an eigenvalue, which by definition would be no eigenvalue if it had no eigenvector to scale, and for which (14.19) necessarily admits at least one nonzero solution $\mathbf{v}$ because its matrix $A - \lambda I_n$ is degenerate.

14.6 Diagonalization

Any $n \times n$ matrix with n independent eigenvectors (which class per § 14.5 includes, but is not limited to, every $n \times n$ matrix with n distinct eigenvalues)

[13] [73]

can be *diagonalized* as

$$A = V\Lambda V^{-1}, \tag{14.22}$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} & 0 \\ 0 & 0 & \cdots & 0 & \lambda_n \end{bmatrix}$$

is an otherwise empty $n \times n$ matrix with the eigenvalues of A set along its main diagonal and where

$$V = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_{n-1} & \mathbf{v}_n \end{bmatrix}$$

is an $n \times n$ matrix whose columns are the eigenvectors of A. This is so because the identity $A\mathbf{v}_j = \mathbf{v}_j\lambda_j$ holds for all $1 \le j \le n$; or, expressed more concisely, because the identity

$$AV = V\Lambda \tag{14.23}$$

holds (reason: the jth column of the product AV is $A\mathbf{v}_j$, whereas the jth column of Λ having just the one element acts to scale V's jth column only). The matrix V is invertible because its columns the eigenvectors are independent, from which (14.22) follows. Equation (14.22) is called the *eigenvalue decomposition,* the *diagonal decomposition* or the *diagonalization* of the square matrix A.

One might object that we had shown only how to compose some matrix $V\Lambda V^{-1}$ with the correct eigenvalues and independent eigenvectors, but had failed to show that the matrix was actually A. However, we need not show this, because § 14.5 has already demonstrated that two matrices with the same eigenvalues and independent eigenvectors are in fact the same matrix, whereby the product $V\Lambda V^{-1}$ can be nothing other than A.

An $n \times n$ matrix with n independent eigenvectors (which class, again, includes every $n \times n$ matrix with n distinct eigenvalues and also includes many matrices with fewer) is called a *diagonalizable* matrix. Besides factoring a diagonalizable matrix by (14.22), one can apply the same formula to compose a diagonalizable matrix with desired eigensolutions.

The diagonal matrix $\operatorname{diag}\{\mathbf{x}\}$ of (11.55) is trivially diagonalizable as $\operatorname{diag}\{\mathbf{x}\} = I_n \operatorname{diag}\{\mathbf{x}\} I_n$.

It is a curious and useful fact that

$$A^2 = (V\Lambda V^{-1})(V\Lambda V^{-1}) = V\Lambda^2 V^{-1}$$

and by extension that

$$A^k = V\Lambda^k V^{-1} \tag{14.24}$$

for any diagonalizable matrix A. The diagonal matrix Λ^k is nothing more than the diagonal matrix Λ with each element individually raised to the kth power, such that

$$\left[\Lambda^k\right]_{ij} = \delta_{ij}\lambda_j^k.$$

Changing $z \leftarrow k$ implies the generalization[14]

$$\begin{aligned} A^z &= V\Lambda^z V^{-1}, \\ \left[\Lambda^z\right]_{ij} &= \delta_{ij}\lambda_j^z, \end{aligned} \tag{14.25}$$

good for any diagonalizable A and complex z.

Nondiagonalizable matrices are troublesome and interesting. The nondiagonalizable matrix vaguely resembles the singular matrix in that both represent edge cases and can be hard to handle numerically; but the resemblance ends there, and a matrix can be either without being the other. The $n \times n$ null matrix for example is singular but still diagonalizable. What a nondiagonalizable matrix is, is, in essence, a matrix with a repeated eigensolution: the same eigenvalue with the same eigenvector, twice or more. More formally, a nondiagonalizable matrix is a matrix with an n-fold eigenvalue whose corresponding eigenvector space fewer than n eigenvectors fully characterize. Section 14.10.2 will have more to say about the nondiagonalizable matrix.

14.7 Remarks on the eigenvalue

Eigenvalues and their associated eigenvectors stand among the principal causes that one should go to such considerable trouble to develop matrix theory as we have done in recent chapters. The idea that a matrix resembles a humble scalar in the right circumstance is powerful. Among the reasons for this is that a matrix can represent an iterative process, operating repeatedly on a vector $\mathbf{v}$ to change it first to $A\mathbf{v}$, then to $A^2\mathbf{v}$, $A^3\mathbf{v}$ and so on. The *dominant eigenvalue* of A, largest in magnitude, tends then to transform $\mathbf{v}$ into the associated eigenvector, gradually but relatively eliminating all other components of $\mathbf{v}$. Should the dominant eigenvalue have greater than unit

[14] It may not be clear however according to (5.13) which branch of λ_j^z one should choose at each index j, especially if A has negative or complex eigenvalues.

magnitude, it destabilizes the iteration; thus one can sometimes judge the stability of a physical process indirectly by examining the eigenvalues of the matrix which describes it. Then there is the edge case of the nondiagonalizable matrix, which matrix surprisingly covers only part of its domain with eigenvectors. All this is fairly deep mathematics. It brings an appreciation of the matrix for reasons which were anything but apparent from the outset of chapter 11.

Remarks continue in §§ 14.10.2 and 14.13.

14.8 Matrix condition

The largest in magnitude of the several eigenvalues of a diagonalizable operator A, denoted here λ_{max}, tends to dominate the iteration $A^k\mathbf{x}$. Section 14.7 has named λ_{max} the *dominant eigenvalue* for this reason.

One sometimes finds it convenient to normalize a dominant eigenvalue by defining a new operator $A' \equiv A/\left|\lambda_{\mathrm{max}}\right|$, whose own dominant eigenvalue $\lambda_{\mathrm{max}}/\left|\lambda_{\mathrm{max}}\right|$ has unit magnitude. In terms of the new operator, the iteration becomes $A^k\mathbf{x} = \left|\lambda_{\mathrm{max}}\right|^k A'^k\mathbf{x}$, leaving one free to carry the magnifying effect $\left|\lambda_{\mathrm{max}}\right|^k$ separately if one prefers to do so. However, the scale factor $1/\left|\lambda_{\mathrm{max}}\right|$ scales all eigenvalues equally; thus, if A's eigenvalue of *smallest* magnitude is denoted λ_{min}, then the corresponding eigenvalue of A' is $\lambda_{\mathrm{min}}/\left|\lambda_{\mathrm{max}}\right|$. If zero, then both matrices according to § 14.3 are singular; if nearly zero, then both matrices are ill conditioned.

Such considerations lead us to define the *condition* of a diagonalizable matrix quantitatively as[15]

$$\kappa \equiv \left|\frac{\lambda_{\mathrm{max}}}{\lambda_{\mathrm{min}}}\right|, \tag{14.26}$$

by which

$$\kappa \ge 1 \tag{14.27}$$

is always a real number of no less than unit magnitude. For best invertibility, $\kappa = 1$ would be ideal (it would mean that all eigenvalues had the same magnitude), though in practice quite a broad range of κ is usually acceptable. Could we always work in exact arithmetic, the value of κ might not interest us much as long as it stayed finite; but in computer floating point, or where the elements of A are known only within some tolerance, infinite κ tends to emerge imprecisely rather as large $\kappa \gg 1$. An *ill-conditioned* matrix by

[15] [159]

definition[16] is a matrix of large $\kappa \gg 1$. The applied mathematician handles such a matrix with due skepticism.

Matrix condition so defined turns out to have another useful application. Suppose that a diagonalizable matrix A is precisely known but that the corresponding driving vector $\mathbf{b}$ is not. If

$$A(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b},$$

where $\delta\mathbf{b}$ is the error in $\mathbf{b}$ and $\delta\mathbf{x}$ is the resultant error in $\mathbf{x}$, then one should like to bound the ratio $|\delta\mathbf{x}| / |\mathbf{x}|$ to ascertain the reliability of $\mathbf{x}$ as a solution. Transferring A to the equation's right side,

$$\mathbf{x} + \delta\mathbf{x} = A^{-1}(\mathbf{b} + \delta\mathbf{b}).$$

Subtracting $\mathbf{x} = A^{-1}\mathbf{b}$ and taking the magnitude,

$$|\delta\mathbf{x}| = \left|A^{-1}\,\delta\mathbf{b}\right|.$$

Dividing by $|\mathbf{x}| = \left|A^{-1}\mathbf{b}\right|$,

$$\frac{|\delta\mathbf{x}|}{|\mathbf{x}|} = \frac{\left|A^{-1}\,\delta\mathbf{b}\right|}{|A^{-1}\mathbf{b}|}.$$

The quantity $\left|A^{-1}\,\delta\mathbf{b}\right|$ cannot exceed $\left|\lambda_{\mathrm{min}}^{-1}\,\delta\mathbf{b}\right|$. The quantity $\left|A^{-1}\mathbf{b}\right|$ cannot fall short of $\left|\lambda_{\mathrm{max}}^{-1}\mathbf{b}\right|$. Thus,

$$\frac{|\delta\mathbf{x}|}{|\mathbf{x}|} \le \frac{\left|\lambda_{\mathrm{min}}^{-1}\,\delta\mathbf{b}\right|}{\left|\lambda_{\mathrm{max}}^{-1}\mathbf{b}\right|} = \left|\frac{\lambda_{\mathrm{max}}}{\lambda_{\mathrm{min}}}\right| \frac{|\delta\mathbf{b}|}{|\mathbf{b}|}.$$

That is,

$$\frac{|\delta\mathbf{x}|}{|\mathbf{x}|} \le \kappa \frac{|\delta\mathbf{b}|}{|\mathbf{b}|}. \tag{14.28}$$

Condition, incidentally, might technically be said to apply to scalars as well as to matrices, but ill condition remains a property of matrices alone. According to (14.26), the condition of every nonzero scalar is happily $\kappa = 1$.

[16] There is of course no definite boundary, no particular edge value of κ, less than which a matrix is well conditioned, at and beyond which it turns ill-conditioned; but you knew that already. If I tried to claim that a matrix with a fine $\kappa = 3$ were ill conditioned, for instance, or that one with a wretched $\kappa = 2^{0\mathrm{x}18}$ were well conditioned, then you might not credit me—but the mathematics nevertheless can only give the number; it remains to the mathematician to interpret it.

14.9 The similarity transformation

Any collection of vectors assembled into a matrix can serve as a *basis* by which other vectors can be expressed. For example, if the columns of

$$B = \begin{bmatrix} 1 & -1 \\ 0 & 2 \\ 0 & 1 \end{bmatrix}$$

are regarded as a basis, then the vector

$$B \begin{bmatrix} 5 \\ 1 \end{bmatrix} = 5 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 1 \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \\ 1 \end{bmatrix}$$

is $(5, 1)$ in the basis B: five times the first basis vector plus once the second. The basis provides the units from which other vectors can be built.

Particularly interesting is the $n \times n$, invertible *complete basis* B, in which the n basis vectors are independent and address the same full space the columns of I_n address. If

$$\mathbf{x} = B\mathbf{u}$$

then $\mathbf{u}$ represents $\mathbf{x}$ in the basis B. Left-multiplication by B evidently converts out of the basis. Left-multiplication by B^{-1},

$$\mathbf{u} = B^{-1}\mathbf{x},$$

then does the reverse, converting into the basis. One can therefore convert any operator A to work within a complete basis B by the successive steps

$$\begin{aligned} A\mathbf{x} &= \mathbf{b}, \\ AB\mathbf{u} &= \mathbf{b}, \\ [B^{-1}AB]\mathbf{u} &= B^{-1}\mathbf{b}, \end{aligned}$$

by which the operator $B^{-1}AB$ is seen to be the operator A, only transformed to work within the basis[17,18] B.

[17] The reader may need to ponder the basis concept a while to grasp it, but the concept is simple once grasped and little purpose would be served by dwelling on it here. Basically, the idea is that one can build the same vector from alternate building blocks, not only from the standard building blocks $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$, etc.—except that the right word for the relevant "building block" is *basis vector.* The books [75] and [106] introduce the basis more gently; one might consult one of those if needed.

[18] The professional matrix literature sometimes distinguishes by typeface between the matrix B and the basis B its columns represent. Such semantical distinctions seem a little too fine for applied use, though. This book just uses B.

The conversion from A into $B^{-1}AB$ is called a *similarity transformation.* If B happens to be unitary (§ 13.12), then the conversion is also called a *unitary transformation.* The matrix $B^{-1}AB$ the transformation produces is said to be *similar* (or, if B is unitary, *unitarily similar*) to the matrix A. We have already met the similarity transformation in §§ 11.5 and 12.2. Now we have the theory to appreciate it properly.

Probably the most important property of the similarity transformation is that it alters no eigenvalues. That is, if

$$A\mathbf{x} = \lambda\mathbf{x},$$

then, by successive steps,

$$\begin{aligned} B^{-1}A(BB^{-1})\mathbf{x} &= \lambda B^{-1}\mathbf{x}, \\ [B^{-1}AB]\mathbf{u} &= \lambda\mathbf{u}. \end{aligned} \tag{14.29}$$

The eigenvalues of A and the similar $B^{-1}AB$ are the same for any square, $n \times n$ matrix A and any invertible, square, $n \times n$ matrix B.

14.10 The Schur decomposition

The *Schur decomposition* of an arbitrary, $n \times n$ square matrix A is

$$A = QU_SQ^*, \tag{14.30}$$

where Q is an $n \times n$ unitary matrix whose inverse, as for any unitary matrix (§ 13.12), is $Q^{-1} = Q^*$; and where U_S is a general upper triangular matrix which can have any values (even zeros) along its main diagonal. The Schur decomposition is slightly obscure, is somewhat tedious to derive and is of limited use in itself, but serves a theoretical purpose.[19] We derive it here for this reason.

[19] The alternative is to develop the interesting but difficult *Jordan canonical form,* which for brevity's sake this chapter prefers to omit.

14.10.1 Derivation

Suppose that[20] (for some reason, which will shortly grow clear) we have a matrix B of the form

$$B = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \cdots \\ \cdots & 0 & \ast & \ast & \ast & \ast & \ast & \ast & \cdots \\ \cdots & 0 & 0 & \ast & \ast & \ast & \ast & \ast & \cdots \\ \cdots & 0 & 0 & 0 & \ast & \ast & \ast & \ast & \cdots \\ \cdots & 0 & 0 & 0 & 0 & \ast & \ast & \ast & \cdots \\ \cdots & 0 & 0 & 0 & 0 & \ast & \ast & \ast & \cdots \\ \cdots & 0 & 0 & 0 & 0 & \ast & \ast & \ast & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{14.31}$$

where the ith row and ith column are depicted at center. Suppose further that we wish to transform B not only similarly but unitarily into

$$C \equiv W^* B W = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \cdots \\ \cdots & 0 & \ast & \ast & \ast & \ast & \ast & \ast & \cdots \\ \cdots & 0 & 0 & \ast & \ast & \ast & \ast & \ast & \cdots \\ \cdots & 0 & 0 & 0 & \ast & \ast & \ast & \ast & \cdots \\ \cdots & 0 & 0 & 0 & 0 & \ast & \ast & \ast & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & \ast & \ast & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & \ast & \ast & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{14.32}$$

where W is an $n \times n$ unitary matrix, and where we do not mind if any or all of the $\ast$ elements change in going from B to C but we require zeros in the indicated spots. Let B_o and C_o represent the $(n-i) \times (n-i)$ submatrices in the lower right corners respectively of B and C, such that

$$\begin{aligned} B_o &\equiv I_{n-i} H_{-i} B H_i I_{n-i}, \\ C_o &\equiv I_{n-i} H_{-i} C H_i I_{n-i}, \end{aligned} \tag{14.33}$$

[20] This subsection assigns various capital Roman letters to represent the several matrices and submatrices it manipulates. Its choice of letters except in (14.30) is not standard and carries no meaning elsewhere. The writer had to choose some letters and these are ones he chose.

This footnote mentions the fact because good mathematical style avoid assigning letters that already bear a conventional meaning in a related context (for example, this book avoids writing $A\mathbf{x} = \mathbf{b}$ as $T\mathbf{e} = \mathbf{i}$, not because the latter is wrong but because it would be extremely confusing). The Roman alphabet provides only twenty-six capitals, though, of which this subsection uses too many to be allowed to reserve any. See appendix B.

where H_k is the shift operator of § 11.9. Pictorially,

$$B_o = \begin{bmatrix} * & * & * & \cdots \\ * & * & * & \cdots \\ * & * & * & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad C_o = \begin{bmatrix} * & * & * & \cdots \\ 0 & * & * & \cdots \\ 0 & * & * & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Equation (14.32) seeks an $n \times n$ unitary matrix W to transform the matrix B into a new matrix $C \equiv W^*BW$ such that C fits the form (14.32) stipulates. The question remains as to whether a unitary W exists that satisfies the form and whether for general B we can discover a way to calculate it. To narrow the search, because we need not find every W that satisfies the form but only one such W, let us look first for a W that fits the restricted template

$$W = I_i + H_i W_o H_{-i} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 1 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 1 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & * & * & * & \cdots \\ \cdots & 0 & 0 & 0 & 0 & * & * & * & \cdots \\ \cdots & 0 & 0 & 0 & 0 & * & * & * & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{14.34}$$

which contains a smaller, $(n-i) \times (n-i)$ unitary submatrix W_o in its lower right corner and resembles I_n elsewhere. Beginning from (14.32), we have by successive, reversible steps that

$$\begin{aligned} C &= W^*BW \\ &= (I_i + H_i W_o^* H_{-i})(B)(I_i + H_i W_o H_{-i}) \\ &= I_i B I_i + I_i B H_i W_o H_{-i} + H_i W_o^* H_{-i} B I_i \\ &\qquad + H_i W_o^* H_{-i} B H_i W_o H_{-i}. \end{aligned}$$

The unitary submatrix W_o has only $n-i$ columns and $n-i$ rows, so $I_{n-i}W_o =$

$W_o = W_o I_{n-i}$. Thus,[21]

$$\begin{aligned} C &= I_i B I_i + I_i B H_i W_o I_{n-i} H_{-i} + H_i I_{n-i} W_o^* H_{-i} B I_i \\ &\qquad + H_i I_{n-i} W_o^* I_{n-i} H_{-i} B H_i I_{n-i} W_o I_{n-i} H_{-i} \\ &= I_i [B] I_i + I_i [B H_i W_o H_{-i}](I_n - I_i) + (I_n - I_i)[H_i W_o^* H_{-i} B] I_i \\ &\qquad + (I_n - I_i)[H_i W_o^* B_o W_o H_{-i}](I_n - I_i), \end{aligned}$$

where the last step has used (14.33) and the identity (11.76). The four terms on the equation's right, each term with rows and columns neatly truncated, represent the four quarters of $C \equiv W^* B W$—upper left, upper right, lower left and lower right, respectively. The lower left term is null because

$$\begin{aligned} (I_n - I_i)[H_i W_o^* H_{-i} B] I_i &= (I_n - I_i)[H_i W_o^* I_{n-i} H_{-i} B I_i] I_i \\ &= (I_n - I_i)[H_i W_o^* H_{-i}][(I_n - I_i) B I_i] I_i \\ &= (I_n - I_i)[H_i W_o^* H_{-i}][0] I_i = 0, \end{aligned}$$

leaving

$$\begin{aligned} C &= I_i [B] I_i + I_i [B H_i W_o H_{-i}](I_n - I_i) \\ &\qquad + (I_n - I_i)[H_i W_o^* B_o W_o H_{-i}](I_n - I_i). \end{aligned}$$

But the upper left term makes the upper left areas of B and C the same, and the upper right term does not bother us because we have not restricted the content of C's upper right area. Apparently any $(n-i) \times (n-i)$ unitary submatrix W_o whatsoever obeys (14.32) in the lower left, upper left and upper right.

That leaves the lower right. Left- and right-multiplying (14.32) by the truncator $(I_n - I_i)$ to focus solely on the lower right area, we have the reduced requirement that

$$(I_n - I_i) C (I_n - I_i) = (I_n - I_i) W^* B W (I_n - I_i). \tag{14.35}$$

Further left-multiplying by H_{-i}, right-multiplying by H_i, and applying the identity (11.76) yields that

$$I_{n-i} H_{-i} C H_i I_{n-i} = I_{n-i} H_{-i} W^* B W H_i I_{n-i};$$

[21] The algebra is so thick that, even if one can logically follow it, one might nonetheless wonder how the writer had thought to write it. However, much of the algebra consists of crop-and-shift operations like $H_i I_{n-i}$ which, when a sample matrix is sketched on a sheet of paper, are fairly easy to visualize. Indeed, the whole derivation is more visual than the inscrutable symbols let on. The writer had the visuals in mind.

or, substituting from (14.33), that

$$C_o = I_{n-i}H_{-i}W^*BWH_iI_{n-i}.$$

Expanding W per (14.34),

$$C_o = I_{n-i}H_{-i}(I_i + H_iW_o^*H_{-i})B(I_i + H_iW_oH_{-i})H_iI_{n-i};$$

or, since $I_{n-i}H_{-i}I_i = 0 = I_iH_iI_{n-i}$,

$$\begin{aligned} C_o &= I_{n-i}H_{-i}(H_iW_o^*H_{-i})B(H_iW_oH_{-i})H_iI_{n-i} \\ &= I_{n-i}W_o^*H_{-i}BH_iW_oI_{n-i} \\ &= W_o^*I_{n-i}H_{-i}BH_iI_{n-i}W_o. \end{aligned}$$

Per (14.33), this has that

$$C_o = W_o^*B_oW_o. \tag{14.36}$$

The steps from (14.35) to (14.36) are reversible, so the latter is as good a way to state the reduced requirement as the former is. To achieve a unitary transformation of the form (14.32), therefore, it suffices to satisfy (14.36).

The increasingly well-stocked armory of matrix theory we now have to draw from makes satisfying (14.36) possible as follows. Observe per § 14.5 that every square matrix has at least one eigensolution. Let $(\lambda_o, \mathbf{v}_o)$ represent an eigensolution of B_o—*any* eigensolution of B_o—with $\mathbf{v}_o$ normalized to unit magnitude. Form the broad, $(n-i) \times (n-i+1)$ matrix

$$F \equiv \begin{bmatrix} \mathbf{v}_o & \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 & \cdots & \mathbf{e}_{n-i} \end{bmatrix}.$$

Decompose F by the Gram-Schmidt technique of § 13.11.2, choosing $p = 1$ during the first instance of the algorithm's step 3 (though choosing any permissible p thereafter), to obtain

$$F = Q_FR_F.$$

Noting that the Gram-Schmidt algorithm orthogonalizes only rightward, observe that the first column of the $(n-i) \times (n-i)$ unitary matrix Q_F remains simply the first column of F, which is the unit eigenvector $\mathbf{v}_o$:

$$[Q_F]_{*1} = Q_F\mathbf{e}_1 = \mathbf{v}_o.$$

Transform B_o unitarily by Q_F to define the new matrix

$$G \equiv Q_F^*B_oQ_F,$$

then transfer factors to reach the equation

$$Q_F G Q_F^* = B_o.$$

Right-multiplying by $Q_F\mathbf{e}_1 = \mathbf{v}_o$ and noting that $B_o\mathbf{v}_o = \lambda_o\mathbf{v}_o$, observe that

$$Q_F G\mathbf{e}_1 = \lambda_o\mathbf{v}_o.$$

Left-multiplying by Q_F^*,

$$G\mathbf{e}_1 = \lambda_o Q_F^*\mathbf{v}_o.$$

Noting that the Gram-Schmidt process has rendered orthogonal to $\mathbf{v}_o$ all columns of Q_F but the first, which is $\mathbf{v}_o$, observe that

$$G\mathbf{e}_1 = \lambda_o Q_F^*\mathbf{v}_o = \lambda_o\mathbf{e}_1 = \begin{bmatrix} \lambda_o \\ 0 \\ 0 \\ \vdots \end{bmatrix},$$

which means that

$$G = \begin{bmatrix} \lambda_o & * & * & \cdots \\ 0 & * & * & \cdots \\ 0 & * & * & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

which fits the very form (14.33) the submatrix C_o is required to have. Conclude therefore that

$$\begin{aligned} W_o &= Q_F, \\ C_o &= G, \end{aligned} \tag{14.37}$$

where Q_F and G are as this paragraph develops, together constitute a valid choice for W_o and C_o, satisfying the reduced requirement (14.36) and thus also the original requirement (14.32).

Equation (14.37) completes a failsafe technique to transform unitarily any square matrix B of the form (14.31) into a square matrix C of the form (14.32). Naturally the technique can be applied recursively as

$$B|_{i=i'} = C|_{i=i'-1}, \quad 1 \le i' \le n, \tag{14.38}$$

because the form (14.31) of B at $i = i'$ is nothing other than the form (14.32) of C at $i = i' - 1$. Therefore, if we let

$$B|_{i=0} = A, \tag{14.39}$$

then it follows by induction that

$$B|_{i=n} = U_S, \tag{14.40}$$

where per (14.31) the matrix U_S has the general upper triangular form the Schur decomposition (14.30) requires. Moreover, because the product of unitary matrices according to (13.66) is itself a unitary matrix, we have that

$$Q = \coprod_{i'=0}^{n-1} \left(W|_{i=i'} \right), \tag{14.41}$$

which along with (14.40) accomplishes the Schur decomposition.

14.10.2 The nondiagonalizable matrix

The characteristic equation (14.20) of the general upper triangular matrix U_S is

$$\det(U_S - \lambda I_n) = 0.$$

Unlike most determinants, this determinant brings only the one term

$$\det(U_S - \lambda I_n) = \prod_{i=1}^{n} (u_{Sii} - \lambda) = 0$$

whose factors run straight down the main diagonal, where the determinant's $n! - 1$ other terms are all zero because each of them includes at least one zero factor from below the main diagonal.[22] Hence no element above the main diagonal of U_S even influences the eigenvalues, which apparently are

$$\lambda_i = u_{Sii}, \tag{14.42}$$

the main-diagonal elements.

[22] The determinant's definition in § 14.1 makes the following two propositions equivalent: (i) that a determinant's term which includes one or more factors above the main diagonal also includes one or more factors below; (ii) that the only permutor that marks no position below the main diagonal is the one which also marks no position above. In either form, the proposition's truth might seem less than obvious until viewed from the proper angle.

Consider a permutor P. If P marked no position below the main diagonal, then it would necessarily have $p_{nn} = 1$, else the permutor's bottom row would be empty which is not allowed. In the next-to-bottom row, $p_{(n-1)(n-1)} = 1$, because the nth column is already occupied. In the next row up, $p_{(n-2)(n-2)} = 1$; and so on, thus affirming the proposition.

According to (14.29), similarity transformations preserve eigenvalues. The Schur decomposition (14.30) is in fact a similarity transformation; and, as we have seen, every matrix A has a Schur decomposition. If therefore

$$A = QU_SQ^*,$$

then *the eigenvalues of A are just the values along the main diagonal of U_S.*[23]

One might think that the Schur decomposition offered an easy way to calculate eigenvalues, but it is less easy than it first appears because one must calculate eigenvalues to reach the Schur decomposition in the first place. Whatever practical merit the Schur decomposition might have or lack, however, it brings at least the theoretical benefit of (14.42): every square matrix without exception has a Schur decomposition, whose triangular factor U_S openly lists all eigenvalues along its main diagonal.

This theoretical benefit pays when some of the n eigenvalues of an $n \times n$ square matrix A repeat. By the Schur decomposition, one can construct a second square matrix A', as near as desired to A but having n distinct eigenvalues, simply by perturbing the main diagonal of U_S to[24]

$$\begin{aligned} U_S' &\equiv U_S + \epsilon \operatorname{diag}\{\mathbf{u}\}, \\ u_{i'} &\neq u_i \text{ if } \lambda_{i'} = \lambda_i, \end{aligned} \tag{14.43}$$

where $|\epsilon| \ll 1$ and where $\mathbf{u}$ is an arbitrary vector that meets the criterion given. Though infinitesimally near A, the modified matrix $A' = QU_S'Q^*$ unlike A has n (maybe infinitesimally) distinct eigenvalues. With sufficient toil, one might analyze such perturbed eigenvalues and their associated eigenvectors similarly as § 9.7.2 has analyzed perturbed poles.

Equation (14.43) brings us to the nondiagonalizable matrix of the subsection's title. Section 14.6 and its diagonalization formula (14.22) diagonalize

[23] An unusually careful reader might worry that A and U_S had the same eigenvalues with different multiplicities. It would be surprising if it actually were so; but, still, one would like to give a sounder reason than the participle "surprising." Consider however that

$$\begin{aligned} A - \lambda I_n = QU_SQ^* - \lambda I_n &= Q[U_S - Q^*(\lambda I_n)Q]Q^* \\ &= Q[U_S - \lambda(Q^*I_nQ)]Q^* = Q[U_S - \lambda I_n]Q^*. \end{aligned}$$

According to (14.11) and (14.13), this equation's determinant is

$$\det[A - \lambda I_n] = \det\{Q[U_S - \lambda I_n]Q^*\} = \det Q \det[U_S - \lambda I_n] \det Q^* = \det[U_S - \lambda I_n],$$

which says that A and U_S have not only the same eigenvalues but also the same characteristic polynomials, and thus further the same eigenvalue multiplicities.

[24] Equation (11.55) defines the diag$\{\cdot\}$ notation.

any matrix with distinct eigenvalues and even any matrix with repeated eigenvalues but distinct eigenvectors, but fail where eigenvectors repeat. Equation (14.43) separates eigenvalues, and thus also eigenvectors—for according to § 14.5 eigenvectors of distinct eigenvalues never depend on one another—permitting a nonunique but still sometimes usable form of diagonalization in the limit $\epsilon \rightarrow 0$ even when the matrix in question is strictly nondiagonalizable.

The finding that every matrix is arbitrarily nearly diagonalizable illuminates a question the chapter has evaded up to the present point. The question: does a p-fold root in the characteristic polynomial (14.20) necessarily imply a p-fold eigenvalue in the corresponding matrix? The existence of the nondiagonalizable matrix casts a shadow of doubt until one realizes that every nondiagonalizable matrix is arbitrarily nearly diagonalizable—and, better, is arbitrarily nearly diagonalizable with distinct eigenvalues. If you claim that a matrix has a triple eigenvalue and someone disputes the claim, then you can show him a nearly identical matrix with three infinitesimally distinct eigenvalues. That is the essence of the idea. We will leave the answer in that form.

Generalizing the nondiagonalizability concept leads one eventually to the ideas of the *generalized eigenvector*[25] (which solves the higher-order linear system $[A - \lambda I]^k \mathbf{v} = 0$) and the *Jordan canonical form,*[26] which together roughly track the sophisticated conventional pole-separation technique of § 9.7.6. Then there is a kind of sloppy Schur form called a Hessenberg form which allows content in U_S along one or more subdiagonals just beneath the main diagonal. One could profitably propose and prove any number of useful theorems concerning the nondiagonalizable matrix and its generalized eigenvectors, or concerning the eigenvalue problem[27] more broadly, in more and less rigorous ways, but for the time being we will let the matter rest there.

14.11 The Hermitian matrix

An $m \times m$ square matrix A that is its own adjoint,

$$A^* = A, \tag{14.44}$$

[25] [63, chapter 7]

[26] [59, chapter 5]

[27] [184]

is called a *Hermitian* or *self-adjoint* matrix. Properties of the Hermitian matrix include that

- its eigenvalues are real,
- its eigenvectors corresponding to distinct eigenvalues lie orthogonal to one another, and
- it is unitarily diagonalizable (§§ 13.12 and 14.6) such that

$$A = V\Lambda V^*. \tag{14.45}$$

That the eigenvalues are real is proved by letting $(\lambda, \mathbf{v})$ represent an eigensolution of A and constructing the product $\mathbf{v}^*A\mathbf{v}$, for which

$$\lambda^*\mathbf{v}^*\mathbf{v} = (A\mathbf{v})^*\mathbf{v} = \mathbf{v}^*A\mathbf{v} = \mathbf{v}^*(A\mathbf{v}) = \lambda\mathbf{v}^*\mathbf{v}.$$

That is,

$$\lambda^* = \lambda,$$

which naturally is possible only if λ is real.

That eigenvectors corresponding to distinct eigenvalues lie orthogonal to one another is proved[28] by letting $(\lambda_1, \mathbf{v}_1)$ and $(\lambda_2, \mathbf{v}_2)$ represent eigensolutions of A and constructing the product $\mathbf{v}_2^*A\mathbf{v}_1$, for which

$$\lambda_2^*\mathbf{v}_2^*\mathbf{v}_1 = (A\mathbf{v}_2)^*\mathbf{v}_1 = \mathbf{v}_2^*A\mathbf{v}_1 = \mathbf{v}_2^*(A\mathbf{v}_1) = \lambda_1\mathbf{v}_2^*\mathbf{v}_1.$$

That is,

$$\lambda_2^* = \lambda_1 \quad \text{or} \quad \mathbf{v}_2^*\mathbf{v}_1 = 0.$$

But according to the last paragraph all eigenvalues are real; the eigenvalues λ_1 and λ_2 are no exceptions. Hence,

$$\lambda_2 = \lambda_1 \quad \text{or} \quad \mathbf{v}_2^*\mathbf{v}_1 = 0.$$

To prove the last hypothesis of the three needs first some definitions as follows. Given an $m \times m$ matrix A, let the s columns of the $m \times s$ matrix V_o represent the s independent eigenvectors of A such that (i) each column has unit magnitude and (ii) columns whose eigenvectors share the same eigenvalue lie orthogonal to one another. Let the $s \times s$ diagonal matrix Λ_o carry the eigenvalues on its main diagonal such that

$$AV_o = V_o\Lambda_o,$$

[28] [106, § 8.1]

where the distinction between the matrix Λ_o and the full eigenvalue matrix Λ of (14.22) is that the latter always includes a p-fold eigenvalue p times, whereas the former includes a p-fold eigenvalue only as many times as the eigenvalue enjoys independent eigenvectors. Let the $m-s$ columns of the $m\times(m-s)$ matrix $V_o^{\perp}$ represent the complete orthogonal complement (§ 13.10) to V_o—perpendicular to all eigenvectors, each column of unit magnitude—such that

$$V_o^{\perp *}V_o = 0 \text{ and } V_o^{\perp *}V_o^{\perp} = I_{m-s}.$$

Recall from § 14.5 that $s \neq 0$ but $0 < s \leq m$ because every square matrix has at least one eigensolution. Recall from § 14.6 that $s = m$ if and only if A is diagonalizable.[29]

With these definitions in hand, we can now prove by contradiction that all Hermitian matrices are diagonalizable, falsely supposing a nondiagonalizable Hermitian matrix A, whose $V_o^{\perp}$ (since A is supposed to be nondiagonalizable, implying that $s < m$) would have at least one column. For such a matrix A, $s \times (m-s)$ and $(m-s)\times(m-s)$ auxiliary matrices F and G necessarily would exist such that

$$AV_o^{\perp} = V_oF + V_o^{\perp}G,$$

not due to any unusual property of the product $AV_o^{\perp}$ but for the mundane reason that the columns of V_o and $V_o^{\perp}$ together by definition addressed

[29] A concrete example: the invertible but nondiagonalizable matrix

$$A = \begin{bmatrix} -1 & 0 & 0 & 0 \\ -6 & 5 & \frac{5}{2} & -\frac{5}{2} \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}$$

has a single eigenvalue at $\lambda = -1$ and a triple eigenvalue at $\lambda = 5$, the latter of whose eigenvector space is fully characterized by two eigenvectors rather than three such that

$$V_o = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}} & 1 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}, \quad \Lambda_o = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix}, \quad V_o^{\perp} = \begin{bmatrix} 0 \\ 0 \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

The orthogonal complement $V_o^{\perp}$ supplies the missing vector, not an eigenvector but perpendicular to them all.

In the example, $m = 4$ and $s = 3$.

All vectors in the example are reported with unit magnitude. The two $\lambda = 5$ eigenvectors are reported in mutually orthogonal form, but notice that eigenvectors corresponding to distinct eigenvalues need not be orthogonal when A is not Hermitian.

the space of *all* m-element vectors—including the columns of $AV_o^{\perp}$. Left-multiplying by V_o^*, we would have by successive steps that

$$\begin{aligned}
V_o^* A V_o^{\perp} &= V_o^* V_o F + V_o^* V_o^{\perp} G, \\
(AV_o)^* V_o^{\perp} &= I_s F + V_o^* V_o^{\perp} G, \\
(V_o \Lambda_o)^* V_o^{\perp} &= F + V_o^* V_o^{\perp} G, \\
\Lambda_o^* V_o^* V_o^{\perp} &= F + V_o^* V_o^{\perp} G, \\
\Lambda_o^*(0) &= F + (0)G, \\
0 &= F,
\end{aligned}$$

where we had relied on the assumption that A were Hermitian and thus that, as proved above, its distinctly eigenvalued eigenvectors lay orthogonal to one another; in consequence of which $A^* = A$ and $V_o^* V_o = I_s$.

The finding that $F = 0$ reduces the $AV_o^{\perp}$ equation above to read

$$AV_o^{\perp} = V_o^{\perp} G.$$

In the reduced equation the matrix G would have at least one eigensolution, not due to any unusual property of G but because according to § 14.5 every square matrix, 1×1 or larger, has at least one eigensolution. Let $(\mu, \mathbf{w})$ represent an eigensolution of G. Right-multiplying by the $(m - s)$-element vector $\mathbf{w} \neq 0$, we would have by successive steps that

$$\begin{aligned}
AV_o^{\perp}\mathbf{w} &= V_o^{\perp} G \mathbf{w}, \\
A(V_o^{\perp}\mathbf{w}) &= \mu(V_o^{\perp}\mathbf{w}).
\end{aligned}$$

The last equation claims that $(\mu, V_o^{\perp}\mathbf{w})$ were an eigensolution of A, when we had supposed that all of A's eigenvectors lay in the space addressed by the columns of V_o, and thus by construction did not lie in the space addressed by the columns of $V_o^{\perp}$. The contradiction proves false the assumption that gave rise to it. The assumption: that a nondiagonalizable Hermitian A existed. We conclude that all Hermitian matrices are diagonalizable—and conclude further that they are *unitarily* diagonalizable on the ground that their eigenvectors lie orthogonal to one another—as was to be demonstrated.

Having proven that all Hermitian matrices are diagonalizable and have real eigenvalues and orthogonal eigenvectors, one wonders whether the converse holds: are all diagonalizable matrices with real eigenvalues and orthogonal eigenvectors Hermitian? To show that they are, one can construct the matrix described by the diagonalization formula (14.22),

$$A = V \Lambda V^*,$$

where $V^{-1} = V^*$ because this V is unitary (§ 13.12). The equation's adjoint is

$$A^* = V\Lambda^* V^*.$$

But all the eigenvalues here are real, which means that $\Lambda^* = \Lambda$ and the right sides of the two equations are the same. That is, $A^* = A$ as was to be demonstrated. *All diagonalizable matrices with real eigenvalues and orthogonal eigenvectors are Hermitian.*

This section brings properties that simplify many kinds of matrix analysis. The properties demand a Hermitian matrix, which might seem a severe and unfortunate restriction—except that one can left-multiply any exactly determined linear system $C\mathbf{x} = \mathbf{d}$ by C^* to get the equivalent Hermitian system

$$[C^*C]\mathbf{x} = [C^*\mathbf{d}], \tag{14.46}$$

in which $A = C^*C$ and $\mathbf{b} = C^*\mathbf{d}$, for which the properties obtain.[30]

14.12 The singular-value decomposition

Occasionally an elegant idea awaits discovery, overlooked, almost in plain sight. If the unlikely thought occurred to you to take the square root of a matrix, then the following idea is one you might discover.[31]

Consider the $n \times n$ product A^*A of a tall or square, $m \times n$ matrix A of full column rank

$$r = n \leq m$$

and its adjoint A^*. The product A^*A is invertible according to § 13.6.2; is positive definite according to § 13.6.3; and, since $(A^*A)^* = A^*A$, is clearly Hermitian according to § 14.11; thus is unitarily diagonalizable according to (14.45) as

$$A^*A = V\Lambda V^*. \tag{14.47}$$

Here, the $n \times n$ matrices Λ and V represent respectively the eigenvalues and eigenvectors not of A but of the product A^*A. Though nothing requires the product's eigenvectors to be real, because the product is positive definite § 14.5 does require all of its eigenvalues to be real and moreover positive—which means among other things that the eigenvalue matrix Λ has full rank. That the eigenvalues, the diagonal elements of Λ, are real and positive is

[30] The device (14.46) worsens a matrix's condition and may be undesirable for this reason, but it works in theory at least.

[31] [182, "Singular value decomposition," 14:29, 18 Oct. 2007]

a useful fact; for just as a real, positive scalar has a real, positive square root, so equally has Λ a real, positive square root under these conditions. Let the symbol $\Sigma = \sqrt{\Lambda}$ represent the $n \times n$ real, positive square root of the eigenvalue matrix Λ such that

$$\begin{aligned} \Lambda &= \Sigma^*\Sigma, \qquad (14.48) \\ \Sigma^* = \Sigma &= \begin{bmatrix} +\sqrt{\lambda_1} & 0 & \cdots & 0 & 0 \\ 0 & +\sqrt{\lambda_2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & +\sqrt{\lambda_{n-1}} & 0 \\ 0 & 0 & \cdots & 0 & +\sqrt{\lambda_n} \end{bmatrix}, \end{aligned}$$

where the *singular values* of A populate Σ's diagonal. Applying (14.48) to (14.47) then yields that

$$\begin{aligned} A^*A &= V\Sigma^*\Sigma V^*, \\ V^*A^*AV &= \Sigma^*\Sigma. \end{aligned} \qquad (14.49)$$

Now consider the $m \times m$ matrix U such that

$$\begin{aligned} AV\Sigma^{-1} &= UI_n, \\ AV &= U\Sigma, \\ A &= U\Sigma V^*. \end{aligned} \qquad (14.50)$$

Substituting (14.50)'s second line into (14.49)'s second line gives the equation

$$\Sigma^*U^*U\Sigma = \Sigma^*\Sigma;$$

but $\Sigma\Sigma^{-1} = I_n$, so left- and right-multiplying respectively by Σ^{-*} and Σ^{-1} leaves that

$$I_nU^*UI_n = I_n,$$

which says neither more nor less than that the first n columns of U are orthonormal. Equation (14.50) does not constrain the last $m - n$ columns of U, leaving us free to make them anything we want. Why not use Gram-Schmidt to make them orthonormal, too, thus making U a unitary matrix? If we do this, then the surprisingly simple (14.50) constitutes the *singular-value decomposition* of A.

If A happens to have broad shape then we can decompose A^*, instead, so this case poses no special trouble. Apparently every full-rank matrix has a singular-value decomposition.

But what of the matrix of less than full rank $r < n$? In this case the product A^*A is singular and has only $s < n$ nonzero eigenvalues (it may be that $s = r$, but this is irrelevant to the proof at hand). However, if the s nonzero eigenvalues are arranged first in Λ, then (14.50) becomes

$$\begin{aligned} AV\Sigma^{-1} &= UI_s, \\ AV &= U\Sigma, \\ A &= U\Sigma V^*. \end{aligned} \tag{14.51}$$

The product A^*A is nonnegative definite in this case and $\Sigma\Sigma^{-1} = I_s$, but the reasoning is otherwise the same as before. Apparently every matrix of less than full rank has a singular-value decomposition, too.

If A happens to be an invertible square matrix, then the singular-value decomposition evidently inverts it as

$$A^{-1} = V\Sigma^{-1}U^*. \tag{14.52}$$

14.13 General remarks on the matrix

Chapters 11 through 14 have derived the uncomfortably bulky but—incredibly—approximately minimal knot of theory one needs to grasp the matrix properly and to use it with moderate versatility. As far as the writer knows, no one has yet discovered a satisfactory way to untangle the knot. The choice to learn the basic theory of the matrix is almost an all-or-nothing choice; and how many scientists and engineers would rightly choose the "nothing" if the matrix did not serve so very many applications as it does? Since it does serve so very many, the "all" it must be.[32] Applied mathematics brings nothing else quite like it.

These several matrix chapters have not covered every topic they might. The topics they omit fall roughly into two classes. One is the class of more advanced and more specialized matrix theory, about which we will have more to say in a moment. The other is the class of basic matrix theory these chapters do not happen to use. The essential agents of matrix analysis—multiplicative associativity, rank, inversion, pseudoinversion, the kernel, the orthogonal complement, orthonormalization, the eigenvalue, diagonalization and so on—are the same in practically all books on the subject, but the way the agents are developed differs. This book has chosen a way that needs some tools like truncators other books omit, but does not need other tools like

[32] Of course, one might avoid true understanding and instead work by memorized rules. That is not always a bad plan, really; but if that were *your* plan then it seems spectacularly unlikely that you would be reading a footnote buried beneath the further regions of the hinterland of Chapter 14 in such a book as this.

projectors other books[33] include. What has given these chapters their hefty bulk is not so much the immediate development of the essential agents as the preparatory development of theoretical tools used to construct the essential agents, yet most of the tools are of limited interest in themselves; it is the agents that matter. Tools like the projector not used here tend to be omitted here or deferred to later chapters, not because they are altogether useless but because they are not used *here* and because the present chapters are already too long. The reader who understands the Moore-Penrose pseudoinverse and/or the Gram-Schmidt process reasonably well can after all pretty easily figure out how to construct a projector without explicit instructions thereto, should the need arise.[34]

Paradoxically and thankfully, more advanced and more specialized matrix theory though often harder tends to come in smaller, more manageable increments: the Cholesky decomposition, for instance; or the conjugate-gradient algorithm. The theory develops endlessly. From the present pause one could proceed directly to such topics. However, since this is the *first* proper pause these several matrix chapters have afforded, since the book is *Derivations of Applied Mathematics* rather than *Derivations of Applied Matrices,* maybe we ought to take advantage to change the subject.

[33] Such as [75, § 3.VI.3], a lengthy but well-knit tutorial this writer recommends.

[34] Well, since we have brought it up (though only as an example of tools these chapters have avoided bringing up), briefly: a projector is a matrix that flattens an arbitrary vector $\mathbf{b}$ into its nearest shadow $\tilde{\mathbf{b}}$ within some restricted subspace. If the columns of A represent the subspace, then $\mathbf{x}$ represents $\tilde{\mathbf{b}}$ in the subspace basis iff $A\mathbf{x} = \tilde{\mathbf{b}}$, which is to say that $A\mathbf{x} \approx \mathbf{b}$, whereupon $\mathbf{x} = A^\dagger\mathbf{b}$. That is, per (13.33),

$$\tilde{\mathbf{b}} = A\mathbf{x} = AA^\dagger\mathbf{b} = [BC][C^*(CC^*)^{-1}(B^*B)^{-1}B^*]\mathbf{b} = B(B^*B)^{-1}B^*\mathbf{b},$$

in which the matrix $B(B^*B)^{-1}B^*$ is the projector. Thence it is readily shown that the deviation $\mathbf{b} - \tilde{\mathbf{b}}$ lies orthogonal to the shadow $\tilde{\mathbf{b}}$. More broadly defined, any matrix M for which $M^2 = M$ is a projector. One can approach the projector in other ways, but there are two ways at least.

Chapter 15

Vector analysis

Leaving the matrix, this chapter and the next turn to an agent of applied mathematics that, though ubiquitous in some fields of study like physics, remains curiously underappreciated in other fields that should use it more. This agent is the *three-dimensional geometrical vector,* first met in §§ 3.3, 3.4 and 3.9. Seen from one perspective, the three-dimensional geometrical vector is the $n = 3$ special case of the general, n-dimensional vector of chapters 11 through 14. Because its three elements represent the three dimensions of the physical world, however, the three-dimensional geometrical vector merits closer attention and special treatment.[1]

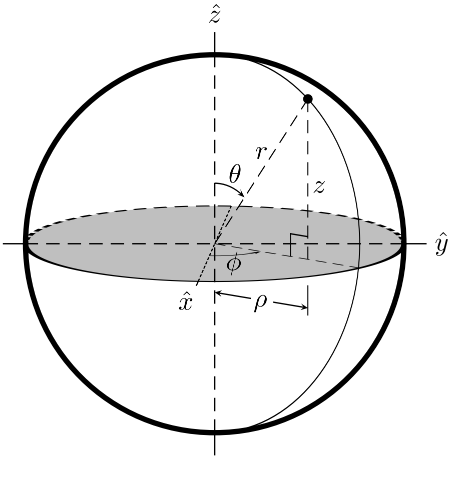
It also merits a shorter name. Where the geometrical context is clear—as it is in this chapter and the next—we will call the three-dimensional geometrical vector just a *vector.* A name like "matrix vector" or "n-dimensional vector" can disambiguate the vector of chapters 11 through 14 where necessary but, since the three-dimensional geometrical vector is in fact a vector in the broader sense, to disambiguate is usually unnecessary. The lone word *vector* serves.

In the present chapter's context and according to § 3.3, a vector consists of an amplitude of some kind plus a direction. Per § 3.9, three scalars called *coordinates* suffice together to specify the amplitude and direction and thus the vector, the three being (x, y, x) in the rectangular coordinate system, $(\rho; \phi, z)$ in the cylindrical coordinate system, or $(r; \theta; \phi)$ in the spherical spherical coordinate system—as Fig. 15.1 illustrates and Table 3.4 on page 90 interrelates—among other, more exotic possibilities (§ 15.7).

The vector brings an elegant notation. This chapter and chapter 16

[1] [36, chapter 2]

Figure 15.1: A point on a sphere, in spherical $(r; \theta; \phi)$ and cylindrical $(\rho; \phi, z)$ coordinates. (The axis labels bear circumflexes in this figure only to disambiguate the $\hat{z}$ axis from the cylindrical coordinate z. See also Fig. 15.5.)



detail it. Without the notation, one would write an expression like

$$\frac{(z-z') - [\partial z'/\partial x]_{x=x',y=y'}\,(x-x') - [\partial z'/\partial y]_{x=x',y=y'}\,(y-y')}{\sqrt{[1+(\partial z'/\partial x)^2+(\partial z'/\partial y)^2]_{x=x',y=y'}\,[(x-x')^2+(y-y')^2+(z-z')^2]}}$$

for the aspect coefficient relative to a local surface normal (and if the sentence's words do not make sense to you yet, don't worry; just look the symbols over and appreciate the expression's bulk). The same coefficient in standard vector notation is

$$\hat{\mathbf{n}} \cdot \Delta\hat{\mathbf{r}}.$$

Besides being more evocative (once one has learned to read it) and much more compact, the standard vector notation brings the major advantage of freeing a model's geometry from reliance on any particular coordinate system. Reorienting axes (§ 15.1) for example knots the former expression like spaghetti but does not disturb the latter expression at all.

Two-dimensional geometrical vectors arise in practical modeling about as often as three-dimensional geometrical vectors do. Fortunately, the two-dimensional case needs little special treatment, for it is just the three-dimensional with $z = 0$ or $\theta = 2\pi/4$ (see however § 15.6).

Here at the outset, a word on complex numbers seems in order. Unlike most of the rest of the book this chapter and the next will work chiefly in real numbers, or at least in real coordinates. Notwithstanding, complex coordinates are possible. Indeed, in the rectangular coordinate system complex coordinates are perfectly appropriate and are straightforward enough to handle. The cylindrical and spherical systems however, which these chapters also treat, were not conceived with complex coordinates in mind; and, although it might with some theoretical subtlety be possible to treat complex radii, azimuths and elevations consistently as three-dimensional coordinates, these chapters will not try to do so.[2] (This is not to say that you cannot have a complex vector like, say, $\hat{\boldsymbol{\rho}}[3+i2] - \hat{\boldsymbol{\phi}}[1/4]$ in a nonrectangular basis. You can have such a vector, it is fine, and these chapters will not avoid it. What these chapters will avoid are complex nonrectangular *coordinates* like $[3+i2; -1/4, 0]$.)

Vector addition will already be familiar to the reader from chapter 3 or (quite likely) from earlier work outside this book. This chapter therefore begins with the reorientation of axes in § 15.1 and vector multiplication in § 15.2.

[2] The author would be interested to learn if there existed an uncontrived scientific or engineering application that actually used complex, nonrectangular coordinates.

15.1 Reorientation

Matrix notation expresses the rotation of axes (3.5) as

$$\begin{bmatrix} \hat{\mathbf{x}}' \\ \hat{\mathbf{y}}' \\ \hat{\mathbf{z}}' \end{bmatrix} = \begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \\ \hat{\mathbf{z}} \end{bmatrix}.$$

In three dimensions however one can do more than just to rotate the x and y axes about the z. One can reorient the three axes generally as follows.

15.1.1 The Tait-Bryan rotations

With a *yaw* and a *pitch* to point the x axis in the desired direction plus a *roll* to position the y and z axes as desired about the new x axis,[3] one can reorient the three axes generally:

$$\begin{bmatrix} \hat{\mathbf{x}}' \\ \hat{\mathbf{y}}' \\ \hat{\mathbf{z}}' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & \sin\psi \\ 0 & -\sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \\ \hat{\mathbf{z}} \end{bmatrix}; \tag{15.1}$$

or, inverting per (3.6),

$$\begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \\ \hat{\mathbf{z}} \end{bmatrix} = \begin{bmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & -\sin\psi \\ 0 & \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}' \\ \hat{\mathbf{y}}' \\ \hat{\mathbf{z}}' \end{bmatrix}. \tag{15.2}$$

These are called the *Tait-Bryan rotations,* or alternately the *Cardan rotations.*[4,5]

[3] The English maritime verbs *to yaw, to pitch* and *to roll* describe the rotational motion of a vessel at sea. For a vessel to yaw is for her to rotate about her vertical axis, so her bow (her forwardmost part) yaws from side to side. For a vessel to pitch is for her to rotate about her "beam axis," so her bow pitches up and down. For a vessel to roll is for her to rotate about her "fore-aft axis" such that she rocks or lists (leans) without changing the direction she points [182, "Glossary of nautical terms," 23:00, 20 May 2008]. In the Tait-Bryan rotations as explained in this book, to yaw is to rotate about the z axis, to pitch about the y, and to roll about the x [90]. In the Euler rotations as explained in this book later in the present section, however, the axes are assigned to the vessel differently such that to yaw is to rotate about the x axis, to pitch about the y, and to roll about the z. This implies that the Tait-Bryan vessel points x-ward whereas the Euler vessel points z-ward. The reason to shift perspective so is to maintain the semantics of the symbols θ and ϕ (though not ψ) according to Fig. 15.1.

If this footnote seems confusing, then read (15.1) and (15.7) which are correct.

[4] The literature seems to agree on no standard order among the three Tait-Bryan rotations; and, though the rotational angles are usually named ϕ, θ and ψ, which angle gets which name admittedly depends on the author. If unsure, prefer the names given here.

[5] [35]

Notice in (15.1) and (15.2) that the transpose (though curiously not the adjoint) of each 3×3 Tait-Bryan factor is also its inverse.

In concept, the Tait-Bryan equations (15.1) and (15.2) say nearly all one needs to say about reorienting axes in three dimensions; but, still, the equations can confuse the uninitiated. Consider a vector

$$\mathbf{v} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y + \hat{\mathbf{z}}z. \tag{15.3}$$

It is not the vector one reorients but rather the axes used to describe the vector. Envisioning the axes as in Fig. 15.1 with the z axis upward, one first yaws the x axis through an angle ϕ toward the y then pitches it downward through an angle θ away from the z. Finally, one rolls the y and z axes through an angle ψ about the new x, all the while maintaining the three axes rigidly at right angles to one another. These three Tait-Bryan rotations can orient axes any way. Yet, even once one has clearly visualized the Tait-Bryan sequence, the prospect of applying (15.2) (which inversely represents the sequence) to (15.3) can still seem daunting until one rewrites the latter equation in the form

$$\mathbf{v} = \left[\begin{array}{ccc} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \end{array}\right] \left[\begin{array}{c} x \\ y \\ z \end{array}\right], \tag{15.4}$$

after which the application is straightforward. There results

$$\mathbf{v}' = \hat{\mathbf{x}}'x' + \hat{\mathbf{y}}'y' + \hat{\mathbf{z}}'z',$$

where

$$\left[\begin{array}{c} x' \\ y' \\ z' \end{array}\right] \equiv \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & \cos\psi & \sin\psi \\ 0 & -\sin\psi & \cos\psi \end{array}\right] \left[\begin{array}{ccc} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{array}\right] \left[\begin{array}{ccc} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{array}\right] \left[\begin{array}{c} x \\ y \\ z \end{array}\right], \tag{15.5}$$

and where Table 3.4 converts to cylindrical or spherical coordinates if and as desired. Since (15.5) resembles (15.1), it comes as no surprise that its inverse,

$$\left[\begin{array}{c} x \\ y \\ z \end{array}\right] = \left[\begin{array}{ccc} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{array}\right] \left[\begin{array}{ccc} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{array}\right] \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & \cos\psi & -\sin\psi \\ 0 & \sin\psi & \cos\psi \end{array}\right] \left[\begin{array}{c} x' \\ y' \\ z' \end{array}\right], \tag{15.6}$$

resembles (15.2).

15.1.2 The Euler rotations

A useful alternative to the Tait-Bryan rotations are the *Euler rotations,* which view the problem of reorientation from the perspective of the z axis rather than of the x. The Euler rotations consist of a roll and a pitch followed by another roll, without any explicit yaw:[6]

$$\begin{bmatrix} \hat{\mathbf{x}}' \\ \hat{\mathbf{y}}' \\ \hat{\mathbf{z}}' \end{bmatrix} = \begin{bmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \\ \hat{\mathbf{z}} \end{bmatrix}; \tag{15.7}$$

and inversely

$$\begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \\ \hat{\mathbf{z}} \end{bmatrix} = \begin{bmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}' \\ \hat{\mathbf{y}}' \\ \hat{\mathbf{z}}' \end{bmatrix}. \tag{15.8}$$

Whereas the Tait-Bryan point the x axis first, the Euler tactic is to point first the z.

So, that's it. One can reorient three axes arbitrarily by rotating them in pairs about the z, y and x or the z, y and z axes in sequence—or, generalizing, in pairs about any of the three axes so long as the axis of the middle rotation differs from the axes (Tait-Bryan) or axis (Euler) of the first and last. A firmer grasp of the reorientation of axes in three dimensions comes with practice, but those are the essentials of it.

15.2 Multiplication

One can multiply a vector in any of three ways. The first, scalar multiplication, is trivial: if a vector $\mathbf{v}$ is as defined by (15.3), then

$$\psi\mathbf{v} = \hat{\mathbf{x}}\psi x + \hat{\mathbf{y}}\psi y + \hat{\mathbf{z}}\psi z. \tag{15.9}$$

Such scalar multiplication evidently scales a vector's length without diverting its direction. The other two forms of vector multiplication involve multiplying a vector by another vector and are the subjects of the two subsections that follow.

[6] As for the Tait-Bryan, for the Euler also the literature agrees on no standard sequence. What one author calls a pitch, another might call a yaw, and some prefer to roll twice about the x axis rather than the z. What makes a reorientation an Euler rather than a Tait-Bryan is that the Euler rolls twice.

15.2.1 The dot product

We first met the dot product in § 13.8. It works similarly for the geometrical vectors of this chapter as for the matrix vectors of chapter 13:

$$\mathbf{v}_1 \cdot \mathbf{v}_2 = x_1x_2 + y_1y_2 + z_1z_2, \tag{15.10}$$

which, if the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are real, is the product of the two vectors to the extent to which they run in the same direction. It is the product to the extent to which the vectors run in the same direction because one can reorient axes to point $\hat{\mathbf{x}}'$ in the direction of $\mathbf{v}_1$, whereupon $\mathbf{v}_1 \cdot \mathbf{v}_2 = x_1'x_2'$ since y_1' and z_1' have vanished.

Naturally, to be valid, the dot product must not vary under a reorientation of axes; and indeed if we write (15.10) in matrix notation,

$$\mathbf{v}_1 \cdot \mathbf{v}_2 = \begin{bmatrix} x_1 & y_1 & z_1 \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix}, \tag{15.11}$$

and then expand each of the two factors on the right according to (15.6), we see that the dot product does not in fact vary. As in (13.44) of § 13.8, here too the relationship

$$\begin{aligned} \mathbf{v}_1^* \cdot \mathbf{v}_2 &= v_1^* v_2 \cos\theta, \\ \hat{\mathbf{v}}_1^* \cdot \hat{\mathbf{v}}_2 &= \cos\theta, \end{aligned} \tag{15.12}$$

gives the angle θ between two vectors according Fig. 3.1's cosine if the vectors are real, by definition hereby if complex. Consequently, the two vectors are mutually orthogonal—that is, the vectors run at right angles $\theta = 2\pi/4$ to one another—if and only if

$$\mathbf{v}_1^* \cdot \mathbf{v}_2 = 0.$$

That the dot product is commutative,

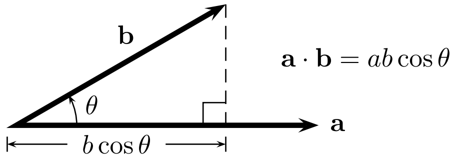
$$\mathbf{v}_2 \cdot \mathbf{v}_1 = \mathbf{v}_1 \cdot \mathbf{v}_2, \tag{15.13}$$

is obvious from (15.10). Fig. 15.2 illustrates the dot product.

15.2.2 The cross product

The dot product of two vectors according to § 15.2.1 is a scalar. One can also multiply two vectors to obtain a vector, however, and it is often useful to do so. As the dot product is the product of two vectors to the extent to

Figure 15.2: The dot product.



which they run in the same direction, the *cross product* is the product of two vectors to the extent to which they run in different directions. Unlike the dot product the cross product is a vector, defined in rectangular coordinates as

$$\begin{aligned}\mathbf{v}_1 \times \mathbf{v}_2 &= \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{vmatrix} \\ &\equiv \hat{\mathbf{x}}(y_1z_2 - z_1y_2) + \hat{\mathbf{y}}(z_1x_2 - x_1z_2) + \hat{\mathbf{z}}(x_1y_2 - y_1x_2),\end{aligned} \tag{15.14}$$

where the $|\cdot|$ notation is a mnemonic (actually a pleasant old determinant notation § 14.1 could have but did not happen to use) whose semantics are as shown.

As the dot product, the cross product too is invariant under reorientation. One could demonstrate this fact by multiplying out (15.2) and (15.6) then substituting the results into (15.14): a lengthy, unpleasant exercise. Fortunately, it is also an unnecessary exercise, forasmuch as an arbitrary reorientation consists of three rotations (§ 15.1) in sequence it suffices merely that rotation about one axis not alter the cross product. One proves the proposition in the latter form by setting any two of ϕ, θ and ψ to zero before multiplying out and substituting. For instance, setting θ and ψ to zero,

$$\begin{aligned}
\mathbf{v}_1 \times \mathbf{v}_2 &= \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{vmatrix} \\
&= \begin{vmatrix} \hat{\mathbf{x}}'\cos\phi - \hat{\mathbf{y}}'\sin\phi & \hat{\mathbf{x}}'\sin\phi + \hat{\mathbf{y}}'\cos\phi & \hat{\mathbf{z}} \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{vmatrix} \\
&= \hat{\mathbf{x}}'\left[y_1 z_2\cos\phi - z_1 y_2\cos\phi + z_1 x_2\sin\phi - x_1 z_2\sin\phi\right] \\
&\qquad + \hat{\mathbf{y}}'\left[-y_1 z_2\sin\phi + z_1 y_2\sin\phi + z_1 x_2\cos\phi - x_1 z_2\cos\phi\right] \\
&\qquad + \hat{\mathbf{z}}\left[x_1 y_2 - y_1 x_2\right] \\
&= \hat{\mathbf{x}}'\left[(-x_1\sin\phi + y_1\cos\phi)z_2 - z_1(-x_2\sin\phi + y_2\cos\phi)\right] \\
&\qquad + \hat{\mathbf{y}}'\left[z_1(y_2\sin\phi + x_2\cos\phi) - (y_1\sin\phi + x_1\cos\phi)z_2\right] \\
&\qquad + \hat{\mathbf{z}}\left[x_1 y_2 - y_1 x_2\right] \\
&= \hat{\mathbf{x}}'\left[y_1' z_2 - z_1 y_2'\right] + \hat{\mathbf{y}}'\left[z_1 x_2' - z_2 x_1'\right] \\
&\qquad + \hat{\mathbf{z}}[(x_1'\cos\phi - y_1'\sin\phi)(x_2'\sin\phi + y_2'\cos\phi) \\
&\qquad\qquad - (x_1'\sin\phi + y_1'\cos\phi)(x_2'\cos\phi - y_2'\sin\phi)].
\end{aligned}$$

Since according to Pythagoras in Table 3.1 $\cos^2\phi + \sin^2\phi = 1$,

$$\begin{aligned}
\mathbf{v}_1 \times \mathbf{v}_2 &= \hat{\mathbf{x}}'\left[y_1' z_2 - z_1 y_2'\right] + \hat{\mathbf{y}}'\left[z_1 x_2' - z_2 x_1'\right] + \hat{\mathbf{z}}[x_1' y_2' - y_1' x_2'] \\
&= \begin{vmatrix} \hat{\mathbf{x}}' & \hat{\mathbf{y}}' & \hat{\mathbf{z}} \\ x_1' & y_1' & z_1 \\ x_2' & y_2' & z_2 \end{vmatrix}
\end{aligned}$$

as was to be demonstrated.

Several facets of the cross product draw attention to themselves.

- The cyclic progression

$$\cdots \to x \to y \to z \to x \to y \to z \to x \to y \to \cdots \tag{15.15}$$

of (15.14) arises again and again in vector analysis. Where the progression is honored, as in $\hat{\mathbf{z}}x_1y_2$, the associated term bears a + sign, otherwise a − sign, due to § 11.6's parity principle and the right-hand rule.

- The cross product is not commutative. In fact,

$$\mathbf{v}_2 \times \mathbf{v}_1 = -\mathbf{v}_1 \times \mathbf{v}_2, \tag{15.16}$$

which is a direct consequence of the previous point regarding parity, or which can be seen more prosaically in (15.14) by swapping the places of $\mathbf{v}_1$ and $\mathbf{v}_2$.

- The cross product is not associative. That is,

  $$(\mathbf{v}_1 \times \mathbf{v}_2) \times \mathbf{v}_3 \neq \mathbf{v}_1 \times (\mathbf{v}_2 \times \mathbf{v}_3),$$

  as is proved by a suitable counterexample like $\mathbf{v}_1 = \mathbf{v}_2 = \hat{\mathbf{x}}$, $\mathbf{v}_3 = \hat{\mathbf{y}}$.

- The cross product runs perpendicularly to each of its two factors if the vectors involved are real. That is,

  $$\mathbf{v}_1 \cdot (\mathbf{v}_1 \times \mathbf{v}_2) = 0 = \mathbf{v}_2 \cdot (\mathbf{v}_1 \times \mathbf{v}_2), \tag{15.17}$$

  as is seen by substituting (15.14) into (15.10) with an appropriate change of variables and simplifying.

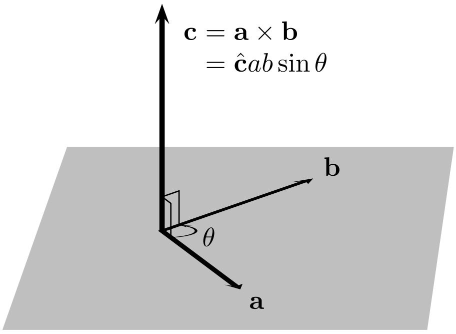
- Unlike the dot product, the cross product is closely tied to three-dimensional space. Two-dimensional space (a plane) can have a cross product so long as one does not mind that the product points off into the third dimension, but to speak of a cross product in four-dimensional space would require arcane definitions and would otherwise make little sense. Fortunately, the physical world is three-dimensional (or, at least, the space in which we model all but a few, exotic physical phenomena is three-dimensional), so the cross product's limitation as defined here to three dimensions will seldom disturb us.

- Section 15.2.1 has related the cosine of the angle between vectors to the dot product. One can similarly relate the angle's sine to the cross product if the vectors involved are real, as

  $$\begin{aligned} |\mathbf{v}_1 \times \mathbf{v}_2| &= v_1 v_2 \sin\theta, \\ |\hat{\mathbf{v}}_1 \times \hat{\mathbf{v}}_2| &= \sin\theta, \end{aligned} \tag{15.18}$$

  demonstrated by reorienting axes such that $\hat{\mathbf{v}}_1 = \hat{\mathbf{x}}'$, that $\hat{\mathbf{v}}_2$ has no component in the $\hat{\mathbf{z}}'$ direction, and that $\hat{\mathbf{v}}_2$ has only a nonnegative component in the $\hat{\mathbf{y}}'$ direction; by remembering that reorientation cannot alter a cross product; and finally by applying (15.14) and comparing the result against Fig. 3.1's sine. (If the vectors involved are complex then nothing prevents the operation $|\mathbf{v}_1^* \times \mathbf{v}_2|$ by analogy with

Figure 15.3: The cross product.


$\mathbf{c} = \mathbf{a} \times \mathbf{b}$
$= \hat{\mathbf{c}} ab \sin\theta$
$\mathbf{b}$
θ
$\mathbf{a}$

> eqn. 15.12—in fact the operation $\mathbf{v}_1^* \times \mathbf{v}_2$ without the magnitude sign is used routinely to calculate electromagnetic power flow[7]—but each of the cross product's three rectangular components has its own complex phase which the magnitude operation flattens, so the result's relationship to the sine of an angle is not immediately clear.)

Fig. 15.3 illustrates the cross product.

15.3 Orthogonal bases

A vector exists independently of the components by which one expresses it, for, whether $\mathbf{q} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y + \hat{\mathbf{z}}z$ or $\mathbf{q} = \hat{\mathbf{x}}'x' + \hat{\mathbf{y}}'y' + \hat{\mathbf{z}}'z'$, it remains the same vector $\mathbf{q}$. However, where a model involves a circle, a cylinder or a sphere, where a model involves a contour or a curved surface of some kind, to choose $\hat{\mathbf{x}}'$, $\hat{\mathbf{y}}'$ and $\hat{\mathbf{z}}'$ wisely can immensely simplify the model's analysis. Normally one requires that $\hat{\mathbf{x}}'$, $\hat{\mathbf{y}}'$ and $\hat{\mathbf{z}}'$ each retain unit length, run perpendiclarly to one another, and obey the right-hand rule (§ 3.3), but otherwise any $\hat{\mathbf{x}}'$, $\hat{\mathbf{y}}'$ and $\hat{\mathbf{z}}'$ can serve. Moreover, various parts of a model can specify various $\hat{\mathbf{x}}'$, $\hat{\mathbf{y}}'$ and $\hat{\mathbf{z}}'$, or various substitutes therefor, under various conditons.

Recalling the constants and variables of § 2.7, such a concept is flexible enough to confuse the uninitiated severely and soon. As in § 2.7, here too an

[7] [72, eqn. 1-51]

example affords perspective. Imagine driving your automobile down a winding road, where q represented your speed[8] and $\hat{\boldsymbol{\ell}}$ represented the direction the road ran, not generally, but just at the spot along the road at which your automobile momentarily happened to be. That your velocity were $\hat{\boldsymbol{\ell}}q$ meant that you kept skilfully to your lane; on the other hand, that your velocity were $(\hat{\boldsymbol{\ell}}\cos\psi + \hat{\mathbf{v}}\sin\psi)q$—where $\hat{\mathbf{v}}$, at right angles to $\hat{\boldsymbol{\ell}}$, represented the direction right-to-left across the road—would have you drifting out of your lane at an angle ψ. A headwind had velocity $-\hat{\boldsymbol{\ell}}q_{\text{wind}}$; a crosswind, $\pm\hat{\mathbf{v}}q_{\text{wind}}$. A car a mile ahead of you had velocity $\hat{\boldsymbol{\ell}}_2 q_2 = (\hat{\boldsymbol{\ell}}\cos\beta + \hat{\mathbf{v}}\sin\beta)q_2$, where β represented the difference (assuming that the other driver kept skilfully to his own lane) between the road's direction a mile ahead and its direction at your spot. For all these purposes the unit vector $\hat{\boldsymbol{\ell}}$ would remain constant. However, fifteen seconds later, after you had rounded a bend in the road, the symbols $\hat{\boldsymbol{\ell}}$ and $\hat{\mathbf{v}}$ would by definition represent different vectors than before, with respect to which one would express your new velocity as $\hat{\boldsymbol{\ell}}q$ but would no longer express the headwind's velocity as $-\hat{\boldsymbol{\ell}}q_{\text{wind}}$ because, since the road had turned while the wind had not, the wind would no longer be a headwind. And this is where confusion can arise: your own velocity had changed while the expression representing it had not; whereas the wind's velocity had not changed while the expression representing *it* had. This is not because $\hat{\boldsymbol{\ell}}$ differs from place to place at a given moment, for like any other vector the vector $\hat{\boldsymbol{\ell}}$ (as defined in this particular example) is the same vector everywhere. Rather, it is because $\hat{\boldsymbol{\ell}}$ is defined relative to the road at your automobile's location, which location changes as you drive.

If a third unit vector $\hat{\mathbf{w}}$ were defined, perpendicular both to $\hat{\boldsymbol{\ell}}$ and to $\hat{\mathbf{v}}$ such that $[\hat{\boldsymbol{\ell}}\ \hat{\mathbf{v}}\ \hat{\mathbf{w}}]$ obeyed the right-hand rule, then the three together would constitute an *orthogonal basis.* Any three real,[9] right-handedly mutually perpendicular unit vectors $[\hat{\mathbf{x}}'\ \hat{\mathbf{y}}'\ \hat{\mathbf{z}}']$ in three dimensions, whether constant or variable, for which

$$\begin{array}{lll}\hat{\mathbf{y}}'\cdot\hat{\mathbf{z}}' = 0, & \hat{\mathbf{y}}'\times\hat{\mathbf{z}}' = \hat{\mathbf{x}}', & \Im(\hat{\mathbf{x}}') = 0,\\ \hat{\mathbf{z}}'\cdot\hat{\mathbf{x}}' = 0, & \hat{\mathbf{z}}'\times\hat{\mathbf{x}}' = \hat{\mathbf{y}}', & \Im(\hat{\mathbf{y}}') = 0,\\ \hat{\mathbf{x}}'\cdot\hat{\mathbf{y}}' = 0, & \hat{\mathbf{x}}'\times\hat{\mathbf{y}}' = \hat{\mathbf{z}}', & \Im(\hat{\mathbf{z}}') = 0,\end{array} \tag{15.19}$$

[8] Conventionally one would prefer the letter v to represent speed, with velocity as $\mathbf{v}$ which in the present example would happen to be $\mathbf{v} = \hat{\boldsymbol{\ell}}v$. However, this section will require the letter v for an unrelated purpose.

[9] A complex orthogonal basis is also theoretically possible but is normally unnecessary in geometrical applications and involves subtleties in the cross product. This chapter, which specifically concerns three-dimensional geometrical vectors rather than the general, n-dimensional vectors of chapter 11, is content to consider real bases only. Note that one can express a complex vector in a real basis.

constitutes such an orthogonal basis, from which other vectors can be built. The geometries of some models suggest no particular basis, when one usually just uses a constant $[\hat{\mathbf{x}}\ \hat{\mathbf{y}}\ \hat{\mathbf{z}}]$. The geometries of other models however do suggest a particular basis, often a variable one.

- Where the model features a contour like the example's winding road, an $[\hat{\boldsymbol{\ell}}\ \hat{\mathbf{v}}\ \hat{\mathbf{w}}]$ basis (or a $[\hat{\mathbf{u}}\ \hat{\mathbf{v}}\ \hat{\boldsymbol{\ell}}]$ basis or even a $[\hat{\mathbf{u}}\ \hat{\boldsymbol{\ell}}\ \hat{\mathbf{w}}]$ basis) can be used, where $\hat{\boldsymbol{\ell}}$ locally follows the contour. The variable unit vectors $\hat{\mathbf{v}}$ and $\hat{\mathbf{w}}$ (or $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$, etc.) can be defined in any convenient way so long as they remain perpendicular to one another and to $\hat{\boldsymbol{\ell}}$—such that $(\hat{\mathbf{z}} \times \hat{\boldsymbol{\ell}}) \cdot \hat{\mathbf{w}} = 0$ for instance (that is, such that $\hat{\mathbf{w}}$ lay in the plane of $\hat{\mathbf{z}}$ and $\hat{\boldsymbol{\ell}}$)—but if the geometry suggests a particular $\hat{\mathbf{v}}$ or $\hat{\mathbf{w}}$ (or $\hat{\mathbf{u}}$), like the direction right-to-left across the example's road, then that $\hat{\mathbf{v}}$ or $\hat{\mathbf{w}}$ should probably be used. The letter ℓ here stands for "longitudinal."[10]

- Where the model features a curved surface like the surface of a wavy sea,[11] a $[\hat{\mathbf{u}}\ \hat{\mathbf{v}}\ \hat{\mathbf{n}}]$ basis (or a $[\hat{\mathbf{u}}\ \hat{\mathbf{n}}\ \hat{\mathbf{w}}]$ basis, etc.) can be used, where $\hat{\mathbf{n}}$ points locally perpendicularly to the surface. The letter n here stands for "normal," a synonym for "perpendicular." Observe, incidentally but significantly, that such a unit normal $\hat{\mathbf{n}}$ tells one everything one needs to know about its surface's local orientation.

- Combining the last two, where the model features a contour along a curved surface, an $[\hat{\boldsymbol{\ell}}\ \hat{\mathbf{v}}\ \hat{\mathbf{n}}]$ basis can be used. One need not worry about choosing a direction for $\hat{\mathbf{v}}$ in this case since necessarily $\hat{\mathbf{v}} = \hat{\mathbf{n}} \times \hat{\boldsymbol{\ell}}$.

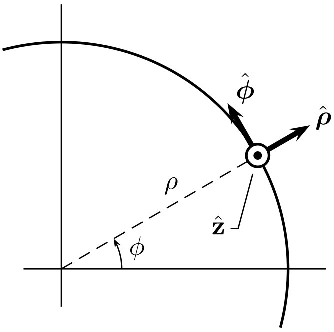
- Where the model features a circle or cylinder, a $[\hat{\boldsymbol{\rho}}\ \hat{\boldsymbol{\phi}}\ \hat{\mathbf{z}}]$ basis can be used, where $\hat{\mathbf{z}}$ is constant and runs along the cylinder's axis (or perpendicularly through the circle's center), $\hat{\boldsymbol{\rho}}$ is variable and points locally away from the axis, and $\hat{\boldsymbol{\phi}}$ is variable and runs locally along the circle's perimeter in the direction of increasing azimuth ϕ. Refer to § 3.9 and Fig. 15.4.

- Where the model features a sphere, an $[\hat{\mathbf{r}}\ \hat{\boldsymbol{\theta}}\ \hat{\boldsymbol{\phi}}]$ basis can be used, where $\hat{\mathbf{r}}$ is variable and points locally away from the sphere's center, $\hat{\boldsymbol{\theta}}$ is variable and runs locally tangentially to the sphere's surface in the direction of increasing elevation θ (that is, though not usually in the $-\hat{\mathbf{z}}$ direction itself, as nearly as possible to the $-\hat{\mathbf{z}}$ direction without departing from the sphere's surface), and $\hat{\boldsymbol{\phi}}$ is variable and

[10] The assertion wants a citation, which the author lacks.

[11] [130]

Figure 15.4: The cylindrical basis. (The conventional symbols ⊙ and ⊗ respectively represent vectors pointing out of the page toward the reader and into the page away from the reader. Thus, this figure shows the constant basis vector $\hat{\mathbf{z}}$ pointing out of the page toward the reader. The dot in the middle of the ⊙ is supposed to look like the tip of an arrowhead.)



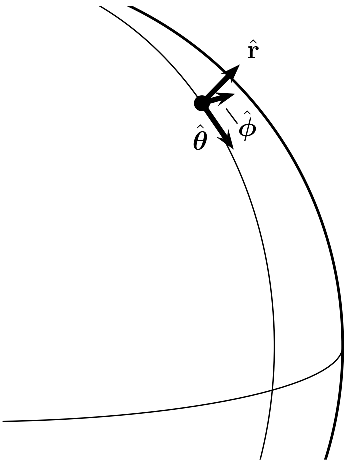
runs locally tangentially to the sphere's surface in the direction of increasing azimuth ϕ (that is, along the sphere's surface perpendicularly to $\hat{\mathbf{z}}$). Standing on the earth's surface, with the earth as the sphere, $\hat{\mathbf{r}}$ would be up, $\hat{\boldsymbol{\theta}}$ south, and $\hat{\boldsymbol{\phi}}$ east. Refer to § 3.9 and Fig. 15.5.

- Occasionally a model arises with two circles that share a center but whose axes stand perpendicular to one another. In such a model one conventionally establishes $\hat{\mathbf{z}}$ as the direction of the principal circle's axis but then is left with $\hat{\mathbf{x}}$ or $\hat{\mathbf{y}}$ as the direction of the secondary circle's axis, upon which an $[\hat{\mathbf{x}}\ \hat{\boldsymbol{\rho}}^x\ \hat{\boldsymbol{\phi}}^x]$, $[\hat{\boldsymbol{\phi}}^x\ \hat{\mathbf{r}}\ \hat{\boldsymbol{\theta}}^x]$, $[\hat{\boldsymbol{\phi}}^y\ \hat{\mathbf{y}}\ \hat{\boldsymbol{\rho}}^y]$ or $[\hat{\boldsymbol{\theta}}^y\ \hat{\boldsymbol{\phi}}^y\ \hat{\mathbf{r}}]$ basis can be used locally as appropriate. Refer to § 3.9.

Many other orthogonal bases are possible (as in § 15.7, for instance) but the foregoing are the most common. Whether listed here or not, each orthogonal basis orders its three unit vectors by the right-hand rule (15.19).

Quiz: what does the vector expression $\hat{\boldsymbol{\rho}}3 - \hat{\boldsymbol{\phi}}(1/4) + \hat{\mathbf{z}}2$ mean? Wrong answer: it meant the cylindrical coordinates $(3; -1/4, 2)$; or, it meant the position vector $\hat{\mathbf{x}}3\cos(-1/4) + \hat{\mathbf{y}}3\sin(-1/4) + \hat{\mathbf{z}}2$ associated with those coordinates. Right answer: the expression means nothing certain in itself but acquires a definite meaning only when an azimuthal coordinate ϕ is also supplied, after which the expression indicates the ordinary rectangular

Figure 15.5: The spherical basis (see also Fig. 15.1).



vector $\hat{\mathbf{x}}'3 - \hat{\mathbf{y}}'(1/4) + \hat{\mathbf{z}}'2$, where $\hat{\mathbf{x}}' = \hat{\boldsymbol{\rho}} = \hat{\mathbf{x}}\cos\phi + \hat{\mathbf{y}}\sin\phi$, $\hat{\mathbf{y}}' = \hat{\boldsymbol{\phi}} = -\hat{\mathbf{x}}\sin\phi + \hat{\mathbf{y}}\cos\phi$, and $\hat{\mathbf{z}}' = \hat{\mathbf{z}}$. But, if this is so—if the cylindrical basis $[\hat{\boldsymbol{\rho}}\ \hat{\boldsymbol{\phi}}\ \hat{\mathbf{z}}]$ is used solely to express *rectangular* vectors—then why should we name this basis "cylindrical"? Answer: only because cylindrical coordinates (supplied somewhere) determine the actual directions of its basis vectors. Once directions are determined, such a basis is used rectangularly like any other orthogonal basis.

This can seem confusing until one has grasped what the so-called nonrectangular bases are for. Consider the problem of air flow in a jet engine. It may suit such a problem that instantaneous local air velocity within the engine cylinder be expressed in cylindrical coordinates, with the z axis oriented along the engine's axle; but this does not mean that the air flow within the engine cylinder were everywhere $\hat{\mathbf{z}}$-directed. On the contrary, a local air velocity of $\mathbf{q} = [-\hat{\boldsymbol{\rho}}5.0 + \hat{\boldsymbol{\phi}}30.0 - \hat{\mathbf{z}}250.0]$ m/s would have air moving through the point in question at 250.0 m/s aftward along the axle, 5.0 m/s inward toward the axle and 30.0 m/s circulating about the engine cylinder.

In this model, it is true that the basis vectors $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{\phi}}$ indicate different directions at different positions within the cylinder, but at a particular position the basis vectors are still used rectangularly to express $\mathbf{q}$, the instantaneous local air velocity at that position. It's just that the "rectangle" is rotated locally to line up with the axle.

Naturally, you cannot make full sense of an air-velocity vector $\mathbf{q}$ unless you have also the coordinates $(\rho; \phi, z)$ of the position within the engine cylinder at which the air has the velocity the vector specifies—yet this is when confusion can arise, for besides the air-velocity vector there is also, separately, a position vector $\mathbf{r} = \hat{\mathbf{x}}\rho \cos\phi + \hat{\mathbf{y}}\rho \sin\phi + \hat{\mathbf{z}}z$. One may denote the air-velocity vector as[12] $\mathbf{q}(\mathbf{r})$, a function of position; yet, though the position vector is as much a vector as the velocity vector is, one nonetheless handles it differently. One will not normally express the position vector $\mathbf{r}$ in the cylindrical basis.

It would make little sense to try to express the position vector $\mathbf{r}$ in the cylindrical basis because the position vector is the very thing that *determines* the cylindrical basis. In the cylindrical basis, after all, the position vector is necessarily $\mathbf{r} = \hat{\boldsymbol{\rho}}\rho + \hat{\mathbf{z}}z$ (and consider: in the spherical basis it is the even more cryptic $\mathbf{r} = \hat{\mathbf{r}}r$), and how useful is that, really? Well, maybe it is useful in some situations, but for the most part to express the position vector in the cylindrical basis would be as to say, "My house is zero miles away from home." Or, "The time is presently now." Such statements may be tautologically true, perhaps, but they are confusing because they only seem to give information. The position vector $\mathbf{r}$ determines the basis, after which one expresses things other than position, like instantaneous local air velocity $\mathbf{q}$, in that basis. In fact, the only basis normally suitable to express a position vector is a fixed rectangular basis like $[\hat{\mathbf{x}}\ \hat{\mathbf{y}}\ \hat{\mathbf{z}}]$. Otherwise, one uses cylindrical *coordinates* $(\rho; \phi, z)$, but not a cylindrical *basis* $[\hat{\boldsymbol{\rho}}\ \hat{\boldsymbol{\phi}}\ \hat{\mathbf{z}}]$, to express a position $\mathbf{r}$ in a cylindrical geometry.

Maybe the nonrectangular bases were more precisely called "rectangular bases of the nonrectangular coordinate systems," but those are too many words and, anyway, that is not how the usage has evolved. Chapter 16 will elaborate the story by considering spatial derivatives of quantities like air velocity, when one must take the variation in $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{\phi}}$ from point to point into account, but the foregoing is the basic story nevertheless.

15.4 Notation

The vector notation of §§ 15.1 and 15.2 is correct, familiar and often expedient but sometimes inconveniently prolix. This admittedly difficult section

[12] Conventionally, one is much more likely to denote a velocity vector as $\mathbf{u}(\mathbf{r})$ or $\mathbf{v}(\mathbf{r})$, except that the present chapter is (as footnote 8 has observed) already using the letters u and v for an unrelated purpose. To denote position as $\mathbf{r}$ however is entirely standard.

augments the notation to render it much more concise.

15.4.1 Components by subscript

The notation

$$\begin{aligned} a_x &\equiv \hat{\mathbf{x}} \cdot \mathbf{a}, & a_\rho &\equiv \hat{\boldsymbol{\rho}} \cdot \mathbf{a}, \\ a_y &\equiv \hat{\mathbf{y}} \cdot \mathbf{a}, & a_r &\equiv \hat{\mathbf{r}} \cdot \mathbf{a}, \\ a_z &\equiv \hat{\mathbf{z}} \cdot \mathbf{a}, & a_\theta &\equiv \hat{\boldsymbol{\theta}} \cdot \mathbf{a}, \\ a_n &\equiv \hat{\mathbf{n}} \cdot \mathbf{a}, & a_\phi &\equiv \hat{\boldsymbol{\phi}} \cdot \mathbf{a}, \end{aligned}$$

and so forth abbreviates the indicated dot product. That is to say, the notation represents the component of a vector $\mathbf{a}$ in the indicated direction. Generically,

$$a_\alpha \equiv \hat{\boldsymbol{\alpha}} \cdot \mathbf{a}. \tag{15.20}$$

Applied mathematicians use subscripts for several unrelated or vaguely related purposes, so the full dot-product notation $\hat{\boldsymbol{\alpha}} \cdot \mathbf{a}$ is often clearer in print than the abbreviation a_α is, but the abbreviation especially helps when several such dot products occur together in the same expression.

Since[13]

$$\begin{aligned} \hat{\mathbf{a}} &= \hat{\mathbf{x}} a_x + \hat{\mathbf{y}} a_y + \hat{\mathbf{z}} a_z, \\ \hat{\mathbf{b}} &= \hat{\mathbf{x}} b_x + \hat{\mathbf{y}} b_y + \hat{\mathbf{z}} b_z, \end{aligned}$$

the abbreviation lends a more amenable notation to the dot and cross products of (15.10) and (15.14):

$$\mathbf{a} \cdot \mathbf{b} = a_x b_x + a_y b_y + a_z b_z; \tag{15.21}$$

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ a_x & a_y & a_z \\ b_x & b_y & b_z \end{vmatrix}. \tag{15.22}$$

In fact—because, as we have seen, reorientation of axes cannot alter the dot and cross products—any orthogonal basis $[\hat{\mathbf{x}}' \; \hat{\mathbf{y}}' \; \hat{\mathbf{z}}']$ (§ 15.3) can serve here, so one can write more generally that

$$\mathbf{a} \cdot \mathbf{b} = a_{x'} b_{x'} + a_{y'} b_{y'} + a_{z'} b_{z'}; \tag{15.23}$$

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \hat{\mathbf{x}}' & \hat{\mathbf{y}}' & \hat{\mathbf{z}}' \\ a_{x'} & a_{y'} & a_{z'} \\ b_{x'} & b_{y'} & b_{z'} \end{vmatrix}. \tag{15.24}$$

[13] "Wait!" comes the objection. "I thought that you said that a_x meant $\hat{\mathbf{x}} \cdot \mathbf{a}$. Now you claim that it means the x component of $\mathbf{a}$?"

But there is no difference between $\hat{\mathbf{x}} \cdot \mathbf{a}$ and the x component of $\mathbf{a}$. The two are one and the same.

Because all those prime marks burden the notation and for professional mathematical reasons, the general forms (15.23) and (15.24) are sometimes rendered

$$\begin{aligned}\mathbf{a}\cdot\mathbf{b} &= a_1b_1 + a_2b_2 + a_3b_3,\\ \mathbf{a}\times\mathbf{b} &= \begin{vmatrix}\hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3\\ a_1 & a_2 & a_3\\ b_1 & b_2 & b_3\end{vmatrix},\end{aligned}$$

but you have to be careful about that in applied usage because people are not always sure whether a symbol like a_3 means "the third component of the vector $\mathbf{a}$" (as it does here) or "the third vector's component in the $\hat{\mathbf{a}}$ direction" (as it would in eqn. 15.10). Typically, applied mathematicians will write in the manner of (15.21) and (15.22) with the implied understanding that they really mean (15.23) and (15.24) but prefer not to burden the notation with extra little strokes—that is, with the implied understanding that x, y and z could just as well be ρ, ϕ and z or the coordinates of any other orthogonal, right-handed, three-dimensional basis.

Some pretty powerful confusion can afflict the student regarding the roles of the cylindrical symbols ρ, ϕ and z; or, worse, of the spherical symbols r, θ and ϕ. Such confusion reflects a pardonable but remediable lack of understanding of the relationship between coordinates like ρ, ϕ and z and their corresponding unit vectors $\hat{\boldsymbol{\rho}}$, $\hat{\boldsymbol{\phi}}$ and $\hat{\mathbf{z}}$. Section 15.3 has already written of the matter; but, further to dispel the confusion, one can now ask the student what the cylindrical coordinates of the vectors $\hat{\boldsymbol{\rho}}$, $\hat{\boldsymbol{\phi}}$ and $\hat{\mathbf{z}}$ are. The correct answer: $(1;\phi,0)$, $(1;\phi+2\pi/4,0)$ and $(0;0,1)$, respectively. Then, to reinforce, one can ask the student which cylindrical coordinates the variable vectors $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{\phi}}$ are functions of. The correct answer: both are functions of the coordinate ϕ only ($\hat{\mathbf{z}}$, a constant vector, is not a function of anything). What the student needs to understand is that, among the cylindrical coordinates, ϕ is a different kind of thing than z and ρ are:

- z and ρ are lengths whereas ϕ is an angle;
- but $\hat{\boldsymbol{\rho}}$, $\hat{\boldsymbol{\phi}}$ and $\hat{\mathbf{z}}$ are all the same kind of thing, unit vectors;
- and, separately, a_ρ, a_ϕ and a_z are all the same kind of thing, lengths.

Now to ask a harder question: in the cylindrical basis, what is the vector representation of $(\rho_1;\phi_1,z_1)$? The correct answer: $\hat{\boldsymbol{\rho}}\rho_1\cos(\phi_1-\phi) + \hat{\boldsymbol{\phi}}\rho_1\sin(\phi_1-\phi) + \hat{\mathbf{z}}z_1$. The student that gives this answer probably grasps the cylindrical symbols.

If the reader feels that the notation begins to confuse more than it describes, the writer empathizes but regrets to inform that the rest of the section, far from granting the reader a comfortable respite to absorb the elaborated notation as it stands, will not delay to elaborate the notation yet further! The confusion however is subjective. The trouble with vector work is that one has to learn to abbreviate or the expressions involved grow repetitive and unreadably long. For vectors, the abbreviated notation really is the proper notation. Eventually one accepts the need and takes the trouble to master the conventional vector abbreviation this section presents; and, indeed, the abbreviation is rather elegant once one has grown used to it. So, study closely and take heart![14] The notation is not actually as impenetrable as it at first will seem.

15.4.2 Einstein's summation convention

Einstein's summation convention is this: *that repeated indices are implicitly summed over.*[15] For instance, where the convention is in force, the equation[16]

$$\mathbf{a}\cdot\mathbf{b} = a_i b_i \tag{15.25}$$

means that

$$\mathbf{a}\cdot\mathbf{b} = \sum_i a_i b_i$$

or more fully that

$$\mathbf{a}\cdot\mathbf{b} = \sum_{i=x',y',z'} a_i b_i = a_{x'}b_{x'} + a_{y'}b_{y'} + a_{z'}b_{z'},$$

which is (15.23), except that Einstein's form (15.25) expresses it more succinctly. Likewise,

$$\mathbf{a}\times\mathbf{b} = \hat{\mathbf{\imath}}(a_{i+1}b_{i-1} - b_{i+1}a_{i-1}) \tag{15.26}$$

is (15.24)—although an experienced applied mathematician would probably apply the Levi-Civita epsilon of § 15.4.3, below, to further abbreviate this last equation to the form of (15.27) before presenting it.

[14] What to study? Besides this book, one can study any good, introductory undergraduate textbook in fluid dynamics, electromagnetics, quantum mechanics or the like. For example, [36] is not bad.

[15] [80]

[16] Some professional mathematicians now write a superscript a^i in certain cases in place of a subscript a_i, where the superscript bears some additional semantics [182, "Einstein notation," 05:36, 10 Feb. 2008]. Scientists and engineers however tend to prefer Einstein's original, subscript-only notation.

Einstein's summation convention is also called the *Einstein notation,* a term sometimes taken loosely to include also the Kronecker delta and Levi-Civita epsilon of § 15.4.3.

What is important to understand about Einstein's summation convention is that, in and of itself, it brings no new mathematics. It is rather a notational convenience.[17] It asks a reader to regard a repeated index like the i in "a_ib_i" as a dummy index (§ 2.3) and thus to read "a_ib_i" as "$\sum_i a_ib_i$." It does not magically create a summation where none existed; it just hides the summation sign to keep it from cluttering the page. It is the kind of notational trick an accountant might appreciate. Under the convention, the summational operator $\sum_i$ is implied not written, but the operator is still there. Admittedly confusing on first encounter, the convention's utility and charm are felt after only a little practice.

Incidentally, nothing requires you to invoke Einstein's summation convention everywhere and for all purposes. You can waive the convention, writing the summation symbol out explicitly whenever you like.[18] In contexts outside vector analysis, to invoke the convention at all may make little sense. Nevertheless, you should indeed learn the convention—if only because you must learn it to understand the rest of this chapter—but once having learned it you should naturally use it only where it actually serves to clarify. Fortunately, in vector work, it often does just that.

Quiz:[19] if δ_{ij} is the Kronecker delta of § 11.2, then what does the symbol δ_{ii} represent where Einstein's summation convention is in force?

15.4.3 The Kronecker delta and the Levi-Civita epsilon

Einstein's summation convention expresses the dot product (15.25) neatly but, as we have seen in (15.26), does not by itself wholly avoid unseemly repetition in the cross product. The *Levi-Civita epsilon*[20] ϵ_{ijk} mends this, rendering the cross-product as

$$\mathbf{a} \times \mathbf{b} = \epsilon_{ijk}\hat{\mathbf{\imath}}a_jb_k, \tag{15.27}$$

[17] [177, "Einstein summation"]

[18] [135]

[19] [80]

[20] Also called the *Levi-Civita symbol, tensor,* or *permutor.* For native English speakers who do not speak Italian, the "ci" in Levi-Civita's name is pronounced as the "chi" in "children."

where[21]

$$\epsilon_{ijk} \equiv \begin{cases} +1 & \text{if } (i,j,k) = (x',y',z'),\ (y',z',x') \text{ or } (z',x',y'); \\ -1 & \text{if } (i,j,k) = (x',z',y'),\ (y',x',z') \text{ or } (z',y',x'); \\ 0 & \text{otherwise [for instance if } (i,j,k) = (x',x',y')]. \end{cases} \tag{15.28}$$

In the language of § 11.6, the Levi-Civita epsilon quantifies parity. (Chapters 11 and 14 did not use it, but the Levi-Civita notation applies in any number of dimensions, not only three as in the present chapter. In this more general sense the Levi-Civita is the determinant of the permutor whose ones hold the indicated positions—which is a formal way of saying that it's a + sign for even parity and a − sign for odd. For instance, in the four-dimensional, 4 × 4 case $\epsilon_{1234} = 1$ whereas $\epsilon_{1243} = -1$: refer to §§ 11.6, 11.7.1 and 14.1. Table 15.1, however, as the rest of this section and chapter, concerns the three-dimensional case only.)

Technically, the Levi-Civita epsilon and Einstein's summation convention are two separate, independent things, but a canny reader takes the Levi-Civita's appearance as a hint that Einstein's convention is probably in force, as in (15.27). The two tend to go together.[22]

The Levi-Civita epsilon ϵ_{ijk} relates to the Kronecker delta δ_{ij} of § 11.2 approximately as the cross product relates to the dot product. Both delta and epsilon find use in vector work. For example, one can write (15.25) alternately in the form

$$\mathbf{a} \cdot \mathbf{b} = \delta_{ij} a_i b_j.$$

Table 15.1 lists several relevant properties,[23] each as with Einstein's summation convention in force.[24] Of the table's several properties, the property that $\epsilon_{imn}\epsilon_{ijk} = \delta_{mj}\delta_{nk} - \delta_{mk}\delta_{nj}$ is proved by observing that, in the case that $i = x'$, either $(j,k) = (y',z')$ or $(j,k) = (z',y')$, and also either $(m,n) = (y',z')$ or $(m,n) = (z',y')$; and similarly in the cases that

[21] [132, "Levi-Civita permutation symbol"]

[22] The writer has heard the apocryphal belief expressed that the letter ϵ, a Greek e, stood in this context for "Einstein." As far as the writer knows, ϵ is merely the letter after δ, which represents the name of Paul Dirac—though the writer does not claim his indirected story to be any less apocryphal than the other one (the capital letter Δ has a point on top that suggests the pointy nature of the Dirac delta of Fig. 7.11, which makes for yet another plausible story). In any event, one sometimes hears Einstein's summation convention, the Kronecker delta and the Levi-Civita epsilon together referred to as "the Einstein notation," which though maybe not quite terminologically correct is hardly incorrect enough to argue over and is clear enough in practice.

[23] [135]

[24] The table incidentally answers § 15.4.2's quiz.

Table 15.1: Properties of the Kronecker delta and the Levi-Civita epsilon, with Einstein's summation convention in force.

$$
\begin{aligned}
\delta_{jk} &= \delta_{kj} \\
\delta_{ij}\delta_{jk} &= \delta_{ik} \\
\delta_{ii} &= 3 \\
\delta_{jk}\epsilon_{ijk} &= 0 \\
\delta_{nk}\epsilon_{ijk} &= \epsilon_{ijn} \\
\epsilon_{ijk} = \epsilon_{jki} = \epsilon_{kij} &= -\epsilon_{ikj} = -\epsilon_{jik} = -\epsilon_{kji} \\
\epsilon_{ijk}\epsilon_{ijk} &= 6 \\
\epsilon_{ijn}\epsilon_{ijk} &= 2\delta_{nk} \\
\epsilon_{imn}\epsilon_{ijk} &= \delta_{mj}\delta_{nk} - \delta_{mk}\delta_{nj}
\end{aligned}
$$

$i = y'$ and $i = z'$ (more precisely, in each case the several indices can take any values, but combinations other than the ones listed drive ϵ_{imn} or ϵ_{ijk}, or both, to zero, thus contributing nothing to the sum). This implies that either $(j, k) = (m, n)$ or $(j, k) = (n, m)$—which, when one takes parity into account, is exactly what the property in question asserts. The property that $\epsilon_{ijn}\epsilon_{ijk} = 2\delta_{nk}$ is proved by observing that, in any given term of the Einstein sum, i is either x' or y' or z' and that j is one of the remaining two, which leaves the third to be shared by both k and n. The factor 2 appears because, for $k = n = x'$, an $(i, j) = (y', z')$ term and an $(i, j) = (z', y')$ term both contribute positively to the sum; and similarly for $k = n = y'$ and again for $k = n = z'$.

Unfortunately, the last paragraph likely makes sense to few who do not already know what it means. A concrete example helps. Consider the compound product $\mathbf{c} \times (\mathbf{a} \times \mathbf{b})$. In this section's notation and with the use of (15.27), the compound product is

$$\begin{aligned}
\mathbf{c}\times(\mathbf{a}\times\mathbf{b}) &= \mathbf{c}\times(\epsilon_{ijk}\hat{\mathbf{\imath}}a_jb_k) \\
&= \epsilon_{mni}\hat{\mathbf{m}}c_n(\epsilon_{ijk}\hat{\mathbf{\imath}}a_jb_k)_i \\
&= \epsilon_{mni}\epsilon_{ijk}\hat{\mathbf{m}}c_na_jb_k \\
&= \epsilon_{imn}\epsilon_{ijk}\hat{\mathbf{m}}c_na_jb_k \\
&= (\delta_{mj}\delta_{nk}-\delta_{mk}\delta_{nj})\hat{\mathbf{m}}c_na_jb_k \\
&= \delta_{mj}\delta_{nk}\hat{\mathbf{m}}c_na_jb_k-\delta_{mk}\delta_{nj}\hat{\mathbf{m}}c_na_jb_k \\
&= \hat{\mathbf{\jmath}}c_ka_jb_k-\hat{\mathbf{k}}c_ja_jb_k \\
&= (\hat{\mathbf{\jmath}}a_j)(c_kb_k)-(\hat{\mathbf{k}}b_k)(c_ja_j).
\end{aligned}$$

That is, in light of (15.25),

$$\mathbf{c}\times(\mathbf{a}\times\mathbf{b})=\mathbf{a}(\mathbf{c}\cdot\mathbf{b})-\mathbf{b}(\mathbf{c}\cdot\mathbf{a}), \tag{15.29}$$

a useful vector identity. Written without the benefit of Einstein's summation convention, the example's central step would have been

$$\begin{aligned}
\mathbf{c}\times(\mathbf{a}\times\mathbf{b}) &= \sum_{i,j,k,m,n}\epsilon_{imn}\epsilon_{ijk}\hat{\mathbf{m}}c_na_jb_k \\
&= \sum_{j,k,m,n}(\delta_{mj}\delta_{nk}-\delta_{mk}\delta_{nj})\hat{\mathbf{m}}c_na_jb_k,
\end{aligned}$$

which makes sense if you think about it hard enough,[25] and justifies the

[25] If thinking about it hard enough does not work, then here it is in interminable detail:

$$
\begin{aligned}
&\sum_{i,j,k,m,n} \epsilon_{imn}\epsilon_{ijk} f(j,k,m,n) \\
&\quad= \epsilon_{x'y'z'}\epsilon_{x'y'z'} f(y',z',y',z') + \epsilon_{x'y'z'}\epsilon_{x'z'y'} f(y',z',z',y') \\
&\qquad + \epsilon_{x'z'y'}\epsilon_{x'y'z'} f(z',y',y',z') + \epsilon_{x'z'y'}\epsilon_{x'z'y'} f(z',y',z',y') \\
&\qquad + \epsilon_{y'z'x'}\epsilon_{y'z'x'} f(z',x',z',x') + \epsilon_{y'z'x'}\epsilon_{y'x'z'} f(z',x',x',z') \\
&\qquad + \epsilon_{y'x'z'}\epsilon_{y'z'x'} f(x',z',z',x') + \epsilon_{y'x'z'}\epsilon_{y'x'z'} f(x',z',x',z') \\
&\qquad + \epsilon_{z'x'y'}\epsilon_{z'x'y'} f(x',y',x',y') + \epsilon_{z'x'y'}\epsilon_{z'y'x'} f(x',y',y',x') \\
&\qquad + \epsilon_{z'y'x'}\epsilon_{z'x'y'} f(y',x',x',y') + \epsilon_{z'y'x'}\epsilon_{z'y'x'} f(y',x',y',x') \\
&\quad= f(y',z',y',z') - f(y',z',z',y') - f(z',y',y',z') + f(z',y',z',y') \\
&\qquad + f(z',x',z',x') - f(z',x',x',z') - f(x',z',z',x') + f(x',z',x',z') \\
&\qquad + f(x',y',x',y') - f(x',y',y',x') - f(y',x',x',y') + f(y',x',y',x') \\
&\quad= \big[f(y',z',y',z') + f(z',x',z',x') + f(x',y',x',y') \\
&\qquad\qquad + f(z',y',z',y') + f(x',z',x',z') + f(y',x',y',x')\big] \\
&\qquad - \big[f(y',z',z',y') + f(z',x',x',z') + f(x',y',y',x') \\
&\qquad\qquad + f(z',y',y',z') + f(x',z',z',x') + f(y',x',x',y')\big] \\
&\quad= \big[f(y',z',y',z') + f(z',x',z',x') + f(x',y',x',y') \\
&\qquad\qquad + f(z',y',z',y') + f(x',z',x',z') + f(y',x',y',x') \\
&\qquad\qquad + f(x',x',x',x') + f(y',y',y',y') + f(z',z',z',z')\big] \\
&\qquad - \big[f(y',z',z',y') + f(z',x',x',z') + f(x',y',y',x') \\
&\qquad\qquad + f(z',y',y',z') + f(x',z',z',x') + f(y',x',x',y') \\
&\qquad\qquad + f(x',x',x',x') + f(y',y',y',y') + f(z',z',z',z')\big] \\
&\quad= \sum_{j,k,m,n} (\delta_{mj}\delta_{nk} - \delta_{mk}\delta_{nj}) f(j,k,m,n).
\end{aligned}
$$

That is for the property that $\epsilon_{imn}\epsilon_{ijk} = \delta_{mj}\delta_{nk} - \delta_{mk}\delta_{nj}$. For the property that $\epsilon_{ijn}\epsilon_{ijk} = 2\delta_{nk}$, the corresponding calculation is

$$
\begin{aligned}
&\sum_{i,j,k,n} \epsilon_{ijn}\epsilon_{ijk} f(k,n) \\
&\quad= \epsilon_{y'z'x'}\epsilon_{y'z'x'} f(x',x') + \epsilon_{z'y'x'}\epsilon_{z'y'x'} f(x',x') \\
&\qquad + \epsilon_{z'x'y'}\epsilon_{z'x'y'} f(y',y') + \epsilon_{x'z'y'}\epsilon_{x'z'y'} f(y',y') \\
&\qquad + \epsilon_{x'y'z'}\epsilon_{x'y'z'} f(z',z') + \epsilon_{y'x'z'}\epsilon_{y'x'z'} f(z',z') \\
&\quad= f(x',x') + f(x',x') + f(y',y') + f(y',y') + f(z',z') + f(z',z') \\
&\quad= 2\big[f(x',x') + f(y',y') + f(z',z')\big] \\
&\quad= 2\sum_{k,n} \delta_{nk} f(k,n).
\end{aligned}
$$

For the property that $\epsilon_{ijk}\epsilon_{ijk} = 6$,

$$
\sum_{i,j,k} \epsilon_{ijk}\epsilon_{ijk} = \epsilon^2_{x'y'z'} + \epsilon^2_{y'z'x'} + \epsilon^2_{z'x'y'} + \epsilon^2_{x'z'y'} + \epsilon^2_{y'x'z'} + \epsilon^2_{z'y'x'} = 6.
$$

table's claim that $\epsilon_{imn}\epsilon_{ijk} = \delta_{mj}\delta_{nk} - \delta_{mk}\delta_{nj}$. (Notice that the compound Kronecker operator $\delta_{mj}\delta_{nk}$ includes nonzero terms for the case that $j = k = m = n = x'$, for the case that $j = k = m = n = y'$ and for the case that $j = k = m = n = z'$, whereas the compound Levi-Civita operator $\epsilon_{imn}\epsilon_{ijk}$ does not. However, the compound Kronecker operator $-\delta_{mk}\delta_{nj}$ includes canceling terms for these same three cases. This is why the table's claim is valid as written.)

To belabor the topic further here would serve little purpose. The reader who does not feel entirely sure that he understands what is going on might work out the table's several properties with his own pencil, in something like the style of the example, until he is satisfied that he adequately understands the several properties and their correct use.

Section 16.7 will refine the notation for use when derivatives with respect to angles come into play but, before leaving the present section, we might pause for a moment to appreciate (15.29) in the special case that $\mathbf{b} = \mathbf{c} = \hat{\mathbf{n}}$:

$$-\,\hat{\mathbf{n}} \times (\hat{\mathbf{n}} \times \mathbf{a}) = \mathbf{a} - \hat{\mathbf{n}}(\hat{\mathbf{n}} \cdot \mathbf{a}). \tag{15.30}$$

The difference $\mathbf{a} - \hat{\mathbf{n}}(\hat{\mathbf{n}} \cdot \mathbf{a})$ evidently projects a vector $\mathbf{a}$ onto the plane whose unit normal is $\hat{\mathbf{n}}$. Equation (15.30) reveals that the double cross product $-\hat{\mathbf{n}} \times (\hat{\mathbf{n}} \times \mathbf{a})$ projects the same vector onto the same plane. Figure 15.6 illustrates.

15.5 Algebraic identities

Vector algebra is not in principle very much harder than scalar algebra is, but with three distinct types of product it has more rules controlling the way its products and sums are combined. Table 15.2 lists several of these.[26,27] Most of the table's identities are plain by the formulas (15.9), (15.21) and (15.22) respectively for the scalar, dot and cross products, and two were proved as (15.29) and (15.30). The remaining identity is proved in the notation of § 15.4 as

$$\begin{array}{rclclcl}
\epsilon_{ijk}c_ia_jb_k & = & \epsilon_{ijk}c_ia_jb_k & = & \epsilon_{kij}c_ka_ib_j & = & \epsilon_{jki}c_ja_kb_i \\
 & = & \epsilon_{ijk}c_ia_jb_k & = & \epsilon_{ijk}a_ib_jc_k & = & \epsilon_{ijk}b_ic_ja_k \\
 & = & \mathbf{c} \cdot (\epsilon_{ijk}\hat{\mathbf{i}}a_jb_k) & = & \mathbf{a} \cdot (\epsilon_{ijk}\hat{\mathbf{i}}b_jc_k) & = & \mathbf{b} \cdot (\epsilon_{ijk}\hat{\mathbf{i}}c_ja_k).
\end{array}$$

It is precisely to encapsulate such interminable detail that we use the Kronecker delta, the Levi-Civita epsilon and the properties of Table 15.1.

[26] [157, appendix II][72, appendix A]

[27] Nothing in any of the table's identities requires the vectors involved to be real. The table is equally as valid when vectors are complex.

Figure 15.6: A vector projected onto a plane.

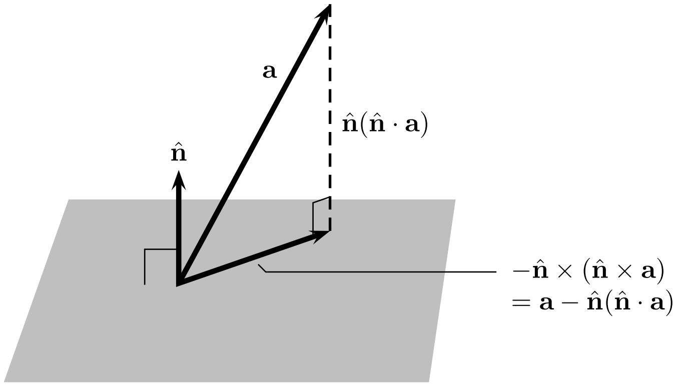

$\mathbf{a}$
$\hat{\mathbf{n}}(\hat{\mathbf{n}}\cdot\mathbf{a})$
$\hat{\mathbf{n}}$
$-\hat{\mathbf{n}}\times(\hat{\mathbf{n}}\times\mathbf{a})$
$= \mathbf{a}-\hat{\mathbf{n}}(\hat{\mathbf{n}}\cdot\mathbf{a})$

Table 15.2: Algebraic vector identities.

$$\begin{array}{rclrclrcl}
\psi\mathbf{a} & = & \hat{\mathbf{\imath}}\psi a_i & \mathbf{a}\cdot\mathbf{b} & \equiv & a_i b_i & \mathbf{a}\times\mathbf{b} & \equiv & \epsilon_{ijk}\hat{\mathbf{\imath}}a_j b_k
\end{array}$$

$$\begin{array}{rclrcl}
\mathbf{a}^*\cdot\mathbf{a} & = & |a|^2 & (\psi)(\mathbf{a}+\mathbf{b}) & = & \psi\mathbf{a}+\psi\mathbf{b} \\
\mathbf{b}\cdot\mathbf{a} & = & \mathbf{a}\cdot\mathbf{b} & \mathbf{b}\times\mathbf{a} & = & -\mathbf{a}\times\mathbf{b} \\
\mathbf{c}\cdot(\mathbf{a}+\mathbf{b}) & = & \mathbf{c}\cdot\mathbf{a}+\mathbf{c}\cdot\mathbf{b} & \mathbf{c}\times(\mathbf{a}+\mathbf{b}) & = & \mathbf{c}\times\mathbf{a}+\mathbf{c}\times\mathbf{b} \\
\mathbf{a}\cdot(\psi\mathbf{b}) & = & (\psi)(\mathbf{a}\cdot\mathbf{b}) & \mathbf{a}\times(\psi\mathbf{b}) & = & (\psi)(\mathbf{a}\times\mathbf{b})
\end{array}$$

$$\begin{array}{rclcl}
\mathbf{c}\cdot(\mathbf{a}\times\mathbf{b}) & = & \mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}) & = & \mathbf{b}\cdot(\mathbf{c}\times\mathbf{a}) \\
\mathbf{c}\times(\mathbf{a}\times\mathbf{b}) & = & \mathbf{a}(\mathbf{c}\cdot\mathbf{b})-\mathbf{b}(\mathbf{c}\cdot\mathbf{a}) & & \\
-\hat{\mathbf{n}}\times(\hat{\mathbf{n}}\times\mathbf{a}) & = & \mathbf{a}-\hat{\mathbf{n}}(\hat{\mathbf{n}}\cdot\mathbf{a}) & &
\end{array}$$

That is,

$$\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}). \tag{15.31}$$

Besides the several vector identities, the table also includes the three vector products in Einstein notation.[28]

Each definition and identity of Table 15.2 is invariant under reorientation of axes.

15.6 Isotropy

A real,[29] three-dimensional coordinate system[30] $(\alpha; \beta; \gamma)$ is *isotropic* at a point $\mathbf{r} = \mathbf{r}_1$ if and only if

$$\begin{aligned} \hat{\boldsymbol\beta}(\mathbf{r}_1) \cdot \hat{\boldsymbol\gamma}(\mathbf{r}_1) &= 0, \\ \hat{\boldsymbol\gamma}(\mathbf{r}_1) \cdot \hat{\boldsymbol\alpha}(\mathbf{r}_1) &= 0, \\ \hat{\boldsymbol\alpha}(\mathbf{r}_1) \cdot \hat{\boldsymbol\beta}(\mathbf{r}_1) &= 0, \end{aligned} \tag{15.32}$$

and

$$\left|\frac{\partial \mathbf{r}}{\partial \alpha}\right|_{\mathbf{r}=\mathbf{r}_1} = \left|\frac{\partial \mathbf{r}}{\partial \beta}\right|_{\mathbf{r}=\mathbf{r}_1} = \left|\frac{\partial \mathbf{r}}{\partial \gamma}\right|_{\mathbf{r}=\mathbf{r}_1}. \tag{15.33}$$

That is, a three-dimensional system is isotropic if its three coordinates advance locally at right angles to one another but at the same rate.

Of the three basic three-dimensional coordinate systems—indeed, of all the three-dimensional coordinate systems this book treats—only the rectangular is isotropic according to (15.32) and (15.33).[31] Isotropy admittedly would not be a very interesting property if that were all there were to it. However, there is also *two-dimensional isotropy,* more interesting because it arises oftener.

[28] If the reader's native language is English, then he is likely to have heard of the unfortunate "back cab rule," which actually is not a rule but an unhelpful mnemonic for one of Table 15.2's identities. The mnemonic is mildly orthographically clever but, when learned, significantly impedes real understanding of the vector. The writer recommends that the reader forget the rule if he has heard of it for, in mathematics, spelling-based mnemonics are seldom if ever a good idea.

[29] The reader is reminded that one can licitly express a complex vector in a real basis.

[30] This chapter's footnote 32 and chapter 16's footnote 21 explain the usage of semicolons as coordinate delimiters.

[31] Whether it is even possible to construct an isotropic, nonrectangular coordinate system in three dimensions is a question we will leave to the professional mathematician. The author has not encountered such a system.

A real, two-dimensional coordinate system $(\alpha;\beta)$ is isotropic at a point $\boldsymbol{\rho}^\gamma = \boldsymbol{\rho}^\gamma_1$ if and only if

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\rho}^\gamma_1)\cdot\hat{\boldsymbol{\beta}}(\boldsymbol{\rho}^\gamma_1) = 0 \tag{15.34}$$

and

$$\left|\frac{\partial\boldsymbol{\rho}^\gamma}{\partial\alpha}\right|_{\boldsymbol{\rho}^\gamma=\boldsymbol{\rho}^\gamma_1} = \left|\frac{\partial\boldsymbol{\rho}^\gamma}{\partial\beta}\right|_{\boldsymbol{\rho}^\gamma=\boldsymbol{\rho}^\gamma_1}, \tag{15.35}$$

where $\boldsymbol{\rho}^\gamma = \hat{\boldsymbol{\alpha}}\alpha + \hat{\boldsymbol{\beta}}\beta$ represents position in the α-β plane. (If the α-β plane happens to be the x-y plane, as is often the case, then $\boldsymbol{\rho}^\gamma = \boldsymbol{\rho}^z = \boldsymbol{\rho}$ and per eqn. 3.20 one can omit the superscript.) The two-dimensional rectangular system (x, y) naturally is isotropic. Because $|\partial\boldsymbol{\rho}/\partial\phi| = (\rho)\,|\partial\boldsymbol{\rho}/\partial\rho|$ the standard two-dimensional cylindrical system $(\rho;\phi)$ as such is nonisotropic, but the change of coordinate

$$\lambda \equiv \ln\frac{\rho}{\rho_o}, \tag{15.36}$$

where ρ_o is some arbitrarily chosen reference radius, converts the system straightforwardly into the *logarithmic cylindrical* system $(\lambda;\phi)$ which is isotropic everywhere in the plane except at the origin $\rho = 0$. Further two-dimensionally isotropic coordinate systems include the parabolic system of § 15.7.2, to follow.

15.7 Parabolic coordinates

Scientists and engineers find most spatial-geometrical problems they encounter in practice to fall into either of two categories. The first category comprises problems of simple geometry conforming to any one of the three basic coordinate systems—rectangular, cylindrical or spherical. The second category comprises problems of complicated geometry, analyzed in the rectangular system not because the problems' geometries fit that system but rather because they fit no system and thus give one little reason to depart from the rectangular. One however occasionally encounters problems of a third category, whose geometries are simple but, though simple, nevertheless fit none of the three basic coordinate systems. Then it may fall to the scientist or engineer to devise a special coordinate system congenial to the problem.

This section will treat the *parabolic* coordinate systems which, besides being arguably the most useful of the various special systems, serve as good examples of the kind of special system a scientist or engineer might be called upon to devise. The two three-dimensional parabolic systems are the *parabolic cylindrical* system (σ, τ, z) of § 15.7.4 and the *circular paraboloidal* system[32] $(\eta; \phi, \xi)$ of § 15.7.5, where the angle ϕ and the length z are familiar to us but σ, τ, η and ξ—neither angles nor lengths but root-lengths (that is, coordinates having dimensions of $[\text{length}]^{1/2}$)—are new.[33] Both three-dimensional parabolic systems derive from the two-dimensional parabolic system (σ, τ) of § 15.7.2.[34]

However, before handling any parabolic system we ought formally to introduce the parabola itself, next.

15.7.1 The parabola

Parabolic coordinates are based on a useful geometrical curve called the *parabola,* which many or most readers will have met long before opening this book's covers. The parabola, simple but less obvious than the circle, may however not be equally well known to all readers, and even readers already acquainted with it (as from §§ 7.4.1 and 7.4.2) might appreciate a reëxamination. This subsection reviews the parabola.

Given a point, called the focus, *and a line, called the* directrix,[35] *plus the plane in which the focus and the directrix both lie, the associated* parabola *is that curve which lies in the plane everywhere equidistant from both focus and directrix.*[36] See Fig. 15.7.

Referring to the figure, if rectangular coordinates are established such that $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ lie in the plane, that the parabola's focus lies at $(x, y) = (0, k)$, and that the equation $y = k - \sigma^2$ describes the parabola's directrix, then the equation

$$x^2 + (y - k)^2 = (y - k + \sigma^2)^2$$

[32] The reader will probably think nothing of it now, but later may wonder why the circular paraboloidal coordinates are $(\eta; \phi, \xi)$ rather than $(\xi; \phi, \eta)$ or $(\eta, \xi; \phi)$. The peculiar ordering is to honor the right-hand rule (§ 3.3 and eqn. 15.19), since $\hat{\boldsymbol{\eta}} \times \hat{\boldsymbol{\xi}} = -\hat{\boldsymbol{\phi}}$ rather than $+\hat{\boldsymbol{\phi}}$. See § 15.7.5. (Regarding the semicolon ";" delimiter, it doesn't mean much. This book arbitrarily uses a semicolon when the following coordinate happens to be an angle, which helps to distinguish rectangular coordinates from cylindrical from spherical. Admittedly, such a notational convention ceases to help much when parabolic coordinates arrive, but we will continue to use it for inertia's sake. See also chapter 16's footnote 21.)

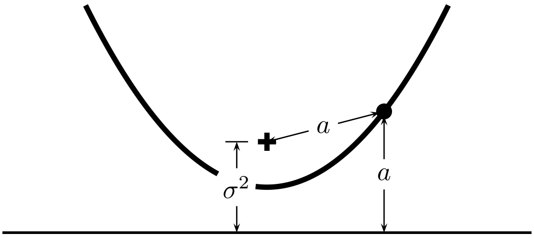
[33] The letters σ, τ, η and ξ are available letters this section happens to use, not necessarily standard parabolic symbols. See appendix B.

[34] [107, § 10.1][182, "Parabolic coordinates," 09:59, 19 July 2008]

[35] Whether the parabola's definition ought to forbid the directrix to pass through the focus is a stylistic question this book will leave unanswered.

[36] [146, § 12-1]

Figure 15.7: The parabola.



evidently expresses the equidistance rule of the parabola's definition. Solving for $y - k$ and then, from that solution, for y, we have that

$$y = \frac{x^2}{2\sigma^2} + \left(k - \frac{\sigma^2}{2}\right). \tag{15.37}$$

With the definitions that

$$\begin{aligned} \mu &\equiv \frac{1}{2\sigma^2}, \\ \kappa &\equiv k - \frac{\sigma^2}{2}, \end{aligned} \tag{15.38}$$

given which

$$\begin{aligned} \sigma^2 &= \frac{1}{2\mu}, \\ k &= \kappa + \frac{1}{4\mu}, \end{aligned} \tag{15.39}$$

eqn. (15.37) becomes

$$y = \mu x^2 + \kappa. \tag{15.40}$$

Equations fitting the general form (15.40) often arise in applications, for example in the equation that describes a projectile's flight in the absence of air resistance. Any equation that fits the form can be plotted as a parabola, which is why projectiles fly in parabolic arcs.

Observe that the parabola's definition does not actually require the directrix to be $\hat{\mathbf{y}}$-oriented: the directrix can be $\hat{\mathbf{x}}$-oriented or, indeed, oriented any way. Observe also the geometrical fact that *the parabola's track necessarily bisects the angle between the two line segments labeled "a" in Fig. 15.7.* One of the consequences of this geometrical fact[37]—is that a parabolic mirror reflects precisely[38] toward its focus all light rays that arrive perpendicularly with respect to the directrix (which for instance is why satellite dish antennas have parabolic cross-sections).

15.7.2 Parabolic coordinates in two dimensions

Parabolic coordinates are most easily first explained in the two-dimensional case that $z = 0$. In two dimensions, the parabolic coordinates (σ, τ) represent the point in the x-y plane that lies equidistant

- from the line $y = -\sigma^2$,
- from the line $y = +\tau^2$, and
- from the point $\rho = 0$,

where the parameter k of § 15.7.1 has been set to $k = 0$. Figure 15.8 depicts the construction described. In the figure are two dotted curves, one of which represents the point's parabolic track if σ were varied while τ were held constant and the other of which represents the point's parabolic track if τ were varied while σ were held constant. Observe according to § 15.7.1's bisection finding that each parabola necessarily bisects the angle between two of the three line segments labeled a in the figure. Observe further that the two angles' sum is the straight angle $2\pi/2$, from which one can conclude, significantly, that *the two parabolas cross at right angles to one another.*

Figure 15.9 lays out the parabolic coordinate grid. Notice in the figure

[37] It seems better merely to let the reader visualize the fact than to try to justify in so many words. If words help nevertheless, some words: consider that the two line segments labeled a in the figure run in the directions of increasing distance respectively from the focus and from the directrix. If you want to draw away from the directrix at the same rate as you draw away from the focus, thus maintaining equal distances, then your track cannot but exactly bisect the angle between the two segments.

Once you grasp the idea, the bisection is obvious, though to grasp the idea can take some thought.

To bisect a thing, incidentally—if the context has not already made the meaning plain—is to divide the thing at its middle into two equal parts.

[38] Well, actually, physically, the ray model of light implied here is valid only insofar as $\lambda \ll \sigma^2$, where λ represents the light's characteristic wavelength. Also, regardless of λ, the ray model breaks down in the immediate neighborhood of the mirror's focus. However, we were not thinking of wave mechanics at the moment. Insofar as rays are concerned, the focusing is precise.

Figure 15.8: Locating a point in two dimensions by parabolic construction.

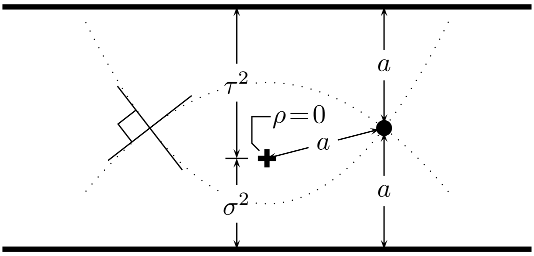
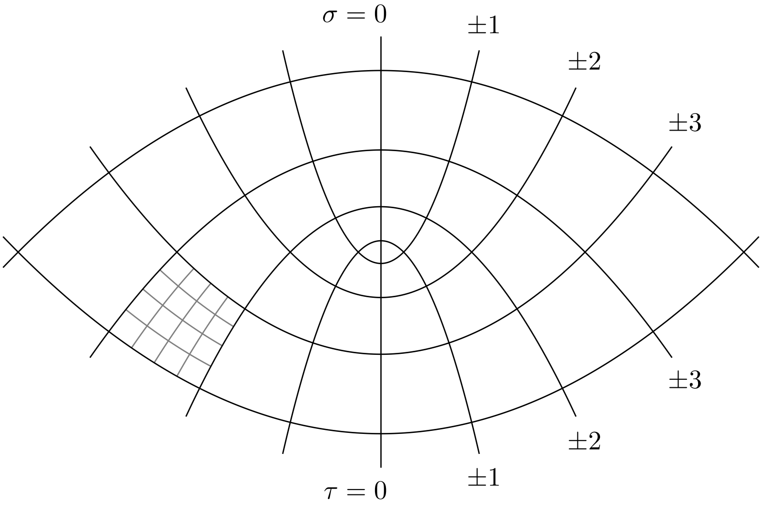


Figure 15.9: The parabolic coordinate grid in two dimensions.



that one of the grid's several cells is subdivided at its quarter-marks for illustration's sake, to show how one can subgrid at need to locate points like, for example, $(\sigma, \tau) = (\frac{7}{2}, -\frac{9}{4})$ visually. (That the subgrid's cells approach square shape implies that the parabolic system is isotropic, a significant fact § 15.7.3 will formally demonstrate.)

Using the Pythagorean theorem, one can symbolically express the equidistant construction rule above as

$$\begin{aligned} a &= \sigma^2 + y = \tau^2 - y, \\ a^2 &= \rho^2 = x^2 + y^2. \end{aligned} \tag{15.41}$$

From the first line of (15.41),

$$y = \frac{\tau^2 - \sigma^2}{2}. \tag{15.42}$$

On the other hand, combining the two lines of (15.41),

$$(\sigma^2 + y)^2 = x^2 + y^2 = (\tau^2 - y)^2,$$

or, subtracting y^2,

$$\sigma^4 + 2\sigma^2 y = x^2 = \tau^4 - 2\tau^2 y.$$

Substituting (15.42)'s expression for y,

$$x^2 = (\sigma\tau)^2.$$

That either $x = +\sigma\tau$ or $x = -\sigma\tau$ would satisfy this equation. Arbitrarily choosing the $+$ sign gives us that

$$x = \sigma\tau. \tag{15.43}$$

Also, since $\rho^2 = x^2 + y^2$, (15.42) and (15.43) together imply that

$$\rho = \frac{\tau^2 + \sigma^2}{2}. \tag{15.44}$$

Combining (15.42) and (15.44) to isolate σ^2 and τ^2 yields that

$$\begin{aligned} \sigma^2 &= \rho - y, \\ \tau^2 &= \rho + y. \end{aligned} \tag{15.45}$$

15.7.3 Properties

The derivatives of (15.43), (15.42) and (15.44) are

$$\begin{aligned} dx &= \sigma\,d\tau + \tau\,d\sigma, \\ dy &= \tau\,d\tau - \sigma\,d\sigma, \\ d\rho &= \tau\,d\tau + \sigma\,d\sigma. \end{aligned} \tag{15.46}$$

Solving the first two lines of (15.46) simultaneously for $d\sigma$ and $d\tau$ and then collapsing the resultant subexpression $\tau^2 + \sigma^2$ per (15.44) yields that

$$\begin{aligned} d\sigma &= \frac{\tau\,dx - \sigma\,dy}{2\rho}, \\ d\tau &= \frac{\sigma\,dx + \tau\,dy}{2\rho}, \end{aligned} \tag{15.47}$$

from which it is apparent that

$$\begin{aligned} \hat{\boldsymbol\sigma} &= \frac{\hat{\mathbf{x}}\tau - \hat{\mathbf{y}}\sigma}{\sqrt{\tau^2+\sigma^2}}, \\ \hat{\boldsymbol\tau} &= \frac{\hat{\mathbf{x}}\sigma + \hat{\mathbf{y}}\tau}{\sqrt{\tau^2+\sigma^2}}; \end{aligned}$$

or, collapsing again per (15.44), that

$$\begin{aligned} \hat{\boldsymbol\sigma} &= \frac{\hat{\mathbf{x}}\tau - \hat{\mathbf{y}}\sigma}{\sqrt{2\rho}}, \\ \hat{\boldsymbol\tau} &= \frac{\hat{\mathbf{x}}\sigma + \hat{\mathbf{y}}\tau}{\sqrt{2\rho}}, \end{aligned} \tag{15.48}$$

of which the dot product

$$\hat{\boldsymbol\sigma}\cdot\hat{\boldsymbol\tau} = 0 \text{ if } \rho \neq 0 \tag{15.49}$$

is null, confirming our earlier finding that the various grid parabolas cross always at right angles to one another. Solving (15.48) simultaneously for $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ then produces

$$\begin{aligned} \hat{\mathbf{x}} &= \frac{\hat{\boldsymbol\tau}\sigma + \hat{\boldsymbol\sigma}\tau}{\sqrt{2\rho}}, \\ \hat{\mathbf{y}} &= \frac{\hat{\boldsymbol\tau}\tau - \hat{\boldsymbol\sigma}\sigma}{\sqrt{2\rho}}. \end{aligned} \tag{15.50}$$

One can express an infinitesimal change in position in the plane as

$$\begin{aligned} d\boldsymbol{\rho} &= \hat{\mathbf{x}}\,dx + \hat{\mathbf{y}}\,dy \\ &= \hat{\mathbf{x}}(\sigma\,d\tau + \tau\,d\sigma) + \hat{\mathbf{y}}(\tau\,d\tau - \sigma\,d\sigma) \\ &= (\hat{\mathbf{x}}\tau - \hat{\mathbf{y}}\sigma)\,d\sigma + (\hat{\mathbf{x}}\sigma + \hat{\mathbf{y}}\tau)\,d\tau, \end{aligned}$$

in which (15.46) has expanded the differentials and from which

$$\begin{aligned} \frac{\partial\boldsymbol{\rho}}{\partial\sigma} &= \hat{\mathbf{x}}\tau - \hat{\mathbf{y}}\sigma, \\ \frac{\partial\boldsymbol{\rho}}{\partial\tau} &= \hat{\mathbf{x}}\sigma + \hat{\mathbf{y}}\tau, \end{aligned}$$

and thus

$$\left|\frac{\partial\boldsymbol{\rho}}{\partial\sigma}\right| = \left|\frac{\partial\boldsymbol{\rho}}{\partial\tau}\right|. \tag{15.51}$$

Equations (15.49) and (15.51) respectively meet the requirements (15.34) and (15.35), implying that *the two-dimensional parabolic coordinate system is isotropic* except at $\rho = 0$.

Table 15.3 summarizes, gathering parabolic coordinate properties from this subsection and § 15.7.2.

15.7.4 The parabolic cylindrical coordinate system

Two-dimensional parabolic coordinates are easily extended to three dimensions by adding a z coordinate, thus constituting the *parabolic cylindrical* coordinate system (σ, τ, z). The surfaces of constant σ and of constant τ in this system are *parabolic cylinders* (and the surfaces of constant z naturally are planes). All the properties of Table 15.3 apply. Observe however that the system is isotropic only in two dimensions not three.

The orthogonal parabolic cylindrical basis is $[\hat{\boldsymbol{\sigma}}\ \hat{\boldsymbol{\tau}}\ \hat{\mathbf{z}}]$.

15.7.5 The circular paraboloidal coordinate system

Sometimes one would like to extend the parabolic system to three dimensions by adding an azimuth ϕ rather than a height z. This is possible, but then one tends to prefer the parabolas, foci and directrices of Figs. 15.8 and 15.9 to run in the ρ-z plane rather than in the x-y. Therefore, one defines the coordinates η and ξ to represent in the ρ-z plane what the letters σ and τ have represented in the x-y. The properties of Table 15.4 result, which are just the properties of Table 15.3 with coordinates changed. The system is

Table 15.3: Parabolic coordinate properties.

$$
\begin{aligned}
x &= \sigma\tau \\
y &= \frac{\tau^2 - \sigma^2}{2} \\
\rho &= \frac{\tau^2 + \sigma^2}{2} \\
\rho^2 &= x^2 + y^2 \\
\sigma^2 &= \rho - y \\
\tau^2 &= \rho + y
\end{aligned}
$$

$$
\begin{aligned}
\hat{\mathbf{x}} &= \frac{\hat{\boldsymbol{\tau}}\sigma + \hat{\boldsymbol{\sigma}}\tau}{\sqrt{2\rho}} \\
\hat{\mathbf{y}} &= \frac{\hat{\boldsymbol{\tau}}\tau - \hat{\boldsymbol{\sigma}}\sigma}{\sqrt{2\rho}} \\
\hat{\boldsymbol{\sigma}} &= \frac{\hat{\mathbf{x}}\tau - \hat{\mathbf{y}}\sigma}{\sqrt{2\rho}} \\
\hat{\boldsymbol{\tau}} &= \frac{\hat{\mathbf{x}}\sigma + \hat{\mathbf{y}}\tau}{\sqrt{2\rho}} \\
\hat{\boldsymbol{\sigma}} \times \hat{\boldsymbol{\tau}} &= \hat{\mathbf{z}} \\
\hat{\boldsymbol{\sigma}} \cdot \hat{\boldsymbol{\tau}} &= 0 \\
\left|\frac{\partial \boldsymbol{\rho}}{\partial \sigma}\right| &= \left|\frac{\partial \boldsymbol{\rho}}{\partial \tau}\right|
\end{aligned}
$$

Table 15.4: Circular paraboloidal coordinate properties.

$$
\begin{aligned}
\rho &= \eta\xi \\
z &= \frac{\xi^2 - \eta^2}{2} \\
r &= \frac{\xi^2 + \eta^2}{2} \\
r^2 &= \rho^2 + z^2 = x^2 + y^2 + z^2 \\
\eta^2 &= r - z \\
\xi^2 &= r + z
\end{aligned}
$$

$$
\begin{aligned}
\hat{\boldsymbol{\rho}} &= \frac{\hat{\boldsymbol{\xi}}\eta + \hat{\boldsymbol{\eta}}\xi}{\sqrt{2r}} \\
\hat{\mathbf{z}} &= \frac{\hat{\boldsymbol{\xi}}\xi - \hat{\boldsymbol{\eta}}\eta}{\sqrt{2r}} \\
\hat{\boldsymbol{\eta}} &= \frac{\hat{\boldsymbol{\rho}}\xi - \hat{\mathbf{z}}\eta}{\sqrt{2r}} \\
\hat{\boldsymbol{\xi}} &= \frac{\hat{\boldsymbol{\rho}}\eta + \hat{\mathbf{z}}\xi}{\sqrt{2r}} \\
\hat{\boldsymbol{\eta}} \times \hat{\boldsymbol{\xi}} &= -\hat{\boldsymbol{\phi}} \\
\hat{\boldsymbol{\eta}} \cdot \hat{\boldsymbol{\xi}} &= 0 \\
\left|\frac{\partial \mathbf{r}}{\partial \eta}\right| &= \left|\frac{\partial \mathbf{r}}{\partial \xi}\right|
\end{aligned}
$$

the *circular paraboloidal system* $(\eta; \phi, \xi)$.

The surfaces of constant η and of constant ξ in the circular paraboloidal system are *paraboloids,* parabolas rotated about the z axis (and the surfaces of constant ϕ are planes, or half planes if you like, just as in the cylindrical system). Like the parabolic cylindrical system, the circular paraboloidal system too is isotropic in two dimensions.

Notice that, given the usual definition of the $\hat{\boldsymbol{\phi}}$ unit basis vector, $\hat{\boldsymbol{\eta}} \times \hat{\boldsymbol{\xi}} = -\hat{\boldsymbol{\phi}}$ rather than $+\hat{\boldsymbol{\phi}}$ as one might first guess. The correct, right-handed sequence of the orthogonal circular paraboloidal basis therefore would be $[\hat{\boldsymbol{\eta}}\ \hat{\boldsymbol{\phi}}\ \hat{\boldsymbol{\xi}}]$.[39]

This concludes the present chapter on the algebra of vector analysis. Chapter 16, next, will venture hence into the larger and even more interesting realm of vector calculus.

[39] See footnote 32.

Chapter 16

Vector calculus

Chapter 15 has introduced the algebra of the three-dimensional geometrical vector. Like the scalar, the vector is a continuous quantity and as such has not only an algebra but also a calculus. This chapter develops the calculus of the vector.

16.1 Fields and their derivatives

A scalar quantity $\sigma(t)$ or vector quantity $\mathbf{f}(t)$ whose value varies over time is "a function of time t." We can likewise call a scalar quantity[1] $\psi(\mathbf{r})$ or vector quantity $\mathbf{a}(\mathbf{r})$ whose value varies over space "a function of position $\mathbf{r}$," but there is a special, alternate name for such a quantity. We call it a *field.*

A field is a quantity distributed over space or, if you prefer, a function in which spatial position serves as independent variable. Air pressure $p(\mathbf{r})$ is an example of a *scalar field,* whose value at a given location $\mathbf{r}$ has amplitude but no direction. Wind velocity[2] $\mathbf{q}(\mathbf{r})$ is an example of a *vector field,* whose value at a given location $\mathbf{r}$ has both amplitude and direction. These are typical examples. Tactically, a vector field can be thought of as composed of three scalar fields

$$\mathbf{q}(\mathbf{r}) = \hat{\mathbf{x}}q_x(\mathbf{r}) + \hat{\mathbf{y}}q_y(\mathbf{r}) + \hat{\mathbf{z}}q_z(\mathbf{r});$$

[1] This $\psi(\mathbf{r})$ is unrelated to the Tait-Bryan and Euler roll angles ψ of § 15.1, an unfortunate but tolerable overloading of the Greek letter ψ in the conventional notation of vector analysis. In the unlikely event of confusion, you can use an alternate letter like η for the roll angle. See appendix B.

[2] As § 15.3, this section also uses the letter q for velocity in place of the conventional v [20, § 18.4], which it needs for another purpose.

but, since

$$\mathbf{q}(\mathbf{r}) = \hat{\mathbf{x}}' q_{x'}(\mathbf{r}) + \hat{\mathbf{y}}' q_{y'}(\mathbf{r}) + \hat{\mathbf{z}}' q_{z'}(\mathbf{r})$$

for any orthogonal basis $[\mathbf{x}' \; \mathbf{y}' \; \mathbf{z}']$ as well, the specific scalar fields $q_x(\mathbf{r})$, $q_y(\mathbf{r})$ and $q_z(\mathbf{r})$ are no more essential to the vector field $\mathbf{q}(\mathbf{r})$ than the specific scalars b_x, b_y and b_z are to a vector $\mathbf{b}$. As we said, the three components come tactically; typically, such components are uninteresting in themselves. The field $\mathbf{q}(\mathbf{r})$ as a whole is the interesting thing.

Scalar and vector fields are of utmost use in the modeling of physical phenomena.

As one can take the derivative $d\sigma/dt$ or $d\mathbf{f}/dt$ with respect to time t of a function $\sigma(t)$ or $\mathbf{f}(t)$, one can likewise take the derivative with respect to position $\mathbf{r}$ of a field $\psi(\mathbf{r})$ or $\mathbf{a}(\mathbf{r})$. However, derivatives with respect to position create a notational problem, for it is not obvious what symbology like $d\psi/d\mathbf{r}$ or $d\mathbf{a}/d\mathbf{r}$ would actually mean. The notation $d\sigma/dt$ means "the rate of σ as time t advances," but if the notation $d\psi/d\mathbf{r}$ likewise meant "the rate of ψ as position $\mathbf{r}$ advances" then it would necessarily prompt one to ask, "advances in which direction?" The notation offers no hint. In fact $d\psi/d\mathbf{r}$ and $d\mathbf{a}/d\mathbf{r}$ mean nothing very distinct in most contexts and we shall avoid such notation. If we will speak of a field's derivative with respect to position $\mathbf{r}$ then we shall be more precise.

Section 15.2 has given the vector three distinct kinds of product. This section gives the field no fewer than four distinct kinds of derivative: the directional derivative; the gradient; the divergence; and the curl.[3]

So many derivatives bring the student a conceptual difficulty one could call "the caveman problem." Imagine a caveman. Suppose that you tried to describe to the caveman a house or building of more than one floor. He might not understand. You and I who already grasp the concept of upstairs and downstairs do not find a building of two floors, or three or even thirty, especially hard to envision, but our caveman is used to thinking of the ground and the floor as more or less the same thing. To try to think of upstairs and downstairs might confuse him with partly false images of sitting in a tree or of clambering onto (and breaking) the roof of his hut. "There are many trees and antelopes but only one sky and floor. How can one speak of many skies or many floors?" The student's principal conceptual difficulty with the several vector derivatives is of this kind.

[3] Vector veterans may notice that the Laplacian is not listed. This is not because the Laplacian were uninteresting but rather because the Laplacian is actually a second-order derivative—a derivative of a derivative. We will address the Laplacian in § 16.4.

16.1.1 The ∇ operator

Consider a vector

$$\mathbf{a} = \hat{\mathbf{x}}a_x + \hat{\mathbf{y}}a_y + \hat{\mathbf{z}}a_z.$$

Then consider a "vector"

$$\mathbf{c} = \hat{\mathbf{x}}[\text{Tuesday}] + \hat{\mathbf{y}}[\text{Wednesday}] + \hat{\mathbf{z}}[\text{Thursday}].$$

If you think that the latter does not look very much like a vector, then the writer thinks as you do, but consider:

$$\mathbf{c} \cdot \mathbf{a} = [\text{Tuesday}]a_x + [\text{Wednesday}]a_y + [\text{Thursday}]a_z.$$

The writer does not know how to interpret a nonsensical term like "[Tuesday]a_x" any more than the reader does, but the point is that **c** behaves as though it were a vector insofar as vector operations like the dot product are concerned. What matters in this context is not that **c** have amplitude and direction (it has neither) but rather that it have the three orthonormal components it needs to participate formally in relevant vector operations. It has these. That the components' amplitudes seem nonsensical is beside the point. Maybe there exists a model in which "[Tuesday]" knows how to operate on a scalar like a_x. (Operate on? Yes. Nothing in the dot product's definition requires the component amplitudes of **c** *to multiply* those of **a**. Multiplication is what the component amplitudes of true vectors do, but **c** is not a true vector, so "[Tuesday]" might do something to a_x other than to multiply it. Section 16.1.2 elaborates the point.) If there did exist such a model, then the dot product $\mathbf{c} \cdot \mathbf{a}$ could be licit in that model. As if this were not enough, the cross product $\mathbf{c} \times \mathbf{a}$ too could be licit in that model, composed according to the usual rule for cross products. The model might allow it. The dot and cross products in and of themselves do not forbid it.

Now consider a "vector"

$$\nabla = \hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}. \tag{16.1}$$

This ∇ is not a true vector any more than **c** is, maybe, but if we treat it as one then we have that

$$\nabla \cdot \mathbf{a} = \frac{\partial a_x}{\partial x} + \frac{\partial a_y}{\partial y} + \frac{\partial a_z}{\partial z}.$$

Such a dot product might or might not prove useful; but, unlike the terms in the earlier dot product, at least we know what this one's terms mean.

Well, days of the week, partial derivatives, ersatz vectors—it all seems rather abstract. What's the point? The answer is that there wouldn't be any point if the only nonvector "vectors" in question were of **c**'s nonsensical kind. The operator ∇ however shares more in common with a true vector than merely having x, y and z components; for, like a true vector, the operator ∇ is amenable to having its axes reoriented by (15.1), (15.2), (15.7) and (15.8). This is easier to see at first with respect the true vector **a**, as follows. Consider rotating the x and y axes through an angle ϕ about the z axis. There ensues

$$\begin{aligned}
\mathbf{a} &= \hat{\mathbf{x}}a_x + \hat{\mathbf{y}}a_y + \hat{\mathbf{z}}a_z \\
&= (\hat{\mathbf{x}}'\cos\phi - \hat{\mathbf{y}}'\sin\phi)(a_{x'}\cos\phi - a_{y'}\sin\phi) \\
&\qquad + (\hat{\mathbf{x}}'\sin\phi + \hat{\mathbf{y}}'\cos\phi)(a_{x'}\sin\phi + a_{y'}\cos\phi) + \hat{\mathbf{z}}'a_{z'} \\
&= \hat{\mathbf{x}}'[a_{x'}\cos^2\phi - a_{y'}\cos\phi\sin\phi + a_{x'}\sin^2\phi + a_{y'}\cos\phi\sin\phi] \\
&\qquad + \hat{\mathbf{y}}'[-a_{x'}\cos\phi\sin\phi + a_{y'}\sin^2\phi + a_{x'}\cos\phi\sin\phi + a_{y'}\cos^2\phi] \\
&\qquad + \hat{\mathbf{z}}'a_{z'} \\
&= \hat{\mathbf{x}}'a_{x'} + \hat{\mathbf{y}}'a_{y'} + \hat{\mathbf{z}}'a_{z'},
\end{aligned}$$

where the final expression has different axes than the original but, relative to those axes, exactly the same form. Further rotation about other axes would further reorient but naturally also would not alter the form. Now consider ∇. The partial differential operators $\partial/\partial x$, $\partial/\partial y$ and $\partial/\partial z$ change no differently under reorientation than the component amplitudes a_x, a_y and a_z do. Hence,

$$\nabla = \hat{\mathbf{\imath}}\frac{\partial}{\partial i} = \hat{\mathbf{x}}'\frac{\partial}{\partial x'} + \hat{\mathbf{y}}'\frac{\partial}{\partial y'} + \hat{\mathbf{z}}'\frac{\partial}{\partial z'}, \tag{16.2}$$

evidently the same operator regardless of the choice of basis $[\hat{\mathbf{x}}'\ \hat{\mathbf{y}}'\ \hat{\mathbf{z}}']$. It is this invariance under reorientation that makes the ∇ operator useful.

If ∇ takes the place of the ambiguous $d/d\mathbf{r}$, then what takes the place of the ambiguous $d/d\mathbf{r}_o$, $d/d\tilde{\mathbf{r}}$, $d/d\mathbf{r}^\dagger$, $d/d\mathbf{r}'$ and so on? Answer: ∇_o, $\tilde{\nabla}$, $\nabla^\dagger$, ∇' and so on. Whatever mark distinguishes the special **r**, the same mark distinguishes the corresponding special ∇. For example, where $\mathbf{r}_o = \hat{\mathbf{\imath}}i_o$, there $\nabla_o = \hat{\mathbf{\imath}}\,\partial/\partial i_o$. That is the convention.[4]

Introduced by William Rowan Hamilton and Oliver Heaviside, informally pronounced "del" (in the author's country at least), the vector differential operator ∇ finds extensive use in the modeling of physical phenomena. After

[4] A few readers not fully conversant with the material of chapter 15, to whom this chapter had been making sense until the last two sentences, may suddenly find the notation

a brief digression to discuss operator notation, the subsections that follow will use the operator to develop and present the four basic kinds of vector derivative.

16.1.2 Operator notation

Section 16.1.1 has introduced *operator notation* without explaining what it is or even what it concerns. This subsection digresses to explain.

Operator notation concerns the representation of unary operators and the operations they specify. Section 7.3 has already broadly introduced the notion of the operator. A *unary operator* is a mathematical agent that transforms a single discrete quantity, a single distributed quantity, a single field, a single function or another single mathematical object in some definite way. For example, $J \equiv \int_0^t dt$ is a unary operator, more fully written as $J \equiv \int_0^t \cdot\, dt$ where the "$\cdot$" holds the place of the thing operated upon,[5] whose effect is such that, for instance, $Jt = t^2/2$ and $J \cos \omega t = (\sin \omega t)/\omega$. Any letter might serve as well as the example's J; but what distinguishes operator notation is that, like the matrix row operator A in matrix notation's product $A\mathbf{x}$ (§ 11.1.1), the operator J in operator notation's operation Jt attacks from the left. Thus, generally, $Jt \neq tJ$ if J is a unary operator, though the notation Jt usually formally resembles multiplication in other respects as we shall see.

The matrix actually is a type of unary operator and matrix notation is a specialization of operator notation, so we have met operator notation before. And, in fact, we have met operator notation much earlier than that. The product $5t$ can if you like be regarded as the unary operator "5 times," operating on t. Perhaps you did not know that 5 was an operator—and, indeed, the scalar 5 itself is no operator but just a number—but where no other operation is defined operator notation implies scalar multiplication by default. Seen in this way, $5t$ and $t5$ actually mean two different things; though naturally in the specific case of scalar multiplication, which happens to be commutative, it is true that $5t = t5$.

The $\mathbf{a}\cdot$ in the dot product $\mathbf{a}\cdot\mathbf{b}$ and the $\mathbf{a}\times$ in the cross product $\mathbf{a}\times\mathbf{b}$ can profitably be regarded as unary operators.

incomprehensible. The notation is Einstein's. It means

$$\hat{\imath} i_o = \sum_{i=x',y',z'} \hat{\imath} i_o = \hat{\mathbf{x}}' x'_o + \hat{\mathbf{y}}' y'_o + \hat{\mathbf{z}}' z'_o,$$

in the leftmost form of which the summation sign is implied not written. Refer to § 15.4.

[5] [30]

Whether operator notation can licitly represent any unary operation whatsoever is a definitional question we will leave for the professional mathematician to answer, but in normal usage operator notation represents only *linear unary operations,* unary operations that honor § 7.3.3's rule of linearity. The operators J and A above are examples of linear unary operators; the operator $K \equiv \cdot + 3$ is not linear and almost certainly should never be represented in operator notation as here, lest an expression like Kt mislead an understandably unsuspecting audience. Linear unary operators often do not commute, so $J_1 J_2 \neq J_2 J_1$ generally; but otherwise linear unary operators follow familiar rules of multiplication like $(J_2 + J_3)J_1 = J_2 J_1 + J_3 J_1$. Linear unary operators obey a definite algebra, the same algebra matrices obey. It is precisely this algebra that makes operator notation so useful.

Operators associate from right to left (§ 2.1.1) so that, in operator notation, $J\omega t = J(\omega t)$, not $(J\omega)t$. Observe however that the perceived need for parentheses comes only of the purely notational ambiguity as to whether ωt bears the semantics of "the product of ω and t" or those of "the unary operator 'ω times,' operating on t." The perceived need and any associated confusion would vanish if $\Omega \equiv (\omega)(\cdot)$ were unambiguously an operator, in which case the product $J\Omega$ would itself be an operator, whereupon $(J\Omega)t = J\Omega t = J(\Omega t) = J(\omega t)$. Indeed, one can compare the distinction in § 11.3.2 between λ and λI against the distinction between ω and Ω here, for a linear unary operator enjoys the same associativity (11.5) a matrix enjoys, and for the same reason. Still, rather than go to the trouble of defining extra symbols like Ω, it is usually easier just to write the parentheses, which take little space on the page and are universally understood; or, better, to rely on the right-to-left convention that $J\omega t = J(\omega t)$. Modern conventional applied mathematical notation though generally excellent remains imperfect; so, notationally, when it matters, operators associate from right to left except where parentheses group otherwise.

One can speak of a unary operator like J, A or Ω without giving it anything in particular to operate upon. One can leave an operation unresolved. For example, tJ is itself a unary operator—it is the operator $t\int_0^t dt$—though one can assign no particular value to it until it actually operates on something. The operator ∇ of (16.2) is an unresolved unary operator of the same kind.

16.1.3 The directional derivative and the gradient

In the calculus of vector fields, the derivative notation $d/d\mathbf{r}$ is ambiguous because, as the section's introduction has observed, the notation gives $\mathbf{r}$

no specific direction in which to advance. In operator notation, however, given (16.2) and accorded a reference vector **b** to supply a direction and a scale, one can compose the *directional derivative* operator

$$(\mathbf{b}\cdot\nabla) = b_i\frac{\partial}{\partial i} \tag{16.3}$$

to express the derivative unambiguously. This operator applies equally to the scalar field,

$$(\mathbf{b}\cdot\nabla)\psi(\mathbf{r}) = b_i\frac{\partial\psi}{\partial i},$$

as to the vector field,

$$(\mathbf{b}\cdot\nabla)\mathbf{a}(\mathbf{r}) = b_i\frac{\partial\mathbf{a}}{\partial i} = \hat{\mathbf{j}}b_i\frac{\partial a_j}{\partial i}. \tag{16.4}$$

For the scalar field the parentheses are unnecessary and conventionally are omitted, as

$$\mathbf{b}\cdot\nabla\psi(\mathbf{r}) = b_i\frac{\partial\psi}{\partial i}. \tag{16.5}$$

In the case (16.4) of the vector field, however, $\nabla\mathbf{a}(\mathbf{r})$ itself means nothing coherent[6] so the parentheses usually are retained. Equations (16.4) and (16.5) define the directional derivative.

Note that the directional derivative is the derivative not of the reference vector **b** but only of the field $\psi(\mathbf{r})$ or $\mathbf{a}(\mathbf{r})$. The vector **b** just directs and scales the derivative; it is not the object of it. Nothing requires **b** to be constant, though. It can be a vector field $\mathbf{b}(\mathbf{r})$ that varies from place to place; the directional derivative does not care.

Within (16.5), the quantity

$$\nabla\psi(\mathbf{r}) = \hat{\imath}\frac{\partial\psi}{\partial i} \tag{16.6}$$

is called the *gradient* of the scalar field $\psi(\mathbf{r})$. Though both scalar and vector fields have directional derivatives, only scalar fields have gradients. The gradient represents the amplitude and direction of a scalar field's locally steepest rate.

Formally a dot product, the directional derivative operator $\mathbf{b}\cdot\nabla$ is invariant under reorientation of axes, whereupon the directional derivative is invariant, too. The result of a ∇ operation, the gradient $\nabla\psi(\mathbf{r})$ is likewise invariant.

[6] Well, it does mean something coherent in *dyadic analysis* [24, appendix B], but this book won't treat that.

16.1.4 Divergence

There exist other vector derivatives than the directional derivative and gradient of § 16.1.3. One of these is divergence. It is not easy to motivate divergence directly, however, so we will approach it indirectly, through the concept of flux as follows.

The *flux* of a vector field $\mathbf{a}(\mathbf{r})$ outward from a region in space is

$$\Phi \equiv \oint_S \mathbf{a}(\mathbf{r}) \cdot d\mathbf{s}, \tag{16.7}$$

where

$$d\mathbf{s} \equiv \hat{\mathbf{n}} \cdot ds \tag{16.8}$$

is a vector infinitesimal of amplitude ds, directed normally outward from the closed surface bounding the region—ds being the area of an infinitesimal element of the surface, the area of a tiny patch. Flux is flow through a surface: in this case, net flow outward from the region in question. (The paragraph says much in relatively few words. If it seems opaque then try to visualize eqn. 16.7's dot product $\mathbf{a}[\mathbf{r}] \cdot d\mathbf{s}$, in which the vector $d\mathbf{s}$ represents the area and orientation of a patch of the region's enclosing surface. When something like air flows through any surface—not necessarily a physical barrier but an imaginary surface like the goal line's vertical plane in a football game[7]—what matters is not the surface's area as such but rather the area the surface presents to the flow. The surface presents its full area to a perpendicular flow but otherwise the flow sees a foreshortened surface, *as though the surface were projected onto a plane perpendicular to the flow.* Refer to Fig. 15.2. Now realize that eqn. 16.7 actually describes flux not through an open surface but through a closed—it could be the imaginary rectangular box enclosing the region of football play to goal-post height; where wind blowing through the region, entering and leaving, would constitute zero net flux; but where a positive net flux would have barometric pressure falling and air leaving the region maybe because a storm is coming—and you've got the idea.)

A region of positive flux is a *source;* of negative flux, a *sink.* One can contemplate the flux $\Phi_{\text{open}} = \int_S \mathbf{a}(\mathbf{r}) \cdot d\mathbf{s}$ through an open surface as well as through a closed, but it is the outward flux (16.7) through a closed surface that will concern us here.

[7] The author has American football in mind but other football games have goal lines and goal posts, too. Pick your favorite brand of football.

The outward flux Φ of a vector field $\mathbf{a}(\mathbf{r})$ through a closed surface bounding some definite region in space is evidently

$$\Phi = \iint \Delta a_x(y,z)\,dy\,dz + \iint \Delta a_y(z,x)\,dz\,dx + \iint \Delta a_z(x,y)\,dx\,dy,$$

where

$$\begin{aligned}
\Delta a_x(y,z) &= \int_{x_{\mathrm{min}}(y,z)}^{x_{\mathrm{max}}(y,z)} \frac{\partial a_x}{\partial x}\,dx, \\
\Delta a_y(z,x) &= \int_{y_{\mathrm{min}}(z,x)}^{y_{\mathrm{max}}(z,x)} \frac{\partial a_y}{\partial y}\,dy, \\
\Delta a_z(x,y) &= \int_{z_{\mathrm{min}}(x,y)}^{z_{\mathrm{max}}(x,y)} \frac{\partial a_z}{\partial z}\,dz
\end{aligned}$$

represent the increase across the region respectively of a_x, a_y or a_z along an $\hat{\mathbf{x}}$-, $\hat{\mathbf{y}}$- or $\hat{\mathbf{z}}$-directed line.[8] If the field has constant derivatives $\partial\mathbf{a}/\partial i$, or equivalently if the region in question is small enough that the derivatives are practically constant through it, then these increases are simply

$$\begin{aligned}
\Delta a_x(y,z) &= \frac{\partial a_x}{\partial x}\,\Delta x(y,z), \\
\Delta a_y(z,x) &= \frac{\partial a_y}{\partial y}\,\Delta y(z,x), \\
\Delta a_z(x,y) &= \frac{\partial a_z}{\partial z}\,\Delta z(x,y),
\end{aligned}$$

upon which

$$\begin{aligned}
\Phi \;=\;& \frac{\partial a_x}{\partial x}\iint \Delta x(y,z)\,dy\,dz + \frac{\partial a_y}{\partial y}\iint \Delta y(z,x)\,dz\,dx \\
&+ \frac{\partial a_z}{\partial z}\iint \Delta z(x,y)\,dx\,dy.
\end{aligned}$$

But each of the last equation's three integrals represents the region's volume V, so

$$\Phi = (V)\left(\frac{\partial a_x}{\partial x} + \frac{\partial a_y}{\partial y} + \frac{\partial a_z}{\partial z}\right);$$

[8] Naturally, if the region's boundary happens to be concave, then some lines might enter and exit the region more than once, but this merely elaborates the limits of integration along those lines. It changes the problem in no essential way.

or, dividing through by the volume,

$$\frac{\Phi}{V} = \frac{\partial a_x}{\partial x} + \frac{\partial a_y}{\partial y} + \frac{\partial a_z}{\partial z} = \frac{\partial a_i}{\partial i} = \nabla \cdot \mathbf{a}(\mathbf{r}). \tag{16.9}$$

We give this ratio of outward flux to volume,

$$\nabla \cdot \mathbf{a}(\mathbf{r}) = \frac{\partial a_i}{\partial i}, \tag{16.10}$$

the name *divergence,* representing the intensity of a source or sink.

Formally a dot product, divergence is invariant under reorientation of axes.

16.1.5 Curl

Curl is to divergence as the cross product is to the dot product. Curl is a little trickier to visualize, though. It needs first the concept of circulation as follows.

The *circulation* of a vector field $\mathbf{a}(\mathbf{r})$ about a closed contour in space is

$$\Gamma \equiv \oint \mathbf{a}(\mathbf{r}) \cdot d\boldsymbol{\ell}, \tag{16.11}$$

where, unlike the $\oint_S$ of (16.7) which represented a double integration over a surface, the $\oint$ here represents only a single integration. One can in general contemplate circulation about any closed contour, but it suits our purpose here to consider specifically a closed contour that happens not to depart from a single, flat plane in space.

Let $[\hat{\mathbf{u}}\ \hat{\mathbf{v}}\ \hat{\mathbf{n}}]$ be an orthogonal basis with $\hat{\mathbf{n}}$ normal to the contour's plane such that travel positively along the contour tends from $\hat{\mathbf{u}}$ toward $\hat{\mathbf{v}}$ rather than the reverse. The circulation Γ of a vector field $\mathbf{a}(\mathbf{r})$ about this contour is evidently

$$\Gamma = \int \Delta a_v(v)\, dv - \int \Delta a_u(u)\, du,$$

where

$$\begin{aligned} \Delta a_v(v) &= \int_{u_{\min}(v)}^{u_{\max}(v)} \frac{\partial a_v}{\partial u}\, du, \\ \Delta a_u(u) &= \int_{v_{\min}(u)}^{v_{\max}(u)} \frac{\partial a_u}{\partial v}\, dv \end{aligned}$$

represent the increase across the contour's interior respectively of a_v or a_u along a $\hat{\mathbf{u}}$- or $\hat{\mathbf{v}}$-directed line. If the field has constant derivatives $\partial\mathbf{a}/\partial i$, or equivalently if the contour in question is short enough that the derivatives are practically constant over it, then these increases are simply

$$\begin{aligned}\Delta a_v(v) &= \frac{\partial a_v}{\partial u}\,\Delta u(v),\\ \Delta a_u(u) &= \frac{\partial a_u}{\partial v}\,\Delta v(u),\end{aligned}$$

upon which

$$\Gamma = \frac{\partial a_v}{\partial u}\int \Delta u(v)\,dv - \frac{\partial a_u}{\partial v}\int \Delta v(u)\,du.$$

But each of the last equation's two integrals represents the area A within the contour, so

$$\Gamma = (A)\left(\frac{\partial a_v}{\partial u} - \frac{\partial a_u}{\partial v}\right);$$

or, dividing through by the area,

$$\begin{aligned}\frac{\Gamma}{A} &= \frac{\partial a_v}{\partial u} - \frac{\partial a_u}{\partial v}\\ &= \hat{\mathbf{n}}\cdot\left[\epsilon_{ijk}\hat{\mathbf{i}}\frac{\partial a_k}{\partial j}\right] = \hat{\mathbf{n}}\cdot\begin{vmatrix}\hat{\mathbf{u}} & \hat{\mathbf{v}} & \hat{\mathbf{n}}\\ \partial/\partial u & \partial/\partial v & \partial/\partial n\\ a_u & a_v & a_n\end{vmatrix}\\ &= \hat{\mathbf{n}}\cdot[\nabla\times\mathbf{a}(\mathbf{r})].\end{aligned} \tag{16.12}$$

We give this ratio of circulation to area,

$$\hat{\mathbf{n}}\cdot[\nabla\times\mathbf{a}(\mathbf{r})] = \hat{\mathbf{n}}\cdot\left[\epsilon_{ijk}\hat{\mathbf{i}}\frac{\partial a_k}{\partial j}\right] = \frac{\partial a_v}{\partial u} - \frac{\partial a_u}{\partial v}, \tag{16.13}$$

the name *directional curl,* representing the intensity of circulation, the degree of twist so to speak, about a specified axis. The cross product in (16.13),

$$\nabla\times\mathbf{a}(\mathbf{r}) = \epsilon_{ijk}\hat{\mathbf{i}}\frac{\partial a_k}{\partial j}, \tag{16.14}$$

we call *curl.*

Curl (16.14) is an interesting quantity. Although it emerges from directional curl (16.13) and although we have developed directional curl with respect to a contour in some specified plane, curl (16.14) itself turns out to be altogether independent of any particular plane. We might have chosen another plane and though $\hat{\mathbf{n}}$ would then be different the same (16.14) would necessarily result. Directional curl, a scalar, is a property of the field and the plane. Curl, a vector, unexpectedly is a property of the field only. Directional curl evidently cannot exceed curl in magnitude, but will equal it when $\hat{\mathbf{n}}$ points in its direction, so it may be said that curl is the locally greatest directional curl, oriented normally to the locally greatest directional curl's plane.

We have needed $\hat{\mathbf{n}}$ and (16.13) to motivate and develop the concept (16.14) of curl. Once developed, however, the concept of curl stands on its own, whereupon one can return to define directional curl more generally than (16.13) has defined it. As in (16.4) here too any reference vector $\mathbf{b}$ or vector field $\mathbf{b}(\mathbf{r})$ can serve to direct the curl, not only $\hat{\mathbf{n}}$. Hence,

$$\mathbf{b}\cdot[\nabla\times\mathbf{a}(\mathbf{r})] = \mathbf{b}\cdot\left[\epsilon_{ijk}\hat{\mathbf{\imath}}\frac{\partial a_k}{\partial j}\right] = \epsilon_{ijk}b_i\frac{\partial a_k}{\partial j}. \tag{16.15}$$

This would be the actual definition of *directional curl.* Note however that directional curl so defined is not a distinct kind of derivative but rather is just curl, dot-multiplied by a reference vector.

Formally a cross product, curl is invariant under reorientation of axes. An ordinary dot product, directional curl is likewise invariant.

16.1.6 Cross-directional derivatives

The several directional derivatives of the $\mathbf{b}\cdot\nabla$ class, including the scalar (16.5) and vector (16.4) directional derivatives themselves and also including directional curl (16.15), compute rates with reference to some direction $\mathbf{b}$. Another class of directional derivatives however is possible, that of the *cross-directional derivatives.*[9] These compute rates perpendicularly to $\mathbf{b}$. Unlike the vector directional derivative (16.4), the cross-directional derivatives are not actually new derivatives but are cross products of $\mathbf{b}$ with derivatives already familiar to us. The cross-directional derivatives are

$$\begin{aligned}\mathbf{b}\times\nabla\psi &= \epsilon_{ijk}\hat{\mathbf{\imath}}b_j\frac{\partial\psi}{\partial k},\\ \mathbf{b}\times\nabla\times\mathbf{a} &= \hat{\mathbf{\jmath}}b_i\left(\frac{\partial a_i}{\partial j}-\frac{\partial a_j}{\partial i}\right).\end{aligned} \tag{16.16}$$

[9] The author is unaware of a conventional name for these derivatives. The name *cross-directional* seems as apt as any.

We call these respectively the *cross-directional derivative* (itself) and *cross-directional curl,* the latter derived as

$$\begin{aligned}
\mathbf{b}\times\nabla\times\mathbf{a} &= \mathbf{b}\times\left(\epsilon_{ijk}\hat{\mathbf{i}}\frac{\partial a_k}{\partial j}\right) \\
&= \epsilon_{mni}\hat{\mathbf{m}}b_n\left(\epsilon_{ijk}\hat{\mathbf{i}}\frac{\partial a_k}{\partial j}\right)_i = \epsilon_{mni}\epsilon_{ijk}\hat{\mathbf{m}}b_n\frac{\partial a_k}{\partial j} \\
&= (\delta_{mj}\delta_{nk}-\delta_{mk}\delta_{nj})\hat{\mathbf{m}}b_n\frac{\partial a_k}{\partial j} \\
&= \hat{\mathbf{j}}b_k\frac{\partial a_k}{\partial j}-\hat{\mathbf{k}}b_j\frac{\partial a_k}{\partial j} = \hat{\mathbf{j}}b_i\frac{\partial a_i}{\partial j}-\hat{\mathbf{j}}b_i\frac{\partial a_j}{\partial i}
\end{aligned}$$

where the Levi-Civita identity that $\epsilon_{mni}\epsilon_{ijk} = \epsilon_{imn}\epsilon_{ijk} = \delta_{mj}\delta_{nk} - \delta_{mk}\delta_{nj}$ comes from Table 15.1.

16.2 Integral forms

The vector field's distinctive maneuver is the shift between integral forms, which we are now prepared to treat. This shift comes in two kinds. The two subsections that follow explain.

16.2.1 The divergence theorem

Section 16.1.4 has contemplated the flux of a vector field $\mathbf{a}(\mathbf{r})$ from a volume small enough that the divergence $\nabla\cdot\mathbf{a}$ were practically constant through the volume. One would like to treat the flux from larger, more general volumes as well. According to the definition (16.7), the flux from any volume is

$$\Phi = \oint_S \mathbf{a}\cdot d\mathbf{s}.$$

If one subdivides a large volume into infinitesimal volume elements dv, then the flux from a single volume element is

$$\Phi_{\text{element}} = \oint_{S_{\text{element}}} \mathbf{a}\cdot d\mathbf{s}.$$

Even a single volume element however can have two distinct kinds of surface area: inner surface area shared with another element; and outer surface area shared with no other element because it belongs to the surface of the larger, overall volume. Interior elements naturally have only the former kind but boundary elements have both kinds of surface area, so one can elaborate the last equation to read

$$\Phi_{\text{element}} = \int_{S_{\text{inner}}} \mathbf{a} \cdot d\mathbf{s} + \int_{S_{\text{outer}}} \mathbf{a} \cdot d\mathbf{s}$$

for a single element, where $\oint_{S_{\text{element}}} = \int_{S_{\text{inner}}} + \int_{S_{\text{outer}}}$. Adding all the elements together, we have that

$$\sum_{\text{elements}} \Phi_{\text{element}} = \sum_{\text{elements}} \int_{S_{\text{inner}}} \mathbf{a} \cdot d\mathbf{s} + \sum_{\text{elements}} \int_{S_{\text{outer}}} \mathbf{a} \cdot d\mathbf{s};$$

but the inner sum is null because it includes each interior surface twice, because each interior surface is shared by two elements such that $d\mathbf{s}_2 = -d\mathbf{s}_1$ (in other words, such that the one volume element's $d\mathbf{s}$ on the surface the two elements share points oppositely to the other volume element's $d\mathbf{s}$ on the same surface), so

$$\sum_{\text{elements}} \Phi_{\text{element}} = \sum_{\text{elements}} \int_{S_{\text{outer}}} \mathbf{a} \cdot d\mathbf{s} = \oint_S \mathbf{a} \cdot d\mathbf{s}.$$

In this equation, the last integration is over the surface of the larger, overall volume, which surface after all consists of nothing other than the several boundary elements' outer surface patches. Applying (16.9) to the equation's left side to express the flux Φ_{element} from a single volume element yields that

$$\sum_{\text{elements}} \nabla \cdot \mathbf{a}\, dv = \oint_S \mathbf{a} \cdot d\mathbf{s}.$$

That is,

$$\int_V \nabla \cdot \mathbf{a}\, dv = \oint_S \mathbf{a} \cdot d\mathbf{s}. \tag{16.17}$$

Equation (16.17) is the *divergence theorem.*[10] The divergence theorem, the vector's version of the fundamental theorem of calculus (7.2), neatly relates the divergence within a volume to the flux from it. It is an important result. The integral on the equation's left and the one on its right each arise in vector analysis more often than one might expect. When they do, (16.17) swaps the one integral for the other, often a profitable maneuver.[11]

[10] [157, eqn. 1.2.8]

[11] Where a wave propagates through a material interface, the associated field can be discontinuous and, consequently, the field's divergence can be infinite, which would seem

16.2.2 Stokes' theorem

Corresponding to the divergence theorem of § 16.2.1 is a second, related theorem for directional curl, developed as follows. If an open surface, whether the surface be confined to a plane or be warped in three dimensions (as for example in bowl shape), is subdivided into infinitesimal surface elements $d\mathbf{s}$—each element small enough not only to experience essentially constant curl but also to be regarded as planar—then according to (16.11) the circulation about the entire surface is

$$\Gamma = \oint \mathbf{a} \cdot d\boldsymbol{\ell}$$

and the circulation about any one surface element is

$$\Gamma_{\text{element}} = \oint_{\text{element}} \mathbf{a} \cdot d\boldsymbol{\ell}.$$

From this equation, reasoning parallel to that of § 16.2.1—only using (16.12) in place of (16.9)—concludes that

$$\int_S (\nabla \times \mathbf{a}) \cdot d\mathbf{s} = \oint \mathbf{a} \cdot d\boldsymbol{\ell}. \tag{16.18}$$

Equation (16.18) is *Stokes' theorem,*[12,13] neatly relating the directional curl over a (possibly nonplanar) surface to the circulation about it. Like the divergence theorem (16.17), Stokes' theorem (16.18) serves to swap one vector integral for another where such a maneuver is needed.

16.3 Summary of definitions and identities of vector calculus

Table 16.1 lists useful definitions and identities of vector calculus,[14] the first several of which it gathers from §§ 16.1 and 16.2, the last several of

to call assumptions underlying (16.17) into question. However, the infinite divergence at a material interface is normally *integrable* in the same way the Dirac delta of § 7.7, though infinite, is integrable. One can integrate finitely through either infinity. If one can conceive of an interface not as a sharp layer of zero thickness but rather as a thin layer of thickness ϵ, through which the associated field varies steeply but continuously, then the divergence theorem necessarily remains valid in the limit $\epsilon \to 0$.

[12] [157, eqn. 1.4.20]

[13] If (16.17) is "the divergence theorem," then should (16.18) not be "the curl theorem"? Answer: maybe it should be, but no one calls it that. Sir George Gabriel Stokes is evidently not to be denied his fame!

[14] [11, appendix II.3][157, appendix II][72, appendix A]

which (exhibiting heretofore unfamiliar symbols like ∇^2) it gathers from § 16.4 to follow. Of the identities in the middle of the table, a few are statements of the ∇ operator's distributivity over summation. The rest are vector derivative product rules (§ 4.5.2).

The product rules resemble the triple products of Table 15.2, only with the ∇ operator in place of the vector $\mathbf{c}$. However, since ∇ is a differential operator for which, for instance, $\mathbf{b}\cdot\nabla \neq \nabla\cdot\mathbf{b}$, its action differs from a vector's in some cases, and there are more distinct ways in which it can act. Among the several product rules the easiest to prove is that

$$\nabla(\psi\omega) = \hat{\mathbf{i}}\frac{\partial(\psi\omega)}{\partial i} = \omega\hat{\mathbf{i}}\frac{\partial\psi}{\partial i} + \psi\hat{\mathbf{i}}\frac{\partial\omega}{\partial i} = \omega\nabla\psi + \psi\nabla\omega.$$

The hardest to prove is that

$$\begin{aligned}
\nabla(\mathbf{a}\cdot\mathbf{b}) &= \nabla(a_ib_i) = \hat{\mathbf{j}}\frac{\partial(a_ib_i)}{\partial j} = \hat{\mathbf{j}}b_i\frac{\partial a_i}{\partial j} + \hat{\mathbf{j}}a_i\frac{\partial b_i}{\partial j}\\
&= \hat{\mathbf{j}}b_i\frac{\partial a_j}{\partial i} + \hat{\mathbf{j}}b_i\left(\frac{\partial a_i}{\partial j} - \frac{\partial a_j}{\partial i}\right) + \hat{\mathbf{j}}a_i\frac{\partial b_j}{\partial i} + \hat{\mathbf{j}}a_i\left(\frac{\partial b_i}{\partial j} - \frac{\partial b_j}{\partial i}\right)\\
&= (\mathbf{b}\cdot\nabla)\mathbf{a} + \mathbf{b}\times\nabla\times\mathbf{a} + (\mathbf{a}\cdot\nabla)\mathbf{b} + \mathbf{a}\times\nabla\times\mathbf{b}\\
&= (\mathbf{b}\cdot\nabla + \mathbf{b}\times\nabla\times\,)\mathbf{a} + (\mathbf{a}\cdot\nabla + \mathbf{a}\times\nabla\times\,)\mathbf{b},
\end{aligned}$$

because to prove it one must recognize in it the cross-directional curl of (16.16). Also nontrivial to prove is that

$$\begin{aligned}
\nabla\times(\mathbf{a}\times\mathbf{b}) &= \nabla\times(\epsilon_{ijk}\hat{\mathbf{i}}a_jb_k)\\
&= \epsilon_{mni}\hat{\mathbf{m}}\frac{\partial(\epsilon_{ijk}\hat{\mathbf{i}}a_jb_k)_i}{\partial n} = \epsilon_{mni}\epsilon_{ijk}\hat{\mathbf{m}}\frac{\partial(a_jb_k)}{\partial n}\\
&= (\delta_{mj}\delta_{nk} - \delta_{mk}\delta_{nj})\hat{\mathbf{m}}\frac{\partial(a_jb_k)}{\partial n}\\
&= \hat{\mathbf{j}}\frac{\partial(a_jb_k)}{\partial k} - \hat{\mathbf{k}}\frac{\partial(a_jb_k)}{\partial j} = \hat{\mathbf{j}}\frac{\partial(a_jb_i)}{\partial i} - \hat{\mathbf{j}}\frac{\partial(a_ib_j)}{\partial i}\\
&= \left(\hat{\mathbf{j}}b_i\frac{\partial a_j}{\partial i} + \hat{\mathbf{j}}a_j\frac{\partial b_i}{\partial i}\right) - \left(\hat{\mathbf{j}}a_i\frac{\partial b_j}{\partial i} + \hat{\mathbf{j}}b_j\frac{\partial a_i}{\partial i}\right)\\
&= (\mathbf{b}\cdot\nabla + \nabla\cdot\mathbf{b})\mathbf{a} - (\mathbf{a}\cdot\nabla + \nabla\cdot\mathbf{a})\mathbf{b}.
\end{aligned}$$

Table 16.1: Definitions and identities of vector calculus (see also Table 15.2 on page 520).

$$
\begin{array}{rclrcl}
\nabla & \equiv & \hat{\mathbf{i}}\dfrac{\partial}{\partial i} & \mathbf{b}\cdot\nabla & = & b_i\dfrac{\partial}{\partial i} \\
\nabla\psi & = & \hat{\mathbf{i}}\dfrac{\partial\psi}{\partial i} & \mathbf{b}\cdot\nabla\psi & = & b_i\dfrac{\partial\psi}{\partial i} \\
\nabla\cdot\mathbf{a} & = & \dfrac{\partial a_i}{\partial i} & (\mathbf{b}\cdot\nabla)\mathbf{a} & = & b_i\dfrac{\partial\mathbf{a}}{\partial i} = \hat{\mathbf{j}}b_i\dfrac{\partial a_j}{\partial i} \\
\nabla\times\mathbf{a} & = & \epsilon_{ijk}\hat{\mathbf{i}}\dfrac{\partial a_k}{\partial j} & \mathbf{b}\cdot\nabla\times\mathbf{a} & = & \epsilon_{ijk}b_i\dfrac{\partial a_k}{\partial j} \\
\mathbf{b}\times\nabla\psi & = & \epsilon_{ijk}\hat{\mathbf{i}}b_j\dfrac{\partial\psi}{\partial k} & \mathbf{b}\times\nabla\times\mathbf{a} & = & \hat{\mathbf{j}}b_i\left(\dfrac{\partial a_i}{\partial j}-\dfrac{\partial a_j}{\partial i}\right) \\
\Phi & \equiv & \displaystyle\int_S \mathbf{a}\cdot d\mathbf{s} & \displaystyle\int_V \nabla\cdot\mathbf{a}\,dv & = & \displaystyle\oint_S \mathbf{a}\cdot d\mathbf{s} \\
\Gamma & \equiv & \displaystyle\int_C \mathbf{a}\cdot d\boldsymbol{\ell} & \displaystyle\int_S (\nabla\times\mathbf{a})\cdot d\mathbf{s} & = & \displaystyle\oint \mathbf{a}\cdot d\boldsymbol{\ell}
\end{array}
$$

$$
\begin{array}{rcl}
\nabla\cdot(\mathbf{a}+\mathbf{b}) & = & \nabla\cdot\mathbf{a}+\nabla\cdot\mathbf{b} \\
\nabla\times(\mathbf{a}+\mathbf{b}) & = & \nabla\times\mathbf{a}+\nabla\times\mathbf{b} \\
\nabla(\psi+\omega) & = & \nabla\psi+\nabla\omega \\
\nabla(\psi\omega) & = & \omega\nabla\psi+\psi\nabla\omega \\
\nabla\cdot(\psi\mathbf{a}) & = & \mathbf{a}\cdot\nabla\psi+\psi\nabla\cdot\mathbf{a} \\
\nabla\times(\psi\mathbf{a}) & = & \psi\nabla\times\mathbf{a}-\mathbf{a}\times\nabla\psi \\
\nabla(\mathbf{a}\cdot\mathbf{b}) & = & (\mathbf{b}\cdot\nabla+\mathbf{b}\times\nabla\times\,)\mathbf{a}+(\mathbf{a}\cdot\nabla+\mathbf{a}\times\nabla\times\,)\mathbf{b} \\
\nabla\cdot(\mathbf{a}\times\mathbf{b}) & = & \mathbf{b}\cdot\nabla\times\mathbf{a}-\mathbf{a}\cdot\nabla\times\mathbf{b} \\
\nabla\times(\mathbf{a}\times\mathbf{b}) & = & (\mathbf{b}\cdot\nabla+\nabla\cdot\mathbf{b})\mathbf{a}-(\mathbf{a}\cdot\nabla+\nabla\cdot\mathbf{a})\mathbf{b}
\end{array}
$$

$$
\begin{array}{rclrcl}
\nabla^2 & \equiv & \dfrac{\partial^2}{\partial i^2} & \nabla\nabla\cdot\mathbf{a} & = & \hat{\mathbf{j}}\dfrac{\partial^2 a_i}{\partial j\,\partial i} \\
\nabla^2\psi & = & \nabla\cdot\nabla\psi=\dfrac{\partial^2\psi}{\partial i^2} & \nabla^2\mathbf{a} & = & \dfrac{\partial^2\mathbf{a}}{\partial i^2}=\hat{\mathbf{j}}\dfrac{\partial^2 a_j}{\partial i^2}=\hat{\mathbf{j}}\nabla^2(\hat{\mathbf{j}}\cdot\mathbf{a}) \\
\nabla\times\nabla\psi & = & 0 & \nabla\cdot\nabla\times\mathbf{a} & = & 0
\end{array}
$$

$$
\begin{array}{rcl}
\nabla\times\nabla\times\mathbf{a} & = & \hat{\mathbf{j}}\dfrac{\partial}{\partial i}\left(\dfrac{\partial a_i}{\partial j}-\dfrac{\partial a_j}{\partial i}\right) \\
\nabla\nabla\cdot\mathbf{a} & = & \nabla^2\mathbf{a}+\nabla\times\nabla\times\mathbf{a}
\end{array}
$$

The others are less hard:[15]

$$\begin{aligned}
\nabla\cdot(\psi\mathbf{a}) &= \frac{\partial(\psi a_i)}{\partial i} = a_i\frac{\partial\psi}{\partial i} + \psi\frac{\partial a_i}{\partial i} = \mathbf{a}\cdot\nabla\psi + \psi\nabla\cdot\mathbf{a};\\
\nabla\times(\psi\mathbf{a}) &= \epsilon_{ijk}\hat{\mathbf{\imath}}\frac{\partial(\psi a_k)}{\partial j} = \epsilon_{ijk}\hat{\mathbf{\imath}}\psi\frac{\partial a_k}{\partial j} + \epsilon_{ijk}\hat{\mathbf{\imath}}a_k\frac{\partial\psi}{\partial j}\\
&= \psi\nabla\times\mathbf{a} - \mathbf{a}\times\nabla\psi;\\
\nabla\cdot(\mathbf{a}\times\mathbf{b}) &= \frac{\partial(\epsilon_{ijk}a_jb_k)}{\partial i} = \epsilon_{ijk}b_k\frac{\partial a_j}{\partial i} + \epsilon_{ijk}a_j\frac{\partial b_k}{\partial i}\\
&= \mathbf{b}\cdot\nabla\times\mathbf{a} - \mathbf{a}\cdot\nabla\times\mathbf{b}.
\end{aligned}$$

Inasmuch as none of the derivatives or products within the table's several product rules vary under rotation of axes, the product rules are themselves invariant. That the definitions and identities at the top of the table are invariant, we have already seen; and § 16.4, next, will give invariance to the definitions and identities at the bottom. The whole table is therefore invariant under rotation of axes.

16.4 The Laplacian and other second-order derivatives

Table 16.1 ends with second-order vector derivatives. Like vector products and first-order vector derivatives, second-order vector derivatives too come in several kinds, the simplest of which is the *Laplacian*[16]

$$\begin{gathered}
\nabla^2 \equiv \frac{\partial^2}{\partial i^2},\\
\nabla^2\psi = \nabla\cdot\nabla\psi = \frac{\partial^2\psi}{\partial i^2},\\
\nabla^2\mathbf{a} = \frac{\partial^2\mathbf{a}}{\partial i^2} = \hat{\mathbf{\jmath}}\frac{\partial^2 a_j}{\partial i^2} = \hat{\mathbf{\jmath}}\nabla^2(\hat{\mathbf{\jmath}}\cdot\mathbf{a}).
\end{gathered} \tag{16.19}$$

[15] And probably should have been left as exercises, except that this book is not actually an instructional textbook. The reader who wants exercises might hide the page from sight and work the three identities out with his own pencil.

[16] Though seldom seen in applied usage in the author's country, the alternate symbol Δ replaces ∇^2 in some books, especially some British books. The author prefers the ∇^2, which better captures the sense of the thing and which leaves Δ free for other uses.

Other second-order vector derivatives include

$$\begin{aligned} \nabla\nabla\cdot\mathbf{a} &= \hat{\mathbf{j}}\frac{\partial^2 a_i}{\partial j\,\partial i}, \\ \nabla\times\nabla\times\mathbf{a} &= \hat{\mathbf{j}}\frac{\partial}{\partial i}\left(\frac{\partial a_i}{\partial j}-\frac{\partial a_j}{\partial i}\right), \end{aligned} \tag{16.20}$$

the latter of which is derived as

$$\begin{aligned} \nabla\times\nabla\times\mathbf{a} &= \nabla\times\left(\epsilon_{ijk}\hat{\mathbf{i}}\frac{\partial a_k}{\partial j}\right) \\ &= \epsilon_{mni}\hat{\mathbf{m}}\frac{\partial}{\partial n}\left(\epsilon_{ijk}\hat{\mathbf{i}}\frac{\partial a_k}{\partial j}\right)_i = \epsilon_{mni}\epsilon_{ijk}\hat{\mathbf{m}}\frac{\partial^2 a_k}{\partial n\,\partial j} \\ &= (\delta_{mj}\delta_{nk}-\delta_{mk}\delta_{nj})\hat{\mathbf{m}}\frac{\partial^2 a_k}{\partial n\,\partial j} \\ &= \hat{\mathbf{j}}\frac{\partial^2 a_k}{\partial k\,\partial j}-\hat{\mathbf{k}}\frac{\partial^2 a_k}{\partial j^2} = \hat{\mathbf{j}}\frac{\partial^2 a_i}{\partial i\,\partial j}-\hat{\mathbf{j}}\frac{\partial^2 a_j}{\partial i^2}. \end{aligned}$$

Combining the various second-order vector derivatives yields the useful identity that

$$\nabla\nabla\cdot\mathbf{a} = \nabla^2\mathbf{a}+\nabla\times\nabla\times\mathbf{a}. \tag{16.21}$$

Table 16.1 summarizes.

The table includes two curious null identities,

$$\begin{aligned} \nabla\times\nabla\psi &= 0, \\ \nabla\cdot\nabla\times\mathbf{a} &= 0. \end{aligned} \tag{16.22}$$

In words, (16.22) states that *gradients do not curl and curl does not diverge.* This is unexpected but is a direct consequence of the definitions of the gradient, curl and divergence:

$$\begin{aligned} \nabla\times\nabla\psi = \nabla\times\left(\hat{\mathbf{i}}\frac{\partial\psi}{\partial i}\right) = \epsilon_{mni}\hat{\mathbf{m}}\frac{\partial^2\psi}{\partial n\,\partial i} &= 0; \\ \nabla\cdot\nabla\times\mathbf{a} = \nabla\cdot\left(\epsilon_{ijk}\hat{\mathbf{i}}\frac{\partial a_k}{\partial j}\right) = \epsilon_{ijk}\frac{\partial^2 a_k}{\partial i\,\partial j} &= 0. \end{aligned}$$

A field like $\nabla\psi$ that does not curl is called an *irrotational* field. A field like $\nabla\times\mathbf{a}$ that does not diverge is called a *solenoidal, source-free* or (prosaically) *divergenceless* field.[17]

[17] In the writer's country, the United States, there has been a mistaken belief afoot

Each of this section's second-order vector derivatives—including the vector Laplacian $\nabla^2\mathbf{a}$, according to (16.21)—is or can be composed of first-order vector derivatives already familiar to us from § 16.1. Therefore, inasmuch as each of those first-order vector derivatives is invariant under reorientation of axes, each second-order vector derivative is likewise invariant.

16.5 Contour derivative product rules

Equation (4.22) gives the derivative product rule for functions of a scalar variable. Fields—that is, functions of a vector variable—obey product rules, too, several of which Table 16.1 lists. The table's product rules however are general product rules that regard full spatial derivatives. What about derivatives along an arbitrary contour? Do they obey product rules, too?

that, if two fields $\mathbf{b}_1(\mathbf{r})$ and $\mathbf{b}_2(\mathbf{r})$ had everywhere the same divergence and curl, then the two fields could differ only by an additive constant. Even at least one widely distributed textbook expresses this belief, naming it *Helmholtz's theorem;* but it is not just the one textbook, for the writer has heard it verbally from at least two engineers, unacquainted with one other, who had earned Ph.D.s in different eras in different regions of the country. So the belief must be correct, mustn't it?

Well, maybe it is, but the writer remains unconvinced. Consider the admittedly contrived counterexample of $\mathbf{b}_1 = \hat{\mathbf{x}}y + \hat{\mathbf{y}}x$, $\mathbf{b}_2 = 0$.

On an applied level, the writer knows of literally no other false theorem so widely believed to be true, which leads the writer to suspect that he himself had somehow erred in judging the theorem false. What the writer really believes however is that Hermann von Helmholtz probably originally had put some appropriate restrictions on $\mathbf{b}_1$ and $\mathbf{b}_2$ which, if obeyed, made his theorem true but which at some time after his death got lost in transcription. That a transcription error would go undetected so many years would tend to suggest that Helmholtz's theorem, though interesting, were not actually very necessary in practical applications. (Some believe the theorem necessary to establish a "gauge" in a wave equation, but if they examine the use of their gauges closely then they will likely discover that one does not logically actually need to invoke the theorem to use the gauges.)

Corrections by readers are invited.

That is, one supposes that[18]

$$\begin{aligned}\frac{\partial}{\partial\ell}(\psi\omega) &= \omega\frac{\partial\psi}{\partial\ell} + \psi\frac{\partial\omega}{\partial\ell},\\ \frac{\partial}{\partial\ell}(\psi\mathbf{a}) &= \mathbf{a}\frac{\partial\psi}{\partial\ell} + \psi\frac{\partial\mathbf{a}}{\partial\ell},\\ \frac{\partial}{\partial\ell}(\mathbf{a}\cdot\mathbf{b}) &= \mathbf{b}\cdot\frac{\partial\mathbf{a}}{\partial\ell} + \mathbf{a}\cdot\frac{\partial\mathbf{b}}{\partial\ell},\\ \frac{\partial}{\partial\ell}(\mathbf{a}\times\mathbf{b}) &= -\mathbf{b}\times\frac{\partial\mathbf{a}}{\partial\ell} + \mathbf{a}\times\frac{\partial\mathbf{b}}{\partial\ell}.\end{aligned} \tag{16.23}$$

where ℓ is the distance along some arbitrary contour in space. As a hypothesis, (16.23) is attractive. But is it true?

That the first line of (16.23) is true is clear, if you think about it in the right way, because, in the restricted case (16.23) represents, one can treat the scalar fields $\psi(\mathbf{r})$ and $\omega(\mathbf{r})$ as ordinary scalar functions $\psi(\ell)$ and $\omega(\ell)$ of the scalar distance ℓ along the contour, whereupon (4.22) applies—for (16.23) never evaluates $\psi(\mathbf{r})$ or $\omega(\mathbf{r})$ but along the contour. The same naturally goes for the vector fields $\mathbf{a}(\mathbf{r})$ and $\mathbf{b}(\mathbf{r})$, which one can treat as vector functions $\mathbf{a}(\ell)$ and $\mathbf{b}(\ell)$ of the scalar distance ℓ; so the second and third lines of (16.23) are true, too, since one can write the second line in the form

$$\hat{\mathbf{i}}\left[\frac{\partial}{\partial\ell}(\psi a_i)\right] = \hat{\mathbf{i}}\left[a_i\frac{\partial\psi}{\partial\ell} + \psi\frac{\partial a_i}{\partial\ell}\right]$$

and the third line in the form

$$\frac{\partial}{\partial\ell}(a_i b_i) = b_i\frac{\partial a_i}{\partial\ell} + a_i\frac{\partial b_i}{\partial\ell},$$

each of which, according to the first line, is true separately for $i = x$, for $i = y$ and for $i = z$.

The truth of (16.23)'s last line is slightly less obvious. Nevertheless, one can reorder factors to write the line as

$$\frac{\partial}{\partial\ell}(\mathbf{a}\times\mathbf{b}) = \frac{\partial\mathbf{a}}{\partial\ell}\times\mathbf{b} + \mathbf{a}\times\frac{\partial\mathbf{b}}{\partial\ell},$$

the Levi-Civita form (§ 15.4.3) of which is

$$\epsilon_{ijk}\hat{\mathbf{i}}\left[\frac{\partial}{\partial\ell}(a_j b_k)\right] = \epsilon_{ijk}\hat{\mathbf{i}}\left[\frac{\partial a_j}{\partial\ell}b_k + a_j\frac{\partial b_k}{\partial\ell}\right].$$

[18] The $-$ sign in (16.23)'s last line is an artifact of ordering the line's factors in the style of Table 16.1. Before proving the line, the narrative will reverse the order to kill the sign. See below.

Table 16.2: The metric coefficients of the rectangular, cylindrical and spherical coordinate systems.

RECT.	CYL.	SPHER.
$h_x = 1$	$h_\rho = 1$	$h_r = 1$
$h_y = 1$	$h_\phi = \rho$	$h_\theta = r$
$h_z = 1$	$h_z = 1$	$h_\phi = \rho = r\sin\theta$

The Levi-Civita form is true separately for $(i,j,k) = (x,y,z)$, for $(i,j,k) = (x,z,y)$, and so forth, so (16.23)'s last line as a whole is true, too, which completes the proof of (16.23).

16.6 Metric coefficients

A scalar field $\psi(\mathbf{r})$ is the same field whether expressed as a function $\psi(x,y,z)$ of rectangular coordinates, $\psi(\rho;\phi,z)$ of cylindrical coordinates or $\psi(r;\theta;\phi)$ of spherical coordinates,[19] or indeed of coordinates in any three-dimensional system. However, cylindrical and spherical geometries normally recommend cylindrical and spherical coordinate systems, systems which make some of the identities of Table 16.1 hard to use.

The reason a cylindrical or spherical system makes some of the table's identities hard to use is that some of the table's identities involve derivatives d/di, notation which per § 15.4.2 stands for d/dx', d/dy' or d/dz' where the coordinates x', y' and z' represent lengths. But among the cylindrical and spherical coordinates are θ and ϕ, angles rather than lengths. Because one cannot use an angle as though it were a length, the notation d/di cannot stand for $d/d\theta$ or $d/d\phi$ and, thus, one cannot use the table in cylindrical or spherical coordinates as the table stands.

We therefore want factors to convert the angles in question to lengths (or, more generally, when special coordinate systems like the parabolic systems of § 15.7 come into play, to convert coordinates other than lengths to lengths). Such factors are called *metric coefficients* and Table 16.2 lists them.[20] The use of the table is this: that for any metric coefficient h_α a change $d\alpha$ in its coordinate α sweeps out a length $h_\alpha\, d\alpha$. For example, in cylindrical coordinates $h_\phi = \rho$ according to table, so a change $d\phi$ in the azimuthal

[19] For example, the field $\psi = x^2 + y^2$ in rectangular coordinates is $\psi = \rho^2$ in cylindrical coordinates. Refer to Table 3.4.

[20] [36, § 2-4]

coordinate ϕ sweeps out a length $\rho\,d\phi$—a fact we have already informally observed as far back as § 7.4.3, which the table now formalizes.

Incidentally, the metric coefficient h_ϕ seems to have two different values in the table, one value in cylindrical coordinates and another in spherical. The two are really the same value, though, since $\rho = r\sin\theta$ per Table 3.4.

16.6.1 Displacements, areas and volumes

In any orthogonal, right-handed, three-dimensional coordinate system $(\alpha;\beta;\gamma)$—whether the symbols $(\alpha;\beta;\gamma)$ stand for (x,y,z), (y,z,x), (z,x,y), (x',y',z'), $(\rho;\phi,z)$, $(r;\theta;\phi)$, $(\phi^x,r;\theta^x)$, etc.,[21] or even something exotic like the parabolic (σ,τ,z) of § 15.7—the product

$$d\mathbf{s} = \hat{\boldsymbol\alpha} h_\beta h_\gamma\,d\beta\,d\gamma \tag{16.24}$$

represents an area infinitesimal normal to $\hat{\boldsymbol\alpha}$. For example, the area infinitesimal on a spherical surface of radius r is $d\mathbf{s} = \hat{\mathbf{r}} h_\theta h_\phi\,d\theta\,d\phi = \hat{\mathbf{r}} r^2 \sin\theta\,d\theta\,d\phi$.

Again in any orthogonal, right-handed, three-dimensional coordinate system $(\alpha;\beta;\gamma)$, the product

$$dv = h_\alpha h_\beta h_\gamma\,d\alpha\,d\beta\,d\gamma \tag{16.25}$$

represents a volumetric infinitesimal. For example, the volumetric infinitesimal in a spherical geometry is $dv = h_r h_\theta h_\phi\,dr\,d\theta\,d\phi = r^2\sin\theta\,dr\,d\theta\,d\phi$.

Notice that § 7.4 has already calculated several integrals involving area and volumetric infinitesimals of these kinds.

A volumetric infinitesimal (16.25) cannot meaningfully in three dimensions be expressed as a vector as an area infinitesimal (16.24) can, since in three dimensions a volume has no orientation. Naturally however, a length or *displacement* infinitesimal can indeed be expressed as a vector, as

$$d\boldsymbol\ell = \hat{\boldsymbol\alpha} h_\alpha\,d\alpha. \tag{16.26}$$

Section 16.10 will have more to say about vector infinitesimals in nonrectangular coordinates.

[21] The book's admittedly clumsy usage of semicolons ";" and commas "," to delimit coordinate triplets, whereby a semicolon precedes an angle (or, in this section's case, precedes a generic coordinate like α that could stand for an angle), serves well enough to distinguish the three principal coordinate systems (x,y,z), $(\rho;\phi,z)$ and $(r;\theta;\phi)$ visually from one another but ceases to help much when further coordinate systems such as $(\phi^x,r;\theta^x)$ come into play. Logically, maybe, it would make more sense to write in the manner of $(,x,y,z)$, but to do so seems overwrought and fortunately no one the author knows of does it in that way. The delimiters just are not that important.

The book adheres to the semicolon convention not for any deep reason but only for lack of a better convention. See also chapter 15's footnote 32.

16.6.2 The vector field and its scalar components

Like a scalar field $\psi(\mathbf{r})$, a vector field $\mathbf{a}(\mathbf{r})$ too is the same field whether expressed as a function $\mathbf{a}(x, y, z)$ of rectangular coordinates, $\mathbf{a}(\rho; \phi, z)$ of cylindrical coordinates or $\mathbf{a}(r; \theta; \phi)$ of spherical coordinates, or indeed of coordinates in any three-dimensional system. A vector field however is the sum of three scalar fields, each scaling an appropriate unit vector. In rectangular coordinates,

$$\mathbf{a}(\mathbf{r}) = \hat{\mathbf{x}} a_x(\mathbf{r}) + \hat{\mathbf{y}} a_y(\mathbf{r}) + \hat{\mathbf{z}} a_z(\mathbf{r});$$

in cylindrical coordinates,

$$\mathbf{a}(\mathbf{r}) = \hat{\boldsymbol{\rho}} a_\rho(\mathbf{r}) + \hat{\boldsymbol{\phi}} a_\phi(\mathbf{r}) + \hat{\mathbf{z}} a_z(\mathbf{r});$$

and in spherical coordinates,

$$\mathbf{a}(\mathbf{r}) = \hat{\mathbf{r}} a_r(\mathbf{r}) + \hat{\boldsymbol{\theta}} a_\theta(\mathbf{r}) + \hat{\boldsymbol{\phi}} a_\phi(\mathbf{r}).$$

The scalar fields $a_\rho(\mathbf{r})$, $a_r(\mathbf{r})$, $a_\theta(\mathbf{r})$ and $a_\phi(\mathbf{r})$ in and of themselves do not differ in nature from $a_x(\mathbf{r})$, $a_y(\mathbf{r})$, $a_z(\mathbf{r})$, $\psi(\mathbf{r})$ or any other scalar field. One does tend to use them differently, though, because constant unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$ exist to combine the scalar fields $a_x(\mathbf{r})$, $a_y(\mathbf{r})$, $a_z(\mathbf{r})$ to compose the vector field $\mathbf{a}(\mathbf{r})$ whereas no such constant unit vectors exist to combine the scalar fields $a_\rho(\mathbf{r})$, $a_r(\mathbf{r})$, $a_\theta(\mathbf{r})$ and $a_\phi(\mathbf{r})$. Of course there are the variable unit vectors $\hat{\boldsymbol{\rho}}(\mathbf{r})$, $\hat{\mathbf{r}}(\mathbf{r})$, $\hat{\boldsymbol{\theta}}(\mathbf{r})$ and $\hat{\boldsymbol{\phi}}(\mathbf{r})$, but the practical and philosophical differences between these and the constant unit vectors is greater than it might seem. For instance, it is true that $\hat{\boldsymbol{\rho}} \cdot \hat{\boldsymbol{\phi}} = 0$, so long as what is meant by this is that $\hat{\boldsymbol{\rho}}(\mathbf{r}) \cdot \hat{\boldsymbol{\phi}}(\mathbf{r}) = 0$. However, $\hat{\boldsymbol{\rho}}(\mathbf{r}_1) \cdot \hat{\boldsymbol{\phi}}(\mathbf{r}_2) \neq 0$, an algebraic error fairly easy to commit. On the other hand, that $\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} = 0$ is always true.

(One might ask why such a subsection as this would appear in a section on metric coefficients. The subsection is here because no obviously better spot for it presents itself, but moreover because we shall need the understanding the subsection conveys to apply metric coefficients consistently and correctly in § 16.9 to come.)

16.7 Nonrectangular notation

Section 15.4 has introduced Einstein's summation convention, the Kronecker delta δ_{ij} and the Levi-Civita epsilon ϵ_{ijk} together as notation for use in the definition of vector operations and in the derivation of vector identities. The notation relies on symbols like i, j and k to stand for unspecified coordinates, and Tables 15.2 and 16.1 use it extensively. Unfortunately, the notation fails in the nonrectangular coordinate systems when derivatives come into play, as they do in Table 16.1, because $\partial/\partial i$ is taken to represent a derivative specifically with respect to a length whereas nonrectangular coordinates like θ and ϕ are not lengths. Fortunately, this failure is not hard to redress.

Whereas the standard Einstein symbols i, j and k can stand only for lengths, the modified Einstein symbols $\tilde{\imath}$, $\tilde{\jmath}$ and $\tilde{k}$, which this section now introduces, can stand for any coordinates, even for coordinates like θ and ϕ that are not lengths. The tilde "$\tilde{}$" atop the symbol $\tilde{\imath}$ warns readers that the coordinate it represents is not necessarily a length and that, if one wants a length, one must multiply $\tilde{\imath}$ by an appropriate metric coefficient $h_{\tilde{\imath}}$ (§ 16.6). The products $h_{\tilde{\imath}}\tilde{\imath}$, $h_{\tilde{\jmath}}\tilde{\jmath}$ and $h_{\tilde{k}}\tilde{k}$ always represent lengths.

The symbols $\hat{\mathbf{\imath}}$, $\hat{\mathbf{\jmath}}$ and $\hat{\mathbf{k}}$ need no modification even when modified symbols like $\tilde{\imath}$, $\tilde{\jmath}$ and $\tilde{k}$ are in use. This is because $\hat{\mathbf{\imath}}$, $\hat{\mathbf{\jmath}}$ and $\hat{\mathbf{k}}$ are taken to represent unit vectors—and $[\hat{\mathbf{\imath}}\,\hat{\mathbf{\jmath}}\,\hat{\mathbf{k}}]$, to represent a proper orthogonal basis—irrespective of the coordinate system; so long, naturally, as the coordinate system is an orthogonal, right-handed coordinate system as are all the coordinate systems in this book.

The modified notation will find use in § 16.9.3.

16.8 Derivatives of the basis vectors

The derivatives of the various unit basis vectors with respect to the several coordinates of their respective coordinate systems are not hard to compute. In fact, looking at Fig. 15.1 on page 496, Fig. 15.4 on page 508, and Fig. 15.5 on page 509, one can just write them down. Table 16.3 records them.

Naturally, one can compute the table's derivatives symbolically, instead, as for example

$$\frac{\partial\hat{\boldsymbol{\rho}}}{\partial\phi} = \frac{\partial}{\partial\phi}(\hat{\mathbf{x}}\cos\phi + \hat{\mathbf{y}}\sin\phi) = -\hat{\mathbf{x}}\sin\phi + \hat{\mathbf{y}}\cos\phi = +\hat{\boldsymbol{\phi}}.$$

Such an approach prospers in special coordinate systems like the parabolic systems of Tables 15.3 and 15.4, but in cylindrical and spherical coordinates it is probably easier just to look at the figures.

16.9 Derivatives in the nonrectangular systems

This section develops vector derivatives in cylindrical and spherical coordinates.

Table 16.3: Derivatives of the basis vectors.

RECTANGULAR

$\frac{\partial \hat{\mathbf{x}}}{\partial x} = 0$	$\frac{\partial \hat{\mathbf{x}}}{\partial y} = 0$	$\frac{\partial \hat{\mathbf{x}}}{\partial z} = 0$
$\frac{\partial \hat{\mathbf{y}}}{\partial x} = 0$	$\frac{\partial \hat{\mathbf{y}}}{\partial y} = 0$	$\frac{\partial \hat{\mathbf{y}}}{\partial z} = 0$
$\frac{\partial \hat{\mathbf{z}}}{\partial x} = 0$	$\frac{\partial \hat{\mathbf{z}}}{\partial y} = 0$	$\frac{\partial \hat{\mathbf{z}}}{\partial z} = 0$

CYLINDRICAL

$\frac{\partial \hat{\boldsymbol{\rho}}}{\partial \rho} = 0$	$\frac{\partial \hat{\boldsymbol{\rho}}}{\partial \phi} = +\hat{\boldsymbol{\phi}}$	$\frac{\partial \hat{\boldsymbol{\rho}}}{\partial z} = 0$
$\frac{\partial \hat{\boldsymbol{\phi}}}{\partial \rho} = 0$	$\frac{\partial \hat{\boldsymbol{\phi}}}{\partial \phi} = -\hat{\boldsymbol{\rho}}$	$\frac{\partial \hat{\boldsymbol{\phi}}}{\partial z} = 0$
$\frac{\partial \hat{\mathbf{z}}}{\partial \rho} = 0$	$\frac{\partial \hat{\mathbf{z}}}{\partial \phi} = 0$	$\frac{\partial \hat{\mathbf{z}}}{\partial z} = 0$

SPHERICAL

$\frac{\partial \hat{\mathbf{r}}}{\partial r} = 0$	$\frac{\partial \hat{\mathbf{r}}}{\partial \theta} = +\hat{\boldsymbol{\theta}}$	$\frac{\partial \hat{\mathbf{r}}}{\partial \phi} = +\hat{\boldsymbol{\phi}} \sin\theta$
$\frac{\partial \hat{\boldsymbol{\theta}}}{\partial r} = 0$	$\frac{\partial \hat{\boldsymbol{\theta}}}{\partial \theta} = -\hat{\mathbf{r}}$	$\frac{\partial \hat{\boldsymbol{\theta}}}{\partial \phi} = +\hat{\boldsymbol{\phi}} \cos\theta$
$\frac{\partial \hat{\boldsymbol{\phi}}}{\partial r} = 0$	$\frac{\partial \hat{\boldsymbol{\phi}}}{\partial \theta} = 0$	$\frac{\partial \hat{\boldsymbol{\phi}}}{\partial \phi} = -\hat{\boldsymbol{\rho}} = -\hat{\mathbf{r}} \sin\theta - \hat{\boldsymbol{\theta}} \cos\theta$

16.9.1 Derivatives in cylindrical coordinates

According to Table 16.1,

$$\nabla\psi = \hat{\mathbf{\imath}}\frac{\partial\psi}{\partial i},$$

but as § 16.6 has observed Einstein's symbol i must stand for a length not an angle, whereas one of the three cylindrical coordinates—the azimuth ϕ—is an angle. The cylindrical metric coefficients of Table 16.2 make the necessary conversion, the result of which is

$$\nabla\psi = \hat{\boldsymbol{\rho}}\frac{\partial\psi}{\partial\rho} + \hat{\boldsymbol{\phi}}\frac{\partial\psi}{\rho\,\partial\phi} + \hat{\mathbf{z}}\frac{\partial\psi}{\partial z}. \tag{16.27}$$

Again according to Table 16.1,

$$(\mathbf{b}\cdot\nabla)\mathbf{a} = b_i\frac{\partial\mathbf{a}}{\partial i}.$$

Applying the cylindrical metric coefficients, we have that

$$(\mathbf{b}\cdot\nabla)\mathbf{a} = b_\rho\frac{\partial\mathbf{a}}{\partial\rho} + b_\phi\frac{\partial\mathbf{a}}{\rho\,\partial\phi} + b_z\frac{\partial\mathbf{a}}{\partial z}. \tag{16.28}$$

Expanding the vector field **a** in the cylindrical basis,

$$(\mathbf{b}\cdot\nabla)\mathbf{a} = \left\{b_\rho\frac{\partial}{\partial\rho} + b_\phi\frac{\partial}{\rho\,\partial\phi} + b_z\frac{\partial}{\partial z}\right\}\left(\hat{\boldsymbol{\rho}}a_\rho + \hat{\boldsymbol{\phi}}a_\phi + \hat{\mathbf{z}}a_z\right).$$

Here are three derivatives of three terms, each term of two factors. Evaluating the derivatives according to the contour derivative product rule (16.23) yields (3)(3)(2) = 0x12 (eighteen) terms in the result. Half the 0x12 terms involve derivatives of the basis vectors, which Table 16.3 computes. Some of the 0x12 terms turn out to be null. The result is that

$$\begin{aligned}(\mathbf{b}\cdot\nabla)\mathbf{a} \;=\;& b_\rho\left[\hat{\boldsymbol{\rho}}\frac{\partial a_\rho}{\partial\rho} + \hat{\boldsymbol{\phi}}\frac{\partial a_\phi}{\partial\rho} + \hat{\mathbf{z}}\frac{\partial a_z}{\partial\rho}\right] \\ &+ \frac{b_\phi}{\rho}\left[\hat{\boldsymbol{\rho}}\left(\frac{\partial a_\rho}{\partial\phi} - a_\phi\right) + \hat{\boldsymbol{\phi}}\left(\frac{\partial a_\phi}{\partial\phi} + a_\rho\right) + \hat{\mathbf{z}}\frac{\partial a_z}{\partial\phi}\right] \\ &+ b_z\left[\hat{\boldsymbol{\rho}}\frac{\partial a_\rho}{\partial z} + \hat{\boldsymbol{\phi}}\frac{\partial a_\phi}{\partial z} + \hat{\mathbf{z}}\frac{\partial a_z}{\partial z}\right].\end{aligned} \tag{16.29}$$

To evaluate divergence and curl wants more care. It also wants a constant basis to work in, whereas $[\hat{\mathbf{x}}\;\hat{\mathbf{y}}\;\hat{\mathbf{z}}]$ is awkward in a cylindrical geometry and $[\hat{\boldsymbol{\rho}}\;\hat{\boldsymbol{\phi}}\;\hat{\mathbf{z}}]$ is not constant. Fortunately, nothing prevents us from defining a constant basis $[\hat{\boldsymbol\rho}_o\ \hat{\boldsymbol\phi}_o\ \hat{\mathbf{z}}]$ such that $[\hat{\boldsymbol\rho}\ \hat{\boldsymbol\phi}\ \hat{\mathbf{z}}] = [\hat{\boldsymbol\rho}_o\ \hat{\boldsymbol\phi}_o\ \hat{\mathbf{z}}]$ *at the point* $\mathbf{r} = \mathbf{r}_o$ *at which the derivative is evaluated.* If this is done, then the basis $[\hat{\boldsymbol\rho}_o\ \hat{\boldsymbol\phi}_o\ \hat{\mathbf{z}}]$ is constant like $[\hat{\mathbf{x}}\ \hat{\mathbf{y}}\ \hat{\mathbf{z}}]$ but not awkward like it.

According to Table 16.1,

$$\nabla \cdot \mathbf{a} = \frac{\partial a_i}{\partial i}$$

In cylindrical coordinates and the $[\hat{\boldsymbol\rho}_o\ \hat{\boldsymbol\phi}_o\ \hat{\mathbf{z}}]$ basis, this is[22]

$$\nabla \cdot \mathbf{a} = \frac{\partial(\hat{\boldsymbol\rho}_o \cdot \mathbf{a})}{\partial \rho} + \frac{\partial(\hat{\boldsymbol\phi}_o \cdot \mathbf{a})}{\rho\, \partial \phi} + \frac{\partial(\hat{\mathbf{z}} \cdot \mathbf{a})}{\partial z}.$$

Applying the contour derivative product rule (16.23),

$$\nabla \cdot \mathbf{a} = \hat{\boldsymbol\rho}_o \cdot \frac{\partial \mathbf{a}}{\partial \rho} + \frac{\partial \hat{\boldsymbol\rho}_o}{\partial \rho} \cdot \mathbf{a} + \hat{\boldsymbol\phi}_o \cdot \frac{\partial \mathbf{a}}{\rho\, \partial \phi} + \frac{\partial \hat{\boldsymbol\phi}_o}{\rho\, \partial \phi} \cdot \mathbf{a} + \hat{\mathbf{z}} \cdot \frac{\partial \mathbf{a}}{\partial z} + \frac{\partial \hat{\mathbf{z}}}{\partial z} \cdot \mathbf{a}.$$

But $[\hat{\boldsymbol\rho}_o\ \hat{\boldsymbol\phi}_o\ \mathbf{z}]$ are constant unit vectors, so

$$\nabla \cdot \mathbf{a} = \hat{\boldsymbol\rho}_o \cdot \frac{\partial \mathbf{a}}{\partial \rho} + \hat{\boldsymbol\phi}_o \cdot \frac{\partial \mathbf{a}}{\rho\, \partial \phi} + \hat{\mathbf{z}} \cdot \frac{\partial \mathbf{a}}{\partial z}.$$

That is,

$$\nabla \cdot \mathbf{a} = \hat{\boldsymbol\rho} \cdot \frac{\partial \mathbf{a}}{\partial \rho} + \hat{\boldsymbol\phi} \cdot \frac{\partial \mathbf{a}}{\rho\, \partial \phi} + \hat{\mathbf{z}} \cdot \frac{\partial \mathbf{a}}{\partial z}.$$

Expanding the field in the cylindrical basis,

$$\nabla \cdot \mathbf{a} = \left\{ \hat{\boldsymbol\rho} \cdot \frac{\partial}{\partial \rho} + \hat{\boldsymbol\phi} \cdot \frac{\partial}{\rho\, \partial \phi} + \hat{\mathbf{z}} \cdot \frac{\partial}{\partial z} \right\} \left(\hat{\boldsymbol\rho} a_\rho + \hat{\boldsymbol\phi} a_\phi + \hat{\mathbf{z}} a_z \right).$$

As above, here again the expansion yields 0x12 (eighteen) terms. Fortunately, this time most of the terms turn out to be null. The result is that

$$\nabla \cdot \mathbf{a} = \frac{\partial a_\rho}{\partial \rho} + \frac{a_\rho}{\rho} + \frac{\partial a_\phi}{\rho\, \partial \phi} + \frac{\partial a_z}{\partial z},$$

[22] Mistakenly to write here that

$$\nabla \cdot \mathbf{a} = \frac{\partial a_\rho}{\partial \rho} + \frac{\partial a_\phi}{\rho\, \partial \phi} + \frac{\partial a_z}{\partial z},$$

which is not true, would be a ghastly error, leading to any number of hard-to-detect false conclusions. Refer to § 16.6.2.

or, expressed more cleverly in light of (4.27), that

$$\nabla\cdot\mathbf{a} = \frac{\partial(\rho a_\rho)}{\rho\,\partial\rho} + \frac{\partial a_\phi}{\rho\,\partial\phi} + \frac{\partial a_z}{\partial z}. \tag{16.30}$$

Again according to Table 16.1,

$$\begin{aligned}
\nabla\times\mathbf{a} &= \epsilon_{ijk}\hat{\mathbf{\imath}}\frac{\partial a_k}{\partial j} \\
&= \hat{\boldsymbol{\rho}}_o\left[\frac{\partial(\hat{\mathbf{z}}\cdot\mathbf{a})}{\rho\,\partial\phi} - \frac{\partial(\hat{\boldsymbol{\phi}}_o\cdot\mathbf{a})}{\partial z}\right] + \hat{\boldsymbol{\phi}}_o\left[\frac{\partial(\hat{\boldsymbol{\rho}}_o\cdot\mathbf{a})}{\partial z} - \frac{\partial(\hat{\mathbf{z}}\cdot\mathbf{a})}{\partial\rho}\right] \\
&\qquad + \hat{\mathbf{z}}\left[\frac{\partial(\hat{\boldsymbol{\phi}}_o\cdot\mathbf{a})}{\partial\rho} - \frac{\partial(\hat{\boldsymbol{\rho}}_o\cdot\mathbf{a})}{\rho\,\partial\phi}\right].
\end{aligned}$$

That is,

$$\begin{aligned}
\nabla\times\mathbf{a} &= \hat{\boldsymbol{\rho}}\left[\hat{\mathbf{z}}\cdot\frac{\partial\mathbf{a}}{\rho\,\partial\phi} - \hat{\boldsymbol{\phi}}\cdot\frac{\partial\mathbf{a}}{\partial z}\right] + \hat{\boldsymbol{\phi}}\left[\hat{\boldsymbol{\rho}}\cdot\frac{\partial\mathbf{a}}{\partial z} - \hat{\mathbf{z}}\cdot\frac{\partial\mathbf{a}}{\partial\rho}\right] \\
&\qquad + \hat{\mathbf{z}}\left[\hat{\boldsymbol{\phi}}\cdot\frac{\partial\mathbf{a}}{\partial\rho} - \hat{\boldsymbol{\rho}}\cdot\frac{\partial\mathbf{a}}{\rho\,\partial\phi}\right].
\end{aligned}$$

Expanding the field in the cylindrical basis,

$$\begin{aligned}
\nabla\times\mathbf{a} &= \left\{\hat{\boldsymbol{\rho}}\left[\hat{\mathbf{z}}\cdot\frac{\partial}{\rho\,\partial\phi} - \hat{\boldsymbol{\phi}}\cdot\frac{\partial}{\partial z}\right] + \hat{\boldsymbol{\phi}}\left[\hat{\boldsymbol{\rho}}\cdot\frac{\partial}{\partial z} - \hat{\mathbf{z}}\cdot\frac{\partial}{\partial\rho}\right]\right. \\
&\qquad \left. + \hat{\mathbf{z}}\left[\hat{\boldsymbol{\phi}}\cdot\frac{\partial}{\partial\rho} - \hat{\boldsymbol{\rho}}\cdot\frac{\partial}{\rho\,\partial\phi}\right]\right\}\left(\hat{\boldsymbol{\rho}}a_\rho + \hat{\boldsymbol{\phi}}a_\phi + \hat{\mathbf{z}}a_z\right).
\end{aligned}$$

Here the expansion yields 0x24 (thirty-six) terms, but fortunately as last time this time most of the terms again turn out to be null. The result is that

$$\nabla\times\mathbf{a} = \hat{\boldsymbol{\rho}}\left[\frac{\partial a_z}{\rho\,\partial\phi} - \frac{\partial a_\phi}{\partial z}\right] + \hat{\boldsymbol{\phi}}\left[\frac{\partial a_\rho}{\partial z} - \frac{\partial a_z}{\partial\rho}\right] + \hat{\mathbf{z}}\left[\frac{\partial a_\phi}{\partial\rho} + \frac{a_\phi}{\rho} - \frac{\partial a_\rho}{\rho\,\partial\phi}\right],$$

or, expressed more cleverly, that

$$\nabla\times\mathbf{a} = \hat{\boldsymbol{\rho}}\left[\frac{\partial a_z}{\rho\,\partial\phi} - \frac{\partial a_\phi}{\partial z}\right] + \hat{\boldsymbol{\phi}}\left[\frac{\partial a_\rho}{\partial z} - \frac{\partial a_z}{\partial\rho}\right] + \frac{\hat{\mathbf{z}}}{\rho}\left[\frac{\partial(\rho a_\phi)}{\partial\rho} - \frac{\partial a_\rho}{\partial\phi}\right]. \tag{16.31}$$

Table 16.4 summarizes.[23]

[23] [11, appendix II.2.2]

Table 16.4: Vector derivatives in cylindrical coordinates.

$$\begin{aligned}
\nabla\psi &= \hat{\boldsymbol{\rho}}\frac{\partial\psi}{\partial\rho} + \hat{\boldsymbol{\phi}}\frac{\partial\psi}{\rho\,\partial\phi} + \hat{\mathbf{z}}\frac{\partial\psi}{\partial z} \\
(\mathbf{b}\cdot\nabla)\mathbf{a} &= b_\rho\frac{\partial\mathbf{a}}{\partial\rho} + b_\phi\frac{\partial\mathbf{a}}{\rho\,\partial\phi} + b_z\frac{\partial\mathbf{a}}{\partial z} \\
&= b_\rho\left[\hat{\boldsymbol{\rho}}\frac{\partial a_\rho}{\partial\rho} + \hat{\boldsymbol{\phi}}\frac{\partial a_\phi}{\partial\rho} + \hat{\mathbf{z}}\frac{\partial a_z}{\partial\rho}\right] \\
&\quad + \frac{b_\phi}{\rho}\left[\hat{\boldsymbol{\rho}}\left(\frac{\partial a_\rho}{\partial\phi} - a_\phi\right) + \hat{\boldsymbol{\phi}}\left(\frac{\partial a_\phi}{\partial\phi} + a_\rho\right) + \hat{\mathbf{z}}\frac{\partial a_z}{\partial\phi}\right] \\
&\quad + b_z\left[\hat{\boldsymbol{\rho}}\frac{\partial a_\rho}{\partial z} + \hat{\boldsymbol{\phi}}\frac{\partial a_\phi}{\partial z} + \hat{\mathbf{z}}\frac{\partial a_z}{\partial z}\right] \\
\nabla\cdot\mathbf{a} &= \frac{\partial(\rho a_\rho)}{\rho\,\partial\rho} + \frac{\partial a_\phi}{\rho\,\partial\phi} + \frac{\partial a_z}{\partial z} \\
\nabla\times\mathbf{a} &= \hat{\boldsymbol{\rho}}\left[\frac{\partial a_z}{\rho\,\partial\phi} - \frac{\partial a_\phi}{\partial z}\right] + \hat{\boldsymbol{\phi}}\left[\frac{\partial a_\rho}{\partial z} - \frac{\partial a_z}{\partial\rho}\right] + \frac{\hat{\mathbf{z}}}{\rho}\left[\frac{\partial(\rho a_\phi)}{\partial\rho} - \frac{\partial a_\rho}{\partial\phi}\right]
\end{aligned}$$

One can compute a second-order vector derivative in cylindrical coordinates as a sequence of two first-order cylindrical vector derivatives. For example, because Table 16.1 gives the scalar Laplacian as $\nabla^2\psi = \nabla\cdot\nabla\psi$, one can calculate $\nabla^2\psi$ in cylindrical coordinates by taking the divergence of ψ's gradient.[24] To calculate the vector Laplacian $\nabla^2\mathbf{a}$ in cylindrical coordinates is tedious but nonetheless can with care be done accurately by means of Table 16.1's identity that $\nabla^2\mathbf{a} = \nabla\nabla\cdot\mathbf{a} - \nabla\times\nabla\times\mathbf{a}$. (This means that to calculate the vector Laplacian $\nabla^2\mathbf{a}$ in cylindrical coordinates takes not just two but actually four first-order cylindrical vector derivatives, for the author regrettably knows of no valid shortcut—the clumsy alternative, less proper, less insightful, even more tedious and not recommended, being to take the Laplacian in rectangular coordinates and then to convert back to the cylindrical domain; for to work cylindrical problems directly in

[24] A concrete example: if $\psi(\mathbf{r}) = e^{i\phi}/\rho$, then $\nabla\psi = (-\hat{\boldsymbol{\rho}} + i\hat{\boldsymbol{\phi}})e^{i\phi}/\rho^2$ per Table 16.4, whereupon

$$\nabla^2\psi = \nabla\cdot\left[\left(-\hat{\boldsymbol{\rho}} + i\hat{\boldsymbol{\phi}}\right)\frac{e^{i\phi}}{\rho^2}\right] = \left(-\hat{\boldsymbol{\rho}} + i\hat{\boldsymbol{\phi}}\right)\cdot\nabla\left(\frac{e^{i\phi}}{\rho^2}\right) + \frac{e^{i\phi}}{\rho^2}\nabla\cdot\left(-\hat{\boldsymbol{\rho}} + i\hat{\boldsymbol{\phi}}\right).$$

To finish the example is left as an exercise.

cylindrical coordinates is almost always advisable.)

16.9.2 Derivatives in spherical coordinates

One can compute vector derivatives in spherical coordinates as in cylindrical coordinates (§ 16.9.1), only the spherical details though not essentially more complicated are messier. According to Table 16.1,

$$\nabla\psi = \hat{\mathbf{i}}\frac{\partial\psi}{\partial i}.$$

Applying the spherical metric coefficients of Table 16.2, we have that

$$\nabla\psi = \hat{\mathbf{r}}\frac{\partial\psi}{\partial r} + \hat{\boldsymbol{\theta}}\frac{\partial\psi}{r\,\partial\theta} + \hat{\boldsymbol{\phi}}\frac{\partial\psi}{(r\sin\theta)\,\partial\phi}. \tag{16.32}$$

Again according to Table 16.1,

$$(\mathbf{b}\cdot\nabla)\mathbf{a} = b_i\frac{\partial\mathbf{a}}{\partial i}.$$

Applying the cylindrical metric coefficients, we have that

$$(\mathbf{b}\cdot\nabla)\mathbf{a} = b_r\frac{\partial\mathbf{a}}{\partial r} + b_\theta\frac{\partial\mathbf{a}}{r\,\partial\theta} + b_\phi\frac{\partial\mathbf{a}}{(r\sin\theta)\,\partial\phi}. \tag{16.33}$$

Expanding the vector field $\mathbf{a}$ in the spherical basis,

$$(\mathbf{b}\cdot\nabla)\mathbf{a} = \left\{b_r\frac{\partial}{\partial r} + b_\theta\frac{\partial}{r\,\partial\theta} + b_\phi\frac{\partial}{(r\sin\theta)\,\partial\phi}\right\}\left(\hat{\mathbf{r}}a_r + \hat{\boldsymbol{\theta}}a_\theta + \hat{\boldsymbol{\phi}}a_\phi\right).$$

Evaluating the derivatives,

$$\begin{aligned}(\mathbf{b}\cdot\nabla)\mathbf{a} \;=\;& b_r\left[\hat{\mathbf{r}}\frac{\partial a_r}{\partial r} + \hat{\boldsymbol{\theta}}\frac{\partial a_\theta}{\partial r} + \hat{\boldsymbol{\phi}}\frac{\partial a_\phi}{\partial r}\right]\\ &+ \frac{b_\theta}{r}\left[\hat{\mathbf{r}}\left(\frac{\partial a_r}{\partial\theta} - a_\theta\right) + \hat{\boldsymbol{\theta}}\left(\frac{\partial a_\theta}{\partial\theta} + a_r\right) + \hat{\boldsymbol{\phi}}\frac{\partial a_\phi}{\partial\theta}\right]\\ &+ \frac{b_\phi}{r\sin\theta}\Bigg[\hat{\mathbf{r}}\left(\frac{\partial a_r}{\partial\phi} - a_\phi\sin\theta\right) + \hat{\boldsymbol{\theta}}\left(\frac{\partial a_\theta}{\partial\phi} - a_\phi\cos\theta\right)\\ &\qquad + \hat{\boldsymbol{\phi}}\left(\frac{\partial a_\phi}{\partial\phi} + a_r\sin\theta + a_\theta\cos\theta\right)\Bigg].\end{aligned} \tag{16.34}$$

According to Table 16.1, reasoning as in § 16.9.1,

$$\nabla\cdot\mathbf{a} = \frac{\partial a_i}{\partial i} = \hat{\mathbf{r}}\cdot\frac{\partial\mathbf{a}}{\partial r} + \hat{\boldsymbol{\theta}}\cdot\frac{\partial\mathbf{a}}{r\,\partial\theta} + \hat{\boldsymbol{\phi}}\cdot\frac{\partial\mathbf{a}}{(r\sin\theta)\,\partial\phi}.$$

Expanding the field in the spherical basis,

$$\nabla\cdot\mathbf{a} = \left\{\hat{\mathbf{r}}\cdot\frac{\partial}{\partial r} + \hat{\boldsymbol{\theta}}\cdot\frac{\partial}{r\,\partial\theta} + \hat{\boldsymbol{\phi}}\cdot\frac{\partial}{(r\sin\theta)\,\partial\phi}\right\}\left(\hat{\mathbf{r}}a_r + \hat{\boldsymbol{\theta}}a_\theta + \hat{\boldsymbol{\phi}}a_\phi\right).$$

Evaluating the derivatives, the result is that

$$\nabla\cdot\mathbf{a} = \frac{\partial a_r}{\partial r} + \frac{2a_r}{r} + \frac{\partial a_\theta}{r\,\partial\theta} + \frac{a_\theta}{r\tan\theta} + \frac{\partial a_\phi}{(r\sin\theta)\,\partial\phi},$$

or, expressed more cleverly, that

$$\nabla\cdot\mathbf{a} = \frac{1}{r}\left[\frac{\partial(r^2a_r)}{r\,\partial r} + \frac{\partial(a_\theta\sin\theta)}{(\sin\theta)\,\partial\theta} + \frac{\partial a_\phi}{(\sin\theta)\,\partial\phi}\right]. \tag{16.35}$$

Again according to Table 16.1, reasoning as in § 16.9.1,

$$\begin{aligned}\nabla\times\mathbf{a} &= \epsilon_{ijk}\hat{\mathbf{i}}\frac{\partial a_k}{\partial j}\\ &= \hat{\mathbf{r}}\left[\hat{\boldsymbol{\phi}}\cdot\frac{\partial\mathbf{a}}{r\,\partial\theta} - \hat{\boldsymbol{\theta}}\cdot\frac{\partial\mathbf{a}}{(r\sin\theta)\,\partial\phi}\right] + \hat{\boldsymbol{\theta}}\left[\hat{\mathbf{r}}\cdot\frac{\partial\mathbf{a}}{(r\sin\theta)\,\partial\phi} - \hat{\boldsymbol{\phi}}\cdot\frac{\partial\mathbf{a}}{\partial r}\right]\\ &\qquad + \hat{\boldsymbol{\phi}}\left[\hat{\boldsymbol{\theta}}\cdot\frac{\partial\mathbf{a}}{\partial r} - \hat{\mathbf{r}}\cdot\frac{\partial\mathbf{a}}{r\,\partial\theta}\right].\end{aligned}$$

Expanding the field in the spherical basis,

$$\begin{aligned}\nabla\times\mathbf{a} &= \left\{\hat{\mathbf{r}}\left[\hat{\boldsymbol{\phi}}\cdot\frac{\partial}{r\,\partial\theta} - \hat{\boldsymbol{\theta}}\cdot\frac{\partial}{(r\sin\theta)\,\partial\phi}\right] + \hat{\boldsymbol{\theta}}\left[\hat{\mathbf{r}}\cdot\frac{\partial}{(r\sin\theta)\,\partial\phi} - \hat{\boldsymbol{\phi}}\cdot\frac{\partial}{\partial r}\right]\right.\\ &\qquad \left.+ \hat{\boldsymbol{\phi}}\left[\hat{\boldsymbol{\theta}}\cdot\frac{\partial}{\partial r} - \hat{\mathbf{r}}\cdot\frac{\partial}{r\,\partial\theta}\right]\right\}\left(\hat{\mathbf{r}}a_r + \hat{\boldsymbol{\theta}}a_\theta + \hat{\boldsymbol{\phi}}a_\phi\right).\end{aligned}$$

Evaluating the derivatives, the result is that

$$\begin{aligned}\nabla\times\mathbf{a} &= \hat{\mathbf{r}}\left[\frac{\partial a_\phi}{r\,\partial\theta} + \frac{a_\phi}{r\tan\theta} - \frac{\partial a_\theta}{(r\sin\theta)\,\partial\phi}\right] + \hat{\boldsymbol{\theta}}\left[\frac{\partial a_r}{(r\sin\theta)\,\partial\phi} - \frac{\partial a_\phi}{\partial r} - \frac{a_\phi}{r}\right]\\ &\qquad + \hat{\boldsymbol{\phi}}\left[\frac{\partial a_\theta}{\partial r} + \frac{a_\theta}{r} - \frac{\partial a_r}{r\,\partial\theta}\right],\end{aligned}$$

or, expressed more cleverly, that

$$\begin{aligned}\nabla\times\mathbf{a} &= \frac{\hat{\mathbf{r}}}{r\sin\theta}\left[\frac{\partial(a_\phi\sin\theta)}{\partial\theta} - \frac{\partial a_\theta}{\partial\phi}\right] + \frac{\hat{\boldsymbol{\theta}}}{r}\left[\frac{\partial a_r}{(\sin\theta)\,\partial\phi} - \frac{\partial(ra_\phi)}{\partial r}\right]\\ &\qquad + \frac{\hat{\boldsymbol{\phi}}}{r}\left[\frac{\partial(ra_\theta)}{\partial r} - \frac{\partial a_r}{\partial\theta}\right].\end{aligned} \tag{16.36}$$

Table 16.5: Vector derivatives in spherical coordinates.

$$
\begin{aligned}
\nabla\psi &= \hat{\mathbf{r}}\frac{\partial\psi}{\partial r} + \hat{\boldsymbol{\theta}}\frac{\partial\psi}{r\,\partial\theta} + \hat{\boldsymbol{\phi}}\frac{\partial\psi}{(r\sin\theta)\,\partial\phi} \\
(\mathbf{b}\cdot\nabla)\mathbf{a} &= b_r\frac{\partial\mathbf{a}}{\partial r} + b_\theta\frac{\partial\mathbf{a}}{r\,\partial\theta} + b_\phi\frac{\partial\mathbf{a}}{(r\sin\theta)\,\partial\phi} \\
&= b_r\left[\hat{\mathbf{r}}\frac{\partial a_r}{\partial r} + \hat{\boldsymbol{\theta}}\frac{\partial a_\theta}{\partial r} + \hat{\boldsymbol{\phi}}\frac{\partial a_\phi}{\partial r}\right] \\
&\quad + \frac{b_\theta}{r}\left[\hat{\mathbf{r}}\left(\frac{\partial a_r}{\partial\theta} - a_\theta\right) + \hat{\boldsymbol{\theta}}\left(\frac{\partial a_\theta}{\partial\theta} + a_r\right) + \hat{\boldsymbol{\phi}}\frac{\partial a_\phi}{\partial\theta}\right] \\
&\quad + \frac{b_\phi}{r\sin\theta}\left[\hat{\mathbf{r}}\left(\frac{\partial a_r}{\partial\phi} - a_\phi\sin\theta\right) + \hat{\boldsymbol{\theta}}\left(\frac{\partial a_\theta}{\partial\phi} - a_\phi\cos\theta\right)\right. \\
&\qquad\qquad \left. + \hat{\boldsymbol{\phi}}\left(\frac{\partial a_\phi}{\partial\phi} + a_r\sin\theta + a_\theta\cos\theta\right)\right] \\
\nabla\cdot\mathbf{a} &= \frac{1}{r}\left[\frac{\partial(r^2 a_r)}{r\,\partial r} + \frac{\partial(a_\theta\sin\theta)}{(\sin\theta)\,\partial\theta} + \frac{\partial a_\phi}{(\sin\theta)\,\partial\phi}\right] \\
\nabla\times\mathbf{a} &= \frac{\hat{\mathbf{r}}}{r\sin\theta}\left[\frac{\partial(a_\phi\sin\theta)}{\partial\theta} - \frac{\partial a_\theta}{\partial\phi}\right] + \frac{\hat{\boldsymbol{\theta}}}{r}\left[\frac{\partial a_r}{(\sin\theta)\,\partial\phi} - \frac{\partial(r a_\phi)}{\partial r}\right] \\
&\quad + \frac{\hat{\boldsymbol{\phi}}}{r}\left[\frac{\partial(r a_\theta)}{\partial r} - \frac{\partial a_r}{\partial\theta}\right]
\end{aligned}
$$

Table 16.5 summarizes.[25]

One can compute a second-order vector derivative in spherical coordinates as in cylindrical coordinates, as a sequence of two first-order vector derivatives. Refer to § 16.9.1.

16.9.3 Finding the derivatives geometrically

The method of §§ 16.9.1 and 16.9.2 is general, reliable and correct, but there exists an alternate, arguably neater method to derive nonrectangular formulas for most vector derivatives. Adapting the notation to this subsection's purpose we can write (16.9) as

$$\nabla \cdot \mathbf{a}(\mathbf{r}) \equiv \lim_{\Delta V \to 0} \frac{\Phi}{\Delta V}, \tag{16.37}$$

thus defining a vector's divergence fundamentally as in § 16.1.4, geometrically, as the ratio of flux Φ from a vanishing test volume ΔV to the volume itself; where per (16.7) $\Phi = \oint_S \mathbf{a}(\mathbf{r}') \cdot d\mathbf{s}$, where $\mathbf{r}'$ is a position on the test volume's surface, and where $d\mathbf{s} = d\mathbf{s}(\mathbf{r}')$ is the corresponding surface patch. So long as the test volume ΔV includes the point $\mathbf{r}$ and is otherwise infinitesimal in extent, we remain free to shape the volume as we like,[26] so let us give it six sides and shape it as an almost rectangular box that conforms precisely to the coordinate system $(\alpha; \beta; \gamma)$ in use:

$$\begin{aligned}
\alpha - \frac{\Delta\alpha}{2} \le \alpha' \le \alpha + \frac{\Delta\alpha}{2};\\
\beta - \frac{\Delta\beta}{2} \le \beta' \le \beta + \frac{\Delta\beta}{2};\\
\gamma - \frac{\Delta\gamma}{2} \le \gamma' \le \gamma + \frac{\Delta\gamma}{2}.
\end{aligned}$$

[25] [11, appendix II.2.3]

[26] A professional mathematician would probably enjoin the volume's shape to obey certain technical restrictions, such as that it remain wholly enclosed within a sphere of vanishing radius, but we will not try for such a level of rigor here.

The fluxes outward through the box's $+\alpha$- and $-\alpha$-ward sides will then be[27]

$$\begin{aligned}\Phi_{+\alpha} &= (+a_\alpha)(h_\beta h_\gamma\,\Delta\beta\,\Delta\gamma)|_{\mathbf{r}'=\mathbf{r}(\alpha+\Delta\alpha/2;\beta;\gamma)}\\ &= +a_\alpha h_\beta h_\gamma|_{\mathbf{r}'=\mathbf{r}(\alpha+\Delta\alpha/2;\beta;\gamma)}\,\Delta\beta\,\Delta\gamma,\\ \Phi_{-\alpha} &= (-a_\alpha)(h_\beta h_\gamma\,\Delta\beta\,\Delta\gamma)|_{\mathbf{r}'=\mathbf{r}(\alpha-\Delta\alpha/2;\beta;\gamma)}\\ &= -a_\alpha h_\beta h_\gamma|_{\mathbf{r}'=\mathbf{r}(\alpha-\Delta\alpha/2;\beta;\gamma)}\,\Delta\beta\,\Delta\gamma,\end{aligned}$$

products of the outward-directed field components and the areas (16.24) of the sides through which the fields pass. Thence by successive steps, the net flux outward through the pair of opposing sides will be

$$\begin{aligned}\Phi_\alpha &= \Phi_{+\alpha}+\Phi_{-\alpha}\\ &= \left[a_\alpha h_\beta h_\gamma|_{\mathbf{r}'=\mathbf{r}(\alpha+\Delta\alpha/2;\beta;\gamma)} - a_\alpha h_\beta h_\gamma|_{\mathbf{r}'=\mathbf{r}(\alpha-\Delta\alpha/2;\beta;\gamma)}\right]\Delta\beta\,\Delta\gamma\\ &= \left[\frac{\partial(a_\alpha h_\beta h_\gamma)}{\partial\alpha}\,\Delta\alpha\right]\Delta\beta\,\Delta\gamma = \Delta\alpha\,\Delta\beta\,\Delta\gamma\,\frac{\partial(a_\alpha h_\beta h_\gamma)}{\partial\alpha}\\ &= \frac{\Delta V\,\partial(a_\alpha h_\beta h_\gamma)}{h_\alpha h_\beta h_\gamma\,\partial\alpha}.\end{aligned}$$

Naturally, the same goes for the other two pairs of sides:

$$\begin{aligned}\Phi_\beta &= \frac{\Delta V\,\partial(a_\beta h_\gamma h_\alpha)}{h_\alpha h_\beta h_\gamma\,\partial\beta};\\ \Phi_\gamma &= \frac{\Delta V\,\partial(a_\gamma h_\alpha h_\beta)}{h_\alpha h_\beta h_\gamma\,\partial\gamma}.\end{aligned}$$

[27] More rigorously, one might digress from this point to expand the field in a three-dimensional Taylor series (§ 8.16) to account for the field's variation over a single side of the test volume. So lengthy a digression however would only formalize what we already knew; namely, that one can approximate to first order the integral of a well-behaved quantity over an infinitesimal domain by the quantity's value at the domain's midpoint. If you will believe that $\lim_{\Delta\tau\to 0}\int_{\tau-\Delta\tau/2}^{\tau+\Delta\tau/2} f(\tau')\,d\tau' = f(\tau)\,\Delta\tau$ for any τ in the neighborhood of which $f(\tau)$ is well behaved, then you will probably also believe its three-dimensional analog in the narrative. (If the vagueness in this context of the adjective "well-behaved" deeply troubles any reader then that reader may possess the worthy temperament of a professional mathematician; he might review chapter 8 and then seek further illumination in the professional mathematical literature. Other readers, of more practical temperament, are advised to visualize test volumes in rectangular, cylindrical and spherical coordinates and to ponder the matter a while. Consider: if the field grows in strength across a single side of the test volume and if the test volume is small enough that second-order effects can be ignored, then what single value ought one to choose to represent the field over the whole side but its value at the side's midpoint? Such visualization should soon clear up any confusion and is what the writer recommends. Incidentally, the contrast between the two modes of thought this footnote reveals is exactly the sort of thing Courant and Hilbert were talking about in § 1.2.1.)

The three equations are better written

$$\begin{aligned}\Phi_\alpha &= \frac{\Delta V}{h^3}\frac{\partial}{\partial\alpha}\left(\frac{h^3 a_\alpha}{h_\alpha}\right),\\ \Phi_\beta &= \frac{\Delta V}{h^3}\frac{\partial}{\partial\beta}\left(\frac{h^3 a_\beta}{h_\beta}\right),\\ \Phi_\gamma &= \frac{\Delta V}{h^3}\frac{\partial}{\partial\gamma}\left(\frac{h^3 a_\gamma}{h_\gamma}\right),\end{aligned}$$

where

$$h^3 \equiv h_\alpha h_\beta h_\gamma. \tag{16.38}$$

The total flux from the test volume then is

$$\begin{aligned}\Phi &= \Phi_\alpha + \Phi_\beta + \Phi_\gamma\\ &= \frac{\Delta V}{h^3}\left[\frac{\partial}{\partial\alpha}\left(\frac{h^3 a_\alpha}{h_\alpha}\right) + \frac{\partial}{\partial\beta}\left(\frac{h^3 a_\beta}{h_\beta}\right) + \frac{\partial}{\partial\gamma}\left(\frac{h^3 a_\gamma}{h_\gamma}\right)\right];\end{aligned}$$

or, invoking Einstein's summation convention in § 16.7's modified style,

$$\Phi = \frac{\Delta V}{h^3}\frac{\partial}{\partial\tilde{\imath}}\left(\frac{h^3 a_{\tilde{\imath}}}{h_{\tilde{\imath}}}\right).$$

Finally, substituting the last equation into (16.37),

$$\nabla\cdot\mathbf{a} = \frac{\partial}{h^3\,\partial\tilde{\imath}}\left(\frac{h^3 a_{\tilde{\imath}}}{h_{\tilde{\imath}}}\right). \tag{16.39}$$

An analogous formula for curl is not much harder to derive but is harder to approach directly, so we will approach it by deriving first the formula for $\hat{\boldsymbol{\gamma}}$-directed directional curl. Equation (16.12) has it that[28]

$$\hat{\boldsymbol{\gamma}}\cdot\nabla\times\mathbf{a}(\mathbf{r}) \equiv \lim_{\Delta A\to 0}\frac{\Gamma}{\Delta A}, \tag{16.40}$$

where per (16.11) $\Gamma = \oint_\gamma \mathbf{a}(\mathbf{r}')\cdot d\boldsymbol{\ell}$ and the notation $\oint_\gamma$ reminds us that the contour of integration lies in the α-β plane, perpendicular to $\hat{\boldsymbol{\gamma}}$. In this case the contour of integration bounds not a test volume but a test surface, which

[28] The appearance of both $\mathbf{a}$ and A in (16.40) is unfortunate but coïncidental, as is the appearance of both $\hat{\boldsymbol{\gamma}}$ and Γ. The capital and minuscule symbols here represent unrelated quantities.

we give four edges and an almost rectangular shape that conforms precisely to the coordinate system $(\alpha;\beta;\gamma)$ in use:

$$\begin{gathered}
\alpha - \frac{\Delta\alpha}{2} \le \alpha' \le \alpha + \frac{\Delta\alpha}{2};\\
\beta - \frac{\Delta\beta}{2} \le \beta' \le \beta + \frac{\Delta\beta}{2};\\
\gamma' = \gamma.
\end{gathered}$$

The circulations along the $+\alpha$- and $-\alpha$-ward edges will be

$$\begin{aligned}
\Gamma_{+\alpha} &= +h_\beta a_\beta\big|_{\mathbf{r}'=\mathbf{r}(\alpha+\Delta\alpha/2;\beta;\gamma)}\,\Delta\beta,\\
\Gamma_{-\alpha} &= -h_\beta a_\beta\big|_{\mathbf{r}'=\mathbf{r}(\alpha-\Delta\alpha/2;\beta;\gamma)}\,\Delta\beta,
\end{aligned}$$

and likewise the circulations along the $-\beta$- and $+\beta$-ward edges will be

$$\begin{aligned}
\Gamma_{-\beta} &= +h_\alpha a_\alpha\big|_{\mathbf{r}'=\mathbf{r}(\alpha;\beta-\Delta\beta/2;\gamma)}\,\Delta\alpha,\\
\Gamma_{+\beta} &= -h_\alpha a_\alpha\big|_{\mathbf{r}'=\mathbf{r}(\alpha;\beta+\Delta\beta/2;\gamma)}\,\Delta\alpha,
\end{aligned}$$

whence the total circulation about the contour is

$$\begin{aligned}
\Gamma &= \frac{\partial(h_\beta a_\beta)}{\partial\alpha}\,\Delta\alpha\,\Delta\beta - \frac{\partial(h_\alpha a_\alpha)}{\partial\beta}\,\Delta\beta\,\Delta\alpha\\
&= \frac{h_\gamma\,\Delta A}{h^3}\left[\frac{\partial(h_\beta a_\beta)}{\partial\alpha} - \frac{\partial(h_\alpha a_\alpha)}{\partial\beta}\right].
\end{aligned}$$

Substituting the last equation into (16.40), we have that

$$\hat{\boldsymbol{\gamma}}\cdot\nabla\times\mathbf{a} = \frac{h_\gamma}{h^3}\left[\frac{\partial(h_\beta a_\beta)}{\partial\alpha} - \frac{\partial(h_\alpha a_\alpha)}{\partial\beta}\right].$$

Likewise,

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}\cdot\nabla\times\mathbf{a} &= \frac{h_\alpha}{h^3}\left[\frac{\partial(h_\gamma a_\gamma)}{\partial\beta} - \frac{\partial(h_\beta a_\beta)}{\partial\gamma}\right],\\
\hat{\boldsymbol{\beta}}\cdot\nabla\times\mathbf{a} &= \frac{h_\beta}{h^3}\left[\frac{\partial(h_\alpha a_\alpha)}{\partial\gamma} - \frac{\partial(h_\gamma a_\gamma)}{\partial\alpha}\right].
\end{aligned}$$

But one can split any vector $\mathbf{v}$ into locally rectangular components as $\mathbf{v} =$

$\hat{\boldsymbol\alpha}(\hat{\boldsymbol\alpha}\cdot\mathbf{v})+\hat{\boldsymbol\beta}(\hat{\boldsymbol\beta}\cdot\mathbf{v})+\hat{\boldsymbol\gamma}(\hat{\boldsymbol\gamma}\cdot\mathbf{v})$, so

$$\begin{aligned}
\nabla\times\mathbf{a} &= \hat{\boldsymbol\alpha}(\hat{\boldsymbol\alpha}\cdot\nabla\times\mathbf{a})+\hat{\boldsymbol\beta}(\hat{\boldsymbol\beta}\cdot\nabla\times\mathbf{a})+\hat{\boldsymbol\gamma}(\hat{\boldsymbol\gamma}\cdot\nabla\times\mathbf{a}) \\
&= \frac{\hat{\boldsymbol\alpha}h_\alpha}{h^3}\left[\frac{\partial(h_\gamma a_\gamma)}{\partial\beta}-\frac{\partial(h_\beta a_\beta)}{\partial\gamma}\right]+\frac{\hat{\boldsymbol\beta}h_\beta}{h^3}\left[\frac{\partial(h_\alpha a_\alpha)}{\partial\gamma}-\frac{\partial(h_\gamma a_\gamma)}{\partial\alpha}\right] \\
&\qquad +\frac{\hat{\boldsymbol\gamma}h_\gamma}{h^3}\left[\frac{\partial(h_\beta a_\beta)}{\partial\alpha}-\frac{\partial(h_\alpha a_\alpha)}{\partial\beta}\right] \\
&= \frac{1}{h^3}\begin{vmatrix}\hat{\boldsymbol\alpha}h_\alpha & \hat{\boldsymbol\beta}h_\beta & \hat{\boldsymbol\gamma}h_\gamma \\ \partial/\partial\alpha & \partial/\partial\beta & \partial/\partial\gamma \\ h_\alpha a_\alpha & h_\beta a_\beta & h_\gamma a_\gamma\end{vmatrix};
\end{aligned}$$

or, in Einstein notation,[29]

$$\nabla\times\mathbf{a}=\frac{\epsilon_{\tilde\imath\tilde\jmath\tilde k}\hat{\boldsymbol\imath}h_{\tilde\imath}\,\partial(h_{\tilde k}a_{\tilde k})}{h^3\,\partial\tilde\jmath}. \tag{16.41}$$

Compared to the formulas (16.39) and (16.41) for divergence and curl, the corresponding gradient formula seems almost trivial. It is

$$\nabla\psi=\frac{\hat{\boldsymbol\imath}\,\partial\psi}{h_{\tilde\imath}\,\partial\tilde\imath}. \tag{16.42}$$

One can generate most of the vector-derivative formulas of Tables 16.4 and 16.5 by means of this subsection's (16.39), (16.41) and (16.42). One can generate additional vector-derivative formulas for special coordinate systems like the parabolic systems of § 15.7 by means of the same equations.

[29] What a marvel mathematical notation is! If you can read (16.41) and understand the message it conveys, then let us pause a moment to appreciate a few of the many concepts the notation implicitly encapsulates. There are the vector, the unit vector, the field, the derivative, the integral, circulation, parity, rotational invariance, nonrectangular coordinates, three-dimensional geometry, the dummy variable and so on—each of which concepts itself yet encapsulates several further ideas—not to mention multiplication and division which themselves are not trivial. It is doubtful that one could explain it even tersely to the uninitiated in fewer than fifty pages, and yet to the initiated one can express it all in half a line.

16.10 Vector infinitesimals

To integrate a field over a contour or surface is a typical maneuver of vector calculus. One might integrate in any of the forms

$$\begin{array}{cccc}
\int_C \psi\, d\ell & \int_C \mathbf{a}\, d\ell & \int_S \psi\, ds & \int_S \mathbf{a}\, ds \\
\int_C \psi\, d\boldsymbol{\ell} & \int_C \mathbf{a}\cdot d\boldsymbol{\ell} & \int_S \psi\, d\mathbf{s} & \int_S \mathbf{a}\cdot d\mathbf{s} \\
 & \int_C \mathbf{a}\times d\boldsymbol{\ell} & & \int_S \mathbf{a}\times d\mathbf{s}
\end{array}$$

among others. Where the integration is over a contour, a pair of functions $\alpha(\gamma)$ and $\beta(\gamma)$ typically can serve to specify the contour. Where over a surface, a single function $\gamma(\alpha;\beta)$ can serve. Given such functions and a field integral to compute, one wants an expression for the integrand's infinitesimal $d\boldsymbol{\ell}$ or $d\mathbf{s}$ in terms respectively of the contour functions $\alpha(\gamma)$ and $\beta(\gamma)$ or of the surface function $\gamma(\alpha;\beta)$.

The contour infinitesimal is evidently

$$d\boldsymbol{\ell} = \left(\hat{\boldsymbol{\gamma}} h_\gamma + \hat{\boldsymbol{\alpha}}\frac{h_\alpha\, d\alpha}{d\gamma} + \hat{\boldsymbol{\beta}}\frac{h_\beta\, d\beta}{d\gamma}\right) d\gamma, \tag{16.43}$$

consisting of a step in the $\hat{\boldsymbol{\gamma}}$ direction plus the corresponding steps in the orthogonal $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ directions. This is easy once you see how to do it. Harder is the surface infinitesimal, but one can nevertheless correctly construct it as the cross product

$$\begin{aligned}
d\mathbf{s} &= \left[\left(\hat{\boldsymbol{\alpha}} h_\alpha + \hat{\boldsymbol{\gamma}}\frac{h_\gamma\,\partial\gamma}{\partial\alpha}\right) d\alpha\right] \times \left[\left(\hat{\boldsymbol{\beta}} h_\beta + \hat{\boldsymbol{\gamma}}\frac{h_\gamma\,\partial\gamma}{\partial\beta}\right) d\beta\right] \\
&= \left(\hat{\boldsymbol{\gamma}}\frac{1}{h_\gamma} - \hat{\boldsymbol{\alpha}}\frac{\partial\gamma}{h_\alpha\,\partial\alpha} - \hat{\boldsymbol{\beta}}\frac{\partial\gamma}{h_\beta\,\partial\beta}\right) h^3\, d\alpha\, d\beta
\end{aligned} \tag{16.44}$$

of two vectors that lie on the surface, one vector normal to $\hat{\boldsymbol{\beta}}$ and the other to $\hat{\boldsymbol{\alpha}}$, edges not of a rectangular patch of the surface but of a patch whose projection onto the α-β plane is an $(h_\alpha\, d\alpha)$-by-$(h_\beta\, d\beta)$ rectangle.

So, that's it. Those are the essentials of the three-dimensional geometrical vector—of its analysis and of its calculus. The geometrical vector of chapters 15 and 16 and the matrix of chapters 11 through 14 have in common that they represent well-developed ways of marshaling several quantities together to a common purpose: three quantities in the specialized case of the geometrical vector; n quantities in the generalized case of the matrix. Matrices and vectors have admittedly not been easy for us to treat but after a slow start, it must be said, they have proven unexpectedly interesting. In applications, they are exceedingly significant. Matrices and vectors vastly expand the domain of physical phenomena a scientist or engineer can model. Mathematically, one cannot well manage without them.

The time nevertheless has come to change the subject. Turning the page, we will begin from the start of the next chapter to introduce a series of advanced topics that pick up where chapter 9 has left off, entering first upon the broad topic of the Fourier transform.