Calculus in Context

The Five College Calculus Project

James Callahan
Kenneth Hoffman

David Cox
Donal O'Shea
Harriet Pollatsek
Lester Senechal

DVI file created at 14:09, 31 January 2008

Advisory Committee of the Five College Calculus Project

Peter Lax, *Courant Institute, New York University*, Chairman
Solomon Garfunkel, *COMAP, Inc.*
John Neuberger, *The University of North Texas*
Barry Simon, *California Institute of Technology*
Gilbert Strang, *Massachusetts Institute of Technology*
John Truxal, *State University of New York, Stony Brook*

DVI file created at 14:09, 31 January 2008

Preface: 2008 edition

We are publishing this edition of *Calculus in Context* online to make it freely available to all users. It is essentially unchanged from the 1994 edition.

The continuing support of Five Colleges, Inc., and especially of the Five College Coordinator, Lorna Peterson, has been crucial in paving the way for this new edition. We also wish to thank the many colleagues who have shared with us their experiences in using the book over the last twenty years—and have provided us with corrections to the text.

Preface: 1994 edition

Our point of view We believe that calculus can be for our students what it was for Euler and the Bernoullis: A language and a tool for exploring the whole fabric of science. We also believe that much of the mathematical depth and vitality of calculus lies in these connections to the other sciences. The mathematical questions that arise are compelling in part because the answers matter to other disciplines as well.

The calculus curriculum that this book represents started with a "clean slate;" we made no presumptive commitment to any aspect of the traditional course. In developing the curriculum, we found it helpful to spell out our **starting points**, our **curricular goals**, our **functional goals**, and our view of the **impact of technology**. Our starting points are a summary of what calculus is really about. Our curricular goals are what we aim to convey about the subject in the course. Our functional goals describe the attitudes and behaviors we hope our students will adopt in using calculus to approach scientific and mathematical questions. We emphasize that what is missing from these lists is as significant as what appears. In particular, we did *not* not begin by asking what parts of the traditional course to include or discard.

Starting Points

- Calculus is fundamentally a way of dealing with functional relationships that occur in scientific and mathematical contexts. The techniques of calculus must be subordinate to an overall view of the underlying questions.
- Technology radically enlarges the range of questions we can explore and the ways we can answer them. Computers and graphing calculators are much more than tools for teaching the traditional calculus.

Starting Points—continued

- The concept of a dynamical system is central to science Therefore, differential equations belong at the center of calculus, and technology makes this possible *at the introductory level*.
- The process of successive approximation is a key tool of calculus, even when the outcome of the process—the limit—cannot be explicitly given in closed form.

Curricular Goals

- Develop calculus in the context of scientific and mathematical questions.
- Treat systems of differential equations as fundamental objects of study.
- Construct and analyze mathematical models.
- Use the method of successive approximations to define and solve problems.
- Develop geometric visualization with hand-drawn and computer graphics.
- Give numerical methods a more central role.

Functional Goals

- Encourage collaborative work.
- Empower students to use calculus as a language and a tool.
- Make students comfortable tackling large, messy, ill-defined problems.
- Foster an experimental attitude towards mathematics.
- Help students appreciate the value of approximate solutions.
- Develop the sense that understanding concepts arises out of working on problems, not simply from reading the text and imitating its techniques.

Impact of Technology

- Differential equations can now be solved numerically, so they can take their rightful place in the introductory calculus course.
- The ability to handle data and perform many computations allows us to explore examples containing more of the messiness of real problems.
- As a consequence, we can now deal with credible models, and the role of modelling becomes much more central to our subject.

Impact of Technology—continued

- In particular, introductory calculus (and linear algebra) now have something more substantial to offer to life and social scientists, as well as to physical scientists, engineers and mathematicians.
- The distinction between pure and applied mathematics becomes even less clear (or useful) than it may have been.

By studying the text you can see, quite explicitly, how we have pursued the curricular goals. In particular, every one of those goals is addressed within the very first chapter. It begins with questions about describing and analyzing the spread of a contagious disease. A model is built, and the model is a system of coupled non-linear differential equations. We then begin a numerical assault on those equations, and the door is opened to a solution by successive approximations.

Our implementation of the functional goals is less obvious, but it is still evident. For instance, the text has many more words than the traditional calculus book—it is a book to be read. Also, the exercises make unusual demands on students. Most exercises are not just variants of examples that have been worked in the text. In fact, the text has rather few simple "template" examples.

Shifts in Emphasis It will also become apparent to you that the text reflects substantial shifts in emphasis in comparison to the traditional course. Here are some of the most striking:

How the emphasis shifts:	
increase	decrease
concepts	techniques
geometry	algebra
graphs	formulas
brute force	elegance
numerical solutions	closed-form solutions

Euler's method is a good example of what we mean by "brute force." It is a general method of wide applicability. Of course when we use it to solve a differential equation like $y'(t) = t$, we are using a sledgehammer to crack a peanut. But at least the sledgehammer *does* work. Moreover, it works with coconuts (like $y' = y(1-y/10)$), and it will just as happily knock down a house (like $y' = \cos^2(t)$). Of course, students also see the elegant special methods that can be invoked to solve $y' = t$ and $y' = y(1-y/10)$ (separation of variables and partial fractions are discussed in chapter 11), but they understand that they are fortunate indeed when a real problem will succumb to these special methods.

Audience Our curriculum is not aimed at a special clientele. On the contrary, we think that calculus is one of the great bonds that unifies science, and all students should have an opportunity to see how the language and tools of calculus help forge that bond. We emphasize, though, that this is not a "service" course or calculus "with applications," but rather a course rich in mathematical ideas that will serve all students well, including mathematics majors. The student population in the first semester course is especially diverse. In fact, since many students take only one semester, we have aimed to make the first six chapters stand alone as a reasonably complete course. In particular, we have tried to present contexts that would be more or less broadly accessible. The emphasis on the physical sciences is clearly greater in the later chapters; this is deliberate. By the second semester, our students have gained skill and insight that allows them to tackle this added complexity.

Handbook for Instructors Working toward our curricular and functional goals has stretched us as well as our students. Teaching in this style is substantially different from the calculus courses most of us have learned from and taught in the past. Therefore we have prepared a handbook based on our experiences and those of colleagues at other schools. We urge prospective instructors to consult it.

Origins The Five College Calculus Project has a singular history. It begins almost thirty years ago, when the Five Colleges were only Four: Amherst, Mount Holyoke, Smith, and the large Amherst campus of the University of Massachusetts. These four resolved to create a new institution which would be a site for educational innovation at the undergraduate level; by 1970, Hampshire College was enrolling students and enlisting faculty.

Early in their academic careers, Hampshire students grapple with primary sources in all fields—in economics and ecology, as well as in history and literature. And journal articles don't shelter their readers from home truths: if a mathematical argument is needed, it is used. In this way, students in the life and social sciences found, sometimes to their surprise and dismay, that they needed to know calculus if they were to master their chosen fields. However, the calculus they needed was not, by and large, the calculus that was actually being taught. The journal articles dealt directly with the relation between quantities and their rates of change—in other words, with differential equations.

Confronted with a clear need, those students asked for help. By the mid-1970s, Michael Sutherland and Kenneth Hoffman were teaching a course for those students. The core of the course was calculus, but calculus as it is *used* in contemporary science. Mathematical ideas and techniques grew out of scientific questions. Given a process, students had to recast it as a model; most often, the model was a set of differential equations. To solve the differential equations, they used numerical methods implemented on a computer.

The course evolved and prospered quietly at Hampshire. More than a decade passed before several of us at the other four institutions paid some attention to it. We liked its fundamental premise, that differential equations belong at the center of calculus. What astounded us, though, was the revelation that differential equations could really *be* at the center—thanks to the use of computers.

This book is the result of our efforts to translate the Hampshire course for a wider audience. The typical student in calculus has not been driven to study calculus in order to come to grips with his or her own scientific questions—as those pioneering students had. If calculus is to emerge organically in the minds of the larger student population, a way must be found to involve that population in a spectrum of scientific and mathematical questions. Hence, calculus *in context*. Moreover, those contexts must be understandable to students with no special scientific training, and the mathematical issues they raise must lead to the central ideas of the calculus—to differential equations, in fact.

Coincidentally, the country turned its attention to the undergraduate science curriculum, and it focused on the calculus course. The National Science Foundation created a program to support calculus curriculum development. To carry out our plans we requested funds for a five-year project; we were fortunate to receive the only multi-year curriculum development grant awarded in the first year of the NSF program. This text is the outcome of our effort.

Acknowledgements

Certainly this book would have been possible without the support of the National Science Foundation and of Five Colleges, Inc. We particularly want to thank Louise Raphael who, as the first director of the calculus program at the National Science Foundation, had faith in us and recognized the value of what had already been accomplished at Hampshire College when we began our work. Five College Coordinators Conn Nugent and Lorna Peterson supported and encouraged our efforts, and Five College treasurer and business manager Jean Stabell has assisted us in countless ways throughout the project.

We are very grateful to the members of our Advisory Board: to Peter Lax, for his faith in us and his early help in organizing and chairing the Board; to Solomon Garfunkel, for his advice on politics and publishing; to Barry Simon, for using our text and giving us his thoughtful and imaginative suggestions for improving it; to Gilbert Strang, for his support of a radical venture; to John Truxal, for his detailed commentaries and insights into the world of engineering.

Among our colleagues, James Henle of Smith College deserves special thanks. Besides his many contributions to our discussions of curriculum and pedagogy, he developed the computer programs that have been so valuable for our teaching: Graph, Slinky, Superslinky, and Tint. Jeff Gelbard and Fred Henle ably extended Jim's programs to the MacIntosh and to DOS Windows and X Windows. All of this software is available on anonymous ftp at emmy.smith.edu. Mark Peterson, Robert Weaver, and David Cox also developed software that has been used by our students.

Several of our colleagues made substantial contributions to our frequent editorial conferences and helped with the writing of early drafts. We offer thanks to David Cohen, Robert Currier and James Henle at Smith; David Kelly at Hampshire; and Frank Wattenberg at the University of Massachusetts. Mary Beck, who is now at the University of Virginia, gave heaps of encouragement and good advice as a co-teacher of the earliest version of the course at Smith. Anne Kaufmann, an Ada Comstock Scholar at Smith, assisted us with extensive editorial reviews from the student perspective.

Two of the most significant new contributions to this edition are the appendix for graphing calculators and a complete set of solutions to all the exercises. From the time he first became aware of our project, Benjamin Levy has been telling us how easy and natural it would be to adapt our

Basic programs for graphing calculators. He has always used them when he taught *Calculus in Context*, and he created the appendix which contains translations of our programs for most of the graphing calculators in common use today. Lisa Hodsdon, Diane Jamrog, and Marcia Lazo have worked long hours over an entire summer to solve all the exercises and to prepare the results as LaTeX documents for inclusion in the Handbook for Instructors. We think both these contributions do much to make the course more useful to a wider audience.

We appreciate the contributions of our colleagues who participated in numerous debriefing sessions at semester's end and gave us comments on the evolving text. We thank George Cobb, Giuliana Davidoff, Alan Durfee, Janice Gifford, Mark Peterson, Margaret Robinson, and Robert Weaver at Mount Holyoke; Michael Albertson, Ruth Haas, Mary Murphy, Marjorie Senechal, Patricia Sipe, and Gerard Vinel at Smith. We learned, too, from the reactions of our colleagues in other disciplines who participated in faculty workshops on Calculus in Context.

We profited a great deal from the comments and reactions of early users of the text. We extend our thanks to Marian Barry at Aquinas College, Peter Dolan and Mark Halsey at Bard College, Donald Goldberg and his colleagues at Occidental College, Benjamin Levy at Beverly High School, Joan Reinthaler at Sidwell Friends School, Keith Stroyan at the University of Iowa, and Paul Zorn at St. Olaf College. Later users who have helped us are Judith Grabiner and Jim Hoste at Pitzer College; Allen Killpatrick, Mary Scherer, and Janet Beery at the University of Redlands; and Barry Simon at Caltech.

Dissemination grants from the NSF have funded regional workshops for faculty planning to adopt Calculus in Context. We are grateful to Donald Goldberg, Marian Barry, Janet Beery, and to Henry Warchall of the University of North Texas for coordinating workshops.

We owe a special debt to our students over the years, especially those who assisted us in teaching, but also those who gave us the benefit of their thoughtful reactions to the course and the text. Seeing what they were learning encouraged us at every step.

We continue to find it remarkable that our text is to be published the way we want it, not softened or ground down under the pressure of anonymous reviewers seeking a return to the mean. All of this is due to the bold and generous stance of W. H. Freeman. Robert Biewen, its president, understands—more than we could ever hope—what we are trying to do, and he has given us his unstinting support. Our aquisitions editors, Jeremiah Lyons and Holly Hodder, have inspired us with their passionate conviction that our book has something new and valuable to offer science education. Christine Hastings, our production editor, has shown heroic patience and grace in shaping the book itself against our often contrary views. We thank them all.

To the Student

In a typical high school math text, each section has a "technique" which you practice in a series of exercises very like the examples in the text. This book is different. In this course you will be learning to use calculus both as a tool and as a language in which you can think coherently about the problems you will be studying. As with any other language, a certain amount of time will need to be spent learning and practicing the formal rules. For instance, the conjugation of *être* must be almost second nature to you if you are to be able to read a novel—or even a newspaper—in French. In calculus, too, there are a number of manipulations which must become automatic so that you can focus clearly on the content of what is being said. It is important to realize, however, that becoming good at these manipulations is not the goal of learning calculus any more than becoming good at declensions and conjugations is the goal of learning French.

Up to now, most of the problems you have met in math classes have had definite answers such as "17," or "the circle with radius 1.75 and center at (2,3)." Such definite answers are satisfying (and even comforting). However, many interesting and important questions, like "How far is it to the planet Pluto," or "How many people are there with sickle-cell anemia," or "What are the solutions to the equation $x^5 + x + 1 = 0$" can't be answered exactly. Instead, we have ways to **approximate** the answers, and the more time and/or money we are willing to expend, the better our approximations may be. While many calculus problems do have exact answers, such problems often tend to be special or atypical in some way. Therefore, while you will be learning how to deal with these "nice" problems, you will also be developing ways of making good approximations to the solutions of the less well-behaved (and more common!) problems.

The computer or the graphing calculator is a tool that that you will need for this course, along with a clear head and a willing hand. We don't assume that you know anything about this technology ahead of time. Everything necessary is covered completely as we go along.

You can't learn mathematics simply by reading or watching others. The only way you can internalize the material is to work on problems yourself. It is by grappling with the problems that you will come to see what it is you do understand, and to see where your understanding is incomplete or fuzzy.

One of the most important intellectual skills you can develop is that of exploring questions on your own. Don't simply shut your mind down when you come to the end of an assigned problem. These problems have been designed not so much to capture the essence of calculus as to prod your thinking, to get you wondering about the concepts being explored. See if you can think up and answer variations on the problem. Does the problem suggest other questions? The ability to ask good questions of your own is at least as important as being able to answer questions posed by others.

We encourage you to work with others on the exercises. Two or three of you of roughly equal ability working on a problem will often accomplish much more than would any of you working alone. You will stimulate one another's imaginations, combine differing insights into a greater whole, and keep up each other's spirits in the frustrating times. This is particularly effective if you first spend time individually working on the material. Many students find it helpful to schedule a regular time to get together to work on problems.

Above all, take time to pause and admire the beauty and power of what you are learning. Aside from its utility, calculus is one of the most elegant and richly structured creations of the human mind and deserves to be profoundly admired on those grounds alone. Enjoy!

Contents

Chapter 7

Periodicity

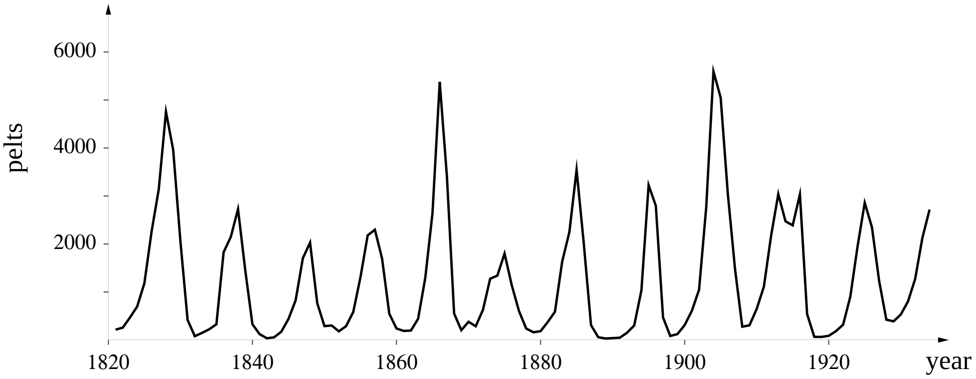
In seeking to describe and understand natural processes, we search for patterns. Patterns that repeat are particularly useful, because we can predict what they will do in the future. The sun rises every day and the seasons repeat every year. These are the most obvious examples of cyclic, or periodic, patterns, but there are many more of scientific interest, too. Periodic behavior is the subject of this chapter. We shall take up the questions of describing and measuring it. To begin, let's look at some intriguing examples of periodic or near-periodic behavior.

Many patterns are periodic

7.1 Periodic Behavior

Example 1: Populations. In chapter 4 we studied several models that describe how interacting populations might change over time. Two of those models—one devised by May and the other by Lotka and Volterra—predict that when one species preys on another, both predator and prey populations will fluctuate periodically over time. How can we tell if that actually happens in nature? Ecologists have examined data for a number of species. Some of the best evidence is found in the records of Hudson's Bay Company, which trapped fur-bearing animals in Canada for almost 200 years. The graph on the next page gives the data for the numbers of lynx pelts harvested in the Mackenzie River region of Canada during the years 1821 to 1934 (Finerty, 1980). (The lynx is a predator; its main prey is the snowshoe hare.) Clearly the numbers go up and down every 10 years in something like a periodic pattern. There is even a more complex pattern, with one large bulge and

Predator and prey populations fluctuate periodically three smaller ones, that repeats about every 40 years. Data sets like this appear frequently in scientific inquiries, and they raise important questions. Here is one: If a quantity we are studying really does fluctuate in a periodic way, why might that happen? Here is another: If there appear to be several periodic influences, what are they, and how strong are they? To explore these questions we will develop a language to describe and analyze periodic functions.



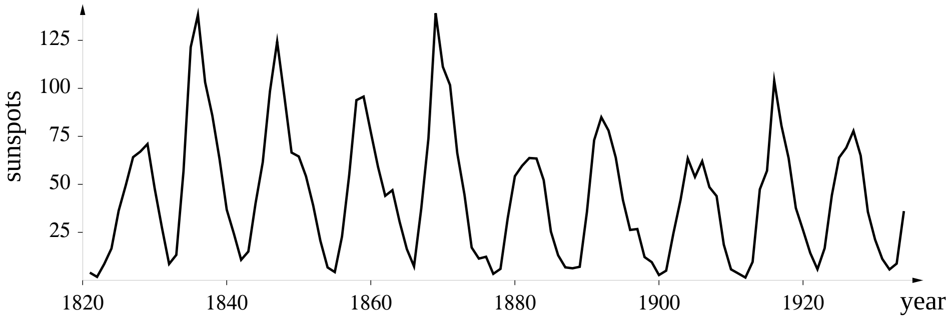
Annual harvest of lynx pelts

The position and the shape of the earth's orbit both fluctuate periodically

Example 2: The earth's orbit. The earth orbits the sun, returning to its original position after one year. This is the most obvious periodic behavior; it explains the cycle of seasons, for example. But there are other, more subtle, periodicities in the earth's orbital motion. The orbit is an ellipse which turns slowly in space, returning to its original position after about 23,000 years. This movement is called **precession**. The orbit fluctuates in other ways that have periods of 41,000 years (the **obliquity** cycle) and 95,000, 123,000, and 413,000 years (the **eccentricity** cycles).

Fluctuations in the climate appear in the geological record

Example 3: The climate. In 1941 the Serbian geophysicist Milutin Milankovitch proposed that all the different periodicities in the earth's orbit affect the climate—that is, the long-term weather patterns over the entire planet. Therefore, he concluded, there should also be periodic fluctuations in the climate, with the same periods as the earth's orbit. In fact, it is possible to test this hypothesis, because there are features of the geological record that tell us about long-term weather patterns. For example, in a year when the weather is warm and wet, rains will fill streams and rivers with mud that is eventually carried to lake bottoms. The result is a thick sediment layer. In a dry year, the sediment layer will be thinner. Over geological time, lakes dry out and their beds turn to clay or shale. By measuring the annual layers over thousands of years, we can see how the climate has varied. Other features that have been analyzed the same way are the thickness of annual ice layers in the Antarctic ice cap, the fluctuations of CO_2 concentrations in the ice caps, changes in the O^{18}/O^{16} ratio in deep-sea sediments and ice caps. In chapter 12 we will look at the results of one such study.



Daily average number of sunspots

Example 4: Sunspot cycles. The number of sunspots fluctuates, reaching a peak every 11 years or so. The graph above shows the average daily number of sunspots during each year from 1821 to 1934. Compare this with the lynx graph which covers the same years. Some earthbound events (e.g., auroras, television interference) seem to follow the same 11-year pattern. According to some scientists, other meteorological phenomena—such as rainfall, average temperature, and CO_2 concentrations in the atmosphere—are also "sunspot cycles," fluctuating with the same 11-year period. It is difficult to get firm evidence, though, because many fluctuations with different possible causes can be found in the data. Even if there is an 11-year cycle, it may be "drowned out" by the effects of these other causes.

Data can have both periodic and random influences

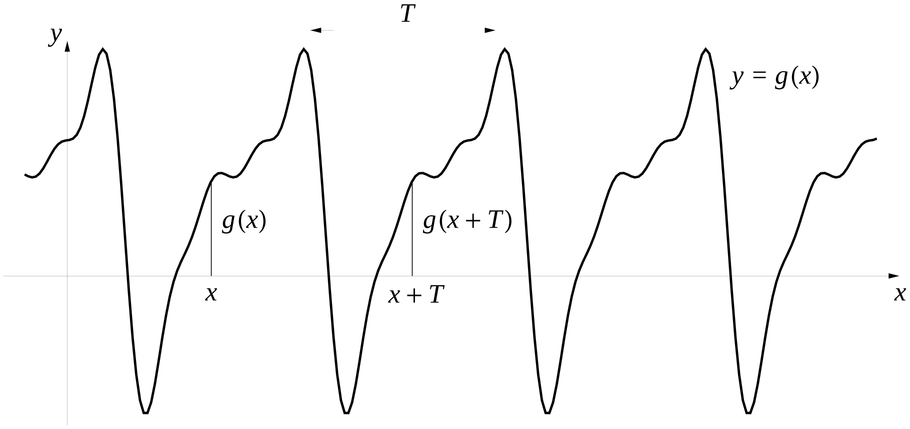
The problem of detecting periodic fluctuations in "noisy" data is one that scientists often face. In chapter 12 we will introduce a mathematical tool called the **power spectrum**, and we will use it to detect and measure periodic behavior—even when it is swamped by random fluctuations.

7.2 Period, Frequency, and the Circular Functions

We are familiar with the notions of period and frequency from everyday experience. For example, a full moon occurs every 28 days, which means that a lunar cycle has a *period* of 28 days and a *frequency* of once per 28 days. Moreover, whatever phase the moon is in today, it will be in the same phase 28 days from now. Let's see how to extend these notions to functions.

Defining a periodic function

The function $y = g(x)$ whose graph is sketched below has a pattern that repeats. The space T between one high point and the next tells us the period of this repeating pattern. There is nothing special about the high point, though. If we take *any* two points x and $x + T$ that are spaced one period apart, we find that g has the same value at those point.



(This is analogous to saying that the moon is in the same phase on any two days that are 28 days apart.) The condition $g(x + T) = g(x)$ for every x guarantees that g will be periodic. We make it the basis of our definition.

Definition. We say that a function $g(x)$ is **periodic** if there is a positive or negative number T for which

$$g(x + T) = g(x) \qquad \text{for all } x.$$

We call T a **period** of $g(x)$.

Since the graph of g repeats after x increases by T, it also repeats after x increases by $2T$, or $-3T$, or any integer multiple (positive or negative) of T. This means that a periodic function always has many periods. (That's why the definition refers to "*a* period" rather than "*the* period.") The same is true of the moon; its phases also repeat after 2×28 days, or 3×28, days. Nevertheless, we think of 28 days as *the* period of the lunar cycle, because we see the entire pattern precisely once. We can say the same for any periodic function:

A periodic function has many periods...

...but we call the smallest positive one *the* period

Definition. The period of a periodic function is its smallest positive period. It is the size of a single cycle.

Another measure of a periodic function is its frequency. Consider first the lunar cycle. Its frequency is the number of cycles—or fractions of a cycle—that occur in unit time. If we measure time in days, then the frequency is 1/28-th of a cycle per day. If we measure time in *years*, though, then the frequency is about 13 cycles per year. Here is the calculation:

Frequency

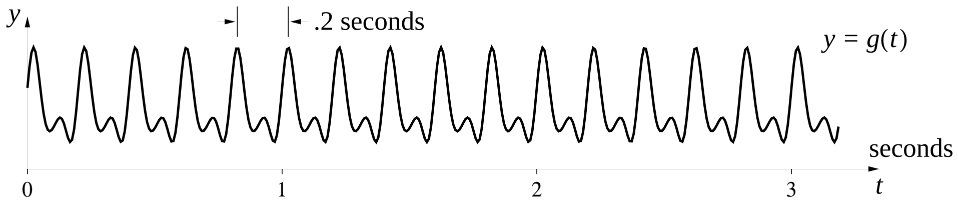
$$\frac{365 \text{ days/year}}{28 \text{ days/cycle}} \approx 13 \text{ cycles/year}.$$

Using this example as a pattern, we make the following definition.

Definition. If the function $g(x)$ is periodic, then its **frequency** is the number of cycles per unit x.

Notice that the period and the frequency of the lunar cycle are reciprocals: the period is 28 days—the time needed to complete one cycle—while the frequency is 1/28-th of a cycle per day. In the example below, t is measured in seconds and g has a period of .2 seconds. Its frequency is therefore 5 cycles per second.

The frequency of a cycle is the reciprocal of its period



Period: .2 *seconds per cycle* Frequency: 5 *cycles per second*

In general, if f is the frequency of a periodic function $g(x)$ and T is its period, then we have

$$f = \frac{1}{T} \qquad \text{and} \qquad T = \frac{1}{f}.$$

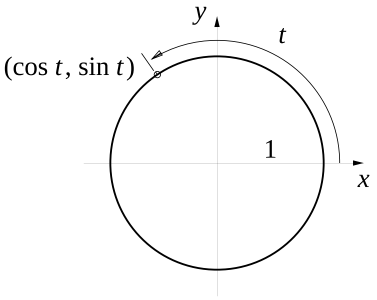
The units are also related in a reciprocal fashion: if the period is measured in seconds, then the frequency is measured in cycles per second.

The units for measuring frequency over time

Because many quantities fluctuate periodically over time, the input variable of a periodic function will often be *time*. If time is measured in seconds, then frequency is measured in "cycles per second." The term **Hertz** is a special unit used to measure time frequencies; it equals one cycle per second. Hertz is abbreviated Hz; thus a **kilohertz** (kHz) and a **megahertz** (MHz) are 1,000 and 1,000,000 cycles per second, respectively. This unit is commonly used to describe sound, light, radio, and television waves. For example, an orchestra tunes to an A at 440 Hz. If an FM radio station broadcasts at 88.5 MHz, this means its carrier frequency is 88,500,000 cycles per second.

Functions can be periodic over other units as well

Quantities may also be periodic in other dimensions. For instance, a scientist studying the phenomenon of ripple formation in a river bed might be interested in the function $h(x)$ measuring the height of a ripple as a function of its distance x along the river bed. This would lead to a function of period, say, 10 inches and corresponding frequency of .1 cycle per inch.



Circular functions. While there are innumerable examples of periodic functions, two in particular are considered basic: the sine and the cosine. They are called circular functions because they are defined by means of a circle. To be specific, take the circle of radius 1 centered at the origin in the x, y-plane. Given any real number t, measure a distance of t units around the circumference of the circle. Start on the positive x-axis, and measure counterclockwise if t is positive, clockwise if t is negative. The coordinates of the point you reach this way are, by definition, the cosine and the sine functions of t, respectively:

$$x = \cos(t),$$
$$y = \sin(t).$$

Why $\cos t$ and $\sin t$ are periodic

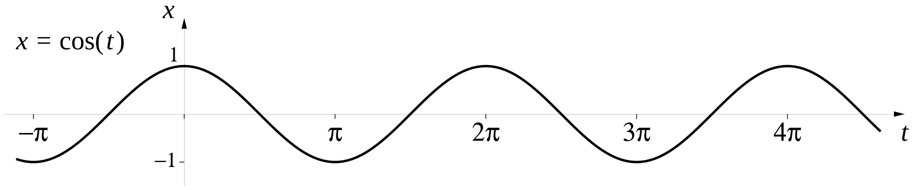
The whole circumference of the circle measures 2π units. Therefore, if we add 2π units to the t units we have already measured, we will arrive back at the same point on the circle. That is, we get to the same point on the circle by measuring either t or $t + 2\pi$ units around the circumference. We can describe the coordinates of this point two ways:

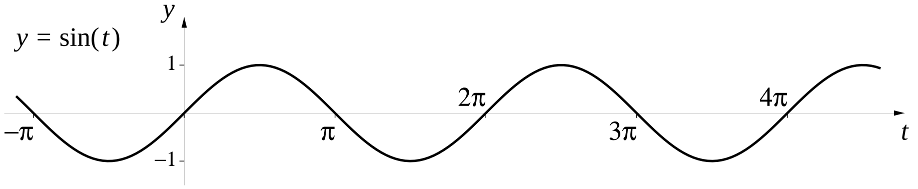
$$(\cos(t), \sin(t)) \qquad \text{or} \qquad (\cos(t + 2\pi), \sin(t + 2\pi)).$$

Thus

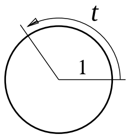
$$\cos(t + 2\pi) = \cos(t) \qquad \sin(t + 2\pi) = \sin(t),$$

so $\cos(t)$ and $\sin(t)$ are both periodic, and they have the same period, 2π. Here are their graphs. By reading their slopes we can see $(\sin t)' = \cos t$ and $(\cos t)' = -\sin t$.





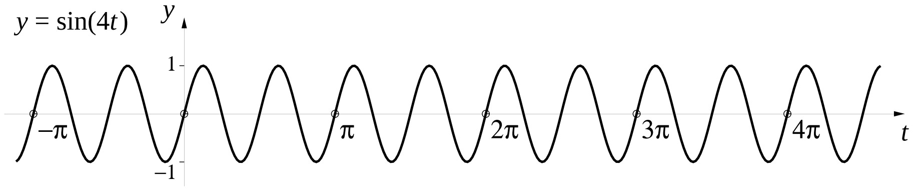
Radian measure



The circular functions are constructed without reference to angles; the variable t is measured around the circumference of a circle (of radius 1). Nevertheless, we *can* think of t as measuring an angle, as shown at the right. In this case, t is called the **radian measure** of the angle. The units are very different from the degree measurement of an angle: an angle of 1 radian is much larger than an angle of 1 degree. The radian measure of a 90° angle is $\pi/2 \approx 1.57$, for instance. If we thought of t as an angle measured in degrees, the slope of $\sin(t)$ would equal $.017 \cos t$! (See the exercises.) Only when we measure t in radians do we get a simple result: $(\sin t)' = \cos t$. This is why we always measure angles in radians in calculus.

Changing the frequency

Compare the graph of $y = \sin(t)$ above with that of $y = \sin(4t)$, below.



Their scales are identical, so it is clear that the frequency of $\sin(4t)$ is four times the frequency of $\sin(t)$. The general pattern is described in the following table.

function		period	frequency
$\sin(t)$	$\cos(t)$	2π	$1/2\pi$
$\sin(4t)$	$\cos(4t)$	$2\pi/4$	$4/2\pi$
$\sin(bt)$	$\cos(bt)$	$2\pi/b$	$b/2\pi$

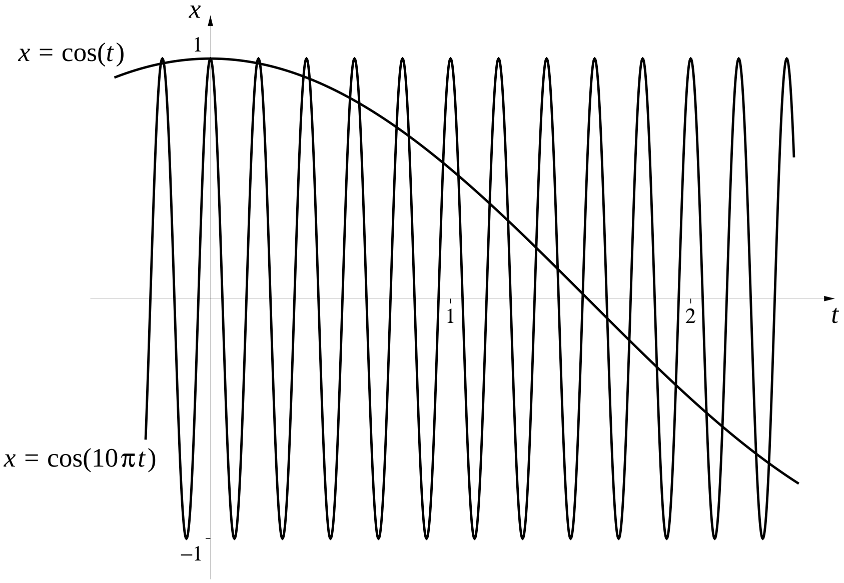
Notice that it is the *frequency*—not the period—that is increased by a factor of b when we multiply the input variable by b.

Constructing a circular function with a given frequency

By using the information in the table, we can construct circular functions with any period or frequency whatsoever. For instance, suppose we wanted a cosine function $x = \cos(bt)$ with a frequency of 5 cycles per unit t. This means

$$5 = \text{frequency} = \frac{b}{2\pi},$$

which implies that we should set $b = 10\pi$ and $x = \cos(10\pi t)$. In order to see the high-frequency behavior of this function better, we magnify the graph a bit. In the figure below, you can compare the graphs of $x = \cos(10\pi t)$ and $x = \cos(t)$ directly. We still have equal scales on the horizontal and vertical axes. Finally, notice that $\cos(10\pi t)$ has exactly 5 cycles on the interval $0 \le t \le 1$.



Frequency ω

We will denote the frequency by ω, the lower case letter *omega* from the Greek alphabet. If

$$\omega = \text{frequency} = \frac{b}{2\pi},$$

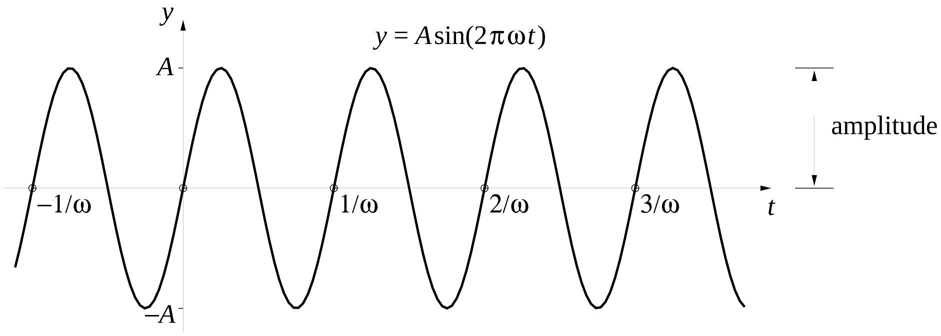
then $b = 2\pi\omega$. Therefore, the basic circular functions of frequency ω are $\cos(2\pi\omega t)$ and $\sin(2\pi\omega t)$.

Amplitude

Suppose we take the basic sine function $\sin(2\pi\omega t)$ of frequency ω and multiply it by a factor A:

$$y = A\sin(2\pi\omega t).$$

The graph of this function oscillates between $y = -A$ and $y = +A$. The number A is called the **amplitude** of the function.



The sine function of amplitude A and frequency ω

Physical interpretations. Sounds are transmitted to our ears as fluctuations in air pressure. Light is transmitted to our eyes as fluctuations in a more abstract medium—the electromagnetic field. Both kinds of fluctuations can be described using circular functions of time t. The amplitude and the frequency of these functions have the physical interpretations given in the following table.

	amplitude	frequency	frequency range
sound	loudness	pitch	10–15000 Hz
light	intensity	color	$4 \times 10^{14} - 7.5 \times 10^{14}$ Hz

Exercises

Circular functions

1. Choose ω so that the function $\cos(2\pi\omega t)$ has each of the following periods.

a) 1 b) 5 c) 2π d) π e) $1/3$

2. Determine the period and the frequency of the following functions.

a) $\sin(x)$, $\sin(2x)$, $\sin(x) + \sin(2x)$

b) $\sin(2x)$, $\sin(3x)$, $\sin(2x) + \sin(3x)$

c) $\sin(6x)$, $\sin(9x)$, $\sin(6x) + \sin(9x)$

3. Suppose a and b are positive integers. Describe how the periods of $\sin(ax)$, $\sin(bx)$, $\sin(ax) + \sin(bx)$ are related. (As the previous exercise shows, the relation between the periods depends on the relation between a and b. Make this clear in your explanation.)

4. a) What are the amplitude and frequency of $g(x) = 5\cos(3x)$?

b) What are the amplitude and frequency of $g'(x)$?

5. a) Is the antiderivative $\displaystyle\int_0^x 5\cos(3t)\,dt$ periodic?

b) If so, what are its amplitude and frequency?

6. Use the definition of the circular functions to explain why

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

hold for all values of t. Describe how these properties are reflected in the graphs of the sine and cosine functions.

7. a) What is the average value of the function $\sin(s)$ over the interval $0 \le s \le \pi$? (This is a half-period.)

b) What is the average value of $\sin(s)$ over $\pi/2 \le s \le 3\pi/2$? (This is also a half-period.)

c) What is the average value of $\sin(s)$ over $0 \le s \le 2\pi$? (This is a full period.)

d) Let c be any number. Find the average value of $\sin(s)$ over the full period $c \le s \le c + 2\pi$.

e) Your work should demonstrate that the average value of $\sin(s)$ over a full period does not depend on the point c where you begin the period. Does it? Is the same true for the average value over a *half* period? Explain.

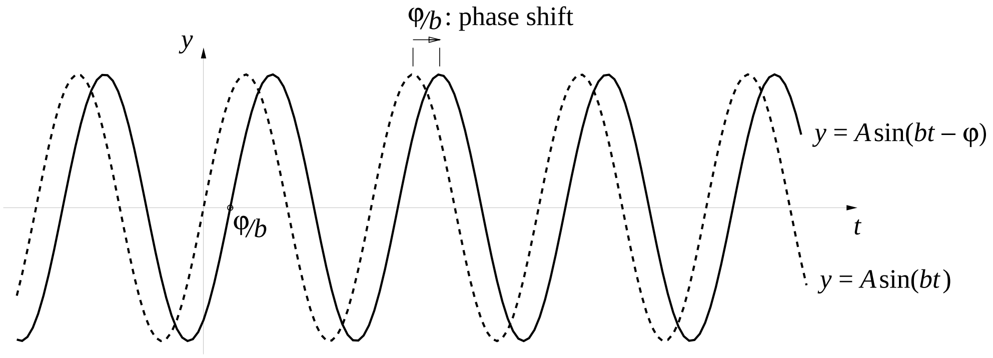
8. a) What is the period T of $P(t) = A\sin(bt)$?

b) Let c be any number. Find the average value of $P(t)$ over the full period $c \le t \le c + T$. Does this value depend on the choice of c?

c) What is the average value of $P(t)$ over the half-period $0 \le t \le T/2$?

Phase

There is still another aspect of circular functions to consider besides amplitude and frequency. It is called *phase difference*. We can illustrate this with the two functions graphed below. They have the same amplitude and frequency, but differ in phase.



Specifically, the variable u in the expression $\sin(u)$ is called the **phase**. In the dotted graph the phase is $u = bt$, while in the solid graph it is $u = bt - \varphi$. They differ in phase by $bt - (bt - \varphi) = \varphi$. In the exercises you will see why a **phase difference** of φ produces a shift—which we call a **phase shift**—of φ/b in the graphs. (φ is the Greek letter *phi*.)

9. The functions $\sin(x)$ and $\cos(x)$ have the same amplitude and frequency; they differ only in phase. In other words,

$$\cos(x) = \sin(x - \varphi)$$

for an appropriately chosen phase difference φ. What is the value of φ?

10. The functions $\sin(x)$ and $-\sin(x)$ *also* differ only in phase. What is their phase difference? In other words, find φ so that

$$\sin(x - \varphi) = -\sin(x).$$

[Note: A circular function and its negative are sometimes said to be "180 degrees out of phase." The value of φ you found here should explain this phrase.]

11. What is the phase difference between $\sin(x)$ and $-\cos(x)$?

12. a) Graph $y = \sin(t)$ and $y = \sin(t - \pi/3)$ on the same plane.

b) What is the phase difference between these two functions?

c) What is the phase shift between their graphs?

13. a) Graph together on the same coordinate plane $y = \cos(t)$ and $y = \cos(t + \pi/4)$.

b) What is the phase difference between these two functions?

c) What is the phase shift between their graphs?

14. We know $y = \cos(t)$ has a maximum at the origin. Determine the point closest to the origin where $y = \cos(t + \pi/4)$ has *its* maximum. Is the second maximum shifted from the first by the amount of the phase shift you identified in the previous question?

15. Repeat the last two exercises for the pair of functions $y = \cos(2t)$ and $\cos(2t + \pi/4)$. Is the phase difference equal to the phase shift in this case?

16. Verify that the graph of $y = A\sin(bt - \varphi)$ crosses the t-axis at the point $t = \varphi/b$. This shows that $A\sin(bt - \varphi)$ is "phase-shifted" by the amount φ/b in relation to $A\sin(bt)$. (Refer to the graph on page 429.)

17. a) At what point nearest the origin does the function $A\cos(bt - \varphi)$ reach its maximum value?

b) Explain why this shows $A\cos(bt - \varphi)$ is "phase-shifted" by the amount φ/b in relation to $A\cos(bt)$.

18. a) Let $f(x) = \sin(x) - .7\cos(x)$. Using a graphing utility, sketch the graph of $f(x)$.

b) The function $f(x)$ is periodic. What is its period? From your graph, estimate its amplitude.

c) In fact, $f(x)$ can be viewed as a "phase-shifted" sine function:

$$f(x) = A\sin(bx - \varphi).$$

From your graph, estimate the phase difference φ and the amplitude A.

19. a) For each of the values $\varphi = 0,\ \pi/4,\ \pi/2,\ 3\pi/4,\ \pi$, sketch the graph $y = \sin(x) \cdot \sin(x - \varphi)$ over the interval $0 \leq x \leq 2\pi$. Put the five graphs on the same coordinate plane.

b) For which graphs is the average value positive, for which is it negative, and for which is it 0? Estimate by eye.

20. The purpose of this exercise is to determine the average value

$$F(\varphi) = \frac{1}{2\pi}\int_0^{2\pi} \sin(x)\sin(x - \varphi)\,dx$$

for an *arbitrary* value of the parameter φ. To stress that the average value is actually a function of φ, we have written it as $F(\varphi)$. Here is one way to determine a formula for $F(\varphi)$ in terms of φ. First, using a "sum of two angles" formula and exercise 6, above, write

$$\sin(x - \varphi) = \cos(\varphi)\sin(x) - \sin(\varphi)\cos(x)$$

Then consider

$$\frac{1}{2\pi}\left[\cos(\varphi)\int_0^{2\pi}(\sin(x))^2 dx - \sin(\varphi)\int_0^{2\pi}\sin(x)\cos(x)\,dx\right],$$

and determine the values of the two integrals separately.

21. a) Sketch the graph of the *average value function* $F(\varphi)$ you found in the previous exercise. Use the interval $0 \leq \varphi \leq \pi$.

b) In exercise 19 you estimated the value of $F(\varphi)$ for five specific values of φ. Compare your estimates with the exact values that you can now calculate using the formula for $F(\varphi)$.

22. Sketch the graph of $y = \cos(x)\sin(x-\varphi)$ for each of the following values of φ: 0, $\pi/2$, $2\pi/3$, π. Use the interval $0 \le x \le 2\pi$. Estimate by eye the average value of each function over that interval.

23. a) Obtain a formula for the average value function

$$G(\varphi) = \frac{1}{2\pi}\int_0^{2\pi} \cos(x)\sin(x-\varphi)\,dx.$$

and sketch the graph of $G(\varphi)$ over the interval $0 \le \varphi \le \pi$.

b) Use your formula for $G(\varphi)$ to compute the average value of the function $\cos(x)\sin(x-\varphi)$ exactly for $\varphi = 0$, $\pi/2$, $2\pi/3$, π. Compare these values with your estimates in the previous exercise.

24. How large a phase difference φ is needed to make the graphs of $y = \sin(3x)$ and $y = \sin(3x-\varphi)$ coincide?

25. Sketch the graphs of the following functions.

a) $y = 3\sin(2x - \pi/6) - 1$

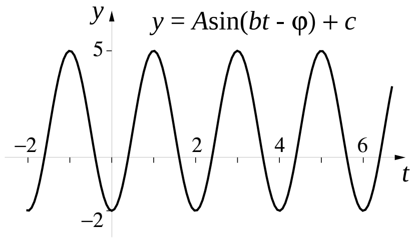
b) $y = 4\sin(2x - \pi) + 2$.

c) $y = 4\sin(2x + \pi) + 2$.

26. The function whose graph is sketched at the right has the form

$$G(x) = A\sin(bx - \varphi) + C.$$

Determine the values of A, b, C, and φ.



27. Write equations for three different functions that all have amplitude 4, period 5, and whose graphs pass through the point $(6,7)$. Be sure the functions are really different—if $g(t)$ is one solution, then $h(t) = g(t+5)$ would really be just the same solution.

Derivatives with degrees

28. a) In this exercise measure the angle θ in degrees. Estimate the derivative of $\sin(\theta)$ at $\theta = 0°$ by calculating $\sin(\theta)/\theta$ for $\theta = 2°$, $1°$, $.5°$.

b) Estimate the derivative of $\sin(\theta)$ at $\theta = 60°$ in a similar way. Is $(\sin(30°))' = \cos(30°)$?

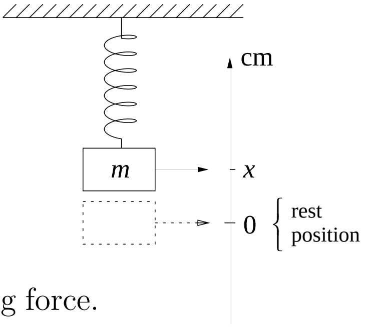
29. a) Your calculations in the previous exercise should support the claim that $(\sin(\theta))' = k\cos(\theta)$ for a particular value of k, when θ is measured in degrees. What is k, approximately?

b) If t is the radian measure of an angle, and if θ is its degree measure, then θ will be a function of t. What is it? Now use the chain rule to get a precise expression for the constant k.

7.3 Differential Equations with Periodic Solutions

The models of predator–prey interactions constructed by Lotka–Volterra and May (see chapter 4) provide us with examples of systems of differential equations that have periodic solutions. Similar examples can be found in many areas of science. We shall analyze some of them in this section. In particular, we will try to understand how the frequency and the amplitude of the periodic solutions depend on the parameters given in the model.

Oscillating Springs



We want to study the motion of a weight that hangs from the end of a spring. First let the weight come to rest. Then pull down on it. You can feel the spring pulling it back up. If you push up on the weight, the spring (and gravity) push it back down. The force you feel is called the **spring force**. Now release the weight; it will move. We'll assume that the only influence on the motion is the spring force. (In particular, we will ignore the force of friction.) With this assumption we can construct a model to describe the motion. We'll suppose the weight has a mass of m grams, and it is x centimeters above its rest position after t seconds. (If the weight goes below the rest position, then x will be negative.)

The linear spring

A linear spring

The simplest assumption we can reasonably make is that the spring force is proportional to the amount x that the spring has been displaced:

$$\text{force} = -c\,x.$$

In this case the spring is said to be **linear**. The multiplier c is called the **spring constant**. It is a positive number that varies from one spring to another. The minus sign tells us the force pushes down if $x > 0$, and it pushes up if $x < 0$. Because this model describes an oscillating spring governed by a linear spring force, it is called the **linear oscillator**.

Newton's laws of motion

To see how the spring force affects the motion of the weight, we use Newton's laws. In their simplest form, they say that the force acting on a body is the product of its mass and its acceleration. Suppose $v = dx/dt$ is the velocity of the weight in cm/sec, and dv/dt is its acceleration in cm/sec^2. Then

$$\text{force} = m\frac{dv}{dt} \quad \text{gm-cm/sec}^2.$$

If we equate our two expressions for the force, we get

$$m\frac{dv}{dt} = -c\,x \qquad \text{or} \qquad \frac{dv}{dt} = -b^2x \quad \text{cm/sec}^2,$$

where we have set $c/m = b^2$. It is more convenient to write c/m as b^2 here, because then $b = \sqrt{c/m}$ itself will be measured in units of 1/sec. (To see why, note that $-b^2x$ is measured in units of cm/sec^2.)

The linear oscillator

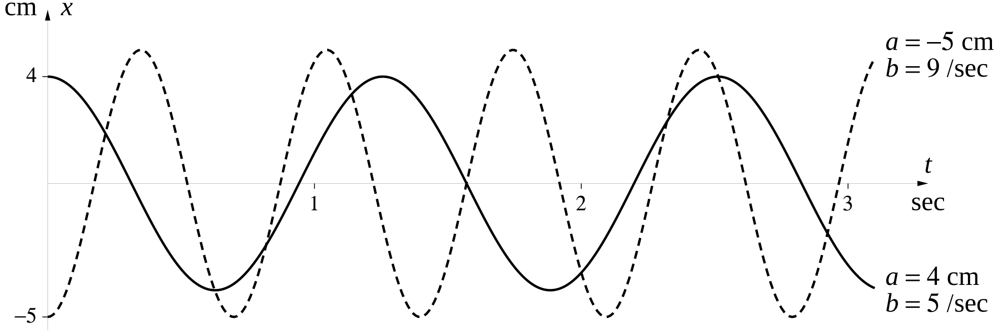
Suppose we move the weight to the point $x = a$ cm on the scale, hold it motionless for a moment, and then release it at time $t = 0$ sec. This gives us the initial value problem

$$\begin{aligned} x' &= v, & x(0) &= a, \\ v' &= -b^2x, & v(0) &= 0. \end{aligned}$$

The solution with fixed parameters

If we give the parameters a and b specific values, we can solve this initial value problem using Euler's method. The figure below shows the solution $x(t)$ for two different sets of parameter values:

$$\begin{aligned} a &= 4 \text{ cm}, & a &= -5 \text{ cm}, \\ b &= 5 \text{ per sec}, & b &= 9 \text{ per sec}. \end{aligned}$$



The graphs were made in the usual way, with the differential equation solver of a computer. They indicate that the weight bounces up and down in a periodic fashion. The amplitude of the oscillation is precisely a, and the frequency appears to be linked directly to the value of b. For instance, when $b = 9$ /sec, the motion completes just under 3 cycles in 2 seconds. This is a frequency of slightly less than 1.5 cycles per second. When $b = 5$ /sec, the motion undergoes roughly 2 cycles in 2.5 seconds, a frequency of about .8 cycles per second. If the frequency is indeed proportional to b, the multiplier must be about 1/6:

$$\text{frequency} \approx \frac{b}{6} \text{ cycles/sec.}$$

The solution for arbitrary parameter values

We can get a better idea how the parameters in a problem affect the solution if we solve the problem with a method that doesn't require us to fix the values of the parameters in advance. This point is discussed in chapter 4.2, pages 214–218. It is particularly useful if we can express the solution by a *formula*, which it turns out we can do in this case. To get a formula, let us begin by noticing that

$$(x')' = v' = -b^2 x.$$

In other words, $x(t)$ is a function whose second derivative is the negative of itself (times the constant b^2). This suggests that we try

$$x(t) = \sin(bt) \qquad \text{or} \qquad x(t) = \cos(bt).$$

You should check that $x'' = -b^2 x$ in both cases.

Turn now to the initial conditions. Since $\sin(0) = 0$, there is no way to modify $\sin(bt)$ to make it satisfy the condition $x(0) = a$. However,

$$x(t) = a\cos(bt)$$

does satisfy it. Finally, we can use the differential equation $x' = v$ to define $v(t)$:

$$v(t) = (a\cos(bt))' = -ab\sin(bt).$$

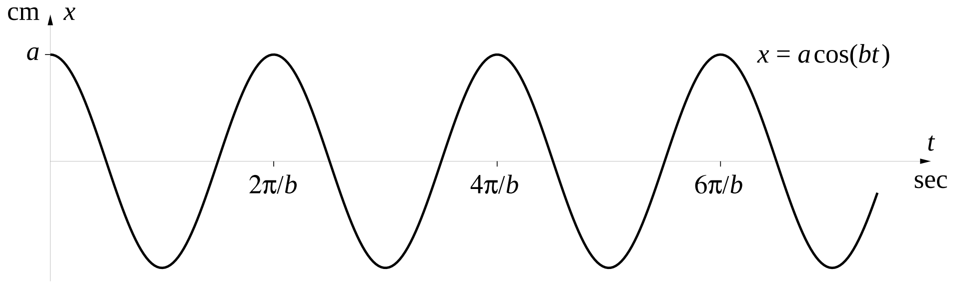
Notice that $v(0) = -ab\sin(0) = 0$, so the second initial condition is satisfied.

The formula *proves* the motion is periodic

In summary, we have a formula for the solution that incorporates the parameters. With this formula we see that the motion is really periodic—a fact that Euler's method could only suggest. Furthermore, the parameters determine the amplitude and frequency of the solution in the following way:

$$\begin{aligned} \text{position}: &\quad a\cos(bt) \text{ cm from rest after } t \text{ sec} \\ \text{amplitude}: &\quad a \text{ cm} \\ \text{frequency}: &\quad b/2\pi \text{ cycles/sec} \end{aligned}$$

We can see the relation between the motion and the parameters in the graph below (in which we take $a > 0$).



Graph of the general linear oscillator

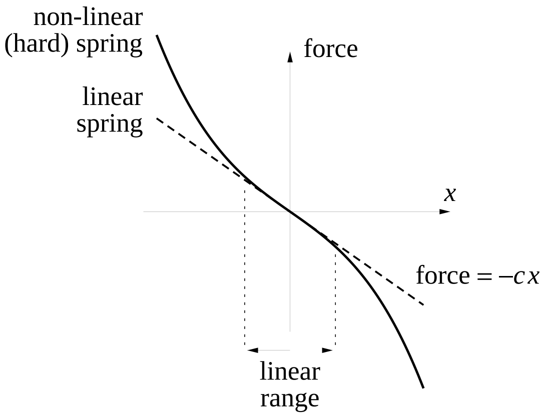
Here are some further properties of the motion that follow from our formula for the solution. Recall that the parameter b depends on the mass m of the weight and the spring constant c: $b^2 = c/m$.

- The amplitude depends only on the initial conditions, not on the mass m or the spring constant c.
- The frequency depends only on the mass and the spring constant, not on the initial amplitude.

These properties are a consequence of the fact that the spring force is *linear*. As we shall see, a non-linear spring and a pendulum move differently.

The non-linear spring

The harder you pull on a spring, the more it stretches. If the stretch is exactly proportional to the pull (i.e., the force), the spring is linear. In other words, to double the stretch you must double the force. Most springs behave this way when they are stretched only a small amount. This is called their **linear range**. Outside that range, the relation is more complicated. One possibility is that, to double the stretch, you must increase the force by *more than* double. A spring that works this way is called a **hard spring**. The graph at the left shows the relation between the applied force and the displacement (or stretch x) of a hard spring.



In a *nonlinear* spring, force is no longer proportional to displacement. Thus, if we write

$$\text{force} = -c\,x$$

we must allow the multiplier c to depend on x. One simple way to achieve this is to replace c by $c + \gamma x^2$. (We use x^2 rather than just x to ensure that $-x$ will have the same effect as $+x$. The multiplier γ is the Greek letter *gamma*.) Then

$$\text{force} = -c\,x - \gamma x^3.$$

Since force $= m\,dv/dt$ as well, we have

$$m\frac{dv}{dt} = -c\,x - \gamma x^3 \qquad \text{or} \qquad \frac{dv}{dt} = -b^2 x - \beta x^3 \quad \text{cm/sec}^2.$$

Here $b^2 = c/m$ and $\beta = \gamma/m$. By taking the same initial conditions as before, we get the following initial value problem:

A non-linear oscillator

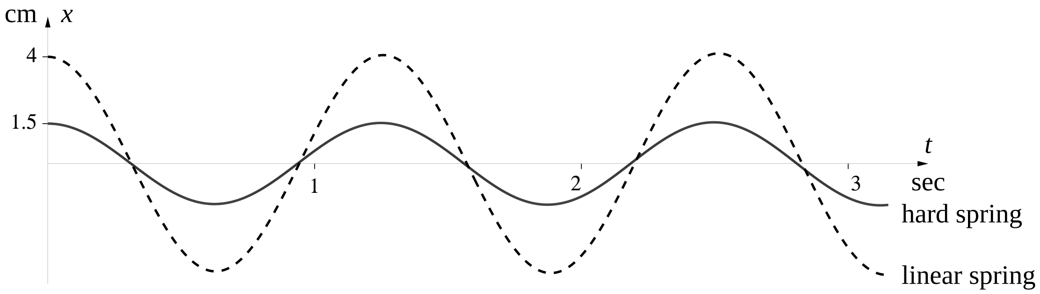
$$\begin{aligned} x' &= v, & x(0) &= a, \\ v' &= -b^2 x - \beta x^3, & v(0) &= 0. \end{aligned}$$

To solve this problem using Euler's method, we must fix the values of the three parameters. For the two parameters that determine the spring force, we choose:

$$b = 5 \text{ per sec} \qquad \beta = .2 \text{ per cm}^2\text{-sec}^2.$$

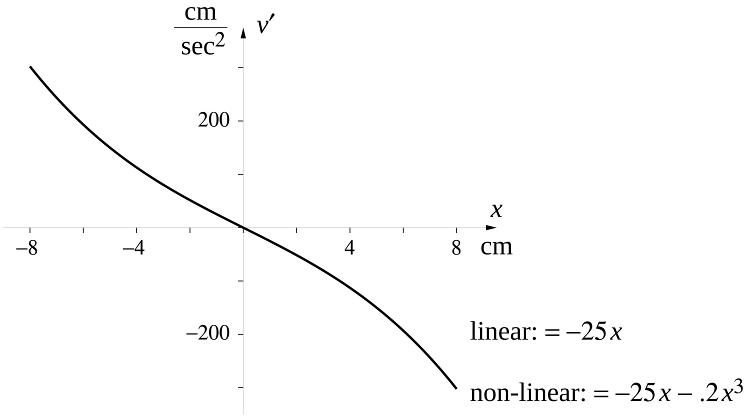
Comparing a hard spring to a linear spring

We have deliberately chosen b to have the same value it did for our first solution to the linear problem. In this way, we can compare the non-linear spring to the linear spring that has the same spring constant. We do this in the figure below. The dashed graph shows the linear spring when its initial amplitude is $a = 4$ cm. The solid graph shows the hard spring when its initial amplitude is $a = 1.5$ cm. Note that the two oscillations have the same frequency.



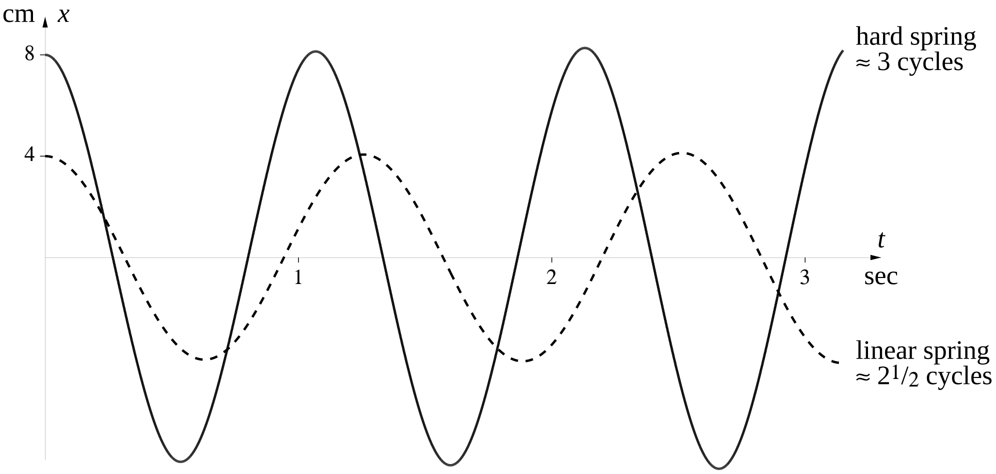
The effect of amplitude on acceleration

The non-linear spring behaves like the linear one because the amplitudes are small. To understand this reason, we must compare the accelerations of the two springs. For the linear spring we have $v' = -25x$, while for the non-linear spring, $v' = -25x - .2x^3$. As the following graph shows, these expressions are approximately equal when the amplitude x lies between $+2$ cm and -2 cm. In other words, the linear range of the hard spring is $-2 \le x \le 2$ cm. Since the initial amplitude was 1.5 cm—well within the linear range—the hard spring acts like a linear one. In particular, its frequency is approximated closely by the formula $b/2\pi$ cycles per second. This is $5/2\pi \approx .8$ Hz.



Large-amplitude oscillations

A different set of circumstances is reflected in the following graph. The hard spring has been given an initial amplitude of 8 cm. As the graph of v' shown above indicates, the hard spring experiences an acceleration about 50% greater than the linear spring at the that amplitude.

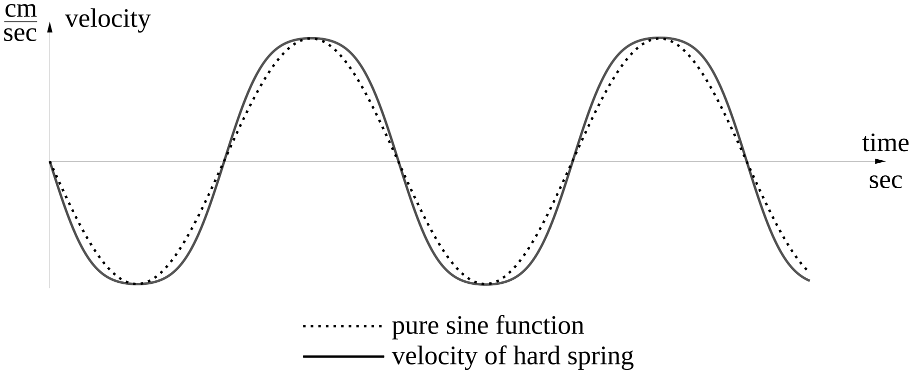


The frequency of a non-linear spring depends on its amplitude

As a consequence, the hard spring oscillates with a noticeably higher frequency! It completes 3 cycles in the time it takes the linear spring to complete $2\frac{1}{2}$—or 6 cycles while the linear spring completes 5. The frequency of the hard spring is therefore about 6/5-th the frequency of the linear spring, or $6/5 \times 5/2\pi = 3/\pi \approx .95$ Hz.

A mathematical comment

The solutions of the non-linear spring problem still look like cosine functions, but they're not. It's easier to see the difference if we take a large amplitude solution, and look at velocity instead of position. In the graph below you can see how the velocity of a hard spring differs from a pure sine function of the same period and amplitude. Since there are no sine or cosine functions here, we can't even yet be sure that the motion of a non-linear spring is truly periodic! We will prove this, though, in the next section by using the notion of a **first integral**.



Other non-linear springs

There are other ways we might have modified the basic equation $v' = -b^2x$ to make the spring non-linear. The formula $v' = -b^2x - \beta x^3$ is only one possibility. Incidentally, our study of a hard spring was based on choosing $\beta > 0$ in this formula. Suppose we choose $\beta < 0$ instead. As you will see in the exercises, this is a **soft** spring: we can double the stretch in a soft spring by using *less than* double the force. The pendulum, which we will study next, behaves like a soft spring.

Although we can use sine and cosine functions to solve the *linear* oscillator problem, there are, in general, no formulas for the solutions to the *non-linear* oscillator problems. We must use numerical methods to find their graphs—as we have done in the last three pages.

The harmonic oscillator

The basic differential equation for a linear spring is also used to model a vibrating string. Think of a tightly-stretched wire, like a piano string or a guitar string. Let x be the distance the center of the string has moved from rest at any instant t. The larger x is, the more strongly the tension on the string will pull it back towards its rest position. Since x is usually very small, it makes sense to assume that this "restoring force" is a linear function of x: $-c\,x$. If v is the velocity of the string, then $mv' = -c\,x$ by Newton's laws of motion. Because of the connection between vibrating strings and music, this differential equation is called the **harmonic oscillator**.

The Sine and Cosine Revisited

The sine and cosine functions first appear in trigonometry, where they are defined for the acute angles of a right triangle. Negative angles and angles larger than 90°, are outside their domain. This is a serious limitation. To overcome it, we redefine the sine and cosine on a circle. The main consequence of this change is that the sine and cosine become *periodic*.

A computable definition of the sine and cosine

However, neither circles nor triangles are particularly useful if we want to *calculate* the values of the sine or the cosine. (How would you use one of them to determine sin(1) to four—or even two—decimal places accuracy?) Our experience with the harmonic oscillator gives yet another way to define the sine and the cosine functions—a way that conveys computational power.

The idea is simple. With hindsight we know that $u = \sin(t)$ and $v = \cos(t)$ are the solutions to the initial value problem

$$\begin{aligned} u' &= v, & u(0) &= 0, \\ v' &= -u & v(0) &= 1. \end{aligned}$$

Now make a fresh start with this initial value problem, and *define* $u = \sin(t)$ and $v = \cos(t)$ to be its solution! Then we can calculate sin(1), for instance, by Euler's method. Here is the result.

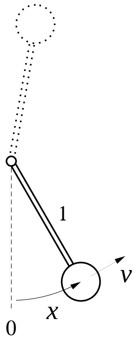
number of steps	estimate of sin(1)
100	.845671
1 000	.841892
10 000	.841513
100 000	.841475
1 000 000	.841471

So we can say sin(1) = .8415 to four decimal places accuracy.

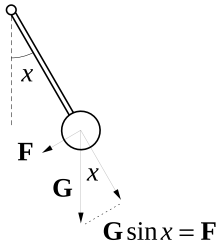
Our point of view here is that *differential equations define functions*. In chapter 10, we shall consider still another method for defining and calculating these important functions, using infinite series.

The Pendulum

We are going to study the motion of a pendulum that can swing in a full 360° circle. To keep the physical details as simple as possible, we'll assume its mass is 1 unit, and that all the mass is concentrated in the center of the pendulum bob, 1 unit from the pivot point. Assume that the pendulum is x units from its rest position at time t, where x is measured around the circular path that the bob traces out. Assume the velocity is v. Take counterclockwise positions and velocities to be positive, clockwise ones to be negative. When the pendulum is at rest we have $x = v = 0$.



When the pendulum is moving, there must be forces at work. Let's ignore friction, as we did with the spring. The force that pulls the pendulum back toward the rest position is gravity. However, gravity itself—**G** in the figure at the right—pulls straight down. Part of the pull of **G** works straight along the arm of the pendulum, and is resisted by the pivot. (If not, the pendulum would be pulled out of the pivot and fall to the floor!) It is the other part, labelled **F**, that moves the pendulum sideways.



The size of **F** depends on the position x of the pendulum. When $x = 0$, the sideways force **F** is zero. When $x = \pi/2$ (the pendulum is horizontal), the entire pull of **G** is "sideways", so $\mathbf{F} = \mathbf{G}$. To see how **F** depends on x in general, note first that we can think of x as the radian measure of the angle between the pendulum and the vertical (because x is measured around a circle of radius 1). In the small right triangle, the hypotenuse is **G** and the side opposite the angle x is exactly as long as **F**. By trigonometry, $\mathbf{F} = \mathbf{G}\sin x$.

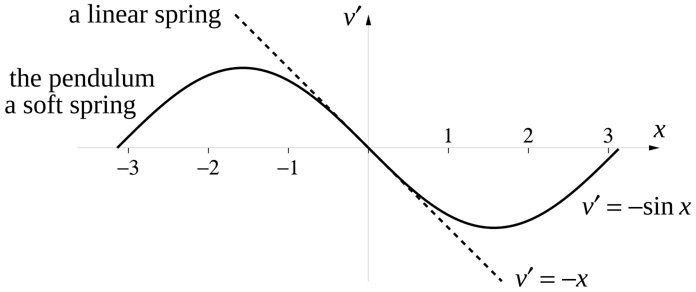
Newton's laws produce a model of the pendulum

Let's choose units which make the size of **G** equal to 1. Then the size of **F** is simply $\sin x$. Since **F** points in the clockwise (or negative) direction when x is positive, we must write $\mathbf{F} = -\sin x$. According to Newton's laws of motion, the force **F** is the product of the mass and the acceleration of the pendulum. Since the mass is 1 unit and the acceleration is $v' = x''$, we finally get $x'' = -\sin x$, or

$$x' = v \qquad v' = -\sin x.$$

The pendulum is a soft spring

Now that we have an explicit description of the restoring force, we can see that the pendulum behaves like a non-linear spring. However, it is true that doubling the displacement x always *less than* doubles the force, as the graph below demonstrates. Thus the pendulum is like a **soft** spring.



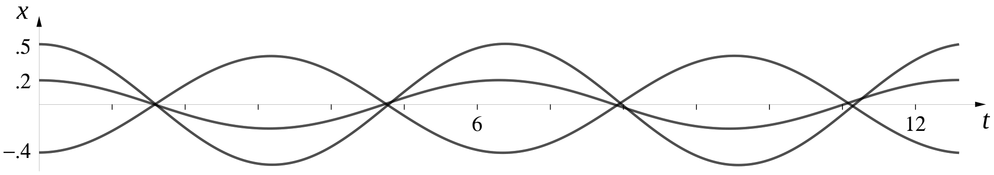
Because a swinging pendulum is used keep time, it is important to control the period of the swing. Physics analyzes how the period depends on the pendulum's length and mass. We will confine ourselves to analyzing how the period depends on its amplitude.

Small-amplitude oscillations

Let's draw on our experience with springs. According to the graph above, the restoring force of the pendulum is essentially linear for small amplitudes—say, for $-.5 \le x \le .5$ radians. Therefore, if the amplitude stays small, it is reasonable to expect that the pendulum will behave like a linear oscillator. As the graph indicates, the differential equation of the linear oscillator is $v' = -x$. This is of the form $v' = -b^2 x$ with $b = 1$. The period of such a linear oscillator is $2\pi/b = 2\pi \approx 6.28$. Let's see if the pendulum has this period when it swings with a small amplitude. We use Euler's method to solve the initial value problem

$$\begin{aligned} x' &= v, & x(0) &= a,\\ v' &= -\sin x, & v(0) &= 0, \end{aligned}$$

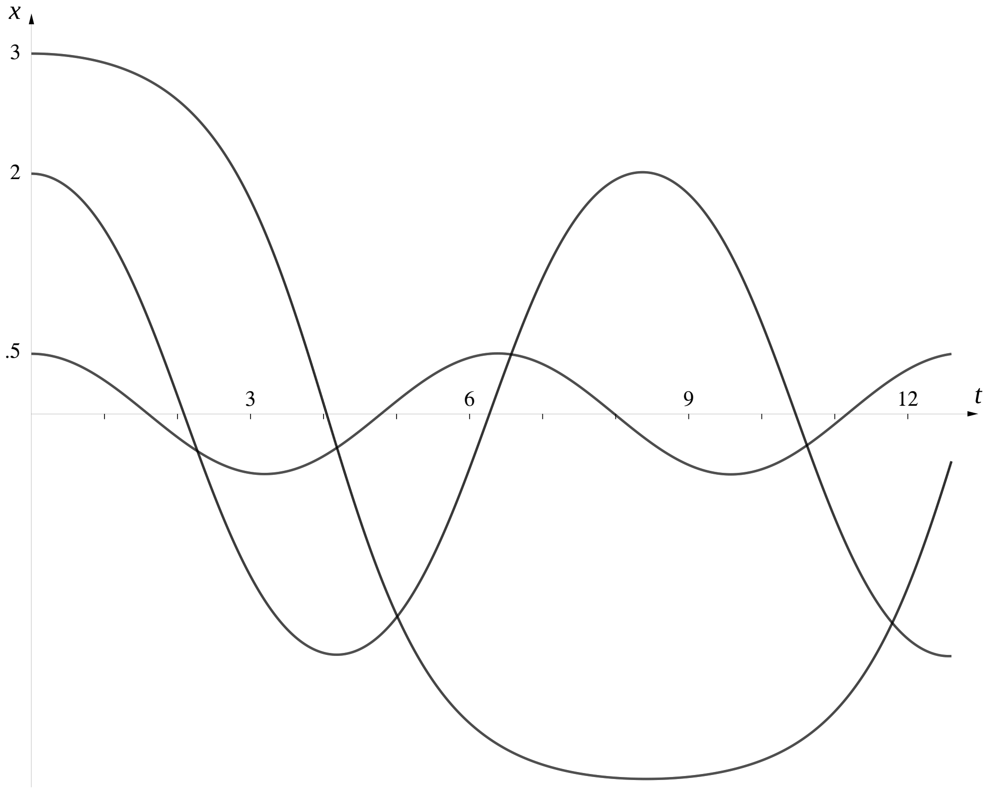
for several small values of a. The results appear in the graph below.

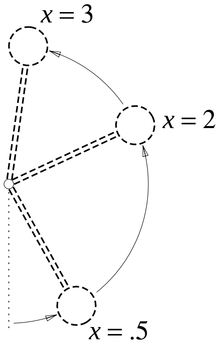


As you can see, small amplitude oscillations have virtually the same period. Thus, we would not expect the fluctuations in the amplitude of the pendulum on a grandfather's clock to affect the timekeeping.

Large-amplitude oscillations

What happens to the period, though, if the pendulum swings in a large arc? The largest possible initial amplitude we can give the pendulum would point it straight up. The pendulum is then 180° from the rest position, corresponding to a value of $x = \pi = 3.14159\ldots$. In the graph below we see the solution that has an initial amplitude of $x = 3$, which is very near the maximum possible. Its period is much larger than the period of the solution with $x = .5$, which has been carried over from the previous graph for comparison. Even the solution with $x = 2$ has a period which is significantly larger that the solution with $x = .5$.





We saw that the period of a hard spring got shorter (its frequency increased) when its amplitude increased. But the pendulum is a soft spring and shows motions of longer period as its initial amplitude is increased. Notice how flat the large-amplitude graph is. This means that the pendulum lingers at the top of its swing for a long time. That's why the period becomes so large. Check the graph now and confirm that the period of the large swing is about 17.

The pendulum at rest

Although we can't get formulas to describe the motion of the pendulum for most initial conditions, there are two special circumstances when we can. Consider a pendulum that is initially at rest: $x = 0$ and $v = 0$ when $t = 0$. It will remain at rest forever: $x(t) = 0$, $v(t) = 0$ for all $t \geq 0$. What we really mean is that the constant functions $x(t) = 0$ and $v(t) = 0$ solve the initial value problem

$$\begin{aligned} x' &= v, & x(0) &= 0, \\ v' &= -\sin x, & v(0) &= 0. \end{aligned}$$

The pendulum balanced on end

There is another way for the pendulum to remain at rest. The key is that v must not change. But $v' = -\sin x$, so v will remain fixed if $v' = -\sin x = 0$. Now, $\sin x = 0$ if $x = 0$. This yields the rest solution we have just identified. But $\sin x$ is also zero if $x = \pi$. You should check that the constant functions $x(t) = \pi$, $v(t) = 0$ solve this initial value problem:

$$\begin{aligned} x' &= v, & x(0) &= \pi, \\ v' &= -\sin x, & v(0) &= 0. \end{aligned}$$

Since the pendulum points straight up when $x = \pi$ radians, this motionless solution corresponds to the pendulum balancing on its end.

Stable and unstable equilibrium solutions

These two solutions are called **equilibrium** solutions (from the Latin *æqui-*, equal + *libra*, a balance scale). If the pendulum is disturbed from its rest position, it tends to return to rest. For this reason, rest is said to be a **stable** equilibrium. Contrast what happens if the pendulum is disturbed when it is balanced upright. This is said to be an **unstable** equilibrium. We will take a longer look at equilibria in chapter 8.

Predator–Prey Ecology

Why do populations fluctuate?

Many animal populations undergo nearly periodic fluctuations in size. It is even more remarkable that the period of those fluctuations varies little from species to species. This fascinates ecologists and frustrates many who hunt, fish, and trap those populations to make their livelihood. Why should there be fluctuations, and can something be done to alter or eliminate them?

There are models of predator-prey interaction that exhibit periodic behavior. Consequently, some researchers have proposed that the fluctuations observed in a real population occur because that species is either the predator or the prey for another species. The models themselves have different properties; we will study one proposed by R. May. As we did with the spring and the pendulum, we will ask how the frequency and amplitude of periodic solutions depend on the initial conditions.

May's model involves two populations that vary in size over time: the predator y and the prey x. The numbers x and y have been set to an arbitrary scale; they lie between 0 and 20. The model also has six adjustable parameters, but we will simply fix their values:

$$\begin{aligned} \text{prey:} \quad x' &= .6\,x\left(1 - \frac{x}{10}\right) - \frac{.5\,xy}{x+1}, \\ \text{predator:} \quad y' &= .1\,y\left(1 - \frac{y}{2x}\right). \end{aligned}$$

These equations will be our starting point. However, if you wish to learn more about the premises behind May's model, you can refer to chapter 4.1 (page 191).

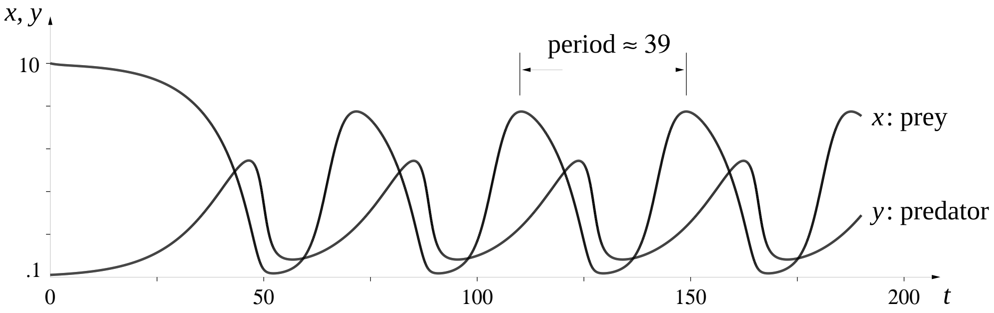
A predator-free equilibrium ...

To begin to explore the model, let's see what happens to the prey population when there are no predators ($y = 0$). Then the size of x is governed by the simpler differential equation $x' = .6x(1 - x/10)$. This is logistic growth, and x will eventually approach the carrying capacity of the environment, which in this case is 10. (See chapter 4.1, pages 183–185.) In fact, you should check that

$$x(t) = 10, \qquad y(t) = 0,$$

... upset by predators

is an equilibrium solution of May's original differential equations. Now suppose we introduce a small number of predators: $y = .1$. Then the equilibrium is lost, and the predator and prey populations fall into cyclic patterns with the same period:

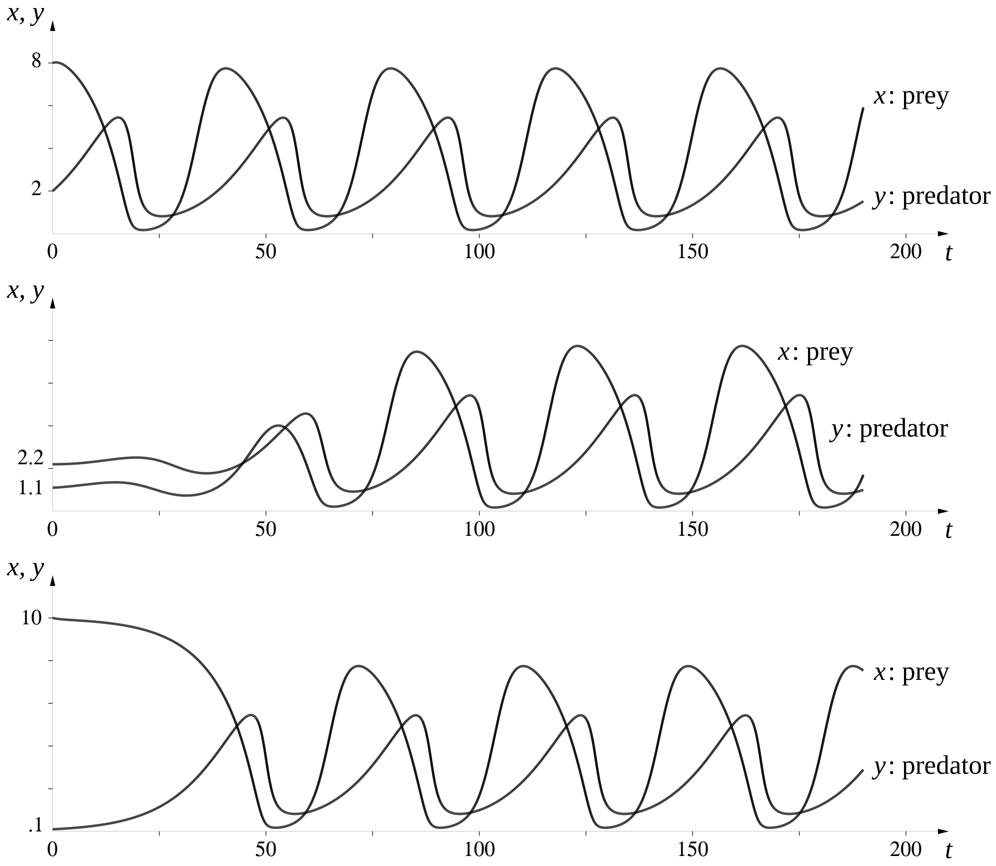


Solutions with various initial conditions . . .

In the other models of periodic behavior we have studied, the frequency and amplitude have depended on the initial conditions. Is the same true here? The following graphs illustrate what happens if the initial populations are either

$$x = 8, \quad y = 2, \qquad \text{or} \qquad x = 1.1, \quad y = 2.2.$$

For the sake of comparison, the solution with $x = 10$, $y = .1$ is also carried over from the previous page.



. . . all have the same amplitude and frequency

In all of these graphs, periodic behavior eventually emerges. What is most striking, though, is that it is the *same* behavior in all cases. The amplitude and the period *do not* depend on the initial conditions. Moreover, even though the populations peak at different times on the three graphs (i.e., the *phases* are different), the y peak always comes about 14 time units after the x peak.

Proving a Solution Is Periodic

Can we *prove* that systems have periodic oscillations?

The graphs in the last ten pages provide strong evidence that non-linear springs, pendulums, and predator-prey systems can oscillate in a periodic way. The evidence is numerical, though. It is based on Euler's method, which gives us only *approximate* solutions to differential equations. Can we now go one step further and *prove* that the solutions to these and other systems are periodic?

The virtue of a formula

Notice that we already have a proof in the case of a linear spring. The solutions are given by formulas that involve sines and cosines, and these are periodic by their very design as circular functions. But we have no formulas for the solutions of the other systems. In particular, we are not able to say anything about the general properties of the solutions (the way we can about sine and cosine functions). The approach we take now does not depend on having a formula for the solution.

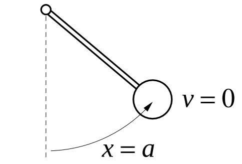
It may seem that what we should do is develop more methods for finding formulas for solutions. In fact, two hundred years of research was devoted to this goal, and much has been accomplished. However, it is now clear that most solutions simply have no representation "in closed form" (that is, as formulas). This isn't a confession that we can't *find* the solutions. It just means the formulas we have are inadequate to describe the the solutions we can find.

The pendulum—a qualitative approach

Let's work with the pendulum and model it by the following initial value problem:

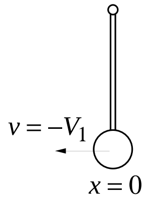
$$\begin{aligned} x' &= v, \qquad & x(0) &= a, \\ v' &= -\sin x, \qquad & v(0) &= 0. \end{aligned}$$

Stage 1:



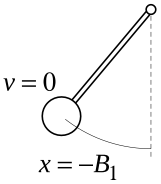
We'll assume $0 < a < \pi$. Thus, at the start the pendulum is motionless and raised to the right. Call this **stage 1**. We'll analyze what happens to x and v in a qualitative sense. That is, we'll pay attention to the *signs* of these quantities, and whether they're increasing or decreasing, but not their exact numerical values.

Stage 2:



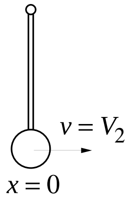
According to the differential equations, v determines the rate at which x changes, and x determines the rate at which v changes. In particular, since we start with $0 < x < \pi$, the expression $-\sin x$ must be negative. Thus v' is negative, so v decreases, becoming more and more negative as time goes on. Consequently, x changes at an ever increasing negative rate, and eventually its value drops to 0. The moment this happens the pendulum is hanging straight down and moving left with some large negative velocity $-V_1$. This is **stage 2**.

Stage 3:



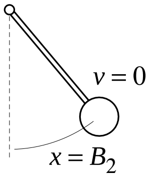
Immediately after the pendulum passes through stage 2, x becomes negative. Consequently, $v' = -\sin x$ now has a positive value (because x is negative). So v stops decreasing and starts increasing. Since x gets more and more negative, v increases more and more rapidly. Eventually v must become 0. Suppose $x = -B_1$ at the moment this happens. The pendulum is then poised motionless and raised up B_1 units to the left. We have reached **stage 3**.

Stage 4:



The situation is now similar to stage 1, because $v = 0$ once again. The difference is that x is now negative instead of positive. This just means that v' is positive. Consequently v becomes more and more positive, implying that x changes at an ever increasing positive rate. Eventually x reaches 0. The moment this happens the pendulum is again hanging straight down (as it was at stage 2), but now it is moving to the right with some large positive velocity V_2. Let's call this **stage 4**. It is similar to stage 2.

Stage 5:



Immediately after the pendulum passes through stage 4, x becomes positive. This makes $v' = -\sin x$ negative, so v stops increasing and starts decreasing. Eventually v becomes 0 again (just as it did in the events that lead up to stage 3). At the moment the pendulum stops, x has reached some positive value B_2. Let's call this **stage 5**.

The "trade-off" between speed and height

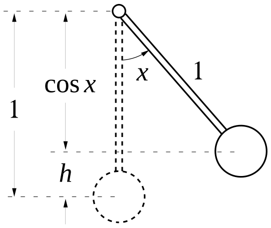
Are we back where we started?

We appear to have gone "full circle." The pendulum has returned to the right and is once again motionless—just as it was at the start. However, we don't know that the *current* position of the pendulum (which is $x = B_2$) is the same as its *initial* position ($x = a$). This is a consequence of working qualitatively instead of quantitatively. But it is also the nub of the problem. For the motion of the pendulum to repeat itself exactly we must have $B_2 = a$. Can we prove that $B_2 = a$?

Since a and B_2 are the successive positive values of x that occur when $v = 0$, it makes sense to explore the connection between x and v. In a real pendulum there is an obvious connection. The higher the pendulum bob rises, the more slowly it moves. If you review the sequence of stages we just went through, you'll see that the same thing is true of our mathematical model. This suggests that we should focus on the height of the pendulum bob and the magnitude of the velocity. This is called the **speed**; it is just the absolute value $|v|$ of the velocity.

A little trigonometry shows us that when the pendulum makes an angle of x radians with the vertical, the height of the pendulum bob is

$$h = 1 - \cos x.$$



When x is a function of time t, then h is too and we have

$$h(t) = 1 - \cos(x(t)).$$

Changes in speed . . . modified

Our intuition about the pendulum tells us that every change in height is offset by a change in speed. (This is the "trade-off.") It makes sense, therefore, to compare the *rates* at which the height and the speed change over time. However, the speed $|v(t)|$ involves an absolute value, and this is difficult to deal with in calculus. (The absolute value function is not differentiable at 0.) Since we are using $|v|$ simply as a way to ignore the difference between positive and negative velocities, we can replace $|v|$ by v^2. Then we find

$$\frac{d}{dt}(v(t))^2 = 2 \cdot v(t) \cdot v'(t) = -2 \cdot v \cdot \sin x.$$

Notice that we needed the chain rule to differentiate $(v(t))^2$. After that we used the differential equations of the pendulum to replace v' by $-\sin x$.

Changes in height

The height of the pendulum changes at this rate:

$$\frac{d}{dt}h(t) = \sin(x(t)) \cdot x'(t) = \sin x \cdot v.$$

We needed the chain rule again, and we used the differential equations of the pendulum to replace x' by v.

The two derivatives are almost exactly the same; except for sign, they differ only by a factor 2. If we use $\frac{1}{2}v^2$ instead of v^2, then the trade-off is exact: every increase in $\frac{1}{2}v^2$ is *exactly* matched by a decrease in h, and *vice versa*. Therefore, if we combine $\frac{1}{2}v^2$ and h to make the new quantity

$$E = \tfrac{1}{2}v^2 + h = \tfrac{1}{2}v^2 + 1 - \cos x,$$

then we can say that the value of E does not change as the pendulum moves.

Showing E is a constant

Since E depends on v and h, and these are functions of the time t, E itself is a function of t. To say that E doesn't change as the pendulum moves is to say that this function is a constant—in other words, that its derivative is 0. This was, in fact, the way we constructed E in the first place. Let's remind ourselves of why this worked. Since $E = \frac{1}{2}v^2 + h$,

$$\frac{dE}{dt} = v \cdot v' + h' = v \cdot (-\sin x) + \sin x \cdot v = 0.$$

To get the second line we used the fact that $v' = -\sin x$ and $x' = v$ when $x(t)$ and $v(t)$ describe pendulum motion.

The quantity E is called the **energy** of the pendulum. The fact that E doesn't change is called **the conservation of energy** of the pendulum. A number of problems in physics can be analyzed starting from the fact that the energy of many systems is constant.

Let's calculate the value of E at the five different stages of our pendulum:

stage	v	x	h	E
1	0	a	$1 - \cos a$	$1 - \cos a$
2	$-V_1$	0	0	$\frac{1}{2}(-V_1)^2$
3	0	$-B_1$	$1 - \cos B_1$	$1 - \cos B_1$
4	V_2	0	0	$\frac{1}{2}{V_2}^2$
5	0	B_2	$1 - \cos B_2$	$1 - \cos B_2$

By the conservation of energy, all the quantities in the right-hand column have the same value. Looking at the value for E in stages 2 and 4, we see that $V_1 = V_2$—*whenever the pendulum is at the bottom of its swing ($x = 0$), it is moving with the same speed*, the velocity being positive when the pendulum is swinging to the right, negative when it is swinging to the left. Similarly, if we look at the value of E at stages 1, 3, and 5, we see that

$$1 - \cos a = 1 - \cos B_1 = 1 - \cos B_2.$$

We can put this another way: *whenever the pendulum is motionless, it must be back at its starting height $h = 1 - \cos a$.*

...and that the oscillations are periodic

In particular, we have thus shown that B_2 (the position of the pendulum after it's gone over and back) $= a$ (the position of the pendulum at the beginning). Thus the value for x and the value for v are the same in stage 5 and in stage 1—the two stages are mathematically indistinguishable. Since the solution to an initial value problem depends only on the differential equation and the initial values, what happens after stage 5 must be identical to what happens after stage 1—the second swing of the pendulum must be identical to the first! Thus the motion is periodic, which completes our proof.

You can also use the fact that the value of E doesn't change to determine the velocities $-V_1$ and V_2 that the pendulum achieves at the bottom of its swing. In the exercises you are asked to show that

$$V_1 = V_2 = \sqrt{2 - 2\cos a}.$$

First Integrals

Notice in what we have just done that we haven't solved the differential equation for the pendulum in the sense of finding explicit formulas giving x and v in terms of t. Instead we found a combination of x and v that remained constant over time and used this to deduce some of the behavior of x and v. Such a combination of the variables that remains constant is called a **first integral** of the differential equation. A surprising amount of information about a system can be inferred from first integrals (when they exist). They play an important role in many branches of physics, giving rise to the basic conservation laws for energy, momentum, and angular momentum. We will have more to say about first integrals and conservation laws in chapter 8.

First integrals

In the exercises you are asked to explore first integrals for linear and non-linear springs—and to prove thereby that (frictionless) non-linear springs have periodic motions.

Exercises

Linear springs

In the text we always assumed that the weight on the spring was motionless at $t = 0$ seconds. The first four exercises explore what happens if the weight is given an initial *impulse*. For example, instead of simply releasing the weight, you could hit it out of your hand with a hammer. This means $v(0) \neq 0$. The general initial value problem is

$$\begin{aligned} x' &= v, & x(0) &= a, \\ v' &= -b^2 x, & v(0) &= p. \end{aligned}$$

The aim is to see how the period, amplitude, and phase of the solution depend on this new condition.

1. **Pure impulse**. Take $b = 5$ per second, as in the first example in the text, but suppose

$$a = 0 \text{ cm}, \qquad p = 20 \text{ cm/sec}.$$

(In other words, you strike the weight with a hammer as it sits motionless at the rest position $x = 0$ cm.)

a) Use the differential equation solver on a computer to solve the initial value problem numerically and graph the result.

b) From the graph, estimate the period and the amplitude of the solution.

c) Find a formula for this solution, using the graph as a guide.

d) From the formula, determine the period and amplitude of the solution. Does the period depend the initial impulse p, or only on the spring constant b? Does the amplitude depend on p?

2. **Impulse and displacement**. Take $a = 4$ cm and $b = 5$ per second, as in the first example on page 434. But assume now that the weight is given an initial *downward* impulse of $p = -20$ cm/sec.

a) Solve the initial value problem numerically and graph the result.

b) From the graph, estimate the period and the amplitude of the solution. Compare these with the period and the amplitude of the solution obtained in the text for $p = 0$ cm/sec.

3. Let a and b have the values they did in the last exercise, but change p to $+20$ cm/sec. Graph the solution, and compare the amplitude and phase of this solution with the solution of the previous exercise.

4. Let a, b, and p have arbitrary values. The last two exercises suggest that the solution to the general initial value problem for a linear spring can be given by the formula $x(t) = A\sin(bt - \varphi)$. The amplitude A and the phase difference φ depend on the initial conditions. Show that the formula for $x(t)$ is correct by expressing A and φ in terms of the initial conditions.

5. **Strength of the spring**. Take two springs, and suppose the second is twice as strong as the first. That is, assume the second spring constant is twice the first. Put equal weights on the ends of the two springs, and use the initial value $v(0) = 0$ in both cases. Which weight oscillates with the higher frequency? How are the frequencies of the two related—e.g., is the frequency of the second equal to twice the frequency of the first, or should the multiplier be a different number?

6. a) **Effect of the weight**. Hang weights from two identical springs (i.e., springs with the same spring constant). Suppose the mass of the second

weight is twice that of the first. Which weight oscillates with the higher frequency? How much higher—twice as high, or some other multiplier?

b) Do this experiment in your head. Measure the frequency of the oscillations of a 200 gram weight on a spring. Suppose a second weight oscillates at twice the frequency; what is *its* mass?

A reality check. Do your results in the last two exercises agree with your intuitions about the way springs operate?

7. a) **First integral**. Show that $E = \frac{1}{2}v^2 + \frac{1}{2}b^2x^2$ is a first integral for the linear spring

$$\begin{aligned} x' &= v, & x(0) &= a, \\ v' &= -b^2x, & v(0) &= p. \end{aligned}$$

In other words, if the functions $x(t)$ and $v(t)$ solve this initial value problem, you must show that the combination

$$E = \tfrac{1}{2}\left(v(t)\right)^2 + \tfrac{1}{2}b^2\left(x(t)\right)^2$$

does not change as t varies.

b) What value does E have in this problem?

c) If x is measured in cm and t in sec, what are the units for E?

8. a) This exercise concerns the initial value problem in the previous question. When $x = 0$, what are the possible values that v can have?

b) At a moment when the weight on the spring is motionless, how far is it from the rest position?

9. You already know that initial value problem in exercise 7 has a solution of the form $x(t) = A\sin(bt - \varphi)$ and therefore must be periodic. Given a different proof of periodicity using the first integral from the same exercise, following the approach used by the book in the case of the pendulum.

Non-linear springs

10. a) Suppose the acceleration v' of the weight on a hard spring depends on the displacement x of the weight according to the formula $v' = -16x - x^3$

cm/sec^2. If you pull the weight down $a = 2$ cm, hold it motionless (so $p = 0$ cm/sec) and then release it, what will its frequency be?

b) How far must you pull the weight so that its frequency will be double the frequency in part (a)? (Assume $p = 0$ cm/sec, so there is still no initial impulse.)

11. Suppose the acceleration of the weight on a hard spring is given by $v' = -16x - .1\,x^3$ cm/sec^2. If the weight is oscillating with very small amplitude, what is the frequency of the oscillation?

12. a) Suppose a weight on a spring accelerates according to the formula

$$\frac{dv}{dt} = -\frac{25x}{1+x^2} \quad \text{cm/sec}^2.$$

This is a soft spring. Explain why. [Graph v' as a function of x.]

b) If the initial amplitude of the weight is $a = 4$ cm, and there is no initial impulse (so $p = 0$ cm/sec), what is the frequency of the oscillation?

c) Double the initial amplitude, making $a = 8$ cm but keeping $p = 0$ cm/sec. What happens to the frequency?

d) Suppose you make the initial amplitude $a = 100$ cm. Now what happens to the frequency?

13. **First integrals**. Suppose the acceleration on a non-linear spring is

$$v' = -b^2x - \beta x^3, \qquad \text{where} \quad v = x'.$$

Show that the function

$$E = \tfrac{1}{2}v^2 + \tfrac{1}{2}b^2x^2 + \tfrac{1}{4}\beta x^4$$

is a first integral. (See the text (page 451) and exercise 7, above.)

14. Suppose the acceleration on a non-linear spring is $v' = -16x - x^3$ cm/sec^2, and initially $x = 2$ cm and $v = 0$ cm/sec.

a) The first integral of the preceding exercise must have a fixed value for this spring. What is that value?

b) How fast is the spring moving when it passes through the rest position?

c) Can the spring ever be more than 2 cm away from the rest position? Explain your answer.

15. Construct a first integral for the initial value problem

$$\begin{aligned} x' &= v, & x(0) &= a, \\ v' &= -b^2 x - \beta x^3, & v(0) &= p, \end{aligned}$$

and use it to show that the solution to the problem is periodic.

16. a) Show that the function

$$E = \tfrac{1}{2}v^2 + \tfrac{25}{2}\ln(1 + x^2)$$

is a first integral for the soft spring in exercise 12.

b) If the initial amplitude is $a = 4$ cm and the initial velocity is 0 cm/sec, what is the speed of the weight as it moves past the rest position?

c) Prove that the motion of this spring is periodic.

17. Suppose the acceleration on a non-linear spring has the general form $v' = -f(x)$. Can you find a first integral for this spring? In other words, you are being asked to show that a first integral always exists whenever the rate of change of the velocity depends only on the position x (and not, for instance, on v itself, or on the time t).

The pendulum

These questions deal with the initial value problem

$$\begin{aligned} x' &= v, & x(0) &= a, \\ v' &= -\sin x, & v(0) &= p. \end{aligned}$$

In particular, we want to allow an initial impulse $p \neq 0$.

18. Take $a = 0$ and given the pendulum three different initial impulses: $p = .05$, $p = .1$, $p = .2$. Use the differential equation solver on a computer to graph the three motions that result. Determine the period of the motion in each case. Are the periods noticeably different?

19. What is the period of the motion if $p = 1$; if $p = 2$?

20. By experiment, find how large an initial impulse p is needed to knock the pendulum "over the top", so it spins around its axis instead of oscillating? Assume $x(0) = 0$. (Note: when the pendulum spins, x just keeps getting larger and larger.) Of course any enormous value for p will guarantee that the pendulum spins. Your task is to find the *threshold*; this is the smallest initial impulse that will cause spinning.

21. a) Suppose the initial position is horizontal: $a = +\pi/2$. If you give the pendulum an initial impulse p in the same direction (that is, $p > 0$), find by experiment how large p must be to cause the pendulum to spin? Once again, the challenge is to find the threshold value.

b) Reverse the direction of the initial impulse: $p < 0$, and choose p so the pendulum spins. What is the smallest $|p|$ that will cause spinning?

22. **First integrals**. Consider the initial value problem described in the text:

$$\begin{aligned} x' &= v, & x(0) &= a, \\ v' &= -\sin x, & v(0) &= 0. \end{aligned}$$

Use the first integral for this problem found on page 449 to show that $v = \sqrt{2 - 2\cos a}$ when $x = 0$.

23. a) Suppose the pendulum described in the previous exercise is at rest ($x(0) = 0$), but given an initial impulse $v(0) = p$. What value does the first integral have in this case?

b) Redo exercise 20 using the information the first integral gives you. You should be able to find the exact threshhold value of the impulse that will push the pendulum "over the top."

24. Redo exercise 21 using an appropriate first integral. Find the threshhold value exactly.

Predator-prey ecology

25. a) **The May model**. The differential equations for this model are on page 445. Show that the constant functions

$$x(t) = 10, \qquad y(t) = 0,$$

are a solution to the equations. This is an equilibrium solution, as defined in the discussion of the pendulum (page 444).

b) Is $x(t) = 0$, $y(t) = 0$ an equilibrium solution?

c) Here is yet another equilibrium solution:

$$x(t) = \frac{-23 \pm \sqrt{889}}{6}, \qquad y(t) = \frac{-23 \pm \sqrt{889}}{3}.$$

Either verify that it *is* an equilibrium, or explain how it was derived.

26. a) Use a computer differential equation solver to graph the solution to the May model that is determined by the initial conditions

$$x(0) = 1.13, \qquad y(0) = 2.27.$$

These initial conditions are very close to the equilibrium solution in part (c) of the previous exercise. Does the solution you've just graphed suggest that this equilibrium is *stable* or that it is *unstable* (as described on page 444).

b) Change the initial conditions to

$$x(0) = 5, \qquad y(0) = 5,$$

and graph the solution. Compare this solution to those determined by the initial conditions used in the text. In particular, compare the shapes of the graphs, their periods, and the time interval between the peak of x and the peak of y.

27. Consider this scenario. Imagine that the prey species x is an agricultural pest, while the predator y does not harm any crops. Farmers would like to eliminate the pest, and they propose to do so by bringing in a large number of predators. Does this strategy work, according to the May model? Suppose that we start with a relatively large number of predators:

$$x(0) = 5, \qquad y(0) = 50.$$

What happens? In particular, does the pest disappear?

28. **The Lotka–Volterra model**. We use the differential equations found in chapter 4, page 193, modified so that relevant values of x and y will be roughly the same size:

$$\begin{aligned} x' &= .1x - .005xy, \\ y' &= .004xy - .04y. \end{aligned}$$

Take $x(0) = 20$ and $y(0) = 10$. Use a computer differential equation solver to graph the solution to this initial value problem. The solutions are periodic. What is the period? Which peaks first, the prey x or the predator y? How much sooner?

29. Solve the Lotka–Volterra model with $x(0) = 10$ and $y(0) = 5$. What is the period of the solutions, and what is the difference between the times when the two populations peak? Compare these results with those of the previous exercise.

30. Show that $x(t) = 0$, $y(t) = 0$ is an equilibrium solution of the Lotka–Volterra equations. Test the stability of this solution, take these nearby initial conditions:

$$x(0) = .1, \qquad y(0) = .1,$$

and find the solution. Does it remain near the equilibrium? If so, the equilibrium is stable; if not, it is unstable.

31. Show that $x(t) = 10$, $y(t) = 20$ is another equilibrium solution of these Lotka–Volterra equations. Is this equilibrium stable? (We will have more to say about stability of equilibria in chapter 8.)

32. This is a repeat of the biological pest control scenario you treated above, using the May model. Solve the Lotka–Volterra model when the initial populations are

$$x(0) = 5, \qquad y(0) = 50.$$

What happens? In particular, does the pest disappear?

33. **First integrals**. As remarkable as it may seem, the Lotka–Volterra model has a first integral. Show that the function

$$E = a \ln y + d \ln x - by - cx$$

is a first integral of Lotka–Volterra model given in the general form

$$\begin{aligned} x' &= ax - bxy, \\ y' &= cxy - dy. \end{aligned}$$

34. Prove that the solutions of the Lotka–Volterra equations are periodic.

The van der Pol oscillator

One of the essential functions of the electronic circuits in a television or radio transmitter is to generate a periodic "signal" that is stable in amplitude and period. One such circuit is described by the van der Pol differential equations. In this circuit $x(t)$ represents the current, and $y(t)$ the voltage, at time t. These functions satisfy the differential equations

$$x' = y, \qquad y' = Ay - By^3 - x, \qquad \text{with } A, B > 0.$$

35. Take $A = 4$, $B = 1$. Make a sketch of the solution whose initial values are $x(0) = .1$, $y(0) = 0$. Your sketch should show that this solution is *not* periodic at the outset, but becomes periodic after some time has passed. Determine the (eventual) period and amplitude of this solution.

36. Obtain the solution whose initial values are $x(0) = 2$, $y(0) = 0$, and then the one whose initial values are $x(0) = 4$, $y(0) = 0$. What are the periods and amplitudes of these solutions? What effect does the initial current $x(0)$ have on the period or the amplitude?

7.4 Chapter Summary

The Main Ideas

- There are many phenomena which exhibit **periodic** and near-periodic behavior. They are modelled by differential equations with periodic solutions.

- A periodic function repeats: the smallest number T for which $g(x + T) = g(x)$ for all x is the **period** of the function g. Its **frequency** is the reciprocal of its period, $\omega = 1/T$.

- The **circular functions** are periodic; they include the sine, cosine and tangent functions. The period of $\sin(t)$ and $\cos(t)$ is 2π and the frequency is $1/2\pi$. The frequency of $A\sin(bt)$ and $A\cos(bt)$ is $b/2\pi$, and the **amplitude** is A. In $A\sin(bt + \varphi)$, the **phase** is shifted by φ.

- A **linear spring** is one for which the spring force is proportional to the amount that the spring has been displaced. The motion of a linear spring is periodic. Its amplitude depends only on the initial conditions, and its frequency only on the mass and the spring constant.

- In a **non-linear spring**, the force is no longer proportional to the displacement. The motion of a non-linear spring can still be periodic, although it is no longer described simply by sines and cosines. Its frequency depends on its amplitude. A pendulum in a non-linear spring. It has two **equilibria**, one **stable** and one **unstable**.

- Many quantities oscillate periodically, or nearly so. Frequently the behavior of these quantities can be modelled by systems of differential equations. Pendulums, electronic components, and animal populations are some examples.

- In some initial value problems, it may still be possible to find a **first integral**—a combination of the variables that remains constant—even when we can't find formulas for the variables separately. We can often derive important properties of the system (such as periodicity) from these constant combinations.

Expectations

- You should be able to find the **period**, **frequency** and **amplitude** of sine and cosine functions.

- You should be able to convert between **radian** measure and degrees.

- You should be able to find a formula for the solution of the differential equation describing a **linear spring**.

- You should be able to use Euler's method to describe the motion of a non-linear spring.

- You should be able to analyze oscillations of various kinds to determine their periodicity.

Chapter 8

Dynamical Systems

A recurring theme in this book is the use of mathematical models consisting of a set of differential equations to explore the behavior of physical systems as they evolve over time. Some examples we have encountered are the S-I-R epidemiological model, predator-prey systems, and the motion of a pendulum. We call such a set of differential equations a **dynamical system**. Dynamical systems play important roles in all branches of science. In this chapter we will develop some general tools for thinking about them, with particular emphasis on the kinds of geometric insight provided by the concepts of **state space** and **vector field**.

8.1 State Spaces and Vector Fields

The standard way to graph solutions

If you look back at the examples we've considered, many of them take the following form: we have two (or more) variable quantities x and y that are functions of time, and we want to find the nature of these functions. What we have to work with is a model for the way the functions $x(t)$ and $y(t)$ are changing—i.e., we are told how to calculate $x'(t)$ and $y'(t)$ whenever we know the values of x and y, and possibly t. From a given starting point, we typically used something like Euler's method to get values for x and y at times on either side of the starting value. We then graphed the solutions as functions of time—x against t and y against t.

In many instances, the rules determining $x'(t)$ and $y'(t)$ depend only on the current values of x and y, but not on the value of t, so that knowing the current state of the system (as specified by its x and y values) is sufficient to determine the future and past states of the system. Such systems are said to be **autonomous**. These are the only systems we will be considering in this chapter.

A new way
to graph solutions:
as trajectories
in state space

In autonomous systems there is another way of visualizing the solutions that can be very powerful. Instead of plotting values of x and y as functions of time, we view these values as coordinates of a point in the x-y plane. As the system changes, the point (x, y) will trace out a curve in this plane. The point (x, y) is called a **state**, and the portion of the plane corresponding to physically possible states is called the **state space** of the system. The solution curves that get traced out in state space are called **trajectories**. By looking at three examples, we will see how this method of analysis can help us understand the overall behavior of a system.

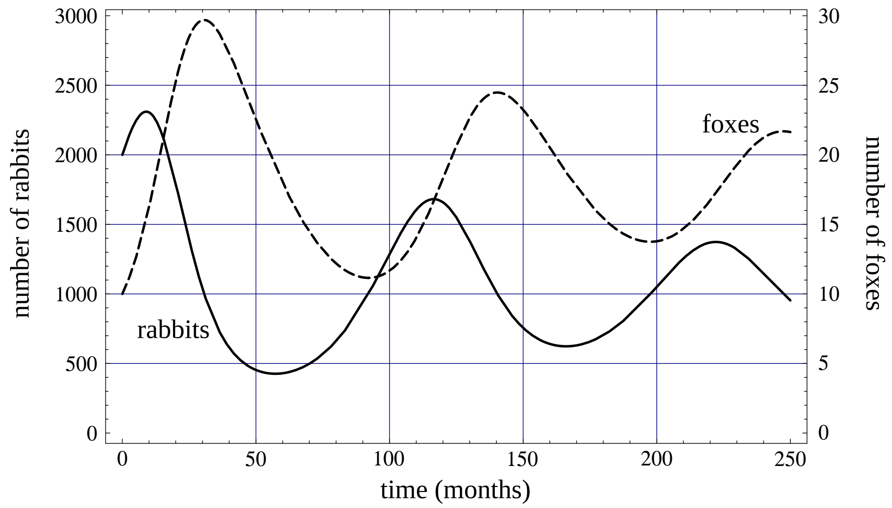
There are a number of effective software packages available which can perform efficiently all the operations we will be considering, and one of them would probably be the most useful tool for exploring the ideas in this chapter. On the other hand, the basic numerical operations are quite simple, and it is easy to modify the programs developed earlier in the text to perform these operations as well. For those of you who enjoy programming, we will from time to time point out some of these modifications. It can be instructive to implement them in your own programs, and we urge you to do so.

Predator–Prey Models

In chapter 4.1, we looked at several models for the dynamics of a simple system consisting of foxes (F) and rabbits (R). Our first model was

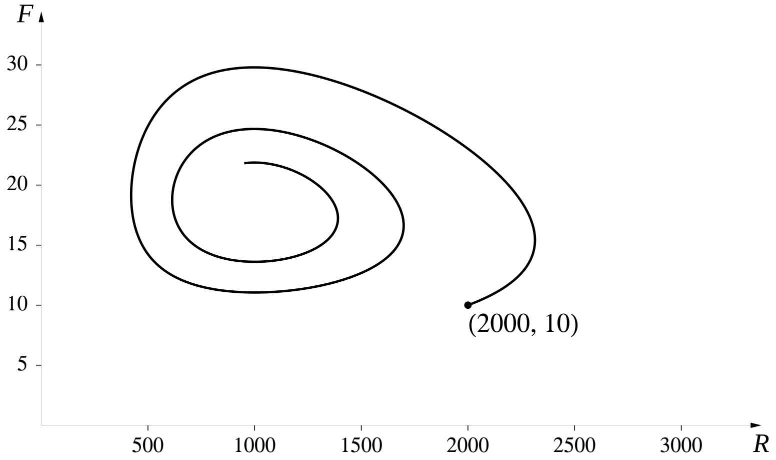
$$\begin{aligned} R' &= .1R\left(1 - \frac{R}{10000}\right) - .005\,RF \quad \text{rabbits per month,} \\ F' &= .00004\,RF - .04\,F \quad \text{foxes per month.} \end{aligned}$$

When we started with the initial values $R(0) = 2000$ rabbits and $F(0) = 10$ foxes, Euler's method produced the following solutions for the first 250 months:



Two ways of representing the same solution graphically

Let's see how the same model looks when we express it in the language of state spaces. The state space consists of points in the R-F plane. For physical reasons our state space consists only of points (R, F) satisfying $R \geq 0$ and $F \geq 0$. That is, our state space is the first quadrant of the R-F plane together with the bounding portions of the R-axis and the F-axis.We can easily modify the program used to obtain the curves in the previous picture to plot the corresponding trajectory in the R-F plane. We only need change the specification of the dimensions of the viewing window and change the `plot` command to plot points with coordinates (R, F) instead of (t, R) and (t, F); all the rest of the calculations are unchanged. Here's what the same solution looks like when we do this:

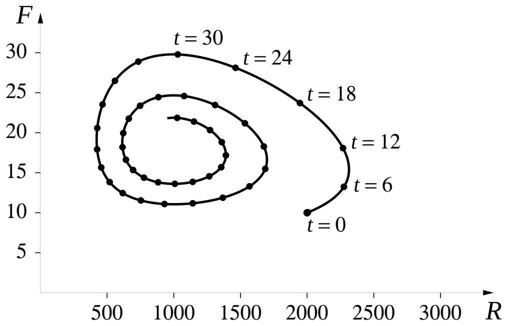


You should notice several things here:

- The trajectory looks like a spiral, moving in towards, but never reaching, some point at its center. We will see later (see page 469) how to determine the coordinates of this limit state.

- If we had started at any other initial state with $R > 0$ and $F > 0$, we would have gotten another spiral converging to the same limit (try it and see).

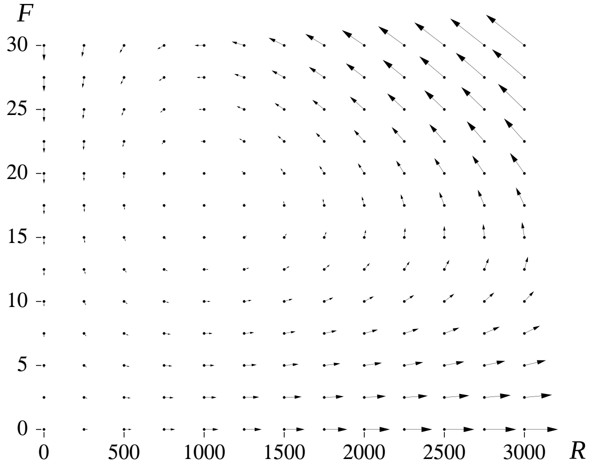
- From the trajectory alone, there is no way of determining the time at which the system passes through the different states. In part, this simply emphasizes that the succession of states the system moves through does not depend on the initial value of t, nor does it depend on the units in which t is measured—if t were measured in days or years, rather than in months, the trajectory would be unchanged.



If we wanted to include some information about time, one way would be to label some points on the trajectory with the associated time value. If we label the points every 6 months, say, we would get the picture at the right. Note that the points are not uniformly spaced along the trajectory: the spacing is largest between points relatively far from the origin, where the values of R and F are largest. Moreover, the closer we come to the limit state, the tighter the spacing becomes.

The differential equations indicate how the state changes

Could we have foreseen some of this behavior by looking at the original differential equations? Since the differential equations give R' and F' as functions of R and F alone, for each point (R, F) in the state space we can calculate the associated values for R' and F'. Knowing these values, we can in turn tell in what direction and with what speed a trajectory would be moving as it passed through the point (R, F). Using our (by now) standard argument, in time Δt the change in R would be $\approx R'\Delta t$, while F would change by $\approx F'\Delta t$. We can convey this information graphically by choosing a number of points in the state space, and from each point (R, F) drawing an arrow to the point $(R + R'\Delta t, F + F'\Delta t)$. We would typically choose a value for Δt that keeps the arrows a reasonable size. Here's what we get in our current example when we choose a 16×16 grid of points in the region $0 \le R \le 3000$ and $0 \le F \le 30$, with $\Delta t = 1$.

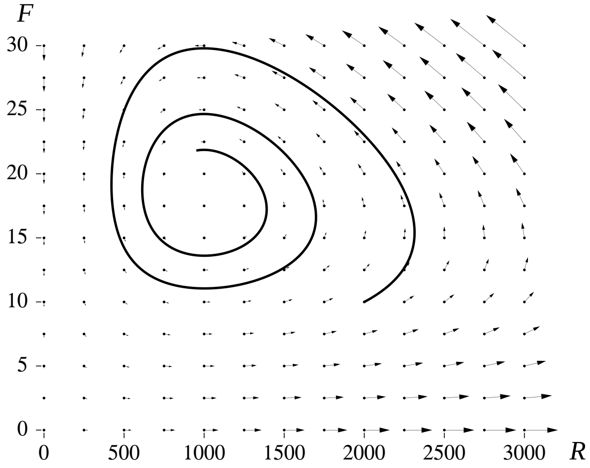


Arrows indicate the change in state...

Several things are immediately clear from this picture: the arrows suggest a general counter-clockwise flow in the plane; change is most rapid in the upper right corner; near the limit point of the flow and near the origin change is so slow that arrows don't even show up there.

...and trajectories are tangent to the arrows

Moreover, since the method used to construct the arrows is exactly the way Euler's method calculates the trajectories themselves, the solution trajectory through a given initial state is a curve in the state space which at every point is tangent to the arrow at that point. For instance, if we superpose the trajectory graphed on page 463 on the picture above, we get the following:



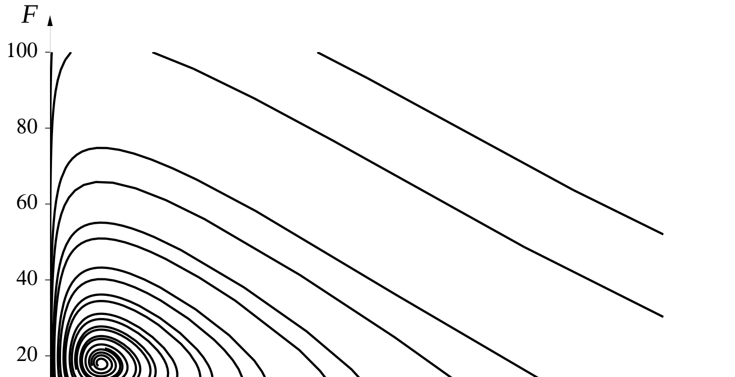
How to construct a geometrical visualization of a dynamical system

The net effect of this construction is thus to transform a problem in *analysis*—findinging a solution to a system of differential equations—into a problem in *geometry*—finding a curve which is tangent everywhere to a prescribed set of arrows. This correspondence between the analytical and the geometrical ways of formulating a problem is very powerful. Let's sum up the way this correspondence was established:

- We set up a **state space** for the system being studied. Each point—called a **state**—in the space corresponds to a possible pair of values the system could have.

- There is a rule which assigns to each point in the state space a **velocity vector**—which can be visualized as an arrow in the space based at the given point—specifying the rates at which the coordinates of the point are changing. The rule itself, which is just our original set of rate equations, is called a **vector field**. Geometrically, we can visualize the vector field as the state space with all the associated arrows.

- Solutions to the dynamical system correspond to **trajectories** in the state space. At every point on a trajectory the associated velocity vector specified by the vector field will be tangent to the trajectory. The existence and uniqueness principle for the solutions of differential equations—there is a unique solution for each set of initial values—is geometrically expressed by the property that every point in the state space lies on exactly one trajectory. The set of all possible trajectories is called the **phase portrait** of the system. For instance, part of the phase portrait of the system we have been considering appears below. We have drawn only a few trajectories—if we had drawn them all, we would have seen only a black rectangle since there is a trajectory through every point.



Simplifying the picture

There is almost too much detail in the picture of the vector field and the phase portrait. One way to see the underlying simplicity is to notice that the space is divided into four regions according to whether F' and R' are positive or negative. The *signs* of F' and R' in turn determine the *direction* of the associated velocity vector. For instance, if F' and R' are both positive, then F and R must both be increasing, which means the velocity vector will be pointing up and to the right, while if $F' > 0$ and $R' < 0$, the velocity vector will be pointing up ($F' > 0$) and to the left ($R' < 0$). Let's see which states correspond to which behaviors. Here are original rate equations:

$$\begin{aligned} R' &= .1R\left(1 - \frac{R}{10000}\right) - .005\,RF \quad \text{rabbits per month,} \\ F' &= .00004\,RF - .04\,F \quad \text{foxes per month.} \end{aligned}$$

The equation for F' is slightly simpler, so we'll start there. We see that $F' = 0$ in exactly two cases:

1. when $F = 0$, or
2. when $.00004R - .04 = 0$, which is equivalent to saying $R = 1000$.

The first case simply says that if we are ever on the R-axis ($F = 0$), then we stay there—a trajectory starting on the R-axis must move horizontally. (If you start with no foxes, you will never have any at a later time.) The second case says that the value of F isn't changing whenever $R = 1000$. The set of points satisfying $R = 1000$ is just a vertical line in the state space. The condition that $F' = 0$ on this line can be expressed geometrically by saying that any trajectory crossing this line must do so horizontally (why?).

Divide the state space into regions

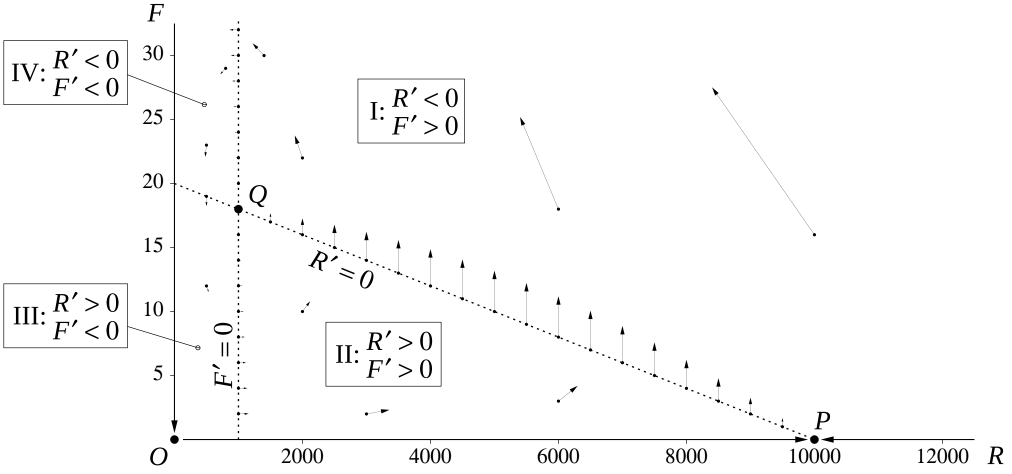
The remainder of the quadrant consists of two regions: one consists of all points (R, F) with $0 \leq R < 1000$ and $F > 0$, the other consists of all points (R, F) with $R > 1000$ and $F > 0$. Moreover, since we've already accounted for all the points where $F' = 0$, it must be true that at every point of these two regions F' must be > 0 or < 0; F' can't equal 0 in either region. Further, within any one region F' must be always positive or always negative. If it were positive at some points and negative at others in a single region, there would have to be transition points where it took on the value 0, which we have just observed can't happen. (Be sure you see why this is so!) Thus to determine the sign of F' in an entire region, we only need to see what the sign is at one point in that region. For instance if we let $R = 2000$ and $F = 1$, we see that $F' = .08 - .04$, which is positive. Therefore we will have $F' > 0$ (fox population increasing) for any other state (R, F) with $R > 1000$. Similarly we can show that $F' < 0$ (fox population decreasing) if $0 \leq R < 1000$—the test point $R = 0$ and $F = 1$ is easy to evaluate. We could, of course, have arrived at the same conclusions through more formal algebraic arguments, which are fairly straightforward in this instance. In other problems, though, the "test point" approach may be the more convenient.

In exactly the same way, if we look at the first rate equation, we find that $R' = 0$ in two cases:

1. when $R = 0$, or
2. when $.1\,(1 - R/10000) - .005\,F = 0$. This is just the equation of a line, which can be rewritten as $F = 20 - .002\,R$.

The interpretations of these two cases are similar to the preceding analysis: any trajectory starting on the F-axis must stay on the F-axis; any trajectory crossing the line $F = 20 - .002\,R$ must cross it vertically, since $R' = 0$—the R-value isn't changing—there. Further, for any other state (R, F) we have $R' > 0$ if the point is below this line (the point $R = 1$ and $F = 0$ is a convenient test point where it's easy to see without doing any arithmetic that $R' > 0$), and $R' < 0$ if the point is above the line.

We can combine all this information into the following picture. We have drawn a number of velocity vectors along the lines where $R' = 0$ and $F' = 0$, with one or two others in each region.



We see that the entire state space is divided into four regions:

1. Region I, above the line $F = 20 - .002\,R$ and to the right of the line $R = 1000$. Here $R' < 0$, and $F' > 0$, so all velocity vectors are pointing up and to the left.
2. Region II, below the line $F = 20 - .002\,R$ and to the right of the line $R = 1000$. Here $R' > 0$, and $F' > 0$, and all velocity vectors are pointing up and to the right.
3. Region III, below the line $F = 20 - .002\,R$ and to the left of the line $R = 1000$. Here $R' > 0$, and $F' < 0$, and all velocity vectors are pointing down and to the right.
4. Region IV, above the line $F = 20 - .002\,R$ and to the left of the line $R = 1000$. Here $R' < 0$, and $F' < 0$, and all velocity vectors are pointing down and to the left.

There are simple trajectories: three are just points. . .

Notice that this diagram makes it clear what the limit state of the spirals is: it is the point $Q = (1000, 18)$ where the line $R = 1000$ and the line $F = 20 - .002\,R$ intersect. Notice that at Q both $R' = 0$ and $F' = 0$, so that if we are ever at Q, we never leave—the point Q is a trajectory all by itself. The points $O = (0, 0)$ and $P = (10\,000, 0)$ are the two other such point trajectories. While the typical trajectory looks like a spiral coming into the point Q, note that this picture contains three other "special" trajectories in addition to the point trajectories:

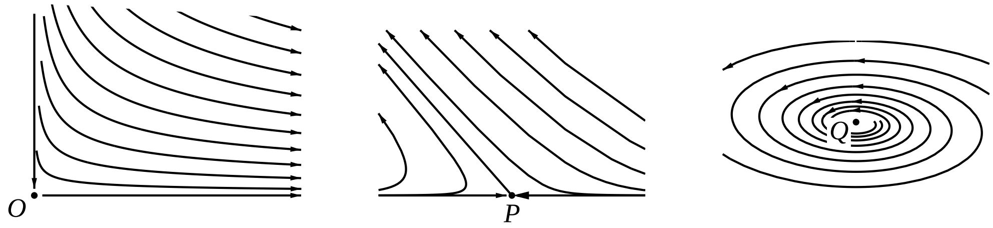
. . . and three are straight line segments

- The F-axis for $F > 0$. The point $(0, 0)$ is *not* part of this trajectory.
- The portion of the R-axis with $0 < R < 10000$. Here the flow is toward the right, towards the point P.
- The portion of the R-axis with $10000 < R$. Flow is to the left, towards P, with movement being slower and slower as P is approached. Note that this is entirely separate from the preceding trajectory—you can't start at any point on one of them and get to any point on the other.

Equilibrium Points

Different kinds of equilibrium points

The three points O, P, and Q in the previous figure—single points which are also trajectories—are called **equilibrium points** for the system. If the ststem ever in such a state, it stays in it forever. Moreover, the system can't reach such a state from any other state (although it may be able to come very close). Nevertheless, the behavior of the system is not the same near the three points. If we zoom in on each of these points and draw some of the nearby trajectories, we get the following pictures:



Points O and P look fairly similar—they would look even more alike if we crossed over into the negative R and negative F regions and included the trajectories there as well (impossible to do in the real world, but elementary in mathematics!). In both cases there is one direction from which trajectories come straight towards the point (in the case of O, this is the F-axis; for P this is the R-axis), and one direction in which trajectories move directly away from the point (the R-axis in the case of point O, and the line of slope $-.0092$ (we'll see how to find this later!) in the case of P). The remaining trajectories look sort of like hyperbolas asymptotic to these two lines. Equilibrium points of this sort are called **saddle points**. They are characterized by the property that there is exactly one direction along which the system can be displaced and still move back towards the equilibrium point. Displacements in any other direction get amplified, with the state eventually moving even further away.

Saddle point equilibrium

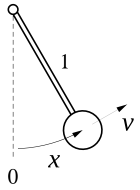
Point Q is quite different. If the state experiences a small displacement away from Q in any direction, over time it will move back towards Q. Such equilibrium points are called **attractors**, and Q is an example of a particular kind of attractor called a **spiral attractor**. In this example, Q is an attractor for almost the entire space—if we start with any point (R, F) with $R > 0$ and $F > 0$, the trajectory through (R, F) will eventually come arbitrarily close to Q and stay there. We will shortly see examples (see page 477, for instance) of attractors that draw from more limited portions of the state space.

Spiral equilibrium

Attractors and repellors

For future reference, we define here the concept of **repellor** and **spiral repellor**. Their vector fields look just like those for the attractors, but with all the arrows reversed. If the state experiences a small displacement from a repellor, over time this displacement will increase. We will see examples of a repellor in 8.4 It turns out that there is a relatively small number of kinds of equilibrium points that a system can have, and we will meet most of them in the next several examples. We will turn more systematically to the problem of identifying the kinds of equilibrium points in 8.2.

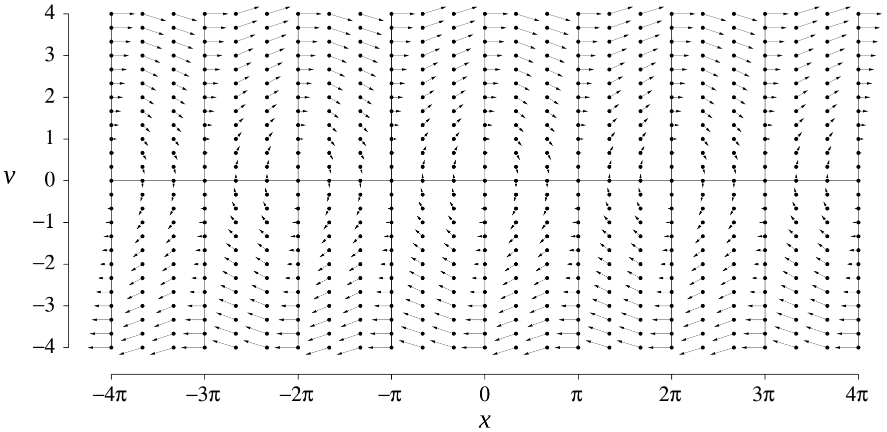
The Pendulum Revisited



In chapter 7 we analyzed the motion of a pendulum. Let's see how this analysis looks when translated into the language of state space. We first need to figure out what the appropriate coordinates are, which means deciding what information we need in order to specify the state of a pendulum. If you look back at the model in the last chapter, you will recall that the two variables we needed were the displacement x and the velocity v. Since x and v can potentially take on any values, our state space will be the entire x-v plane. As before, the dynamical system is specified by the equations

$$x' = v, \qquad v' = -\sin x.$$

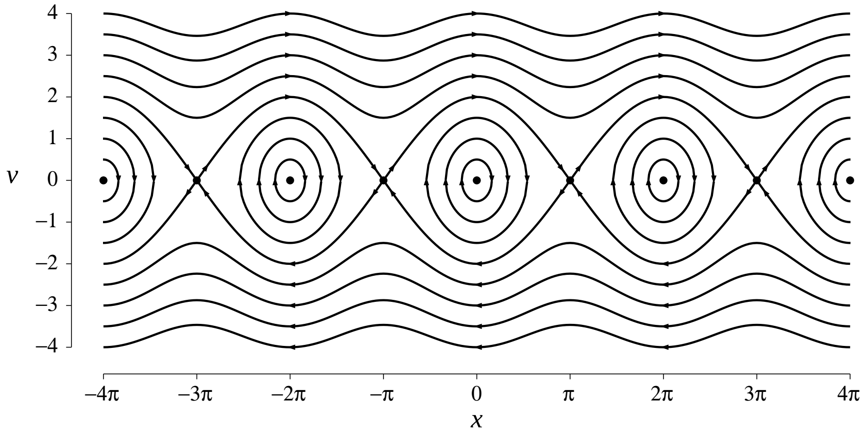
Here is what the vector field for this system looks like:



We have included in this diagram the lines where $v' = 0$ (the vertical lines at every multiple of π) and the line where $x' = 0$ (the horizontal line at $v = 0$). Note that the velocity vectors are horizontal on the lines corresponding to $v' = 0$ and are vertical on the line corresponding to $x' = 0$. The points where these two sets of lines intersect—all points of the form $(k\pi, 0)$ for k an integer—are the equilibrium points of the system. Let's sketch the phase

The equilibrium points

portrait of this system to see more clearly what's going on:



We see that there are several different kinds of trajectories:

Trajectories from left to right in the state plane correspond to $v > 0$

- There are the wavy trajectories moving from left to right across the top of the state space. Note that for these trajectories the value of v is always positive, and x just keeps increasing. These trajectories correspond to the cases where the velocity is great enough that the pendulum can go over the top, continuing to loop around counterclockwise (since x is increasing and x is measured in a counterclockwise direction) forever. Notice that v takes on its minimum value when x is an odd multiple of π, which is what we would expect, since the pendulum is at the top of its arc then. Similarly, v takes on its maximum value at the bottom of its arc—x an even multiple of π.

- The wavy trajectories moving from right to left across the bottom are similar, except that v is always negative. This corresponds to the pendulum wrapping around in a clockwise direction.

Oscillations of the pendulum correspond to closed loops in the state plane

- There are the closed loops. Here x oscillates back and forth between some maximum and minimum value symmetrically placed about an even multiple of π. These trajectories correspond to a pendulum swinging back and forth. The fact that some are centered at x-values other than 0 is due to the fact that the same position of the pendulum can be specified by an infinite number of values of x, all differing from each other by multiples of 2π.

A center: a neutral equilibrium

- There are the equilibrium points $(k\pi, 0)$, with k an even integer. This corresponds to the pendulum hanging straight down. If we perturb the system to a state slightly away from such a point, the pendulum swings back and forth, and the corresponding trajectory loops around the equilibrium point forever. The system neither comes back to the equilibrium point—the condition for an attractor—nor does it go wandering off even further away—the condition for a repellor. Such an equilibrium point is called a **center**. Notice that it is neither an attractor or a repellor; it is said to be a **neutral equilibrium**.

- There are the equilibrium points $(k\pi, 0)$, with k an odd integer, correspond to the pendulum balanced vertically. These are saddle points—if we perturb the system slightly with *exactly* the right v-value for the given x-value, the system will move back toward the vertical position; any other combination, though, will cause the pendulum to wrap around and around forever or to oscillate back and forth forever, depending on whether the v-value is greater than or less than the critical value.

- There are the trajectories connecting the saddle points. These correspond to cases where the pendulum has just enough velocity so that it keeps moving closer to the vertical position without either overshooting and wrapping around, or coming to a stop and reversing direction. In fact, these trajectories divide the state space: on one side of such a trajectory are points corresponding to states where the pendulum will wrap around, and on the other side are points corresponding to states where the pendulum will swing back and forth. Note that the saddle points are *not* part of these trajectories, and that each arc between saddle points is a separate trajectory—you can't get from a point on one of them to a point on another.

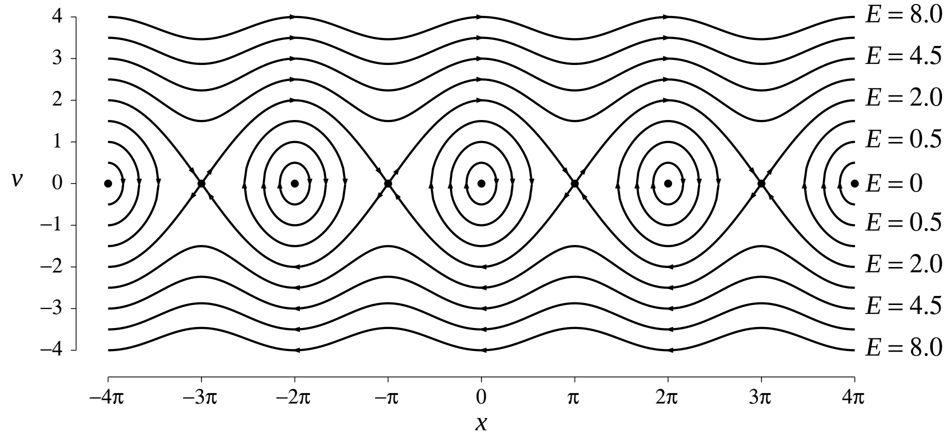
A connected curve in the phase portrait may be composed of more than one trajectory

First Integrals Again

In the case of the pendulum, we have another way of thinking about the trajectories. Recall that in chapter 7.3 we saw that, for any given initial conditions, the quantity $E = \frac{1}{2}v^2 + 1 - \cos x$ was constant over time. In the vocabulary of this chapter, if (x, v) is any state on the trajectory through (x_1, v_1), then it must be true that $\frac{1}{2}v^2 + 1 - \cos x = \frac{1}{2}v_1^2 + 1 - \cos x_1$. But this relation determines a curve in the x-v plane. We thus have an algebraic condition for each of the trajectories, which will depend on the initial values. In this example we could actually get an equation for each trajectory by first using the initial values to determine the energy $E = \frac{1}{2}v_1^2 + 1 - \cos x_1$.

First integrals can give equations of trajectories

We could then solve for v in terms of x by $v = \pm\sqrt{2(E - 1 + \cos x)}$ and plot the resulting function. (Whether we took the plus sign or the minus sign would depend on whether the pendulum was moving counterclockwise or clockwise.) Let's return to our previous sketch of the phase portrait and label some of the trajectories by their corresponding values of E:



Note that for each value of $E \geq 0$ there is more than one trajectory having that value as its energy.

We can now characterize the different kinds of trajectories by their associated energy E:

- If $E > 2$ we get a trajectory extending from $x = \infty$ to $x = -\infty$ (or vice versa).
- If $0 < E < 2$, the trajectory is a closed loop.
- If $E = 0$, we get a neutral equilibrium point.
- If $E = 2$, we get either a saddle point equilibrium, or a trajectory connecting two such saddle points.

A Model for the Acquisition of Immunity

A mathematical model can be used to think about the feasibility of a proposed explanation

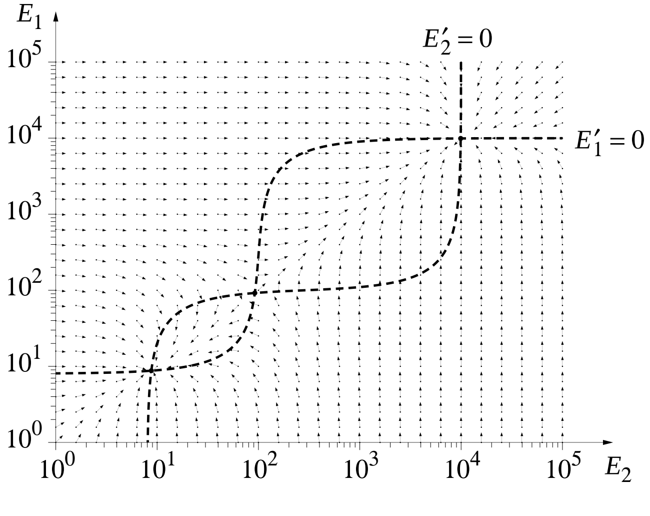
One of the roles of mathematical modelling is to allow researchers to explore possible mechanisms to explain an observed phenomenon. As an example of this, consider the phenomenon of immunity: for many infections, particularly those due to viruses, once you've been exposed to the disease your body continues to produce high levels of antibodies to the disease for the rest of your life, even in the absence of any further stimulation from the virus.

A capsule summary of the immune response: Vertebrates have a wide variety of specialized cells called *lymphocytes* circulating in their blood streams and lymphatic systems at all times. Each lymphocyte has the ability to recognize and bind with a specific kind of invading organism. The invader is called an *antigen*, and the neutralizing molecules produced by the responding lymphocytes are called *antibodies*. Prior to infection, the concentration level of a particular antibody is typically so low as to be undetectable, but the appearance of the antigen causes the system to respond by producing large quantities of the appropriate antibody. If the body can continue to produce high levels of antibodies, it will be immune to reinfection.

In their book *Infectious Diseases of Humans*, Roy Anderson and Robert May propose the following model as a possible mechanism for how antibody levels are sustained. Suppose that there are two kinds of lymphocytes (called *effector cells*) whose densities at time t are denoted by $E_1(t)$ and $E_2(t)$, with the type 2 cells being the potential antibodies for the disease in question. They assume further that new cells of type i ($i = 1$ or $i = 2$) are produced by the bone marrow at constant rates Λ_i and they die at a per capita rates of μ_i. They assume that each cell type is an antigen for the other—that is, contact with cell type 2 triggers cell type 1 to proliferate, and vice versa. They further assume that this proliferation response saturates to a maximum net rate which is dependent on the product of their respective densities. The following equations express this behavior:

$$\begin{aligned} dE_1/dt &= \Lambda_1 - \mu_1 E_1 + a_1 E_1 E_2/(1 + b_1 E_1 E_2), \\ dE_2/dt &= \Lambda_2 - \mu_2 E_2 + a_2 E_1 E_2/(1 + b_2 E_1 E_2). \end{aligned}$$

Here the parameters Λ_i, μ_i, a_i, and b_i would have to be determined by experimental means. At this stage, though, when we are simply exploring to see if such a mechanism might account for the phenomenon of permanent immunity, we can try a range of values for the parameters to see how they affect the behavior of the model.



In the figure above, we have taken $\Lambda_1 = \Lambda_2 = 8000$, $\mu_1 = \mu_2 = 1000$, $a_1 = a_2 = 10$, and $b_1 = b_2 = 10^{-6}$.

There are a couple of features to notice about this graph:

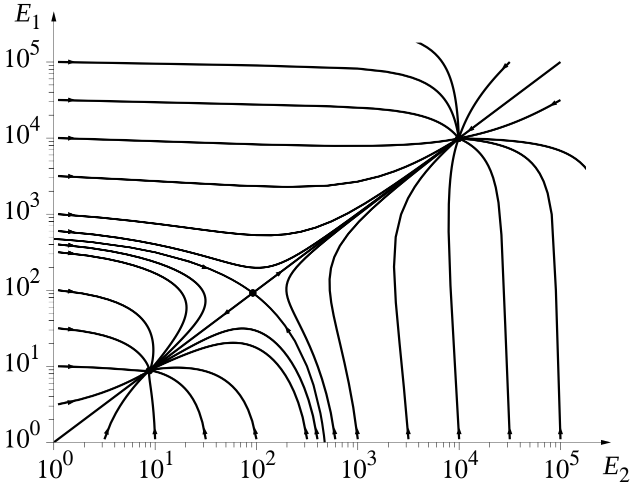
Knowing the direction of the trajectories is often sufficient

1. Since the velocity vectors differ so much in their size, we have recorded only the direction of the velocity vectors, drawing all the arrows to be the same length. Thus we don't really show the vector field, but its close relative, the **direction field**. This is often a useful substitute.

Expressing graphical information over several different orders of magnitude

2. Since the range of values we want to represent is so great we have employed a common device from the sciences of plotting the values on a **log-log scale**. That is, we have plotted the values so that each interval spanning a power of 10—from 10^0 to 10^1, or from 10^3 to 10^4—gets the same space. This is equivalent to plotting the logarithms of the values on ordinary graph paper. This allows us to see effects that take place at different scales. If we hadn't done this, but had plotted this information on regular graph paper with the values running from 0 to 10^5, then some of our most interesting behavior—from 10^0 to 10^2—would be compressed into the lower left-hand corner of the graph, occupying only .001 of the vertical and horizontal scales.

We have included in the graph the two curves corresponding to all points satisfying $E_1' = 0$ and $E_2' = 0$ (note that these curves are *not* trajectories). These curves intersect at the three points $P_1 = (8.7689, 8.7689)$, $P_2 = (92.0869, 92.0869)$, and $P_3 = (9907.14, 9907.14)$, which are then the equilibrium points of this system. The points P_1 and P_3 appear to be attractors, while the point P_2 is a saddle point. In the next section (see page 487) we will see how to zoom in and look at the trajectories near each of these points to confirm this impression. Here is a picture of the phase portrait for this system.



Note that neither P_1 nor P_3 is an attractor for the entire system. The **basin of attraction** for P_1 appears to be a region in the lower left of the graph, while the basin of attraction for P_3 is everything else. The boundary separating these two basins is formed by the two heavily shaded trajectories which come toward the point P_2 (since P_2 is a saddle point, there are only two such trajectories—every other trajectory eventually veers off and heads toward either P_1 or P_3).

We can now interpret this system in the following way. State P_1 represents the **virgin** or **resting state** of the system, with coordinate values on the order of magnitude of $E_i \approx \Lambda_i/\mu_i$, which would just be the steady state values we would have if there were no interactions between the two kinds of cells (why?). Note that after small perturbations (i.e., anything roughly less than a 10-fold increase of type 1 or type 2 cells) from P_1, the system will settle back to this resting state.

Now, though, suppose a viral pathogen appears which possesses an antigen which is identical to that expressed by cell type 1. This has an effect equivalent to moving vertically in the E_1-E_2 plane to a state which is now in the basin for P_3. As a result, the system immediately starts producing large quantities of type 2 cells (which are antibodies for the virus) very rapidly, the virus is wiped out, and the system settles into a new state—the **immune state**—P_3 and remains there. There are now so many type 2 cells permanently floating around the body that no further infection by the viral pathogen is possible. The only way the system can be switched back to state P_1 is if some other agent, such as radiation therapy or infection with an HIV virus, for instance, kills off large numbers of *both* the type 1 and type 2 cells, moving the system back into the basin of attraction for P_1. Just killing off large numbers of one type of cell won't move the system back to stateP_1—do you see why?

Exercises

Two-species interactions

We look at some variations of the predator-prey model. While the original context is given in terms of rabbits and foxes, similar models can be constructed for a variety of interactions between populations—not just predator and prey. The key features of the models are determined by the nature of the **feedback structure** between the populations. In the predator-prey models, the number of foxes has a negative effect on the growth rate of rabbits—the more foxes, the slower the rabbit population grows—while the number of rabbits has a positive effect on the growth rate of foxes. Can you think of other pairs of quantities whose interaction is of this sort? In the first problem we will look at several different models for predator-prey interactions. In the following three problems we will look at models for other kinds of feedback structures.

1. Below are four predator-prey models. In each model all the letters other than R and F are constant parameters. You can perform a general analysis, giving your answers in terms of the unspecified parameters a, b, c, etc., or, if you are more comfortable with specific values, perform the analysis using $a = .1$, $b = .005$, $c = .00004$, $d = g = .04$, $e = .001$, $f = .05$, $h = .004$, and $K = 10,000$. For each model you should carry out the following steps to sketch the vector field for the model in the first quadrant of the R-F plane. Compare your work with the steps that led up to the analysis of the vector field on page 468.

- Write down in words a justification for each rate equation—why is the model a reasonable one? What is it saying about the way rabbit and fox populations change?

- Draw (in red) the set of points where $R' = 0$, and mark the regions where $R' > 0$ and $R' < 0$.

- Draw (in green) the set of points where $F' = 0$, and mark the regions where $F' > 0$ and $F' < 0$.

- Mark the equilibrium points. What color are they?

- Sketch representative vectors of the vector field, and then sketch a couple of trajectories that follow these vectors. You might use a computer to verify your sketches.

- On the basis of your sketches make a conjecture about the stability of the equilibrium points.

a) The original Lotka–Volterra model, proposed independently in the mid-1920's by Lotka and Volterra. This model stimulated much of the subsequent development of mathematical population biology.

$$\begin{aligned} R' &= aR - bR\,F, \\ F' &= cR\,F - dF. \end{aligned}$$

b) The Leslie–Gower model.

$$\begin{aligned} R' &= aR - bRF, \\ F' &= \left(e - f\frac{F}{R}\right)F. \end{aligned}$$

c) Leslie–Gower with carrying capacity for rabbits.

$$\begin{aligned} R' &= aR\left(1 - \frac{R}{K}\right) - bRF, \\ F' &= \left(e - f\frac{F}{R}\right)F. \end{aligned}$$

d) Another combination.

$$\begin{aligned} R' &= aR\left(1 - \frac{R}{K}\right) - bRF, \\ F' &= cRF + gF - hF^2. \end{aligned}$$

2. **Symbiosis and mutualism**. Many flowers cannot pollinate themselves; instead insects like bees transport pollen from one flower to another. For their part, bees collect nectar from flowers and make honey to feed new bees. This sort of feedback structure in which the presence of each element has a positive effect on the growth rate of the other is called **symbiosis** or **mutualism** (there is a distinction made between these two interactions, but mathematically they are similar). Here is a model: B is the number of bees per acre, measured in hundreds of bees, while C is the weight of clover per acre, in thousands of pounds. Assume time to be measured in months.

$$\begin{aligned} B' &= .1(1 - .01B + .005C)B, \\ C' &= .03(1 + .04B - .1C)C. \end{aligned}$$

a) Do these equations describe symbiosis? What terms account for symbiosis?

b) Each equation has a negative term in it. What aspect of reality is this term capturing?

c) Sketch the vector field for this system in the B-C plane. Find the equilibrium points, and mark them on your sketch.

d) Draw some trajectories on your sketch, and use them to determine the stability of the equilibrium points.

e) Suppose an acre of land has 10,000 pounds of clover on it, and a hive of 2,000 bees is introduced. (What are the values of $B(0)$ and $C(0)$ in this case?) What happens? Answer this question both by drawing a trajectory and by describing the situation in words.

f) Let a couple of years pass after the situation in part (e) has stabilized. Suppose the field is now mowed so only 2,000 pounds of clover remain on it. The bee–clover system is now at what point on the B-C plane? What happens now? Does the bee population drop? Does it stay down, or does it recover? Does the clover grow back?

g) This scenario is an alternative to part (f); it is also played out a couple of years after the situation in part (e) has stabilized. Suppose an insecticide applied to the clover field kills two-thirds of the bees. The insecticide is then washed away by rain, leaving the remaining bees unaffected. What happens?

3. **Competition** As a third kind of feedback structure, consider two species X and Y competing for the same food or territory. In this case each has a negative impact on the growth rate of the other. If we let x and y be the number of individuals of species X and Y, respectively, then the larger y is, the less rapidly x increases—and vice versa. Here is a specific model to consider:

$$\begin{aligned} x' &= .15(1 - .005x - .010y)x, \\ y' &= .03(1 - .004x - .005y)y. \end{aligned}$$

The term $-.010y$ in the first equation shows explicitly how an increase in y reduces the growth rate x'. In the second equation $-.004x$ tells us how much X affects the growth of Y. Notice that Y affects X more strongly than X affects Y.

If x and y are both small, then the parenthetical terms are approximately equal to 1, so the equations reduce to

$$\begin{aligned} x' &= .15x, \\ y' &= .03y. \end{aligned}$$

Thus, in these circumstances X's per capita growth rate is five times as large as Y's.

In the competition for resources will the growth rate advantage permit X to win the competition and drive out Y, or will the more adverse effect that Y has on the growth of X permit Y to win? Perhaps the two species will both survive and share the resources for which they compete. The purpose of this exercise is to decide these questions.

a) Suppose we start with $x = y = 10$. What are the two growth rates x' and y'? Is x' about five times as large as y' in this case? What are the approximate values of x and y after .5 time units have elapsed? Is X growing significantly more rapidly than Y?

b) How many equilibrium points does this system have, and where are they?

c) Sketch *and label* in the x-y plane the points where $x' = 0$ and where $y' = 0$. The vector field typically points in one of four directions: up and to the right; up and to the left; down and to the right; or, down and to the left. Indicate on your sketch the zones where these different directions occur and draw representative vectors in each zone.

[Note: Only three of the zones actually occur in the first quadrant; no vectors there point down and to the right.]

d) Sketch on the x-y plane the trajectory that starts at the point $(x, y) = (10, 10)$. Now answer the question: What happens to a population of 10 individuals each from species X and from species Y? In particular, does X gain an early lead? Does X keep its lead? Does either X or Y eventually vanish?

e) Is the outcome of part (d) typical, or is it not? Try several other starting points: $(x, y) = (150, 25)$, $(300, 10)$, $(200, 200)$, $(50, 200)$. Do these starting points lead to the same *eventual* outcome, or are there different outcomes? Use a computer to confirm your analysis.

f) Describe the type of each equilibrium point you found in part (a). Is any equilibrium an attractor?

4. **Fairer competition**. The vector field in question 3 shows that species X didn't have a chance: all trajectories in the first quadrant flow to the equilibrium at $(0, 200)$. We can attribute this to the strength of the adverse effect Y has on X—that is, to the size of the term $-.010y$ in the first equation when compared to the corresponding term $-.004x$ in the second equation.

Let's try to give X a better chance by increasing this term to $-.006x$. The equations become

$$\begin{aligned} x' &= .15(1 - .005x - .010y)x, \\ y' &= .03(1 - .006x - .005y)y. \end{aligned}$$

a) Sketch and label in the x-y plane the points where $x' = 0$ and where $y' = 0$. Sketch representative vectors for the vector field. Mark all equilibrium points.

b) What happens to a population consisting of 10 individuals each from species X and species Y? Is the outcome significantly different from what it was in question 3? To get quantitatively precise results you will probably find a computer helpful.

c) What happens to a population consisting of 150 individuals from species X and 25 individuals from species Y? Is this outcome significantly different from what it was in question 3?

d) Is it possible for X and Y to coexist? What must x and y be? Is that coexistence *stable*; that is, if x and y are changed slightly, will the original values be restored?

e) Sometimes X wins the competition, sometimes Y. Mark in the x-y plane the dividing line between those starting points which lead to X winning and those which lead to Y winning.

f) Identify the type of each equilibrium point.

g) An often-articulated concept in ecology is the *principle of competitive exclusion*, which states that you can't have a stable situation in which two species compete for the same resource—one of them will eventually crowd out the other. Is the model you've been exploring in this problem consistent with such a principle?

5. **More on the Lotka–Volterra model**. The Lotka–Volterra model,

$$\begin{aligned} R' &= aR - bRF, \\ F' &= cRF - dF, \end{aligned}$$

while it had a major impact on the development of mathematical biology, was found to be flawed in several important ways. The chief problem is that the equilibrium point $(d/c, a/b)$ is a neutral equilibrium point—given any starting state, the system would follow a closed trajectory. This in itself was all right and, in fact, stimulated a great many important investigations on whether or not cycles were an intrinsic feature of many populations. The difficulty was that there were so many possible closed trajectories—which one the system followed depended on where it started. A second difficulty, related to the first, is that there is a first integral for the Lotka–Volterra model. What is seen as a virtue in a physical system like the pendulum—since it is equivalent to the conservation of energy—is unrealistic in an ecological system, where there are almost certainly too many outside forces at work for any quantity to be conserved there. In the following exercises we will explore some of these behaviors. As before, you can either perform a general analysis of the model or use the specific parameter values $a = .1$, $b = .005$, $c = .00004$, and $d = .04$.

a) Sketch the vector field, together with some typical trajectories, in the rest of the *R*-*F* plane, including negative values. What happens to any trajectory starting at a state with a negative R or negative F value?

b) For this exercise you will need to go back to a computer program that implements Euler's method of approximating the trajectory by drawing a straight line segment from a point in the direction indicated by the velocity vector (commercial packages use fancier routines which accommodate for the kind of phenomena you are about to see!). Using the specific values for a, b, c, and d suggested above, starting from the point $(2000, 10)$ in the *R*-*F* plane, and using a time step $\Delta t = 1$, draw the first 500 segments of Euler's approximation to the trajectory. What does the trajectory look like? Would you think the trajectory was a closed loop on the basis of this result? How small does Δt have to be before the trajectory looks like it closes? Can you explain this phenomenon?

c) Using the same values for a, b, c, and d as in the preceding part, start at the point $(2000, 1)$ and use $\Delta t = 2$. This time calculate the first 1000 segments of Euler's approximation; what happens? (Your computer will probably give you some sort of overflow message.) Can you explain this? (Think about your answer to part (a).)

d) **Getting a first integral for the system** Show that the Lotka–Volterra equations imply that

$$\frac{R'}{R}(cR - d) = \frac{F'}{F}(a - bF).$$

Integrate this equation and show that the expression

$$cR + bF - d \ln R - a \ln F$$

must be a constant for all points on a given trajectory. If we know one point on the trajectory (such as the starting point), we can evaluate the constant.

e) Show that the function $f(R) = cR - d \ln R$ is decreasing for $0 < R < d/c$ and is increasing for $d/c < R < \infty$. Hence argue that for any given value of F there are at most two values of R giving the same value for the expression $cR+bF-d\ln R-a\ln F$. Hence conclude that the trajectories for the Lotka–Volterra equations can't be spirals, but must then be closed loops.

The pendulum

6. Suppose instead of an idealized frictionless pendulum, we wanted to model a pendulum that "ran down". One approach we might try is to throw in a term for air resistance. Let's see what happens when we add a term to the expression for v' which suggests that there is a drag effect which is proportional to the value of v—the larger v is, the greater will be the drag. Here are equations that do this:

$$\begin{aligned} x' &= v, \\ v' &= -\sin x - .1v. \end{aligned}$$

Perform a vector field analysis of this model, indicating the regions where the velocity vectors are pointing in the various combinations of up, down, right, and left. Try sketching in some trajectories. Where are the equilibrium points? What kinds are they?

The Anderson–May model

7. Consider $dE_1/dt = \Lambda_1 - \mu_1 E_1 + a_1 E_1 E_2/(1 + b_1 E_1 E_2)$. For what values of E_1 is it possible to find a value for E_2 making $dE_1/dt = 0$? Express your answer in terms of the parameters Λ_1, μ_1, a_1, and b_1. Is your answer consistent with the graph on page 475?

8. In the same book—*Infectious Diseases of Humans*—containing the previous model, Anderson and May propose another model to explain the acquisition of (apparently) permanent immunity. In this model there is just the

virus and the lymphocyte cells (effector cells) that kill the virus. We denote their populations at time t by $V(t)$ and $E(t)$. They propose the model

$$\begin{aligned} dE/dt &= \Lambda - \mu E + \varepsilon VE, \\ dV/dt &= rV - \sigma VE. \end{aligned}$$

Here Λ is the (constant) rate of background production of the lymphocytes by the bone marrow, μ is the per capita death rate of such cells, and r is the intrinsic growth rate of the virus if none of the specific lymphocytes was present. Both the increased production of the lymphocytes and the death of the virus are assumed to proceed at rates proportional to the number of their interactions, determined by their product.

a) Show that in the absence of any virus, the effector cells have a stable equilibrium of Λ/μ.

b) Perform a state space analysis of the vector field. Note that there will be two very different cases, depending on whether $\Lambda/\mu > r/\sigma$ or $\Lambda/\mu < r/\sigma$. In each case say what you can about the equilibrium points and the expected long-term behavior of the system.

c) Using parameter values $\Lambda = 1$, $\mu = r = .5$, and $\varepsilon = \sigma = .01$, and starting values $E = V = 1$ find the resulting trajectory. (The trajectory will be a spiral, but it moves in very slowly.)

d) How long, approximately, will it take the spiral to make one revolution? If this time, call it T, is roughly the same length as the lifetime of the infected individual, what will appear to be happening? It might help to plot both E and V as functions of time over the interval $[0, T]$.

8.2 Local Behavior of Dynamical Systems

A Microscopic View

Phase portraits under the microscope

One of the themes of this book has been the concept of the "microscope". When we zoom in on some part of a geometrical object, the structure typically becomes much simpler. In chapter 3 we used this approach to think about the behavior of functions. In this section we will use the same idea to analyze the behavior of a vector field and its phase portrait. There are two parts to this process:

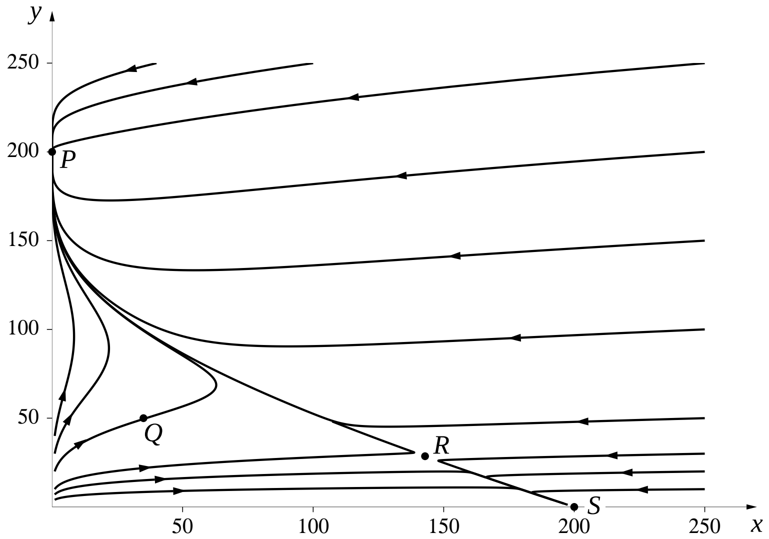
1. We shift the origin of the coordinate system to center on the point we are interested in—we **localize**—and

2. We approximate both the vector field and its phase portrait by suitable linear approximations—we **linearize**.

To get a feel for how this works, let's go back and look at problem 4 on page 481 of the previous section. There we had two species X and Y competing for the same food source. We modeled the dynamics of this system by the equations

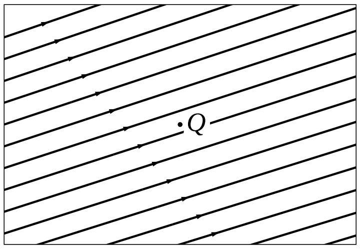
$$x' = .15(1 - .005x - .010y)x,$$
$$y' = .03(1 - .006x - .005y)y.$$

The phase portrait for this system looks like the following figure.



The three equilibrium points—$P = (0, 200)$, $R = (1000/7, 200/7)$, and $S = (0, 200)$—are indicated, together with a generic point $Q = (35, 50)$. Note that P and S are attractors and that R is a saddle point. As was the case with the Anderson–May model, there is a trajectory flowing away from R to each of the attractors. There are also two trajectories (not shown) flowing directly toward R and forming the boundary between the basins of attraction for P and S. We will see how to construct this boundary shortly (page 497).

Let's first zoom in on the point Q and see what the phase portrait looks like there. If we take the region ±1 unit on either side of Q, we get the following phase portrait:



At this level, all the trajectories appear to be parallel straight lines. How could we have anticipated this picture? The first step in analyzing this phase portrait is to observe that since we are interested in its behavior near $Q = (35, 50)$, instead of working with the variables $x(t)$ and $y(t)$, we introduce new variables $r(t)$ and $s(t)$ which measure how far we are from Q:

$$\begin{aligned} r(t) &= x(t) - 35, \\ s(t) &= y(t) - 50. \end{aligned}$$

The effect of this transformation is simply to shift the origin to the point Q—the location of every point in the plane is now measured relative to Q rather than to the x-y origin. A point is close to the point Q if its r-s coordinates are small. Further, if we are given the r-s coordinates of a point, we can always recover the x-y coordinates, and vice versa—we can transform in either direction:

Shifting the origin

$$\begin{aligned} r &= x - 35, &\iff\quad x &= r + 35, \\ s &= y - 50, &\iff\quad y &= s + 50. \end{aligned}$$

Next, note that $r'(t) = x'(t)$ and $s'(t) = y'(t)$ so that the new variables change at the same rates as the old ones. We can now express our original differential equations in terms of the variables r and s by replacing x' by r', x by $r + 35$, y' by s', and y by $s + 50$. When we do this, we get

$$\begin{aligned} r' &= .15(1 - .005(r + 35) - .010(s + 50))(r + 35) \\ &= 1.70625 + .0225r - .0525s - .00075r^2 - .0015rs, \\ s' &= .03(1 - .006(r + 35) - .005(s + 50))(s + 50) \\ &= .81 - .009r + .0087s - .00018rs - .00015s^2. \end{aligned}$$

What we have accomplished by this is to transform a problem about trajectories near the point $(35, 50))$ in the x-y plane into a problem about trajectories near the origin in the r-s plane—we have **localized** the problem to the point we are interested in.

Near an ordinary point, a vector field is almost constant

The second step comes in analyzing the r-s system: since we are only interested in its behavior near the origin, we will be looking at values of r and s that are small. Under these circumstances, the contributions of the constant terms will far outweigh the contributions of any of the terms involving r and s. For instance, in our current example we are looking at a window that is ±1 unit wide and ±1 unit high around Q. In this window, the terms involving r or s are at most 3% of the constant term in the case of r', and a little over 1% in the case of s'. If we had used a smaller window, the contributions of the non-constant terms would be even less significant. This means that *near the r-s origin* the vector field for this system is well-approximated by the behavior of the related **constant linear system**:

$$r' = 1.70625,$$
$$s' = 0.81.$$

Note that 1.70625 and .81 are just the values of x' and y' at Q.

Near an ordinary point all trajectories look the same

In this linearized system, any change Δt in the time produces a change $\Delta r = 1.70625\Delta t$ in r, and a change $\Delta s = .81\Delta t$ in s. Thus the velocity vectors in the vector field near Q would all have the same length and would be pointing in the same direction, with slope $\Delta s/\Delta r = .81/1.70625 = .4747$. This in turn means that near Q all trajectories have the same slope and are traversed at the same speed.

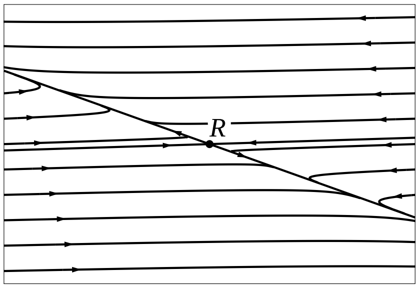
We would see a similar picture—a family of parallel straight lines—whenever we zoom in on the phase portrait near any other ordinary (i.e., non-equilibrium) point (x_*, y_*) The vector field near such a point can always be approximated by a constant linear system of the form

$$r' = e,$$
$$s' = f,$$

where e and f are the values of x' and y' at (x_*, y_*). The trajectories of this approximating linear system will be lines of slope f/e.

Equilibrium points are different

Near an equilibrium point, the picture is more complicated. No matter how far in we zoom, the phase portrait never looks like a family of straight lines. For instance, here's what the picture looks like when we zoom in on $R = (1000/7, 200/7) \approx (142.857, 28.571)$:



—if we zoomed in to a window 1/100-th the size of this one, the picture would be indistinguishable from this one.

Here we see four trajectories that look almost like straight lines—two coming directly towards R and two going directly away. All the other trajectories appear to be asymptotic to these two sets. On page 495 in the next section you will see how to find the equations of these asymptotes.

What happens when we linearize the vector field at R? As before, we first shift the origin so that it is centered at R by changing to coordinates r and s, where

$$\begin{aligned} r(t) &= x(t) - 1000/7, \\ s(t) &= y(t) - 200/7. \end{aligned}$$

When we then write the differential equations in terms of r and s, we get as before that $x' = r'$ and $y' = s'$ and

$$\begin{aligned} r' &= .15(1 - .005(r + 1000/7) - .010(s + 200/7))(r + 1000/7) \\ &= -.107143r - .214286s - .00075r^2 - .0015rs, \\ s' &= .03(1 - .006(r + 1000/7) - .005(s + 200/7))(s + 200/7) \\ &= -.00514286r - .00428571s - .00018rs - .00015s^2. \end{aligned}$$

Linear approximation of the vector field

This time, though, the constant term in the expression for both r' and s' is 0. This is because the point R was an equilibrium point, which meant that both x' and y', and hence r' and s', were 0 there. If we are considering only small values of r and s, though, say much smaller than 1, then the terms involving r^2 or s^2 or rs will be much smaller than the terms involving r and s alone. We can therefore simplify our equations at R by taking only the first powers of r and s, getting for the linearized system

$$\begin{aligned} r' &= -.107143r - .214286s, \\ s' &= -.00514286r - .00428571s. \end{aligned}$$

In a similar fashion we could hope to explore the behavior of any other dynamical system about any of its equilibrium points by approximating the vector field there by a linear system of the form

$$\begin{aligned} r' &= ar + bs, \\ s' &= cr + ds, \end{aligned}$$

for suitable constants a, b, c, and d.

We will see in section 8.3 how to use this linearized form of the vector field to discover many of the properties of equilibrium points.

How can we find values for the constants a, b, c, and d? If the differential equations specifying the rates of change of the variables are polynomials, then we can proceed as above:

- Shift the origin to the point we're interested in;
- Express the rate equations in terms of the new local variables;
- Throw away all the terms except the first degree terms.

This process requires some fairly tedious algebra. Moreover, what if the differential equations are not polynomials? Suppose, for instance, we wanted to study the local behavior of the Anderson–May model (page 475) at the saddle point $P_2 = (92.0869, 92.0869)$. Note that the differential equations are of the form

$$\begin{aligned} dE_1/dt &= f_1(E_1, E_2), \\ dE_2/dt &= f_2(E_1, E_2), \end{aligned}$$

To linearize a vector field, linearize the functions that determine it

where f_1 and f_2 are the functions given in the text. But f_1 and f_2 are just functions, and we learned in chapter 3 how to construct locally linear approximations to them. This was, in fact, how we defined derivatives in the first place. Thus if E_1 changes by a small amount $\Delta E_1 = E_1 - 92.0869$, the function f_i will change by approximately $\partial f_i/\partial E_1 \times \Delta E_1$. Similarly, a small change $\Delta E_2 = E_2 - 92.0869$ will produce a change of approximately

$\partial f_i/\partial E_2 \times \Delta E_2$ in the function f_i. The total change in the function f_i can then be approximated by the sum of these changes:

$$\begin{aligned}\Delta f_1(E_1,E_2) &\approx \frac{\partial f_1}{\partial E_1}\Delta E_1 + \frac{\partial f_1}{\partial E_2}\Delta E_2,\\ \Delta f_2(E_1,E_2) &\approx \frac{\partial f_2}{\partial E_1}\Delta E_1 + \frac{\partial f_2}{\partial E_2}\Delta E_2.\end{aligned}$$

General form for the local linearization at an equilibrium point...

But since P_2 is an equilibrium point, we have by definition that f_1 and f_2 are both zero there, so $\Delta f_i(E_1,E_2) = f_i(E_1,E_2) - f_i(P_2)$ is just $f_i(E_1,E_2)$. Further, if you look closely you will see that the quantity $\Delta E_1 = E_1 - 92.0869$ is identical with what we have been calling the local coordinate r, and $\Delta E_2 = E_2 - 92.0869$ is just the other local coordinate s. Thus, since $E_1' = r'$ and $E_2' = s'$, we have

$$\begin{aligned}r' &= \frac{\partial f_1}{\partial E_1}\,r + \frac{\partial f_1}{\partial E_2}\,s,\\ s' &= \frac{\partial f_2}{\partial E_1}\,r + \frac{\partial f_2}{\partial E_2}\,s,\end{aligned}$$

where the partial derivatives are evaluated at P_2. Notice that there is nothing in this expression which is specific to this particular problem. The local linearization of any vector field at any equilibrium point will be in this form.

Finally, using the values given for the different parameters back on page 475, we can evaluate all the partial derivatives to get the specific local linearization for the point P_2:

$$\begin{aligned}r' &= -94.5525\,r + 905.448\,s,\\ s' &= 905.448\,r - 94.5525\,s,\end{aligned}$$

We will see in the next section how knowing this form will allow us to find the boundary between the two basins of attraction.

...and at a generic point

For completeness, let's remind ourselves of what the local linearization would look like at a nonequilibrium point in the current formulation. The result is immediate and simple, using the analysis we used before. If Q is a generic point, then the local linearization consists of parallel lines, whose slopes are given by the constant rate equations

$$\begin{aligned}r' &= f_1(Q),\\ s' &= f_2(Q).\end{aligned}$$

Exercises

1. Find the local linearizations at all the equilibrium points in exercises 2–4 at the end of the previous section.

2. a) The Lotka–Volterra equations

$$R' = aR - bRF,$$
$$F' = cRF - dF,$$

have an equilibrium point at $(R, F) = 00(d/c, a/b)$.

b) What is the local linearization there?

c) What is a striking feature of this linearization, and what is its physical significance?

d) The trajectories for the local linearizations turn out to be ellipses. If r and f are the local variables, find constants α and β such that the expression $\alpha\, r^2 + \beta\, f^2$ is constant on any trajectory.

3. Find the local linearization at the point P_1 in the Anderson–May model for the acquisition of immunity discussed in the previous section, using the parameter values given in the text on page 476.

4. Go back to the second Anderson–May model analyzed in problem 8 of the previous section (page 484). Using the parameter values given in part (c) there, find the local linearizations at all equilibrium points.

8.3 A Taxonomy of Equilibrium Points

An intuitive classification of equilibrium points

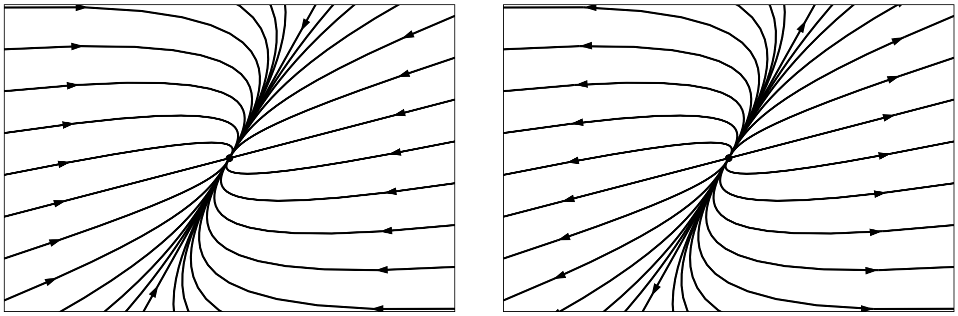
In the exercises and examples we have seen so far in this chapter, there have been several kinds of trajectories near equilibrium points: spirals towards and spirals away from the equilibrium, closed loops about the equilibrium, trajectories that looked vaguely like hyperbolas, and trajectories that seemed to arc more or less directly into or away from the equilibrium. It turns out that this rough classification covers virtually all the equilibrium behaviors we might encounter in a two-dimensional state space. There are many ways to demonstrate this, but we can accomplish almost everything with a couple of simple insights. We begin with a summary of the different kinds of equilibrium points, then turn to the question of devising ways to figure out from the equations what kind we are dealing with.

Suppose, then, that we are studying a two-dimensional dynamical system and that we have linearized the system at an equilibrium point. The point is either an attractor, repellor, saddle point, or neutral point. Attractors and repellors can be further subdivided according to whether they have one or two straight line trajectories, or whether their trajectories are spirals. Note that any attractor can be converted into a repellor simply by reversing the arrows, and vice versa (how do you accomplish this arrow reversal at the level of the defining differential equations?). If you reverse all the arrows at a saddle point, you get another saddle point. If you reverse the arrows at a neutral point, you get the same closed loops, but they are traversed in the opposite direction.

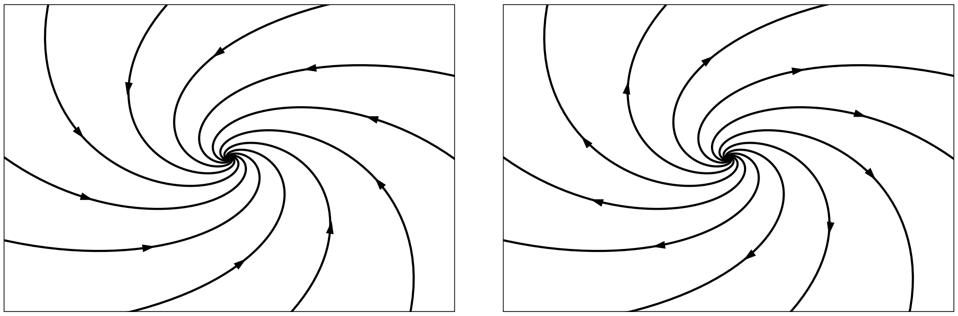
Here, then, is a listing of all the kinds of equilibrium points. There are five generic types. (*Generic* here means “general”; if you generate a random equilibrium point, it will almost certainly be one of these.) They are most easily categorized by whether or not they have **fixed line** trajectories—that is, trajectories which are straight lines going directly toward or directly away from the equilibrium point.

The existence or not of straight line trajectories and how to find them when they do exist is an instance of the so-called *eigenvector* problem. Analogous problems occur elsewhere in many parts of mathematics, physics, and even population biology. Being able to find such eigenvectors efficiently is an important problem in computational mathematics.

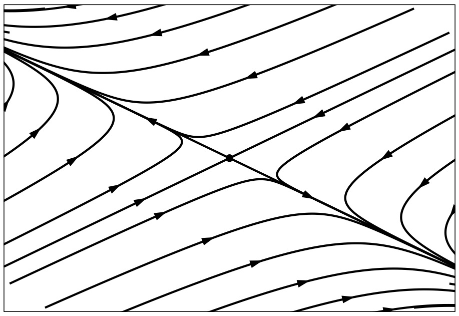
Nodes. Two pairs of fixed lines, all trajectories flowing toward the equilibrium (attractors) or away from it (repellors).



Spirals. No fixed lines, all trajectories spiraling toward the equilibrium (attractors) or away from it (repellors).

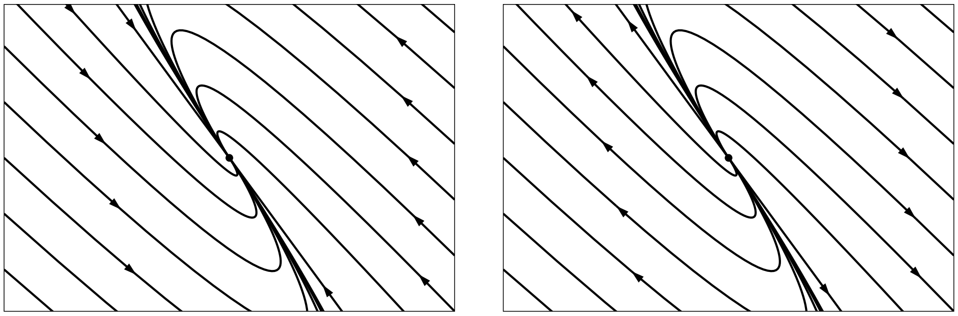


Saddle Point. Two pairs of fixed lines, with the flow along one pair being toward the equilibrium, and the flow along the other pair away from it. All other trajectories are asymptotic to these lines.

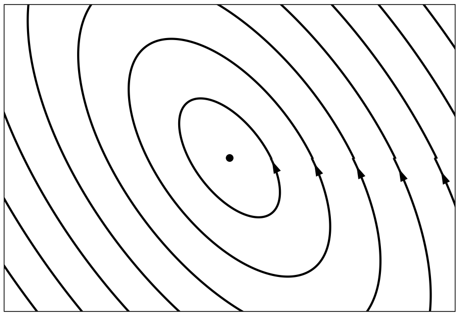


In addition to these five generic cases, there are three more types which arise under more specialized conditions:

Special Nodes. One pair of fixed lines, all trajectories flowing toward the equilibrium (attractors) or away from it (repellors).



Center. No fixed lines, all trajectories flowing around the equilibrium in closed loops.



Except for a variety of highly specialized (or *degenerate*, in mathematical terminology) cases, examples of which are given in the exercises, the region near every equilibrium point will look like one of the above (although the exact shape may vary).

Clearly, it would be helpful to have an efficient way to determine whether or not fixed lines exist, and what their equations are if they do.

Straight-Line Trajectories

Given a dynamical system

$$\begin{aligned} r' &= ar + bs, \\ s' &= cr + ds, \end{aligned}$$

how can we tell whether or not it has any straight-line trajectories? If $b = 0$, then the (vertical) line $r = 0$ is a trajectory. Otherwise, note that the line $s = mr$ will be a trajectory for this system provided the slope of the line—namely m—equals the slope of the vector field at every point (r, s) on the line. But the slope of the vector field at any point (r, s) is just s'/r', which in turn is equal to $(cr + ds)/(ar + bs)$. Since every point on the line of slope m is of the form (r, mr), what we are really asking, then, is whether there are any values of m which satisfy the equation

$$m = \frac{cr + dmr}{ar + bmr} = \frac{c + dm}{a + bm}$$

The condition for a fixed-line trajectory

To see how this works, let's return to the example of two competing species which we last looked at on page 486. There we zoomed in on the saddle point $R = (1000/7, 200/7)$ and found that the local linear approximation was

$$\begin{aligned} r' &= -.1071r - .2143s, \\ s' &= -.0051r - .0043s. \end{aligned}$$

If this system has a straight-line trajectory of slope m, then m must satisfy

$$m = \frac{-.0051 - .0043m}{-.1071 - .2143m},$$

which leads to the quadratic equation

$$.2143m^2 + .1028m - .0051 = 0,$$

which has roots

$$m = .0454 \quad \text{and} \quad m = -.5250.$$

Thus the lines $s = .0454r$ and $s = -.5250r$ are trajectories of the linear system. To be more exact, each of these lines is made up of three distinct trajectories: the portion of the line consisting of all points with $r > 0$, the portion with $r < 0$, and the origin (which is the saddle point R) by itself, which is always a trajectory in any linear system. To see whether flow along these trajectories is towards the origin or away from it, we could look to see where the lines lie in the state plane. It is just as simple, though, to try a test point. For instance, a typical point on the line $s = .0454r$ is $(1, .0454)$. When we substitute these values into the original rate equations, we find that

$$\begin{aligned} r' &= -.1071 \times 1 - .2143 \times .0454, \\ s' &= -.0051 \times 1 - .0043 \times .0454. \end{aligned}$$

We don't even need to do the arithmetic to be able to tell that both r' and s' are negative at this point, hence both r and s are decreasing, which means that on the line of slope .0454 movement is towards the origin. Similarly, on the line of slope $-.5250$ the flow is away from the origin. Finally, it turns out (as is the case with every linear system with straight-line trajectories) that every other trajectory is asymptotic to these lines.

The crux of this approach was the use of the quadratic formula. Of course, it may happen—and we will see examples in the exercises—that when we try the same approach on another system we find there are no real roots to the equation. This means that there are no fixed lines, so that trajectories must be spirals or closed loops.

Attractors and Basins of Attraction

One byproduct of the analysis in the previous section is that it gives us a technique for sketching the boundary separating two basins of attraction. Let's continue with the previous example to illustrate how this is done. We observed that the boundary between the two basins was formed by the two trajectories coming directly into the saddle point R between the two attractors P and S. We have just seen that near R these two trajectories looked like the straight line of slope .0454. We can therefore take a point on this line on each side of R and run the system backward (if we go forward, we simply approach R) in time to reconstruct the trajectories, and hence get the boundary of the basins of attraction.

Exercises

1. In this exercise we look at a number of different linear systems to see what kinds of trajectories we get. In each case you should sketch the trajectories. Do this as before by first identifying the regions in the plane where $r' = 0, r' > 0$, and $r' < 0$, and similarly for s'. Then sketch trajectories consistent with this information. You might want to use a graphing program to check any answer you're unsure of.

a) $r' = 4r + s, \qquad s' = 2r + 3s.$

b) $r' = 4r + s, \qquad s' = -2r + 3s.$

c) $r' = 2r + 3s, \qquad s' = 4r + s.$

d) $r' = -4r + 4s, \qquad s' = 2r + s.$

e) $r' = -.4r - 4s, \qquad s' = 2r - .5s.$

f) $r' = -.4r - 4s, \qquad s' = 2r + .4s.$

g) Make up and analyze four more linear systems.

2. If you start with a given linear system and consider the related system in which all the coefficients are four times as big, how do the trajectories change?

3. If you start with a given linear system and consider the related system in which all the coefficients have their signs reversed, how do the trajectories change?

4. a) Use the quadratic formula to find the general solution to the equation

$$m = \frac{c + dm}{a + bm}.$$

b) In exercises 1, 3, and 4 in the previous section you found local linearizations at the equilibrium points of a number of examples discussed earlier. Determine which of these have straight-line trajectories and which do not. For those that do, find the equations of the lines and determine for each line whether the flow is towards the origin or away from it.

c) What is the general condition for a linear dynamical system to have straight-line trajectories?

5. Make up a system that has the lines of slope ±1 as trajectories.

6. What is the condition for a system to have exactly one fixed line? Construct a couple of systems that have only one fixed line and sketch their phase portraits.

7. **Degeneracy** The analysis developed in this section implicitly assumed that in the local linearization, at least one of the coefficients in each of the expressions for r' and s' was non-zero. If this is not true, then many more possibilities open up. The following two systems have the origin as their only equilibrium point. In each case, write down the local linearization and draw in the trajectory pattern for the linearized system. Notice that the linearized systems have more than one equilibrium point. Then do the standard phase plane analysis for the original system—identify the regions in the plane where $r' = 0$and where $s' = 0$, and specify what the direction field is doing in the rest of the plane, as usual. Sketch in some typical trajectories. Comment on the connections between the linearized and unlinearized forms.

a) $r' = r^2, \qquad s' = -s$. You should see sort of a hybrid between a saddle point and an attractor here.

b) $r' = r^2 + s^2, \qquad s' = r$.

8. a) Use the technique presented at the end of this section (page 497 to graph the boundary between the two basins of attraction.

b) In the same way, construct the boundary between the two basins in the competing species model we've been discussing—problem 4 on page 481.

Distance from the Origin

Another way to distinguish between different kinds of trajectories is to see how their distance from the origin varies over time. For saddle points the distance will first decrease and then increase. For spiral attractors and nodal attractors the distance may be always decreasing, or it may fluctuate, depending on how flat the trajectory is.

Again, let's look at a general linear system

$$\begin{aligned} r' &= ar + bs, \\ s' &= cr + ds. \end{aligned}$$

onsider the system moving along some trajectory in r-s space. At time t it will be at a point $(r(t), s(t)$, situated at a distance $d(t) = \sqrt{r(t)^2 + s(t)^2}$. We would like to know how the function $d(t)$ behaves. Is it always increasing? Always decreasing? Or does it have local maxima and minima? To answer this we need to know if $d'(t)$ is ever $= 0$, or if it is always positive or always negative. We can simplify our calculations if we look at the square of the distance: $D(t) = d(t)^2 = r(t)^2 + s(t)^2$. The function D will be increasing and decreasing at exactly the same points as the function d, and it's easier to work with.

9. a) Show that

$$\begin{aligned} D'(t) &= 2r(t)r'(t) + 2s(t)s'(t) \\ &= 2[r(ar + bs) + s(cr + ds)] \\ &= 2[ar^2 + (b + c)rs + ds^2]. \end{aligned}$$

b) Show that if we look at points on the line of slope m, so that $s = mr$, we will have $D'(t) = 0$ there if and only if

$$a + (b + c)m + dm^2 = 0\,.$$

c) Use the quadratic formula to conclude that this happens precisely where

$$m = \frac{-(b + c) \pm \sqrt{(b + c)^2 - 4ad}}{2d}\,.$$

d) Show in particular, if $(b+c)^2 - 4ad < 0$, there are no solutions to $D'(t) = 0$, and the distance must always be strictly increasing along all trajectories, or strictly decreasing along all trajectories.

e) Return to the example

$$\begin{aligned} r' &= -.1071r - .2143s, \\ s' &= -.0051r - .0043s, \end{aligned}$$

and find the equations of the two lines where trajectories pass closest to the origin. These lines will not be trajectories themselves. Their significance is that the 'vertices' of all the trajectories will lie along them.

10. Choose four of the exercise in the first part of this section and analyze them to see where (and whether) trajectories have a closest approach to the origin.

11. a) Use the results of this section to construct a dynamical system whose trajectories are spirals that are always moving away from the origin.

b) Use the results to construct a dynamical system whose trajectories are flattened spirals, so that the distance from the origin, while increasing overall, has local maxima and minima.

12. It turns out that trajectories which form closed loops should really be considered as a special kind of spiral. In fact, a flattened spiral will close up precisely when the two directions in which the distance is a maximum or minimum are perpendicular to each other. Express this as a condition on the coefficients a, b, c, and d in the dynamical system.

13. Write down the equations of some dynamical systems that will have closed orbits.

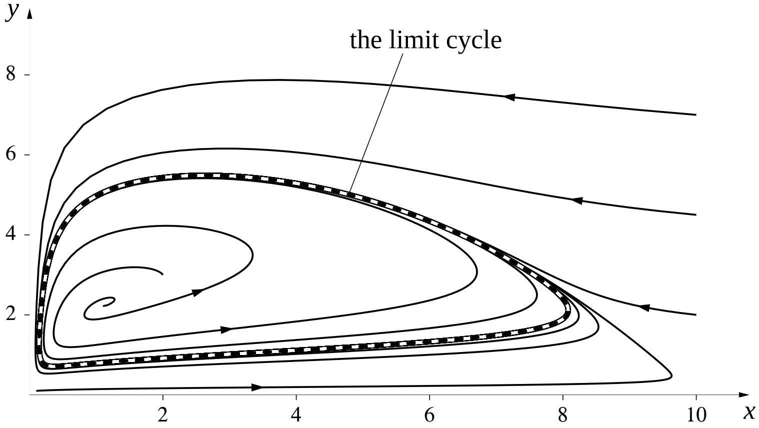
8.4 Limit Cycles

With this analysis of the behavior of vector fields near equilibrium points, we now know most of the possibilities for the long-term behavior of trajectories. The one important phenomenon we haven't discussed is **limit cycles**. To see an example of this, let's return to May's predator–prey model we first encountered in chapter 4. If $x(t)$ and $y(t)$ are the prey and predator populations, respectively, at time t, then the general form of May's model is

$$\begin{aligned} x' &= ax\left(1 - \frac{x}{b}\right) - \frac{cxy}{x + d}, \\ y' &= ey\left(1 - \frac{y}{fx}\right); \end{aligned}$$

the parameters a, b, c, d, e and f are all positive.

Using parameter values of $a = .6$, $b = 10$, $c = .5$, $d = 1$, $e = .1$ and $f = 2$, let's take several different starting values and sketch the resulting trajectories. Here's what we find:



Notice that no matter where we start, the trajectory is apparently always drawn to the closed loop shown in dashes above. This loop is an example of an **attracting limit cycle**. As usual, we could reverse all the arrows in our vector field, in which case this example would be converted to a **repelling limit cycle**.

Limit cycles give models for cyclic behavior

A limit cycle is very different from the kind of behavior we saw in the neighborhood of a neutral equilibrium point called a center. Around a center there is a closed loop trajectory through every point: displace the state slightly, and it would move happily along the new loop. If the state is on an attracting limit cycle, though, and you displace it, it will move back toward the cycle it started from. For this reason limit cycles make very good models for cyclic behavior, whether it is in the firing of neurons or population cycles of mammals.

The size of the limit cycle, and even its very existence, depends on the specific values of the parameters in the model. If you change the parameters, you change the limit cycle. If you change the parameters enough, the limit cycle may disappear all together. (See the exercises.)

A result proved early in this century is the **Poincaré–Bendixson Theorem** which says that equilibrium points and limit cycles are as complicated as dynamical systems in two variables can get. Once we pass to three variables, the situation becomes much more complicated. Many of the phenomena associated with such systems have been discovered only within the past 50 years, and their exploration is a subject of continuing research. In the next section we will give a brief introduction of some of the new behaviors that can arise.

Exercises

1. Using the parameter values given in the text, find the coordinates of the equilibrium point at the center of the limit cycle and show that it is, in fact, a repellor.

2. May's model is interesting in that it exhibits a phenomenon known as **Hopf bifurcation**. Namely, the existence of a limit cycle depends on the values of the parameters. Choose one of the parameters in May's model and try a range of values both larger and smaller than in the example we've worked out. At what value does the limit cycle disappear? When this happens, the equilibrium point inside the cycle has become an attractor. Can you work out analytically when this happens?

8.5 Beyond the Plane: Three-Dimensional Systems

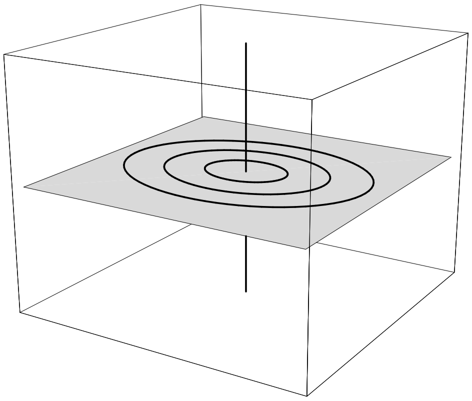
Up to now we have worked with dynamical systems in which there are only two interacting quantities. We have thought of the two quantities as specifying a point in the state space, which we think of as some subset of the plane. The dynamical system defined a vector field on the state space. These geometric notions carry over to dynamical systems involving more than two interacting quantities.

In particular, if we have a dynamical system consisting of three interacting quantities, then we think of the values of the three quantities as specifying a point (or state) in three-dimensional space. So, for instance, if we have an ecological system consisting of three species, then we think of the numbers x, y, z of each of the three species as specifying a point (x, y, z) in space. The set of all possible points or "states" is the set of points

$$\{(x, y, z) : x \geq 0, y \geq 0, z \geq 0\}$$

that constitute the “first octant” in Cartesian 3-space. We think of the dynamical system as a vector field: that is, as a rule which assigns to each point of the state space a vector. As in the case of the plane, we can define equilibrium points, trajectories, limit cycles, attractors, and the like.

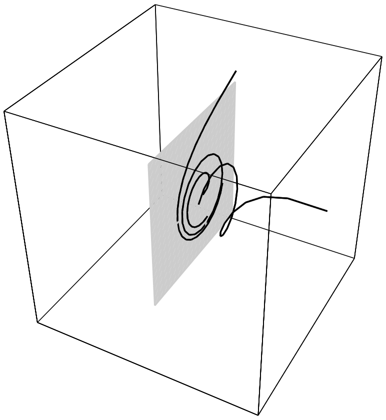
In three-space there is a much wider range of behavior possible, even in the case of equilibrium points. We do, of course, have point attractors and repellors: all trajectories near a point attractor flow towards the attractor and all trajectories near a point repellor flow away from the repellor. However, a greater range of combinations is possible: an equilibrium point can attract all points along some plane, but repel all other points. Or the equilibrium point could be a center, surrounded by closed orbits lying in some plane which attract trajectories off the plane.



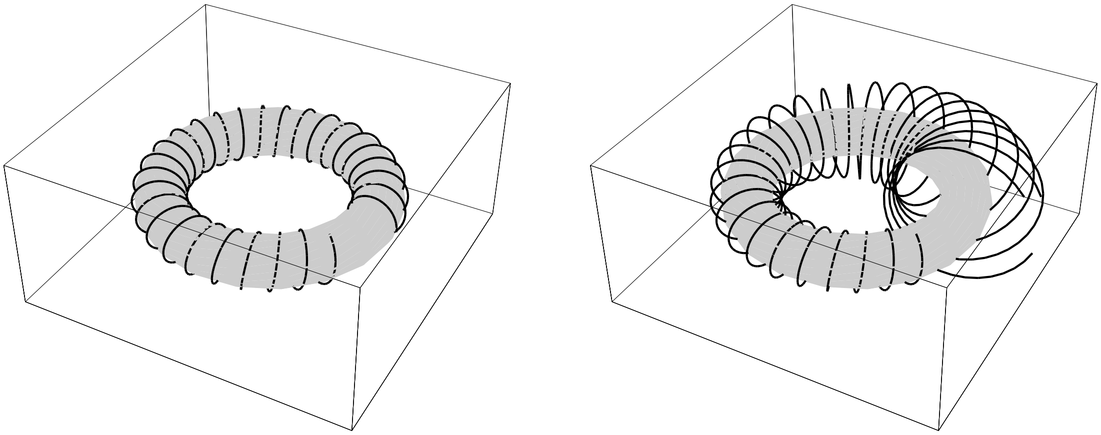
It is worth pointing out that we can represent the two-dimensional systems we’ve been exploring so far in this chapter in three dimensions by introducing a time axis. This has the effect of ‘unwinding” the trajectories by stretching them out in the t-direction: closed trajectories become endless coils, equilibrium points become straight lines parallel to the t-axis, and so on.

The analytic tools we introduced to find and explore the nature of equilibrium points in two-dimensional systems carry over to three dimensions. In particular, in investigating the nature of an equilibrium point analytically, we first localize the system at the equilibrium point and linearize. The behavior of the linearized system can then be explored using analogues of the techniques introduced in the previous section (or using simple linear algebra).

There are also, of course, limit cycles, which can be attractors, repellors of a mix of the two (attracting, for example, all trajectories on a plane, but repelling all trajectories off the plane) in three-dimensional systems. As in the case of dynamical systems in the plane, attracting limit cycles in a three-dimensional system signal stable periodic behavior.



However, more complicated types of periodic behavior are possible in the three-dimensional case: we could, for example, have an attracting torus in the state space.

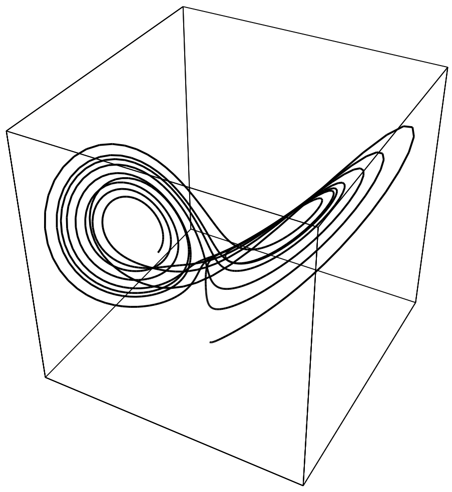


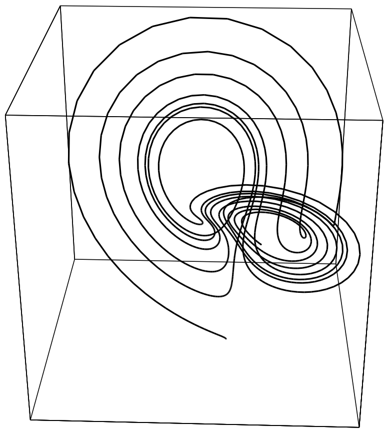
In this case, the behavior of the states does not settle down to periodic behavior, but a behavior which is approximately periodic (often called **quasi-periodic**). In the plane, there is a well studied phenomenon called Hopf bifurcation in which changing the parameters in a dynamical system can cause an attracting fixed point to become a repellor surrounded by a stable attracting limit cycle. Such dynamical systems arise in modelling situations in which a state begins to oscillate. In three dimensions we also sees the same sort of phenomenon in which an attractor can give birth to an attracting limit cycle. However, there are also three-dimensional systems in which varying the parameters results in an attracting limit cycle becoming a repelling limit cycle enclosed by an attracting torus (this is also called Hopf bifurcation and is frequently encountered in applications).

These sorts of behavior are relatively straightforward generalizations of behavior in the plane. At the turn of the century, Poincaré realized that simple three-dimensional systems could have exceedingly complicated trajectories which exhibit behavior totally unlike any two-dimensional trajectory. Discoveries in the last three decades have made it clear that qualitatively new types of *attractors* (not just trajectories) can exist in three-dimensional systems with even very simple equations. The most famous such attractor was discovered by a meteorologist, Edward Lorenz, in the course of using dynamical systems to model weather patterns. He discovered a class of simple systems with an attractor which corresponded to behavior which was in no sense periodic. An example of such a system is

$$\begin{aligned} x' &= -3x - 3y, \\ y' &= -xz + 30x - y, \\ z' &= xy - z. \end{aligned}$$

All trajectories of the system entered a bounded region of the state space and tended towards a clearly defined geometrical object (resembling a butterfly). But along the attractor, nearby points followed trajectories which rapidly diverged from one another. Below, we have sketched two views of a trajectory beginning at (0,1,0) of the system above.





As Lorenz noted in the paper describing his discovery (Deterministic Nonperiodic Flow, *J. Atmos. Sci.*, **20** 130 (1963)), this divergence of trajectories along an attractor has astonishing practical implications. It means that that the trajectories of nearby points in state space could (and would) wind up following very different paths along the attractor. Since we never know initial conditions exactly (and even if we did, a computer truncates decimal expansions of the coordinates of any point, effectively replacing the point with a nearby point), this means that long-term predictions using a model possessing such an attractor are impossible. In other words, although the future is completely determined by a dynamical system given an initial state, it is unknowable in systems of the sort discovered by Lorenz, because initial states are never known exactly in practice. Such systems are called **chaotic** and attractors which are not points, limit cycles or tori are called **strange attractors**. These systems have been intensively studied in the last thirty years and are still far from completely understood. Chaotic systems have been used to attempt to model a wide variety of real situations which exhibit unpredictable behavior: business cycles, turbulence, heart attacks, etc. Although fascinating and philosophically provocative, most of this work is still very speculative and has yet to prove of practical value.

Systems involving more than three variables can still be treated geometrically: we think of the space of states as a higher dimensional space (one dimension for each quantity) and the dynamical system as defining a vector field on the state space. Of course, we cannot visualize such spaces directly, but the geometrical insight we gain in dimensions two and three very frequently allows us to handle such systems.

Exercises

In the next two exercises, we look at some three-dimensional systems which arise in ecology. These questions are challenging and you will probably find it helpful to work them out in a group.

1. a) Consider a system consisting of three species: giant carnivorous reptiles, vegetarian mammals, and plants. Suppose that the populations of these are given by x, y and z respectively. The reptiles eat the mammals, the mammals eat the plants, and the plants compete among themselves. Explain why

the following system is consistent with these hypotheses:

$$\begin{aligned} x' &= -.2x + .0001xy, \\ y' &= -.05y - .001xy + .000001yz, \\ z' &= z - .00001z^2 - .0001yz. \end{aligned}$$

b) Find all equilibrium points of the system. There are five, one of which is physically impossible. Describe the significance of the other four.

c) The most interesting equilibrium is the one in which all three species are present. Localize the system at this equilibrium, using local variables u, v, and w. Linearize. Show that the linearized system has the form

$$\begin{aligned} u' &= \quad .003v, \\ v' &= -2u \quad + .002w, \\ w' &= \quad -8v - .8w. \end{aligned}$$

Can you determine whether the equilibrium is an attractor? This is a hard question—it *is* an attractor. One way to show this is a generalization of the technique we used in the preceding section to examine the distance of points on a trajectory from the origin over time. For the current problem we use a **generalized distance function**

$$D(t) = 8 \cdot 10^6 u^2 + 12000v^2 + 3w^2.$$

Show, using arguments like those we used when we looked at ordinary distance, that as we move along a trajectory, the value of D must decrease. Hence conclude that the equilibrium point must be an attractor.

2. The system of equations

$$\begin{aligned} x' &= x - .001x^2 + .002xy - .1xz, \\ y' &= y - .01y^2 + .001xy, \\ z' &= -z + .001xz. \end{aligned}$$

arises in a general family of models proposed in 1980 by Heithaus, Culver, and Beattie ("Models of Some Ant-Plant Mutualisms," *American Naturalist*, **116** (1980) pp. 347-361) for investigating the interactions three species: violets, ants, and mice. Violets produce seeds with density x (per square meter, say). The ants take some of the seeds and use the seed covering for food. But they leave the remainder, which is still a perfectly good seed, in their refuse piles, which happily turn out to be good sites for germination. The ants have density y. Finally, the seeds are also taken by mice, who use the whole seed for food (destroying both the cover and the seed within). The mice have density z.

a) Explain why these equations are consistent with the hypotheses we made on the interactions between the violets, ants and mice.

b) Find all equilibrium points for the system. Don't forget the points where one or more of the variables equals 0.

$$\begin{aligned} x' &= x - .001x^2 + .002xy - .1xz, \\ y' &= y - .01y^2 + .001xy, \\ z' &= -z + .001xz. \end{aligned}$$

c) Localize the model at each of these equilibria, using local coordinates u, v, and w as before, and linearize.

d) In the case of the equilibrium point (1000, 200, 4) the local linearization is

$$\begin{aligned} u' &= -u + 2v - 100w, \\ v' &= .2u - 2v, \\ w' &= .004u. \end{aligned}$$

As in the preceding problem, show that this point is an attractor by examining the generalized distance function $R(t) = u(t)^2 + v(t)^2 + 25000w(t)^2$ and showing that the value of R decreases as you move along a trajectory.

8.6 Chapter Summary

The Main Ideas

- A dynamical system can be viewed as a geometrical object. The possible values of the dependent variables are then the coordinates of a point—called a **state**. The set of all possible points is called the **state space** for the system.

- The differential equations become a rule assigning a **velocity vector** to each state. Thought of in this way, the equations are called a **vector field**.

- Solutions to the differential equations correspond to **trajectories** in the state space. At every point a trajectory is tangent to the corresponding velocity vector, and is changing at the rate given by the length of the vector. The set of all possible trajectories is called the **phase portrait** of the system.

- **Equilibrium points** are points where the velocity vector is 0. An equilibrium point is a trajectory consisting of a single point. A dynamical system is conveniently analyzed by examining the nature of its equilibrium points—whether they are **attractors**, **repellors**, **saddle points**, or **centers**.

- To study the nature of an equilibrium point it is helpful to look at the **local linearization** of the vector field near the point.

- Determining whether **fixed-line trajectories** exist is a crucial part of analyzing the nature of an equilibrium point.

- In addition to equilibrium points, dynamical systems in two dimensions may also have **limit cycles** that shape the asymptotic behavior of the system.

- In higher dimensional state spaces, there are not only the obvious extensions of point attractors and limit cycles, but it is possible to have an **attracting torus** as well. There are even more complicated attracting objects called **strange attractors**.

Expectations

- You should be able to describe the assumptions embodied in a particular dynamical system modeling the interaction between two (or three) species and evaluate whether the assumptions seem reasonable.

- For a dynamical system with two dependent variables, you should be able to determine the regions where each variable is zero or has a constant sign, find equilibrium points, sketch representative vectors of the vector field, and draw trajectories that are consistent with this information.

- You should be able to determine whether a linear system of differential equations with two dependent variables has **fixed-line** trajectories—that is, trajectories that are straight lines going directly toward or directly away from an equilibrium point.
- You should be able to **localize** and **linearize** a dynamical system in two variables to explore its behavior near an equilibrium point.
- You should be able to recognize the five generic types of equilibrium points: attracting and repelling **nodes**, attracting and repelling **spirals**, and **saddle points**.
- Using a differential equation solver, you should be able to recognize when a dnamical system has a **limit cycle**.
- You should be able to analyze a dynamical system with three dependent variables.

Chapter 9

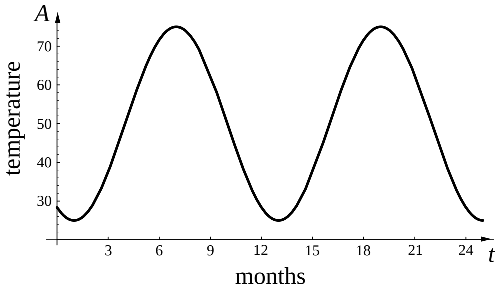
Functions of Several Variables

Functions that depend on several input variables first appeared in the S-I-R model at the beginning of the course. Usually, the number of variables has not been an issue for us. For instance, when we introduced the derivative in chapter 3, we used partial derivatives to treat functions of several variables in a parallel fashion. However, when there are questions of visualization and geometric understanding, the number of variables *does* matter. Every variable adds a dimension to the problem—one way or another. For example, if a function has two input variables instead of one, we will see that its graph is a surface rather than a curve.

The problem of visualizing a function of several variables

This chapter deals with the geometry of functions of two or more variables. We start with graphs and level sets. These are the basic tools for visualization. Then we turn to microscopic views, and see what form the microscope equation takes. Finally, we consider optimization problems using both direct visual methods and dynamical systems.

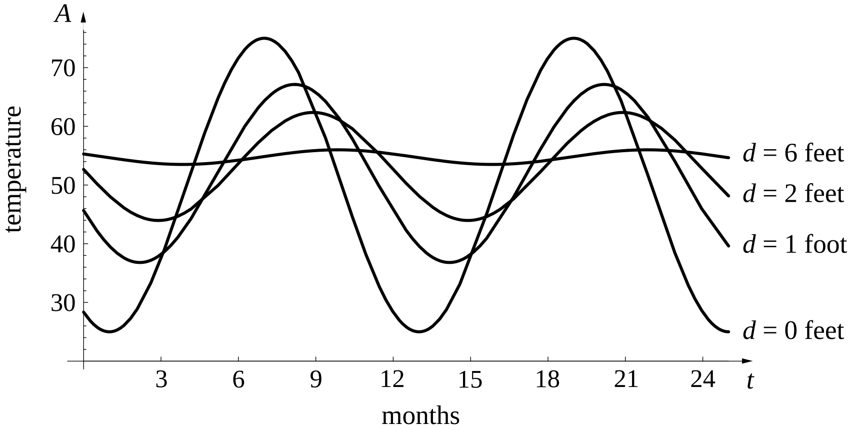
9.1 Graphs and Level Sets



The graph at the right comes from a model that describes how the average daily temperature at one place varies over the course of a year. It shows the temperature A in °F, and the time t in months from January. As we would expect, the temperature is a periodic function (which we can write as $A(t)$), and its period is 12 months. Furthermore, the lowest temperature occurs in February (when $t \approx 1$ or 13) and the highest in July (when $t \approx 7$ or 19). This is about what we would expect.

Underground temperatures fluctuate less

However, all these temperature fluctuations disappear a few feet underground. Below a depth of 6 or 8 feet, the temperature of the soil remains about 55°F year-round! Between ground level and that depth, the temperature still fluctuates, but the range from low to high decreases with the depth. Here is what happens at some specific depths.



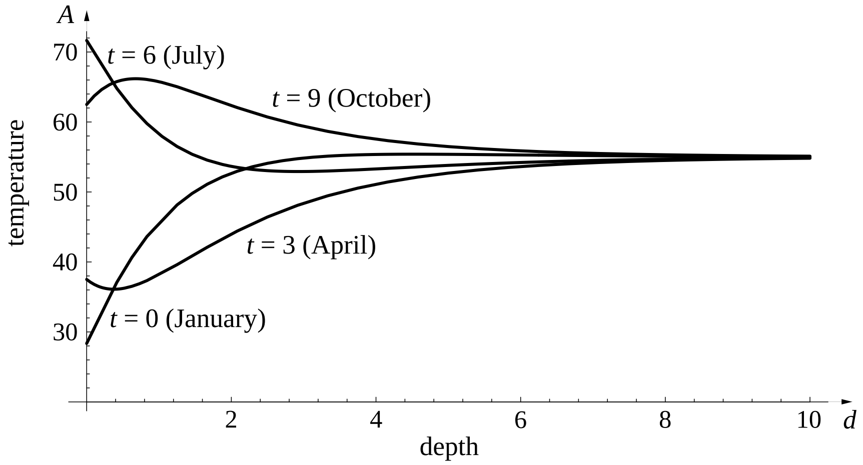
Notice how the time at which the temperature peaks gets later and later as we go farther and farther underground. For example, at $d = 2$ feet the highest temperature occurs in September ($t \approx 9$), not July. It literally takes time for the heat to sink in. In chapter 7 we called this a phase shift. The lowest temperature shifts in just the same way. At a depth of 2 feet, it is colder in March than in January.

The phase shifts with the depth

Thus A is really a function of *two* variables, the depth d as well as the time t. To reflect this addition, let's change our notation for the function to $A(t, d)$. In the figure above, d plays the role of a parameter: it has a fixed value for each graph. We can reverse these roles and make t the parameter. This is done in the figure on the top of the next page. It shows us how the temperature varies with the depth at fixed times of year. Notice that, in April and October, the extreme temperature is not found on the surface. In October, for example, the soil is warmest at a depth of about 9 inches.

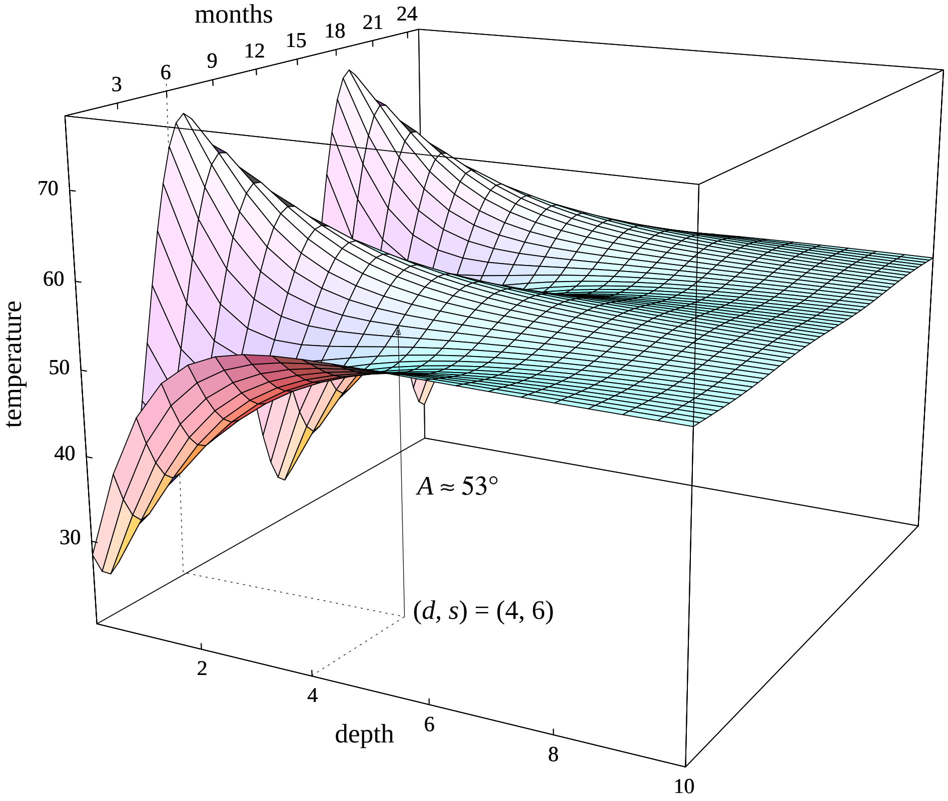
Graphing temperature as a function of depth

The lower figure on the page is a single graph that combines all the information in these two sets of graphs. Each point on the bottom of the box of the box corresponds to a particular depth and a particular time. The height of the surface above that point tells us the temperature at that depth and time. For example, suppose you want to find the temperature 4 feet



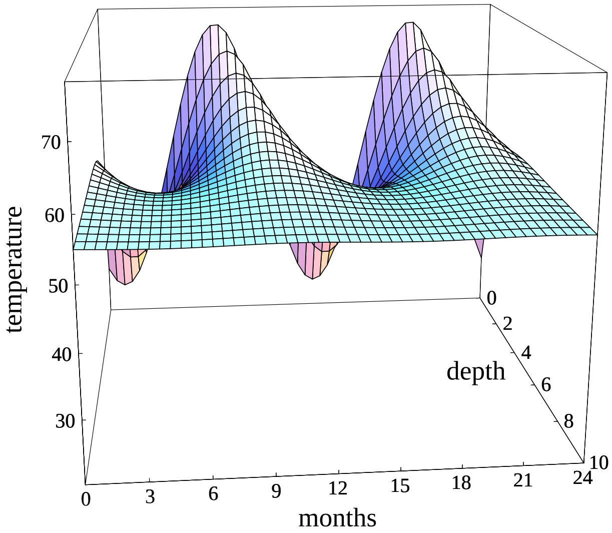
below surface at the beginning of July. Working from the bottom front corner of the box, move 4 feet to the right and then 6 months toward the back. This is the point $(d, s) = (4, 6)$. The height of the graph above this point is the temperature A that we want.

Reading a surface graph



Grid lines are slices of the surface that show how the function depends on each variable separately

There is a definite connection between this surface and the two collections of curves. Imagine that the box containing the surface graph is a loaf of bread. If you slice the loaf parallel to the left or right side, this slice is taken at a fixed depth. The cut face of the slice will look like one of the graphs on page 512. These show how the temperature depends on the time at fixed depths. If you slice the loaf the other way—parallel to the front or back face—then the time is fixed. The cut face will look like one of the graphs on page 513. They show how the temperature depends on the depth at fixed times. The grid lines on the surface are precisely these "slice" marks.

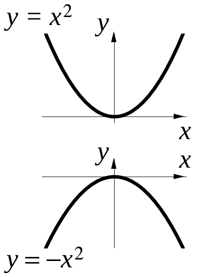


Here is the same graph seen from a different viewpoint. Now time is measured from the left, while depth is measured from the back. The temperature is still the height, though. One advantage of this view is that it shows more clearly how the peak temperature is phase-shifted with the depth.

We now have two ways to visualize how the average daily temperature depends on the time of year and the depth below ground. One is the surface graph itself, and the other is a collection of curves that are slices of the surface. The surface gives us an overall view, but it is not so easy to read the surface graph to determine the temperature at a specific time and depth. Check this yourself: what is the temperature 2 feet underground at the beginning of April? The slices are much more helpful here. You should be able to read from either collection of slices that $A \approx 44°$F.

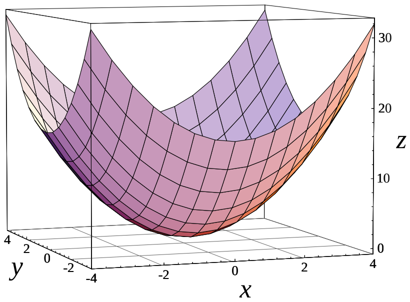
Comparing the surface to its slices

Examples of Graphs



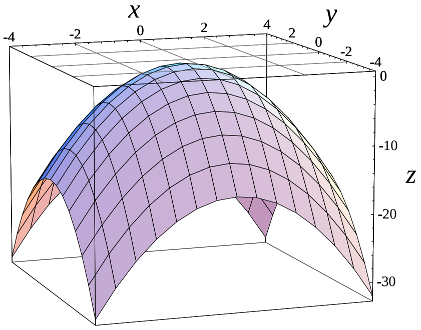
The purpose of this section is to get some experience constructing and interpreting surface graphs. To work in a context, look first at the functions $y = x^2$ and $y = -x^2$. They provide us with standard examples of a minimum and a maximum when there is just one input variable. Let's consider now the corresponding examples for two input variables. Besides an ordinary maximum and an ordinary minimum, we will find a *third* type—called a minimax—that is completely new. It arises because a function can have a minimum with respect to one of its input variables and a maximum with respect to the other.

A minimum: $z = x^2 + y^2$



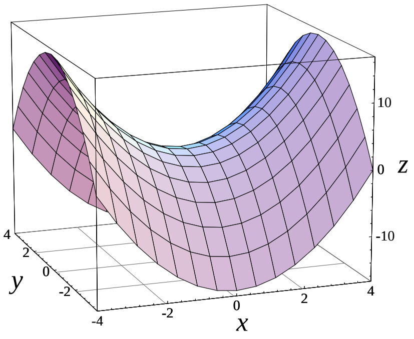
At the origin $(x, y) = (0, 0)$, $z = 0$. At any other point, either x or y is non-zero. Its square is positive, so $z > 0$. Consequently, z has a minimum at the origin. The graph of this function is a parabolic **bowl** whose lowest point sits on the origin. As always, the grid lines are slices, made by fixing the value of x or y. For example, if $y = c$, then the slice is $z = x^2 + c^2$. This is an ordinary parabolic curve.

A maximum: $z = -x^2 - y^2$



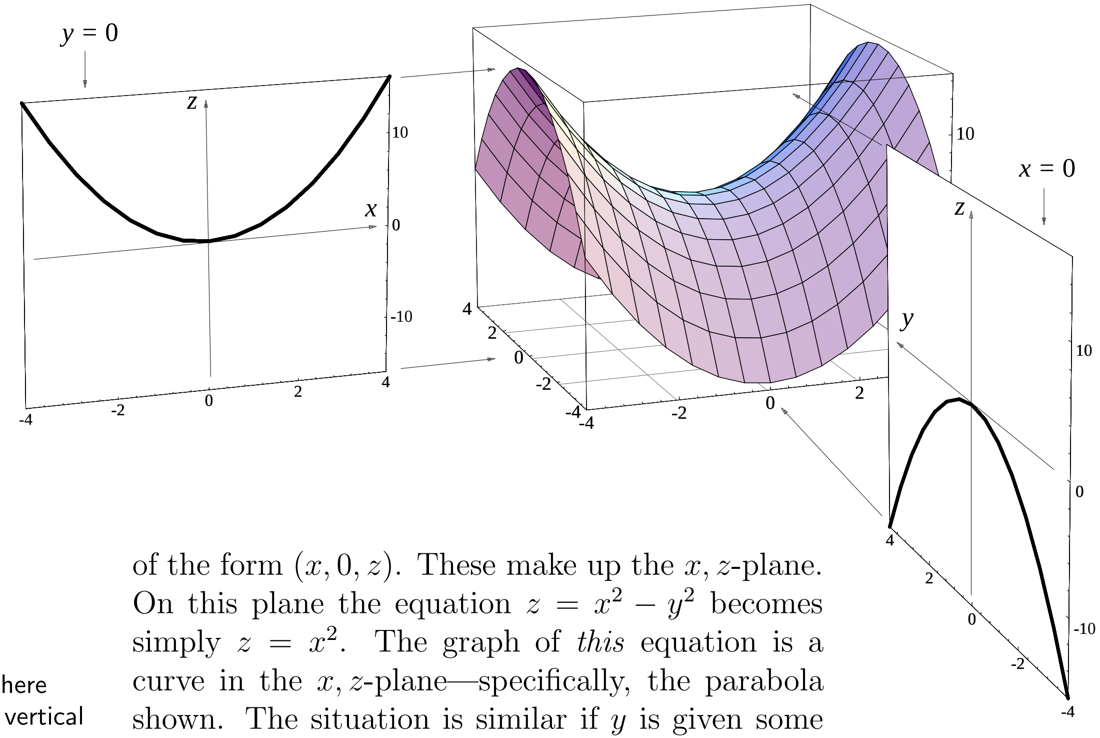
For any x and y, the value of z in this example is the opposite of its value in the previous one. Thus, z is everywhere negative, except at the origin, where its value is 0. Thus z has a maximum at the origin. Its graph is an upside-down bowl, or **peak**, whose highest point reaches up and touches the origin. Grid lines are the curves $z = -x^2 - c^2$ and $z = -c^2 - y^2$. These are parabolic curves that open downward.

A minimax: $z = x^2 - y^2$



Suppose we fix y at $y = 0$. This slice has the equation $z = x^2$, so it is an ordinary parabola (that opens upward). Thus, as far as the input x is concerned, z has a *minimum* at the origin. Suppose, instead, that we fix x at $x = 0$. Then we get a slice whose equation is $z = -y^2$. It is also an ordinary parabola, but this one opens downward. As far as y is concerned, z has a *maximum* at the origin. It is clear from the graph how upward-opening slices in the x-direction fit together with downward-opening slices in the y-direction. Because of the shape of the surface, a minimax is commonly called a **saddle**, or a **saddle point**.

Here are two slices of $z = x^2 - y^2$ shown in more detail. Points in the box have three coordinates: (x, y, z). If we set $y = 0$ we are selecting the points of the form $(x, 0, z)$. These make up the x, z-plane. On this plane the equation $z = x^2 - y^2$ becomes simply $z = x^2$. The graph of *this* equation is a curve in the x, z-plane—specifically, the parabola shown. The situation is similar if y is given some other fixed value. For example, $y = -4$ specifies the points $(x, -4, z)$. These describe the plane that forms the front face of the box. The equation $z = x^2 - y^2$ becomes $z = x^2 - 16$. The curve tracing out the intersection of the saddle with the front of the box is precisely the graph of $z = x^2 - 16$.



The points where $y = c$ form a vertical plane parallel to the x, z-plane

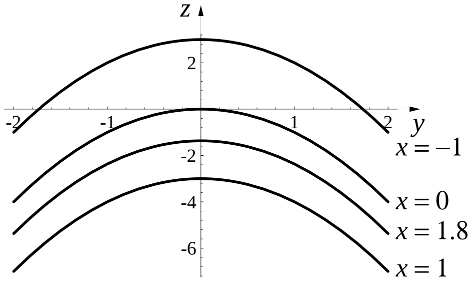
If $x = 0$ we get the points $(0, y, z)$ that make up the y, z-plane. On this plane the equation simplifies to $z = -y^2$, and its graph is the parabolic curve shown. Giving x a different fixed value leads to similar results. A good example is $x = -4$. The points $(-4, y, z)$ lie on the plane that forms the left side of the box. The equation becomes $z = 16 - y^2$ there, and this is the parabolic curve marking the intersection of the saddle with the left side of the box.

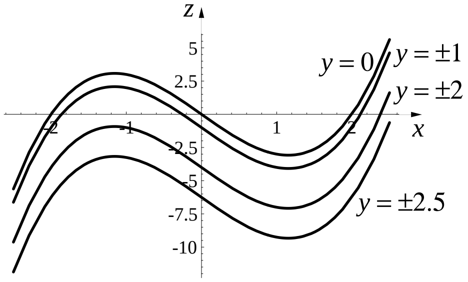
The points where $x = c$ form a vertical plane parallel to the y, z-plane

As you can see, it is valuable for you to be able to generate surface graphs yourself. There are now a number of computer utilities which will do the job. Some can even rotate the surface while you watch, or give you a stereo view. However, even without one of these powerful utilities, you should try to generate the slicing curves that make up the grid lines of the surface.

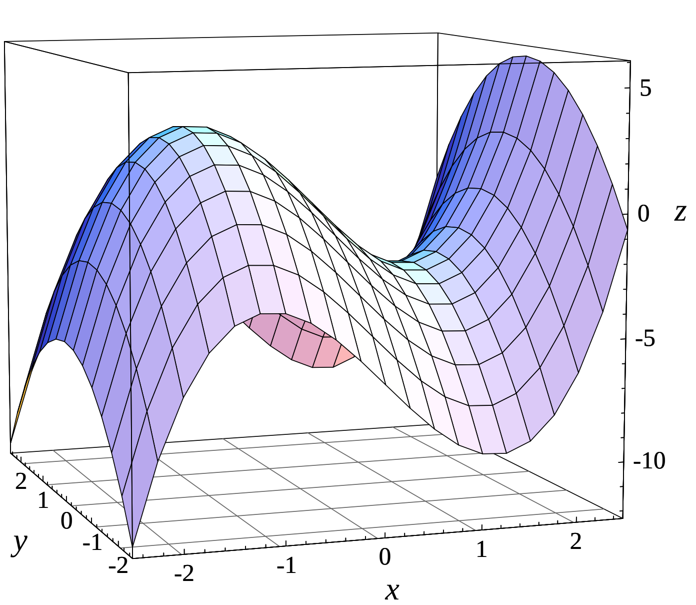
A cubic: $z = x^3 - 4x - y^2$

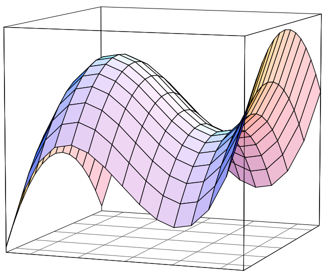
Slices of this graph are downward-opening parabolas (when $x = c$) and are cubic curves that have the same shape (when $y = c$). Notice that each cubic curve has a maximum and a minimum, and each parabola has a maximum. The surface graph itself has a *peak* where the cubics have their maximum, but it has a *saddle* where the cubics have a minimum. Do you see why? The saddle point is a minimax for $z = x^3 - 4x - y^2$: z has a minimum there *as a function of x alone* but a maximum *as a function of y alone*.

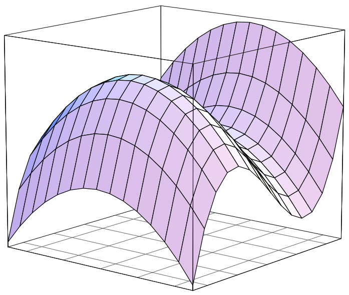




The surface has a saddle



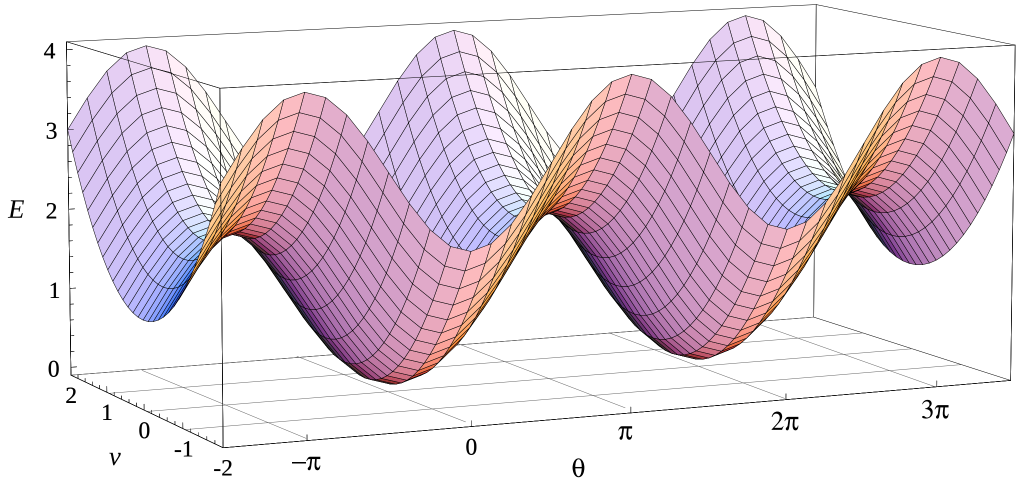


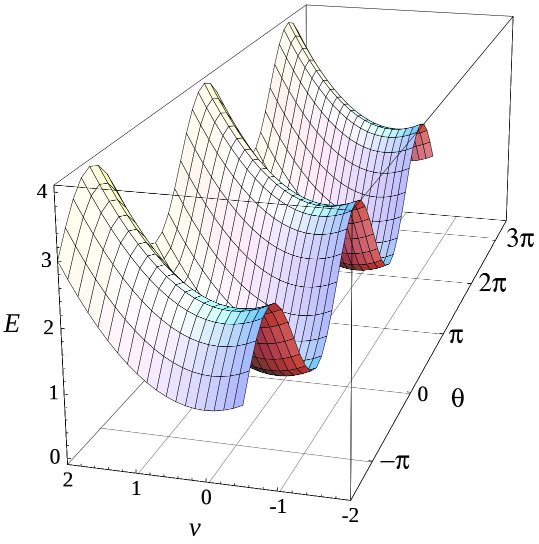


The small figures on the right show the same surface as the large figure; they just show it from different viewpoints. As a practical matter, you should look at these surfaces the way you would look at sculpture: "walk around them" by generating diverse views.

See the graph from different viewpoints

Energy of the pendulum: $E = 1 - \cos\theta + \frac{1}{2}v^2$



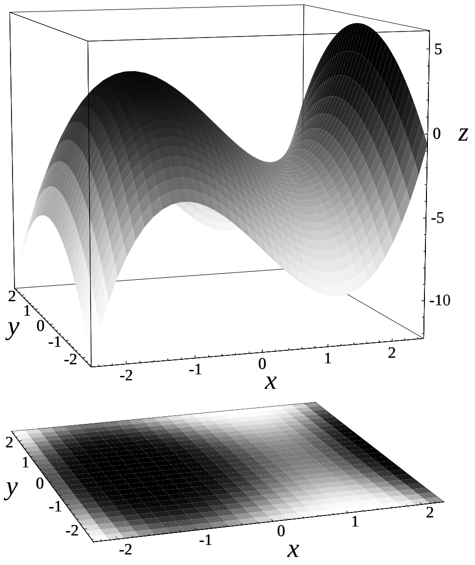


This function first came up in chapter 7, where it was used to demonstrate that a dynamical system describing the motion of a frictionless pendulum had periodic solutions. It was used again in chapter 8 to clarify the phase portrait of that dynamical system. The function E varies periodically with θ, and you can see this in the graph. The minimum at the origin is repeated at $(\theta, v) = (2\pi, 0)$, and so on. The graph also has a saddle at the point $(\theta, v) = (-\pi, 0)$. This too repeats with period 2π in the θ direction.

The figure at the left is the same surface with part cut away by a slice of the form $v = c$. These slices are sine curves: $E = 1 - \cos\theta + \frac{1}{2}c^2$. Slices of the form $\theta = c$ are upward-opening parabolas. From this viewpoint, the saddle points show up clearly.

One way to describe what happens to a real pendulum—that is, one governed by frictional forces as well as gravity—is to say that its energy "runs down" over time. Now, at any moment the pendulum's energy is a point on this graph. As the energy runs down, that point must work its way down the graph. Ultimately, it must reach the bottom of the graph—the minimum energy point at the origin $(\theta, v) = (0, 0)$. This is the stable equilibrium point. The pendulum hangs straight down ($\theta = 0$) and is motionless ($v = 0$). The graph gives us an abstract—but still vivid and concrete—way of thinking of the dissipation of energy.

From Graphs to Levels



There is still another way to picture a function of two variables. To see how it works we can start with an ordinary graph. On the right is the graph of

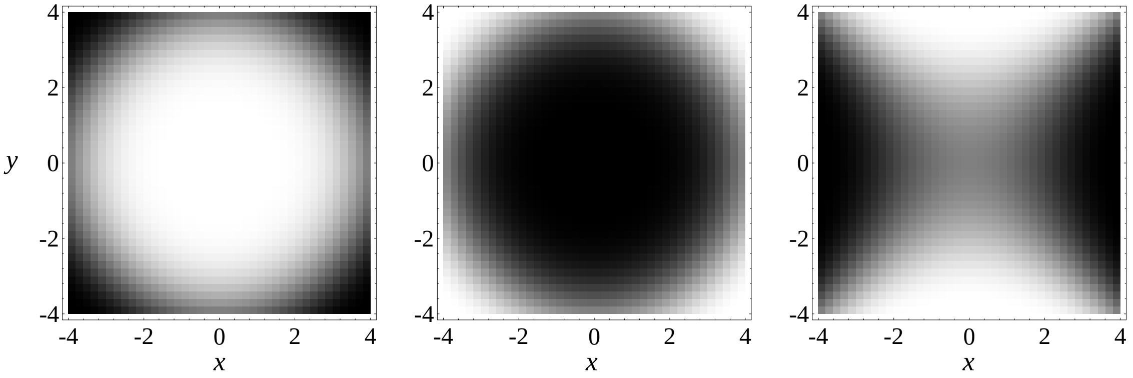
$$z = f(x, y) = x^3 - 4x - y^2,$$

the cubic function we considered on page 517. This graph looks different, though. The difference is that points are shaded according to their height. Points at the bottom are lightest, points at the top are darkest.

Notice that the flat x, y-plane is shaded exactly like the graph above it. For instance, the dark spot centered at the point $(x, y) = (-1, 0)$ is directly under the peak on the graph. The other dark patch, near the right edge of the plane, is under the highest visible part of the surface. Consequently, the shading on the x, y-plane gives us the same information as the graph. In other words, *the intensity of shading at (x, y) is proportional to the value of the function $f(x, y)$.*

Density plots

The figure in the x, y-plane is called a **density plot**. Think of the intensity of shading as the *density* of ink on the page. Here are density plots of the standard minimum, maximum, and minimax. Compare these with the graphs on page 515. The third density plot is the most interesting. From the center of the x, y-plane, the shading increases to the right and left. Therefore,



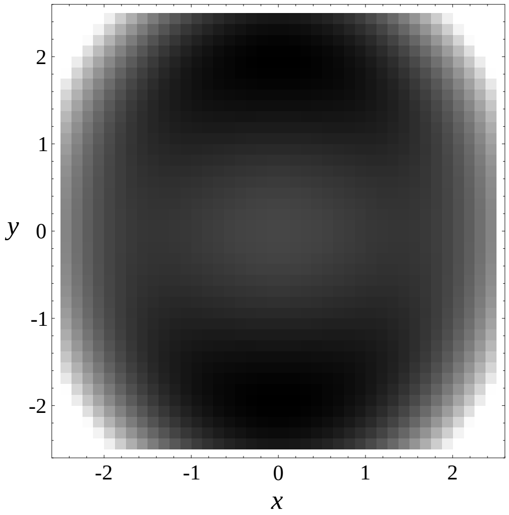
$z = x^2 + y^2$ $z = -x^2 - y^2$ $z = x^2 - y^2$

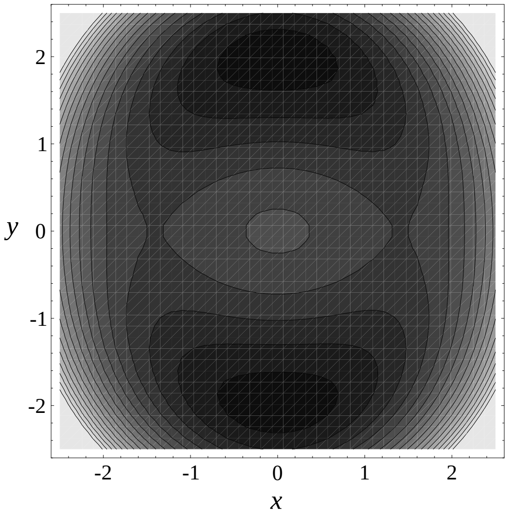
z has a minimum in the horizontal direction. However, the shading *decreases* above and below the center. Therefore, z has a *maximum* in the vertical direction. Thus, you really can see there is a minimax at the origin.

A sample plot

Try your hand at reading the density plot on the left below. You should see two maxima (directly above and below the origin), a minimum (at the origin itself), and two saddles (to the right and the left of the origin). The function defining the plot is

$$f(x, y) = (x^2 + (y - 1)^2 - 3)(3 - x^2 - (y + 1)^2).$$





Can you visualize what the graph looks like? This density plot should help you, and you can also construct slices by setting $x = c$ and $y = c$. The slices $x = 0$ and $y = 0$ are especially useful. With them you could determine the exact coordinates of the maxima and the saddles.

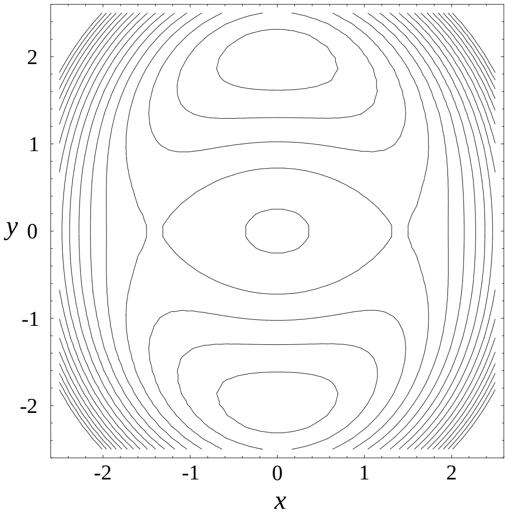
These density plots show a "checkerboard" pattern because *Mathematica* (the computer program that produces them) shades each little square according to the value of the function at the center of the square. This pattern is an artefact; it is not inherent to a density plot.

From densities to contours

In a density plot, the shading varies smoothly with the value of the function. This is accurate, but it may be a bit difficult to read. On the right you see a modified density plot. There is still shading, but there are now just a few distinct shades. This makes a sharp boundary between one shade and the next. The boundary is called a **contour**, or a **level**. The figure itself is called a **contour plot**. The two maxima on the vertical line $x = 0$ stand out more clearly on the contour plot. Also, the contour lines around the two saddles help us see that the function has a minimum in the vertical direction and a maximum in the horizontal direction.

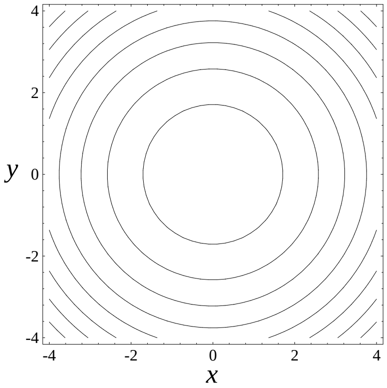
Once we have contour lines to separate one density level from the next, we can even dispense with the shading. The figure on the right is just the contour plot from the opposite page, minus the shading. The contour lines, or level curves, now stand out clearly. On each contour, the value of the function is constant. This is also called a **contour plot**.

There is some loss of information here, however. For example, we can't tell where the value of the function is large and where it is small. Nevertheless, the nested ovals on the vertical line $x = 0$ *do* tell us that there is either a maximum or a minimum at the center of each nest.



Contours of the standard functions

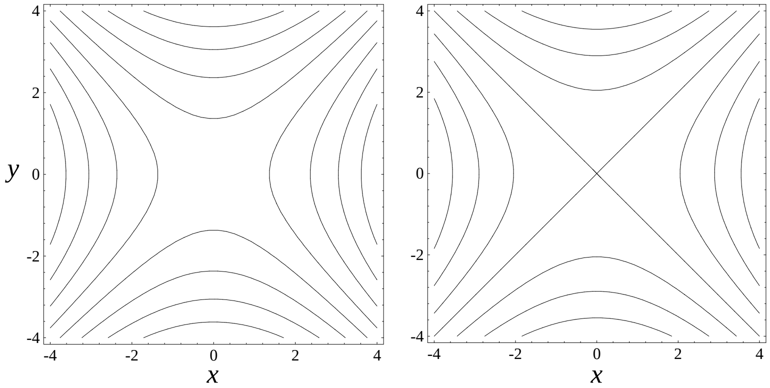
For reference purposes, here are the contour plots for the standard minimum, maximum, and saddle. In the first two cases, the contours are concentric ovals. These look the same, so only one is illustrated. The other two pictures show a saddle. In general, the contours around a saddle are a family of hyperbolas. However, it is possible for one of the contour lines to pass exactly through the minimax point. That contour is a pair of crossed lines, as shown in the version on the right. You should compare these contour plots with the density plots of the same functions on page 519, and with their graphs on page 515.



two functions but one plot

$z = x^2 + y^2$

$z = -x^2 - y^2$

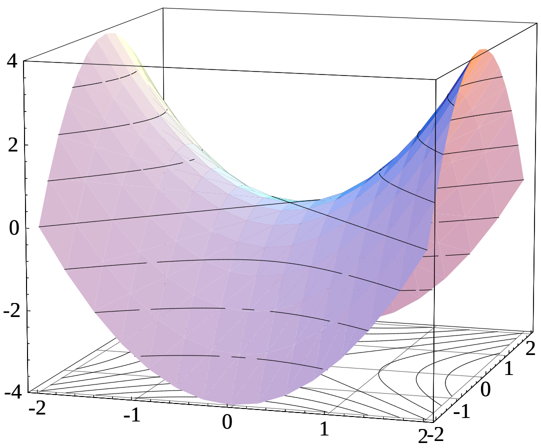


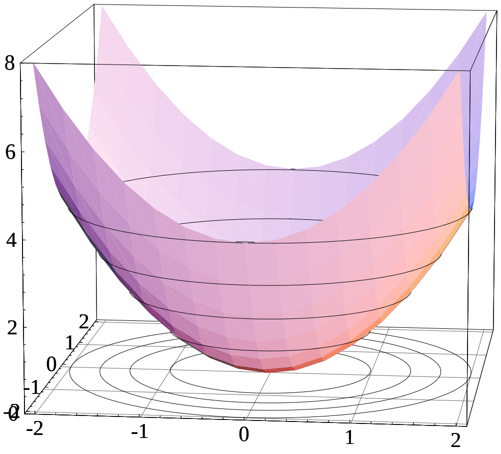
two plots of a single function

$z = x^2 - y^2$

Contours are horizontal slices of a graph

There is a direct connection between the contour plot of a function and its graph. Contours are horizontal slices of the graph, just as grid lines are vertical slices. Below, we use the standard functions $z = x^2 - y^2$ and $z = x^2 + y^2$ to illustrate the connection. Notice that every contour down in the x, y-plane lies exactly below, and has the same shape as, a horizontal slice of the graph. This picture explains why contours are called *level* curves.



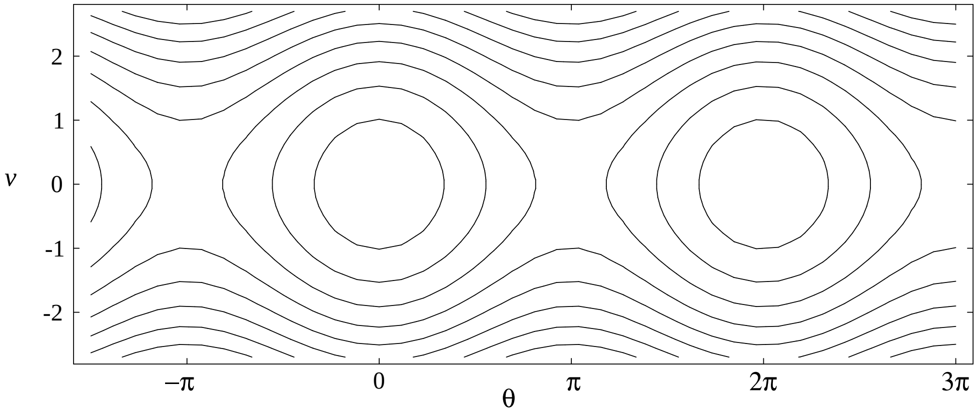


Energy of the pendulum, again

To get some more experience with contour plots, we return to the energy function of the pendulum:

$$E(\theta, v) = 1 - \cos\theta + \tfrac{1}{2}v^2.$$

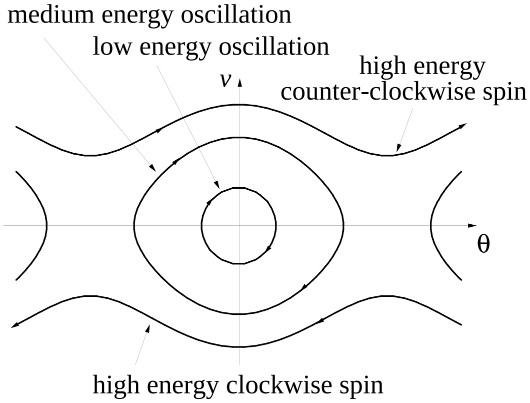
From the contour plot alone you should be able to see that E has either a minimum or a maximum at $(\theta, v) = (0, 0)$, and another at $(2\pi, 0)$. The contours also provide evidence that there is a saddle (minimax) near $(\theta, v) = (-\pi, 0)$ and $(\pi, 0)$. It is also apparent that E is a *periodic* function of θ.



What you should find most striking about this plot, however, is the way it resembles the phase portrait of the pendulum (chapter 8, pages 471–474). Every level curve here looks like a trajectory of the dynamical system. This is no accident. We know from chapter 8 that the energy is a first integral for the dynamics. In other words, energy is constant along each trajectory—this is the law of conservation of energy. But each level curve shows where the energy function has some fixed value. Therefore, each trajectory must lie on a single energy level.

Energy contours are trajectories of the dynamics

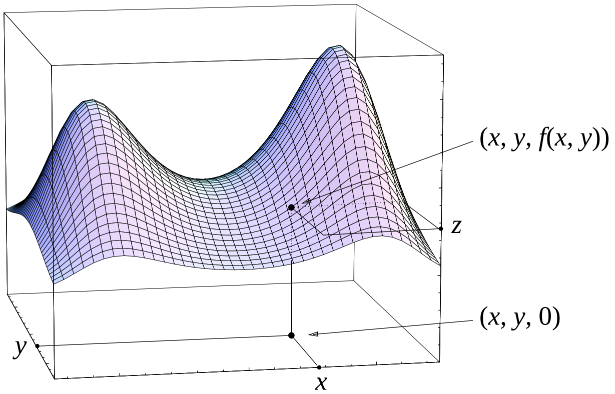
We can carry the connection between contours and trajectories even further. Closed trajectories correspond to *oscillations* of the pendulum. But the closed trajectories are the closed contours, and these are the ones that surround the minimum. In particular, they are *low* energy levels. By contrast, at higher energies ($E > 2$, in fact), the pendulum will just continue to spin in what ever direction it was moving initially. Thus, each high energy level is occupied by *two* trajectories—one for clockwise spinning and one for counter-clockwise.

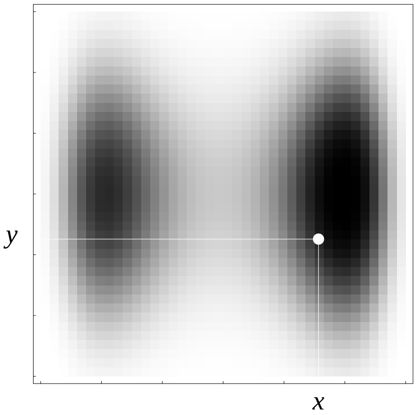


Technical Summary

The examples we have seen so far were meant to introduce some of the common ways of visualizing a function $z = f(x, y)$. To use them most effectively, though, you need to know more precisely how each is defined. We review here the definition of a graph, a density plot, a contour plot, and a terraced density plot.

Graph. The graph of $z = f(x, y)$ lies in the 3-dimensional space with coordinates (x, y, z). To construct it, take any input (x, y). Identify this with the point $(x, y, 0)$ in the x, y-plane (which is defined by the condition $z = 0$). The corresponding point on the graph lies at the height $z = f(x, y)$ above the x, y-plane. This point has coordinates $(x, y, f(x, y))$. **The graph is the set of all points of the form $(x, y, f(x, y))$.** This is a 2-dimensional surface.

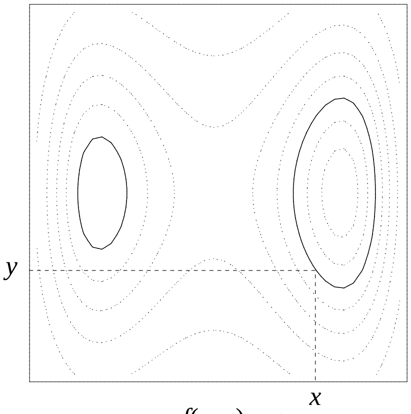




Density plot. The density plot of $z = f(x, y)$ lies in the 2-dimensional x, y-plane. Choose any rectangle where the function is defined, and let m and M be the minimum and maximum values, respectively, of $f(x, y)$ on the rectangle. Define

$$\rho(x, y) = \frac{f(x, y) - m}{M - m}.$$

Then ρ satisfies $0 \leq \rho(x, y) \leq 1$ on the rectangle; it is called a **density function** (ρ is the Greek letter *rho*). **In the density plot, the density of ink—or darkness—at** (x, y) **is** $\rho(x, y)$**.**

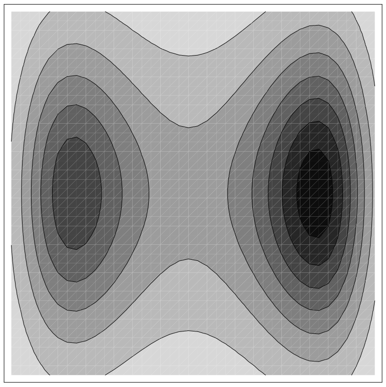


$f(x, y) = c$

Contour plot. A **contour** of $z = f(x, y)$ is the set of points in the x, y-plane where f has some fixed value:

$$f(x, y) = c.$$

That fixed value c is called the **level** of the contour. (The two solid ovals in the figure at the left make up a single contour.) A contour is also called a level curve. **A contour plot of** f **is a collection of curves** $f(x, y) = c_j$ **in the** x, y**-plane.** In the plot it is customary to use constants $c_1, c_2, \ldots$ that are equally spaced; that is, the interval between one c_j and the next always has the same value Δc.



Terraced density plot. This is a contour plot in which the region between two adjacent contours is shaded with ink of a single density. If the contours are at levels c_1 and c_2, then the density that is typically chosen is the one for the level half-way between these two—that is, for their average $(c_1+c_2)/2$. Each region is called a **terrace**. Often, a terraced density plot is drawn in color, using different colors for each terrace. Television weather programs use terraced density plots to describe the temperature forecast for a large region.

We find density plots everywhere. A photograph is a density plot of the light that fell on the film when it was exposed. A newspaper 'half-tone' illustration is also a density plot of an image.

The pros and cons

Each of these modes of visualization has advantages and disadvantages. All are reasonably good at indicating the extremes (the maxima and minima) of a function. A contour plot needs some additional information—for example, a label on each contour to indicate its level—to distinguish between maxima and minima. However, if you want to know the numerical value of $f(x, y)$ at a particular point (x, y), a contour plot with labels offers more precision than a density plot. It's usually better than a graph, too.

Plots are visually economical in comparison to graphs

Overall, a graph has the biggest visual impact, but there is a cost. It takes three dimensions to represent the graph of a function of two variables, but only two to represent a plot. The cost is that extra dimension. It means that we cannot draw the graph of a function of three variables. That would take four mutually perpendicular axes—an impossibility in our three-dimensional space. However, we can produce a contour plot.

Contours of a Function of Three Variables

Contours and levels

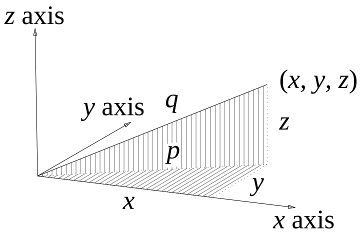
We pause here for a brief glimpse of a large subject. By analogy with the definition for a function of two variables, we say that a **contour** of the function $f(x, y, z)$ is the set of points (x, y, z) that satisfy the equation

$$f(x, y, z) = c,$$

for some fixed number c. We call c the **level** of the contour.

The standard minimum

Let's find the contours of $w = x^2 + y^2 + z^2$. This is completely analogous to the function $x^2 + y^2$ with two inputs. (What do the contours of $x^2 + y^2$ look like?) In particular, w has a minimum when $(x, y, z) = (0, 0, 0)$. As the following diagram shows, $w = x^2 + y^2 + z^2$ is the square of the distance from the origin to (x, y, z). (We use the Pythagorean theorem twice: once for p^2 and once for q^2.) Consequently, all points (x, y, z) where w has a fixed value



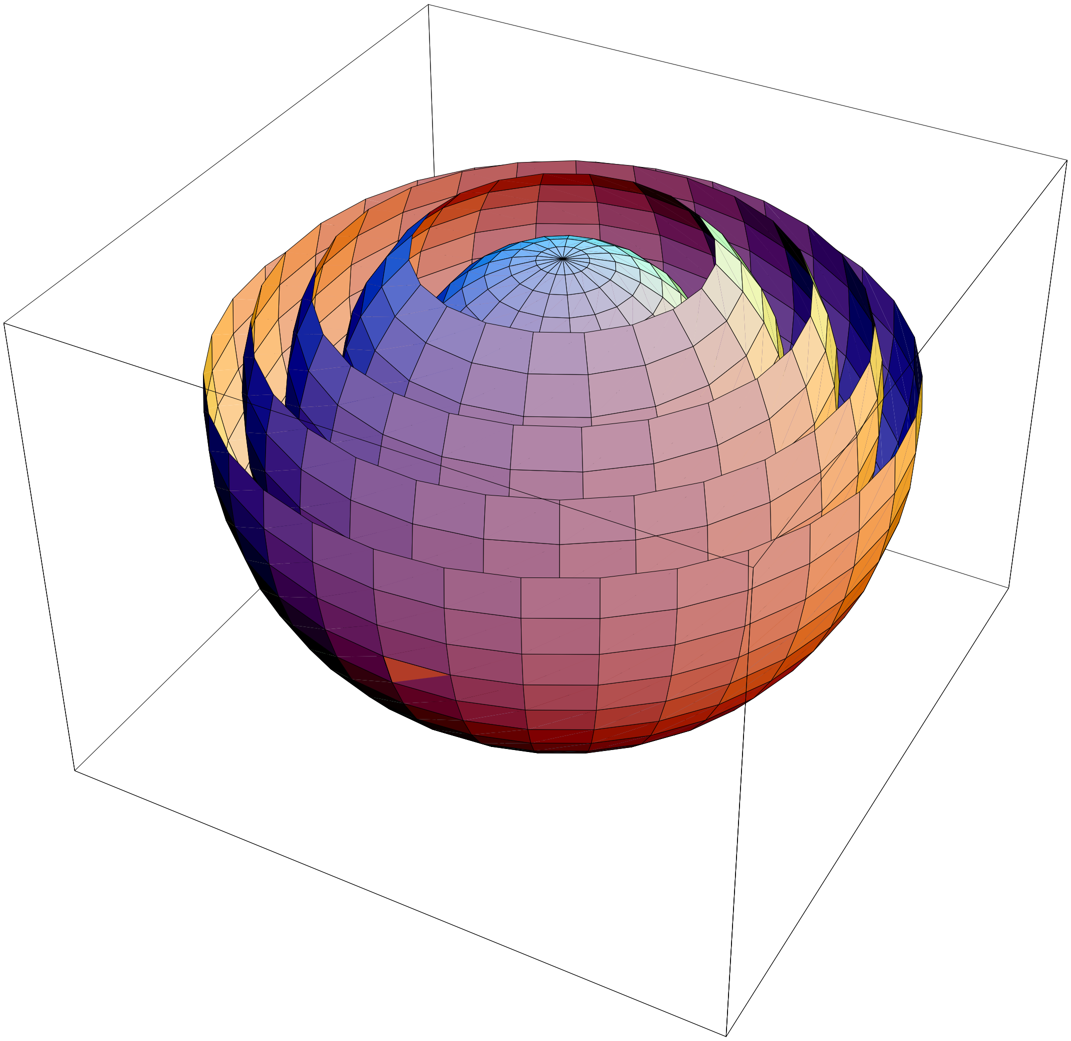
$$\begin{aligned} p^2 &= x^2 + y^2, \\ q^2 &= p^2 + z^2 \\ &= x^2 + y^2 + z^2 \\ &= w. \end{aligned}$$

lie a fixed distance from the origin. Specifically, $w = x^2 + y^2 + z^2 = c$ is the following set:

- $c > 0$: the sphere of radius $\sqrt{c}$ entered at the origin;
- $c = 0$: the origin itself;
- $c < 0$: the empty set.

The contours are spheres

The contour plot of $w = x^2 + y^2 + z^2$ is thus a nest of concentric spheres, as shown in the illustration below. The value of w is constant on each sphere. (The tops of the spheres have been cut away so you can see how the spheres nest; the whole thing resembles an onion.)



A standard minimax

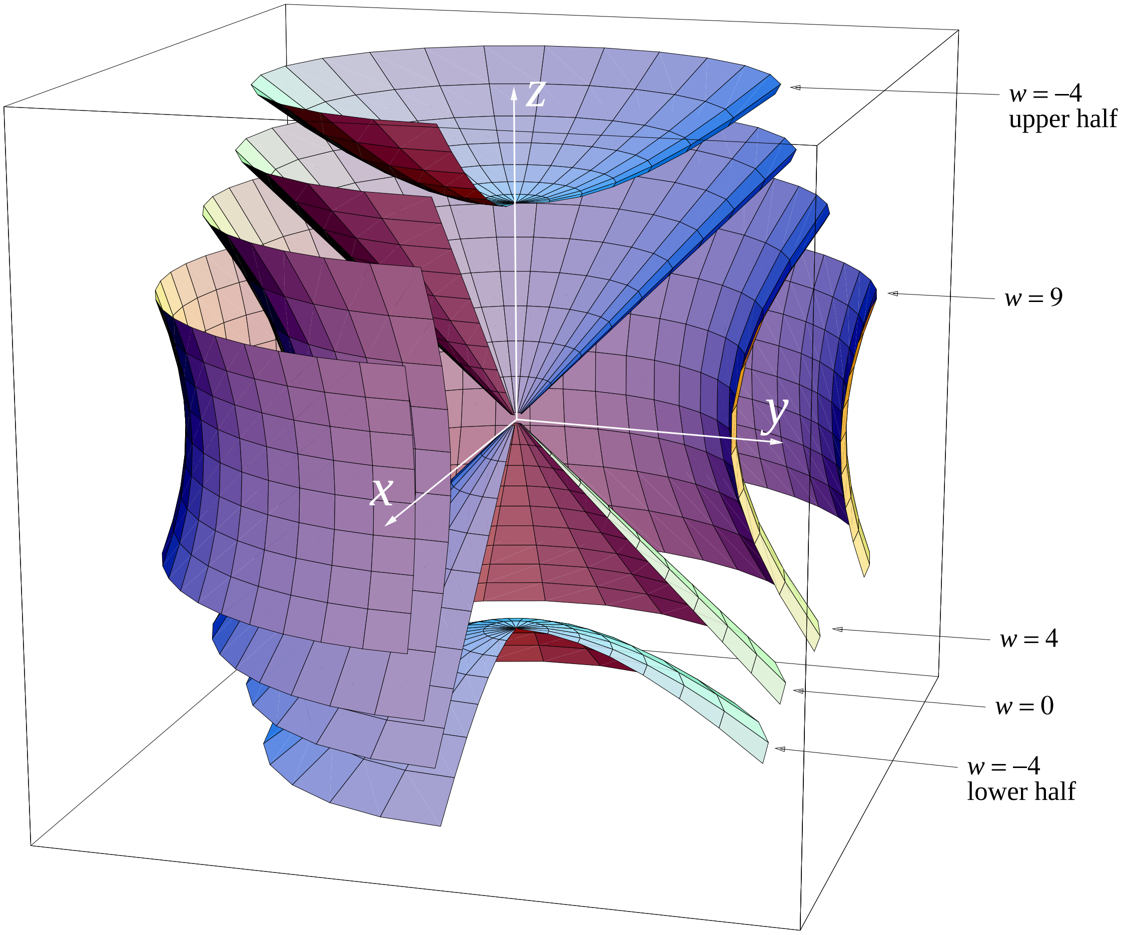
Below is the contour plot of another standard function with three input variables:

$$w = f(x, y, z) = x^2 + y^2 - z^2.$$

A quarter of each surface has been cut away so you can see how the surfaces nest together. Note that $w = 0$ is a cone, and every surface with $w < 0$ consists of two disconnected (but congruent) pieces—an upper half and a lower half.

The contours are hyperbolic shapes

You should compare this function to the standard minimax $x^2 - y^2$ in two variables. The three-variable function w has a minimum with respect to *both* of the variables x and y, while it has a maximum with respect to z. (Do you see why? The arguments are exactly the same as they were for two input variables on page 515.) Furthermore, the contours of $x^2 - y^2$ are a family of hyperbolas, and the contours of $x^2 + y^2 - z^2$ are surfaces obtained by rotating these hyperbolas about a common axis.



When there are three input variables, the contours are surfaces

It is a general fact—and our two examples provide good evidence for it—that a single contour of a function of three variables is a *surface*. Thus a contour is a curve or a surface, depending on the number of input variables. We often use the term **level set** (rather than a level *curve* or a level *surface*) as a generic name for a contour.

Exercises

In many of these exercises it will be essential to have a computer program to make graphs, terraced density plots, and contour plots of functions of two variables.

1. a) Use a computer to obtain a graph of the function $z = \sin x \sin y$ on the domain $0 \leq x \leq 2\pi$, $0 \leq y \leq 2\pi$. How many maximum points do you see? How many minimum points? How many saddles?

b) Determine, as well as you can from the graph, the location of the maximum, minimum, and saddle points.

2. Continuation. Make the domain $-2\pi \leq x \leq 2\pi$, $-2\pi \leq y \leq 2\pi$ and answer the same questions you did in the previous exercise. (Does the graph look like an egg carton?)

3. Obtain a terraced density plot (or a contour plot) of $z = \sin x \sin y$ on the domain $-2\pi \leq x \leq 2\pi$, $-2\pi \leq y \leq 2\pi$. Locate the maximum, minimum, and saddle points of the function. Do these results agree with those from the previous exercise?

4. Obtain the graph of $z = \sin x \cos y$ on the domain $0 \leq x \leq 2\pi$, $0 \leq y \leq 2\pi$. How does this graph differ from the one in exercise 1? In what ways is it similar?

5. Obtain the graph of $z = 2x+4x^2-x^4-y^2$ when $-2 \leq x \leq 2$, $-4 \leq y \leq 4$. Locate all the minimum, maximum, and saddle points in this domain. [Note: the minimum is on the boundary!]

6. Continuation. Obtain a terraced density plot (or contour plot) for the function in the previous exercise, using the same domain. Use the plot to locate all the minimum, maximum, and saddle points. Compare your results with those of the previous exercise.

7. a) Obtain the graph of $z = 2x - y$ on the domain $-2 \leq x \leq 2$, $-2 \leq y \leq 2$. What is the shape of the graph?

b) Graph the same function of the domain $2 \leq x \leq 6$, $0 \leq y \leq 4$. What is the shape of the graph? How does this graph compare to the one in part (a)?

8. a) Continuation. Sketch three different slices of the graph of $z = 2x - y$ in the y-direction. What do the slices have in common? How are they different?

b) Answer the same questions for slices in the x-direction.

9. a) Obtain the graph of $z = .3x + .8y + 2.3$; choose the domain yourself. Where does the graph intercept the z-axis?

b) Describe the vertical slices of this graph in the y-direction and in the x-direction.

10. Describe the vertical slices of the graph of $z = px + qy + r$ in the y-direction and in the x-direction.

11. a) Compare the contours of the function $z = x^2 + 2y^2$ to those of $z = x^2 + y^2$.

b) What is the shape of the graph of $z = x^2 + 2y^2$? Decide this first using only the information you have about the contours. Then use a computer to obtain the graph.

12. a) Compare the contours of the function $z = x^2 - 2y^2$ to those of $z = x^2 - y^2$.

b) What is the shape of the graph of $z = x^2 - 2y^2$? Decide this first using only the information you have about the contours. Then use a computer to obtain the graph.

13. a) Obtain a contour plot of the function $z = x^2 + xy + y^2$.

b) What is the shape of the graph of $z = x^2 + xy + y^2$? Decide this first using only the information you have about the contours. Then use a computer to obtain the graph.

14. a) Obtain a contour plot of the function $z = x^2 + 3xy + y^2$.

b) What is the shape of the graph of $z = x^2+3xy+y^2$? Decide this first using only the information you have about the contours. Then use a computer to obtain the graph.

15. a) Obtain a contour plot of the function $z = x^2 + 2xy + y^2$.

b) What is the shape of the graph of $z = x^2+2xy+y^2$? Decide this first using only the information you have about the contours. Then use a computer to obtain the graph.

16. Complete this statement: The function $f(x, y; p) = x^2 + pxy + 4y^2$, which depends on the parameter p, has a minimum at the origin when _____ and a minimax when _____.

17. a) Obtain the graph and a terraced density plot of the function $z = 3x^2 + 17xy + 12y^2$. What is the shape of the graph?

b) What is the shape of the contours? Indicate how the contours fit on the graph.

18. a) Obtain the graph and a terraced density plot of the function $z = 3x^2 + 7xy + 12y^2$. What is the shape of the graph?

b) What is the shape of the contours? Indicate how the contours fit on the graph.

19. a) Obtain the graph and a terraced density plot of the function $z = 3x^2 + 12xy + 12y^2$. What is the shape of the graph?

b) What is the shape of the contours? Indicate how the contours fit on the graph.

20. Obtain the graph of $z = f(x, y) = xy$ on the domain $-3 \leq x \leq 3$, $-3 \leq y \leq 3$. Does this function have a maximum or a minimum or a saddle point? Where?

21. a) Continuation. Sketch slices of the graph of $z = xy$ in the y-direction, for each of the values $x = -2$, -1, 0, 1, and 2. What is the general shape of each of these slices?

b) Repeat part (a), but make the five slices in the x-direction—that is, fix y instead of x.

22. Continuation. Show how the slices you obtained in the previous exercise fit (or appear) on the graph you obtained in the exercise just before that one.

23. a) Continuation. Let $x = u + v$ and $y = u - v$. Express z in terms of u and v, using the fact that $z = xy$. Then obtain the graph of z as a function of the new variables u and v.

b) What is the shape of the graph you just obtained? Compare it to the graph of $z = xy$ you obtained earlier.

24. Can you draw a network of straight lines on the saddle surface $z = x^2 - y^2$?

25. Obtain a terraced density plot of $z = xy$. How do the contours of this plot fit on the graph of $z = xy$ you obtained in a previous exercise?

26. The graphs of $z = x^2 + 5xy + 10y^2$ and $z = 3$ intersect in a curve. What is the shape of that curve?

27. The graphs of $z = x^2 + y^3$ and $z = 0$ intersect in a curve. What is the shape of that curve?

28. The graphs of $z = 2x - y$ and $z = .3x + .8y + 2.3$ intersect in a curve. What is the shape of that curve?

First integrals

29. A hard spring described by the dynamical system

$$\frac{dx}{dt} = v, \qquad \frac{dv}{dt} = -cx - \beta x^3,$$

has a first integral of the form

$$E(x, v) = \tfrac{1}{2}cx^2 + \tfrac{1}{2}\beta x^4 + \tfrac{1}{2}v^2.$$

This is the **energy** of the spring. (See chapter 7.3, especially exercise 13, page 454.)

a) Let $c = 16$ and $\beta = 1$. Obtain the graph of $E(x, v)$ on a domain that has the origin at its center. Locate all the minimum, maximum, and saddle points in this domain.

b) What is the state of the spring (that is, its position x and its velocity v) when it has minimum energy?

30. A soft spring described by the dynamical system

$$\frac{dx}{dt} = v, \qquad \frac{dv}{dt} = \frac{-25x}{1+x^2},$$

has an energy integral of the form

$$E(x, v) = \tfrac{25}{2}\ln(1 + x^2) + \tfrac{1}{2}v^2.$$

(See exercise 16, page 455.)

a) Obtain the graph of $E(x, v)$. Experiment with different possibilities for the domain until you get a good representation.

b) Obtain a terraced density plot of $E(x, v)$ over the same domain you chose in part (a). Compare the two representations of E.

c) Does the spring have a state of minimum energy? If so, where is it?

d) Does the spring have a state of *maximum* energy? Explain your answer.

31. a) **The Lotka–Volterra equations**. According to exercise 33 of chapter 7.3 (page 458), the function

$$E(x, y) = .1\ln y + .04\ln x - .005\,y - .004\,x$$

is a first integral of the dynamical system

$$\begin{aligned} x' &= .1x - .005\,xy, \\ y' &= .004\,xy - .04\,y. \end{aligned}$$

Obtain the graph of E on the domain $1 \le x \le 50$, $1 \le y \le 50$. (Why not enlarge the domain to $0 \le x \le 50$, $0 \le y \le 50$?)

b) Find all maximum, minimum, and saddle points on this graph. What is the connection between the maximum of E and the equilibrium point of the dynamical system?

c) Obtain a contour plot of E on the same domain as in part (a). Compare the contours of E and the trajectories of the dynamical system. (This reveals a *conservation of "energy"* for the solutions of the Lotka-Volterra equations.)

32. a) Continuation. Here is another first integral of the same dynamical system as in the previous exercise:

$$H(x,y) = \frac{x^{0.04}y^{0.1}}{\exp(.005\,y + .004\,x)}.$$

Obtain the graph of H and compare it to the graph of E in the previous exercise.

b) Obtain a contour plot of H, and compare the contours to the trajectories of the dynamical system.

9.2 Local Linearity

Local linearity is the central idea of chapter 3: it says that a graph looks straight when viewed under a microscope. Using this observation we were able to give a precise meaning to the *rate of change* of a function and, as a consequence, to see why Euler's method produces solutions to differential equations. At the time we concentrated on functions with a single input variable. In this section we explore local linearity for functions with two or more input variables.

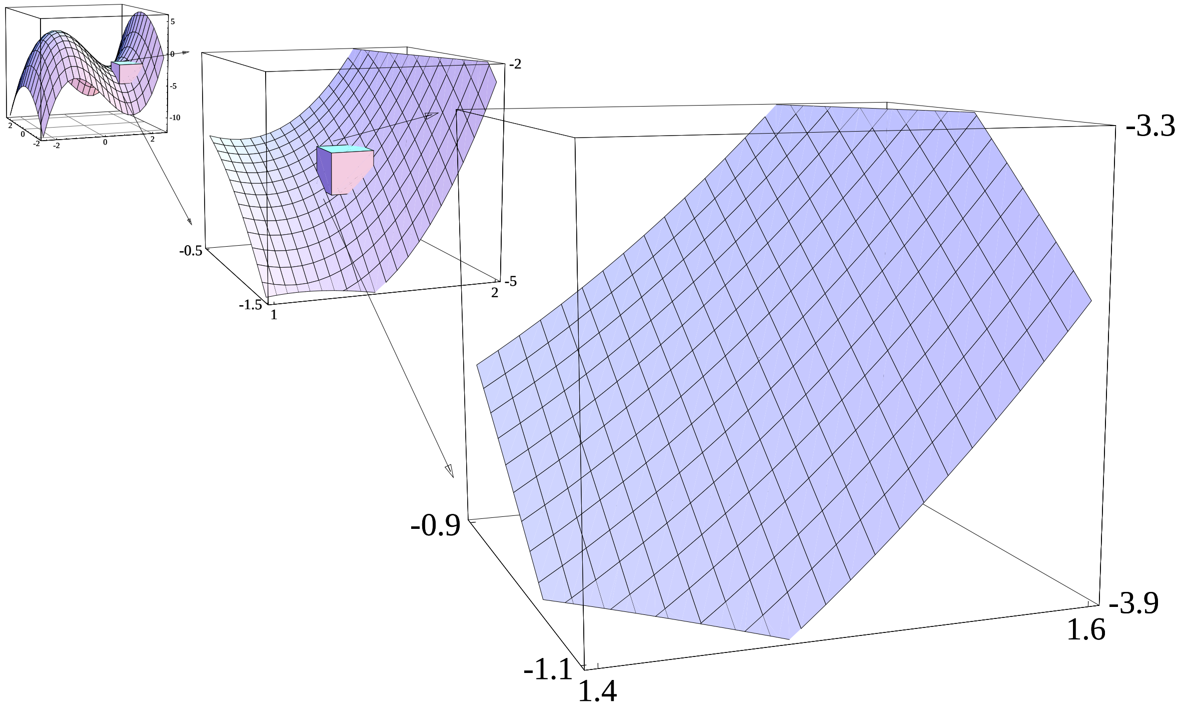
Microscopic Views

Magnifying a graph

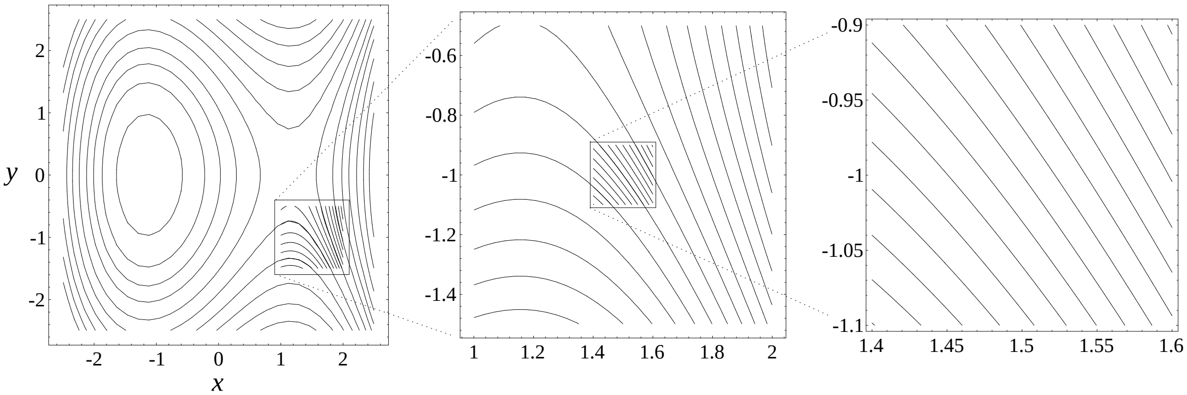
Consider the cubic $f(x, y) = x^3 - 4x - y^2$ that we used as an example in the previous section. We'll examine both the graph and the plot of f under a microscope. In the figure below we see successive magnifications of the graph near the point where $(x, y) = (1.5, -1)$. The initial graph, in the left rear, is drawn over the square

$$-2.5 \leq x \leq 2.5, \qquad -2.5 \leq y \leq 2.5.$$

With each magnification, the portion of the surface we see bends less and less. **The graph approaches the shape of a flat plane**.



Contour plots for $f(x, y) = x^3 - 4x - y^2$ appear below. Again, we magnify near the point $(x, y) = (1.5, -1)$. Each window below is a small part of the window to its left. In the large scale plot, which is the first one on the left, the contours are quite variable in their direction and spacing. With each magnification, that variability decreases. **The contours become straight, parallel, and equally spaced**.



The process of magnification thus leads us to functions whose graphs are flat and whose contours are straight, parallel, and equally-spaced. As we shall now see, these are the linear functions.

Linear Functions

Responses to changes in input

A linear function is defined by the way its output responds to *changes* in the input. Specifically, in chapter 1 we said

$$y = f(x) \quad \text{is linear if} \quad \Delta y = m \cdot \Delta x.$$

This is the simplest possibility: changes in output are strictly proportional to changes in input. The multiplier m is both the *rate* at which y changes with respect to x and the *slope* of the graph of f.

The definition

Exactly same idea defines a linear function of two or more variables: the change in output is strictly proportional to the change in any one of the inputs.

Definition. The function $z = f(x_1, x_2, \ldots, x_n)$ is **linear** if there are multipliers $p_1, p_2, \ldots, p_n$ for which

$$\Delta z = p_1 \cdot \Delta x_1, \quad \Delta z = p_2 \cdot \Delta x_2, \quad \ldots, \quad \Delta z = p_n \cdot \Delta x_n.$$

There is one multiplier for each input variable. The multipliers are constants and they are, in general, all different.

Partial and total changes

The definition describes how z responds to each input separately. We call each $p_j \cdot \Delta x_j$ a **partial change**. The multiplier $p_j = \Delta z/\Delta x_j$ is the corresponding **partial rate of change**. Of course, several input variables may change simultaneously. In that case, the **total change** in z will just be the sum of the individual changes produced by the several variables:

$$\Delta z = p_1 \cdot \Delta x_1 + p_2 \cdot \Delta x_2 + \cdots + p_n \cdot \Delta x_n.$$

Another way to describe a linear function

Of course, if the total change of a function satisfies this condition, then each partial change has the form $p_j \cdot \Delta x_j$. (If only x_j changes, then all the other Δx_k must be 0. So Δz becomes simply $p_j \cdot \Delta x_j$.) Consequently, the function must be linear. In other words, we can use the formula for the total change as another way to define a linear function.

Alternate definition. The function $z = f(x_1, x_2, \ldots, x_n)$ is **linear** if there are multipliers $p_1, p_2, \ldots, p_n$ for which

$$\Delta z = p_1 \cdot \Delta x_1 + p_2 \cdot \Delta x_2 + \cdots + p_n \cdot \Delta x_n.$$

Formulas for linear functions

From the definition to a formula

When $z = f(x_1, x_2, \ldots, x_n)$ is a linear function, we know how Δz depends on the changes Δx_j, but that doesn't tell us explicitly how z itself is related to the input variables x_j. There are several ways to express this relation as a formula, depending on the nature of the information we have about the function. For the sake of clarity, we'll develop these formulas first for a function of two variables: $z = f(x, y)$.

Given the partial rates of change and an initial point

• **The initial-value form**. Suppose we know the value of a linear function at some given point—called the **initial point**—and we also know its partial rates of change. Can we construct a formula for the function? Suppose $z = z_0$ when $(x, y) = (x_0, y_0)$, and suppose the partial rates of change are

$$p = \frac{\Delta z}{\Delta x} \qquad \text{and} \qquad q = \frac{\Delta z}{\Delta y}.$$

If we let

$$\Delta x = x - x_0, \qquad \Delta y = y - y_0, \qquad \Delta z = z - z_0,$$

then we can write

$$\begin{aligned} z - z_0 = \Delta z &= p \cdot \Delta x + q \cdot \Delta y \\ &= p \cdot (x - x_0) + q \cdot (y - y_0). \end{aligned}$$

This is the initial-value form of a linear function. For example, if the initial point is $(x, y) = (4, 3)$, $z = 5$, and the partial rates of change are $\Delta z/\Delta x = -\frac{1}{2}$, $\Delta z/\Delta y = +1$, the equation of the linear function can be written

$$z - 5 = -\tfrac{1}{2}(x - 4) + (y - 3).$$

Given the partial rates of change and the z-intercept

- **The intercept form**. This is a special case of the initial-value form, in which the initial point is the origin: $(x, y) = (0, 0)$, $z = r$. The formula becomes

$$z - r = px + qy, \qquad \text{or} \qquad z = px + qy + r.$$

As we shall see, the graph of this function in x, y, z-space passes through the point $(x, y, z) = (0, 0, r)$ on the z-axis. This point is called the ***z*-intercept** of the graph. Sometimes we simply call the number r itself the z-intercept.

Notice that, with a little algebra, we can convert the previous example to the form $z = -\frac{1}{2}x + y + 4$. This is the intercept form, and the z-intercept is $z = 4$.

If there are n input variables, x_1, x_2, …, x_n, instead of two, and an initial point has coordiantes x_1^0, x_2^0, …, x_n^0, then a linear equation has the following forms:

The form of a linear function of n variables

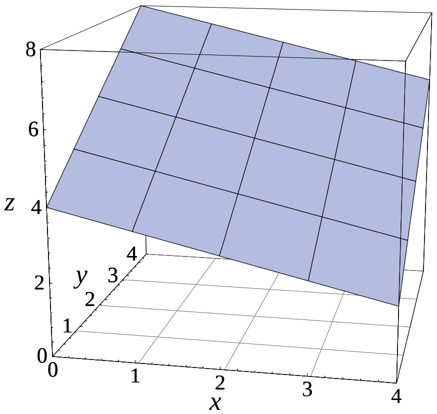
$$\begin{aligned} &\textbf{initial-value:} \quad & z - z_0 &= p_1(x_1 - x_1^0) + p_2(x_2 - x_2^0) + \cdots + p_n(x_n - x_n^0), \\ &\boldsymbol{z}\textbf{-intercept:} \quad & z &= p_1x_1 + p_2x_2 + \cdots + p_nx_n + r. \end{aligned}$$

The graph of a linear function

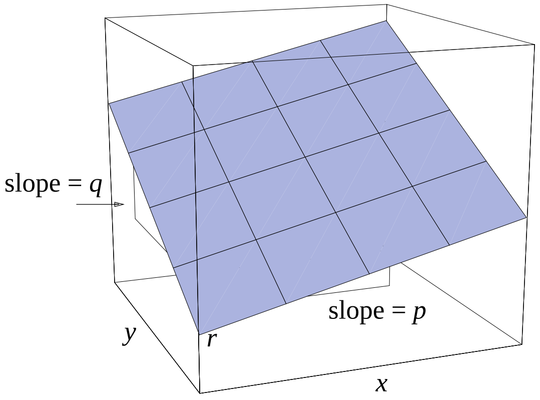
On the left at the top of the next page is the graph of the linear function

$$z = \tfrac{1}{2}x + y + 4.$$

The graph is a flat plane. In particular, grid lines parallel to the x-axis (which represent vertical slices with $y = c$) are all straight lines with the same slope $\Delta z/\Delta x = -\frac{1}{2}$. The other grid lines (with $x = c$) are all straight lines with the same slope $\Delta z/\Delta y = +1$.



$z = -\frac{1}{2}x + y + 4$



$z = px + qy + r$

On the right, above, is the graph of the general linear function written in intercept form: $z = px+qy+r$. The graph is the plane that can be identified by three distinguishing features:

- it has slope p in the x-direction;
- it has slope q in the y-direction.
- it intercepts the z-axis at $z = r$;

The definition of a linear function implies that its graph is a flat plane

Let's see how we can *deduce* that the graph must be this plane. First of all, the partial rate $\Delta z/\Delta x$ tells us how z changes when y is held fixed. But if we fix $y = c$, we get a vertical slice of the graph in the x-direction. The slope if that vertical slice is $\Delta z/\Delta x = p$. Since p is constant, the slice is a straight line. The value of $y = c$ determines which slice we are looking at. Since $\Delta z/\Delta x$ doesn't depend on y, all the slices in the x-direction have the *same* slope. Similarly, all the slices in the y-direction are straight lines with the same slope q. The only surface that can be covered by a grid of straight lines in this way is a flat plane. Finally, since $z = r$ when $(x, y) = (0, 0)$, the graph intercepts the z-axis at $z = r$.

Contours of a linear function

Each contour is a straight line

A contour of *any* function $f(x, y)$ is the set of points in the x, y-plane where $f(x, y) = c$, for some given constant c. If $f = px + qy + r$, then a contour has the equation

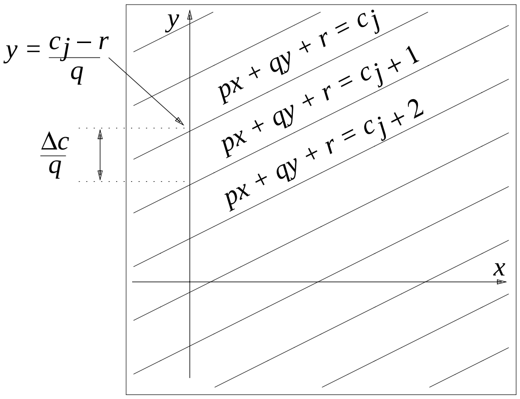
$$px + qy + r = c \qquad \text{or} \qquad y = -\frac{p}{q}x + \frac{c-r}{q}.$$

This is an ordinary straight line in the x, y-plane. Its slope is $-p/q$ and its y-intercept is $(c-r)/q$. (If $q = 0$ we can't do these divisions. However, this causes no problem; you should check that the contour is just the vertical line $x = (c-r)/p$.)

To construct a contour plot, we must give the constant c a sequence of equally-spaced values c_j, with $c_{j+1} = c_j + \Delta c$. This generates a sequence of straight lines

$$px + qy + r = c_j, \quad \text{or} \quad y = -\frac{p}{q}x + \frac{c_j - r}{q}.$$

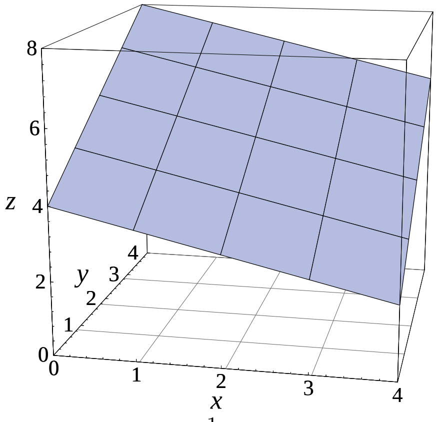
These lines all have the same slope $-p/q$, so they are parallel. (Notice the value of c doesn't affect the slope.) The y-intercept of the j-th contour is $(c_j - r)/q$. Therefore, the distance along the y-axis between one intercept and the next is $\Delta c/q$. The contours are thus straight, parallel, and equally-spaced. (You should check that this is still true if $q = 0$.) Note that the figure at the left, above, is drawn with $\Delta c > 0$ but $q < 0$.



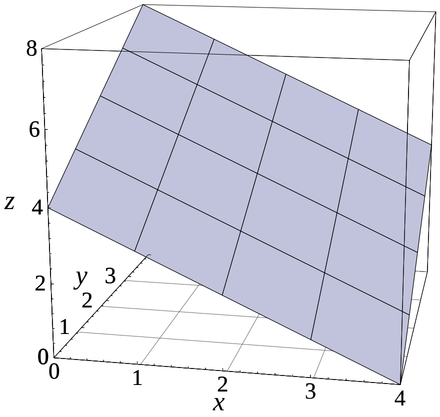
Geometric interpretation of the partial rates

Partial rates and partial slopes

What happens to the graph or the contour plot if you double one of the partial rates of change of a linear function? The graph on the right, below, shows the effect of doubling the partial rate with respect to x of the function $z = -\frac{1}{2}x + y + 4$. As you can see, the slope in the x-direction



$z = -\frac{1}{2}x + y + 4$



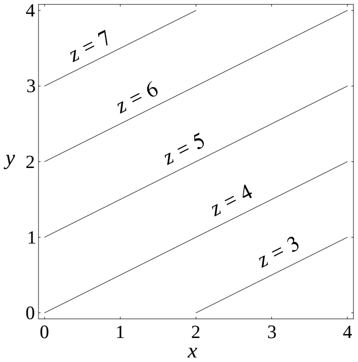
$z = -x + y + 4$

The overall tilt of a graph is altered

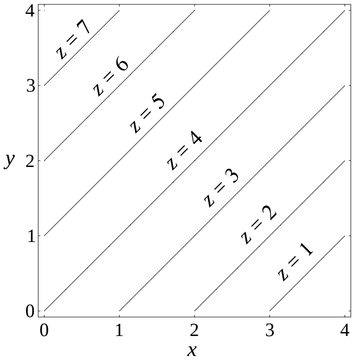
is doubled (from $-\frac{1}{2}$ to -1). Had we increased the partial rate by a factor of 10, the slope would have increased by a factor of 10 as well. Notice that the slope in the y-direction is not affected. Nevertheless, the overall 'tilt' of the graph *has* been altered. We shall have more to say about this feature in a moment, when we introduce the **gradient** of a linear function to describe the overall tilt.

Partial rate and the spacing of contours

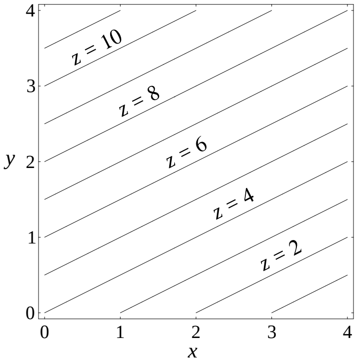
A change in the partial rates has a more complex effect on the contour plot. Perhaps it is more surprising, too. To make valid comparisons, we have constructed all three plots below with the same spacing between levels (namely $\Delta z = 1$). Notice how the levels meet the x- and y-axes in the plot on the left ($z = -tfrac12x + y + 4$). For each unit step we take along the y-axis, the z-value increases by 1. This is the meaning of $\Delta z/\Delta y = +1$. By contrast, we have to take *two* unit steps along the x-axis to produce the same size change in z. Moreover, z *decreases* by 1 when x increases by 2. This is the meaning of $\Delta z/\Delta x = -\frac{1}{2}$. In particular, the relatively wide spacing between z-levels along the x-axis reflects the relative smallness of $\Delta z/\Delta x$.



$z = -\frac{1}{2}x + y + 4$



$z = -x + y + 4$



$z = -x + 2y + 4$

The larger the partial rate, the closer the contours

Therefore, when we double the size of $\Delta z/\Delta x$—as we do in the middle plot—we should cut in half the spacing between z-levels along the x-axis. As you can see, this is exactly what happens. Notice that the spacing along the y-axis is not altered. Consequently, the contours change direction and they get packed more closely together.

Suppose we double *both* partial rates—as we do in the plot on the right. Then the spacing between contours is cut in half along both axes. Because the change is uniform, the contours keep their original direction.

The Gradient of a Linear Function

The vector of partial rates

By making use of the concept of a vector, introduced in the last chapter, we can construct still another geometric interpretation of the partial rates of a linear function. This vector is called the gradient, and it is defined in the following way.

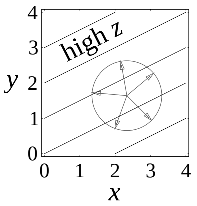
Definition. The **gradient** of a linear function $z = f(x, y)$ is the vector whose components are its partial rates of change:

$$\text{grad } z = \nabla z = \left(\frac{\Delta z}{\Delta x}, \frac{\Delta z}{\Delta y}\right).$$

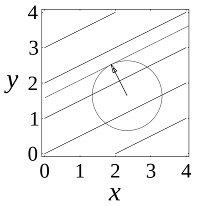
The direction of most rapid growth

The gradient is perhaps the most concise and useful tool for describing the growth of a function of several variables. To get an idea of the role that it plays, consider this question: *In what direction should we move from a given point in the x, y-plane so that the value of a linear function increases most rapidly*?

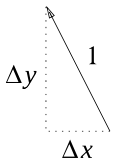
$z = -\frac{1}{2}x + y + 4$



Of course, the answer will depend on the linear function. Let's use $z = -\frac{1}{2}x + y + 4$ and start from the point $(x, y) = (2.4, 1.6)$. We can make z undergo a very large change simply by moving very far from this point. Therefore, to make valid comparisons, we will restrict ourselves to motions that carry us exactly one unit of distance in various directions. The vectors in the figure at the right show some of the possibilities. Their tips lie on a circle of radius 1.



Thus, to choose the direction in which z increases most rapidly, we must simply find the point on this circle where the value of z is largest. The contour line at this level must be tangent to the circle. The vector perpendicular to this contour line (see the second figure) therefore points in the direction of most rapid growth. Since perpendiculars have negative reciprocal slopes, and since all the contour lines have slope $+1/2$, it follows that the vector must have slope $-2/1$.



At the left is a magnified view of this vector. We know

$$\Delta x < 0, \qquad \frac{\Delta y}{\Delta x} = -2, \qquad \text{and} \qquad (\Delta x)^2 + (\Delta y)^2 = 1.$$

Thus $\Delta y = -2 \cdot \Delta x$, so $(\Delta x)^2 + 4\,(\Delta x)^2 = 1$. This implies

$$5\,(\Delta x)^2 = 1, \qquad \text{so} \qquad \Delta x = \frac{-1}{\sqrt{5}}, \quad \Delta y = \frac{2}{\sqrt{5}}.$$

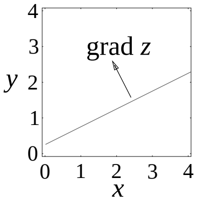
Thus, among all the motions $(\Delta x, \Delta y)$ we have considered, we obtain the greatest change in z by choosing

$$(\Delta x, \Delta y) = \left(\frac{-1}{\sqrt{5}}, \frac{2}{\sqrt{5}}\right).$$

The magnitude of most rapid growth

To determine how large this change is, we can use the alternate definition of a linear function (see page 536)

$$\Delta z = \frac{\Delta z}{\Delta x}\cdot\Delta x + \frac{\Delta z}{\Delta y}\cdot\Delta y = -\frac{1}{2}\cdot\frac{-1}{\sqrt{5}} + 1\cdot\frac{2}{\sqrt{5}} = \frac{5}{2\sqrt{5}} = \frac{\sqrt{5}}{2}.$$



The gradient vector quickly gives us all this information. First of all, the gradient vector has the value

$$\text{grad } z = \left(\frac{\Delta z}{\Delta x}, \frac{\Delta z}{\Delta y}\right) = \left(-\tfrac{1}{2}, 1\right).$$

Information from the gradient

Since its slope is $1/-\frac{1}{2} = -2$, we see that it does indeed point in the direction of most rapid growth. Consequently, it is also perpendicular to the contour line. Furthermore, its *length* gives the maximum growth rate. We can see this by calculating the length using the Pythagorean theorem:

$$\text{length} = \sqrt{\left(\frac{\Delta z}{\Delta x}\right)^2 + \left(\frac{\Delta z}{\Delta y}\right)^2} = \sqrt{\frac{1}{4}+1} = \sqrt{\frac{5}{4}} = \frac{\sqrt{5}}{2}.$$

Our findings with this example point to the following conclusion.

Theorem. The gradient of the linear function $z = px + qy + r$ is perpendicular to its contour lines. It points in the direction in which z increases most rapidly, and its length is equal to the maximum rate of increase.

A proof

Let's see why this is true. According to the observation on the previous page, the direction of most rapid increase will be perpendicular to the contour lines. The gradient of $z = px + qy + r$ is the vector $\nabla z = (p, q)$. Its slope is q/p. On page 539 we saw that the slope of the contour lines is $-p/q$. Since these slopes are negative reciprocals, the gradient is indeed perpendicular to the contour lines.

To determine the maximum rate of increase, we must see how much z increases when we move exactly 1 unit of distance in the gradient direction. The gradient vector is (p, q), and its length is $\sqrt{p^2+q^2}$. Therefore, the vector

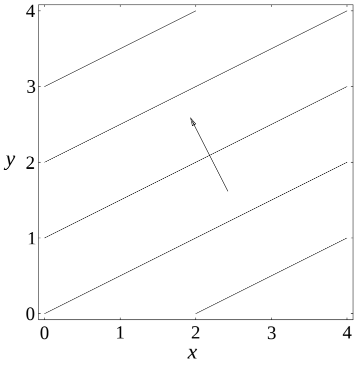
$$(\Delta x, \Delta y) = \left(\frac{p}{\sqrt{p^2+q^2}}, \frac{q}{\sqrt{p^2+q^2}} \right)$$

is 1 unit long and in the same direction as the gradient. The increase in z along this vector is

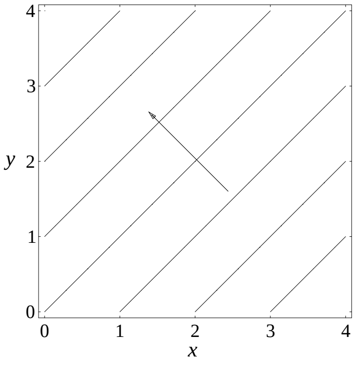
$$\Delta z = p \cdot \Delta x + q \cdot \Delta y = p \cdot \frac{p}{\sqrt{p^2+q^2}} + q \cdot \frac{q}{\sqrt{p^2+q^2}} = \frac{p^2+q^2}{\sqrt{p^2+q^2}} = \sqrt{p^2+q^2}.$$

End of the proof

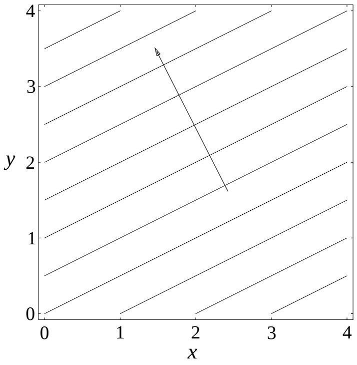
This *is* the length of the gradient vector, so we have confirmed that the length of the gradient is equal to the maximum rate of increase.



$z = -\frac{1}{2}x + y + 4$



$z = -x + y + 4$



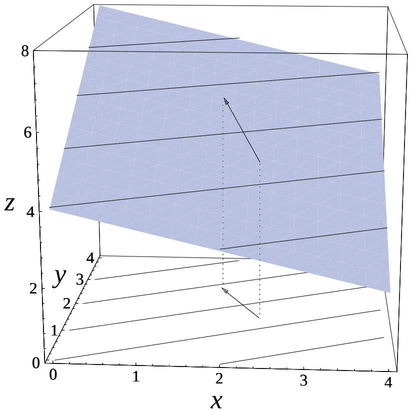
$z = -x + 2y + 4$

Contour spacing and the length of the gradient

Shown above are the three linear functions we've already examined. In each case the gradient vector is perpendicular to the contours, and it gets longer as the space between the contours *decreases*. This is to be expected because the space between contours is also an indicator of the maximum rate of growth of the function. Widely-spaced contours tell us that z changes relatively little as x and y change; closely-spaced contours tell us that z changes a lot as x and y change.

The gradient points directly uphill

The connection between the gradient and the graph is particularly simple. Since the gradient (which is a vector in the x, y-plane) points in the direction of greatest increase, it points in the direction in which the graph is tilted up. If we project the gradient vector onto the graph, as in the figure at the top of the next page, it points directly "uphill". Putting it another way, we can



say that the gradient shows us the "overall tilt" of the graph. There are two parts to this information. First, the direction of the gradient tells us which way the graph is tilted. Second, the length tells us how steep the graph is.

The figure at the left combines all the visual elements we have introduced to analyze a linear function: contours, graph, and gradient. Study it to see how they are related.

The Microscope Equation

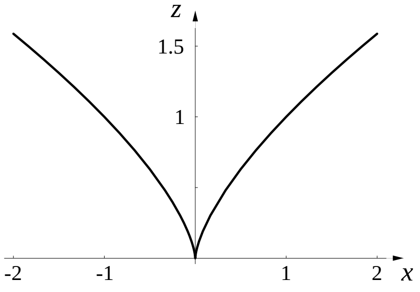
Local linearity

Local linearity

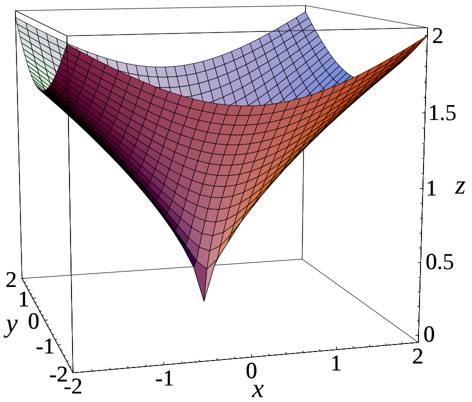
Let's return to arbitrary functions of two variables—that is, ones that are not necessarily linear. First we looked at magnifications of their graphs and contour plots under a microscope. We found that the graph becomes a plane, and the plot becomes a series of parallel, equally-spaced lines. Next, we saw that it is precisely the linear functions which have planar graphs and uniformly parallel contour plots. Hence this function is **locally linear**.

Exceptions

Of course, not *every* function is locally linear, and even a function that *is* locally linear at most points may fail to be so at particular points. We have already seen this with functions of a single variable in chapter 3. For example, $g(x) = x^{2/3}$ is locally linear everywhere *except* the origin. It has a sharp spike there. The two-variable function $f(x, y) = (x^2 + y^2)^{1/3}$ has the same sort of spike at the origin. The two graphs help make it clear that g is just a slice of f



$z = g(x) = x^{2/3}$



$z = f(x, y) = (x^2 + y^2)^{1/3}$

(constructed by taking $y = 0$). (Compare chapter 3.2, pages 113–114.) The spike is just one example; there are many other ways that a function can fail to be locally linear.

The relation between calculus and fractals

Functions that are *nowhere* locally linear are now being used in science to construct what are called **fractal** models. However, calculus does not deal with fractals. On the contrary, we remind you of the stipulation first made in chapter 3:

Calculus studies functions that are locally linear almost everywhere.

The microscope equation with two input variables

The equation of a microscopic view

If the function $z = f(x, y)$ is locally linear, then its graph looks like a plane when we view it under a microscope. The linear equation that describes that plane is the **microscope equation**. Since the plane is part of the graph of f, f itself must determine the form of the microscope equation. Let's see how that happens.

The idea is to reduce f to a function of one variable and then use the microscope equation for one-variable functions (described in chapter 3.3 and 3.7). Suppose the microscope is focused at the point $(x, y) = (a, b)$. If we fix y (at $y = b$), then z depends on x alone: $z = f(x, b)$. The microscope equation for this function at $x = a$ is just

The microscope equation in the x-direction...

$$\Delta z \approx \frac{\partial f}{\partial x}(a, b) \cdot \Delta x.$$

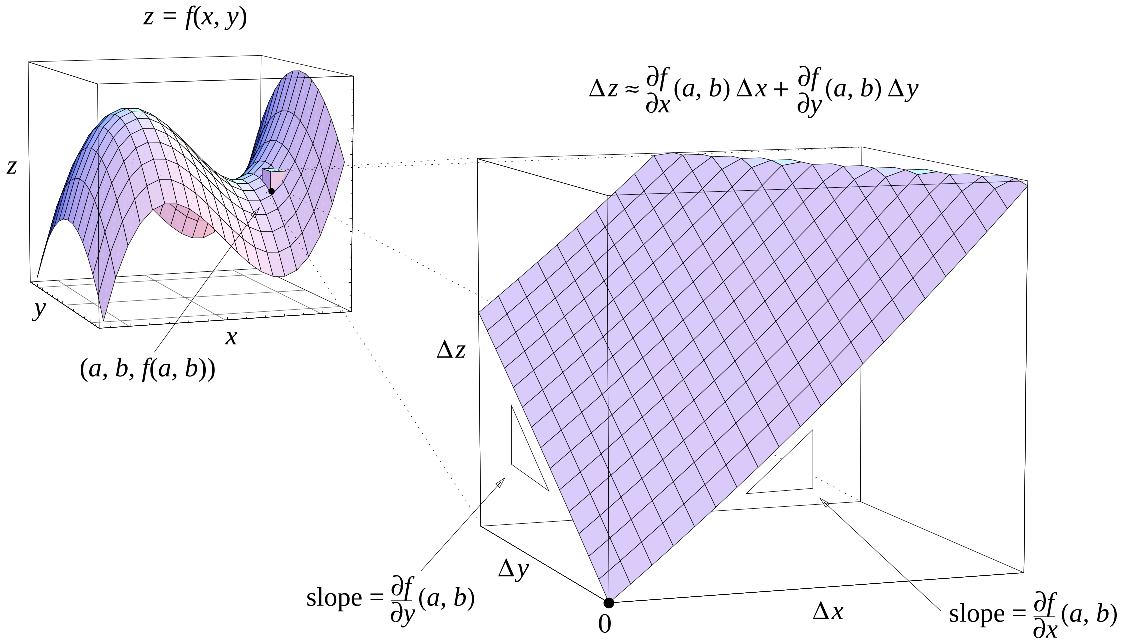
The multiplier $\partial f/\partial x$ is the rate of change of f with respect to x. We need to write it as a partial derivative because f is a function of two variables. Geometrically, the multiplier tells us the slope of a vertical slice of the graph in the x-direction.

Now reverse the roles of x and y, fixing $x = a$. The microscope equation for the function $z = f(a, y)$ at $y = b$ is

...and in the y-direction

$$\Delta z \approx \frac{\partial f}{\partial y}(a, b) \cdot \Delta y.$$

The multiplier $\partial f/\partial y$ in this equation is the slope of a vertical slice of the graph in the y-direction. The slopes of the two vertical slices are indicated in the microscope window that appears in the foreground of the following figure.


$z = f(x, y)$
z
y
x
$(a, b, f(a, b))$
$\Delta z \approx \frac{\partial f}{\partial x}(a, b)\,\Delta x + \frac{\partial f}{\partial y}(a, b)\,\Delta y$
Δz
Δy
slope = $\frac{\partial f}{\partial y}(a, b)$
0
Δx
slope = $\frac{\partial f}{\partial x}(a, b)$

From partial changes to total change

The separate microscope equations for the x- and y-directions give us the partial changes in z. However, as we saw when we were defining linear functions (page 536), when we know all the *partial* changes, we can immediately write down the *total* change:

The microscope equation:

$$\Delta z \approx \frac{\partial f}{\partial x}(a, b) \cdot \Delta x + \frac{\partial f}{\partial y}(a, b) \cdot \Delta y$$

As always, the origin of the microscope window corresponds to the point $(a, b, f(a, b))$ on which the microscope is focused. The microscope coordinates Δx, Δy, and Δz measure distances from this origin. For the sake of clarity in the figure above, we put the origin at one corner of the (three-dimensional) window.

An example

Incidentally, the function shown above is $f(x, y) = x^3 - 4x - y^2$, and the microscope is focused at $(a, b) = (1.5, -1)$. Since

$$\frac{\partial f}{\partial x} = 3x^2 - 4 = 2.75, \qquad \frac{\partial f}{\partial y} = -2y = 2$$

when $(x, y) = (1.5, -1)$, the microscope equation is

$$\Delta z \approx 2.75\,\Delta x + 2\,\Delta y.$$

Linear Approximation

The microscope equation gives a linear approximation

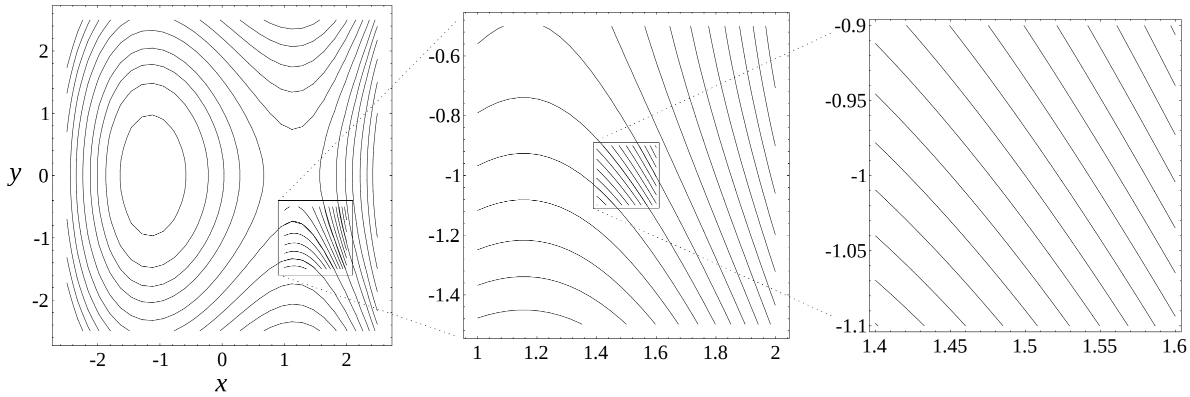
The microscope equation describes a linear function that approximates the original function near the point on which the microscope is focused. It is easy to see exactly how good the approximation is by comparing contour plots of the two functions. This is done below. In the window on the right, which shows the highest magnification, the solid contours belong to the original function $f = x^3 - 4x - y^2$. They are curved, but only slightly so. The dotted contours belong to the linear function

$$(z + 3.625) = 2.75(x - 1.5) + 2(y + 1) \quad \text{or} \quad z = 2.75\,x + 2\,y - 5.75.$$

(This is the microscope equation expressed in terms of the original variables x, y and z instead of the microscope coordinates $\Delta x = x - 1.5$, $\Delta y = y - (-1)$ and $\Delta z = z - (-3.625)$.)

Comparing the contours . . .

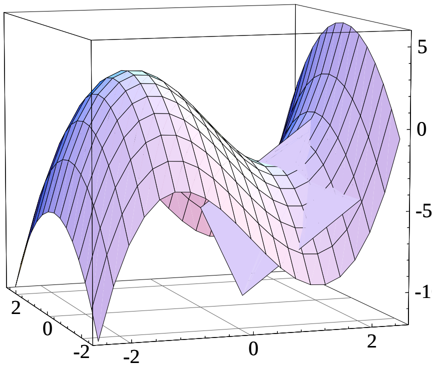
The difference between the two sets of contours shows us just how good the approximation is. As you can see, the two functions are almost indistinguishable near the center of the window, which is the point $(x, y) = (1.5, -1)$. As we look farther from the center, we find the contours of f depart more and more from strict linearity.

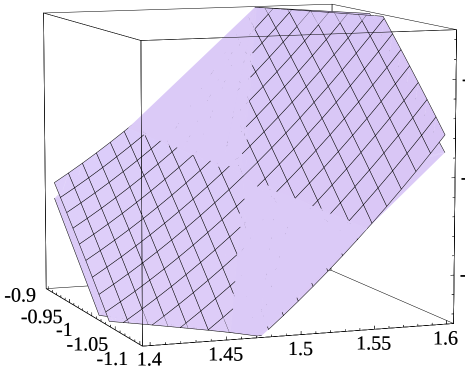


. . . and the graph of the function and its linear approximation

We can also compare the *graphs* of a function and its linear approximation. The graph of the linear approximation is a plane, of course. It is, in fact, the plane that is tangent to the graph of the function.

In the left figure immediately below we see the tangent plane to the graph of $z = f(x, y) = x^3 - 4x - y^2$ at the point where $(x, y) = (1.5, -1)$. On the right is a magnified view at the point of tangency. The graph of f is almost flat.

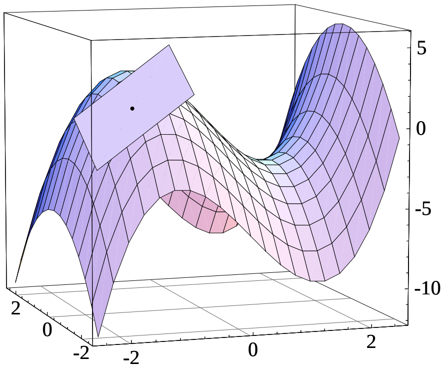


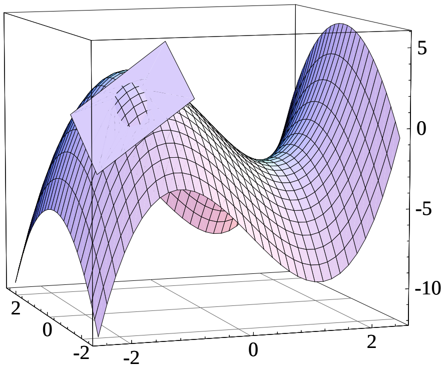


The grid helps us distinguish it from the tangent plane, which is solid gray. Such close agreement between the graph and the plane demonstrates how good the linear approximation is near the point. Notice, however, that the plane diverges from the graph as we move away from the point of tangency.

Tangent planes

At first glance you may not think that the plane in the figures above is tangent to the surface. The word "tangent" comes from the Latin *tangere*, to touch. We sometimes take this to mean "touch at one point", like the plane in the figure at the left, below. More properly, though, two objects are **tangent** if they have the same direction at a point where they meet. The plane in the figures above *does* meet this condition—as the microscopic view helps make clear.

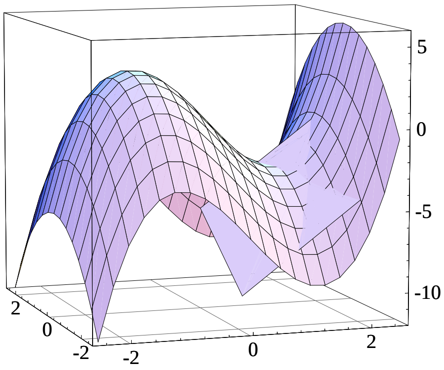


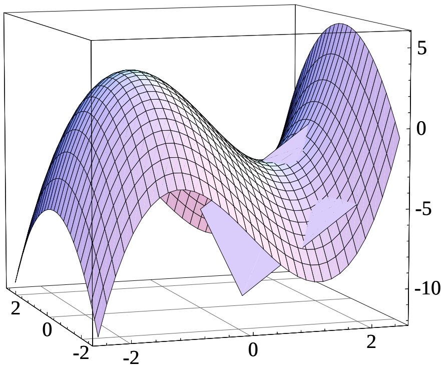


Elliptic points. . .

There are many different ways that a tangent plane can intersect a surface. What happens depends on the shape of the surface at the point of tangency.

The surface could bend the same way in all directions at that point, or it could bend up in some directions and down in others. In the first case, it will bend away from its tangent plane, so the two will meet at only one point. This is called an **elliptic point**, because the intersection turns into an ellipse if we push the plane in a bit. The figure on the right, above, shows what happens.





. . . and hyperbolic points

Suppose, on the other hand, that the surface bends up in some directions but down in others. Then, in some intermediate directions, it will not be bending at all. In those directions it will meet its tangent plane—which doesn't bend, either. Typically, there are two pairs of such directions. The surface and the tangent plane then intersect in an X. This always happens at a minimax (or saddle point) on a graph. and it also happens at the first point we considered on the graph of $z = x^3 - 4x - y^2$. This is shown again in the figure on the left above. It is called a **hyperbolic point**, because the intersection turns into a hyperbola if we push the plane a bit—as on the right. The lines where the tangent plane itself intersects the surface are called **asymptotic lines**, because they are the asymptotes of the hyperbolas. (To make it easier to see the elliptical and hyperbolic intersections we made the surface grid finer.)

> The curve of intersection between a surface and a shifted tangent plane is called the *Dupin indicatrix*. The Dupin indicatrix can take many forms besides the ones we have described here. However, at almost all points on almost all surfaces it turns out to be an ellipse or a hyperbola. More precisely, the indicatrix is *approximately* an ellipse or a hyperbola—in the same way that the surface itself is only approximately flat.

Parabolic points

Most points on a surface fall into one of two regions; one region consists of elliptic points, the other of hyperbolic. Points on the boundary between these two regions are said to be **parabolic**. On the graph of $z = x^3 - 4x - y^2$, if $x < 0$ the point is elliptic; if $x > 0$ it is hyperbolic; and if $x = 0$, it is parabolic. Try to confirm this yourself, just by looking at the surface.

The classification of the points into elliptical, parabolic, and hyperbolic types is one of the first steps in studying the curvature of a surface. This is part of *differential geometry*; calculus provides an essential language and tool. Differential geometry is used to model the physical world at both the cosmic scale (general relativity) and the subatomic (string theory).

The Gradient

Consequences of local linearity

Local linearity is a powerful principle. It says that an arbitrary function looks linear when we view it on a sufficiently small scale. In particular, all these statements are approximately true in a microscope window:

- the contours are straight, parallel, and equally spaced;
- the graph is a flat plane (the tangent plane);
- the function has a linear formula (the microscope equation);
- the partial derivatives are the slopes of the graph in the directions of the axes.

Extending the gradient to non-linear functions

There is one more aspect of a linear function for us to interpret—the gradient. Since partial rates become partial derivatives, we can make the following definition for an arbitrary locally linear function.

Definition. The **gradient** of a function $z = f(x, y)$ is the vector whose components are the partial derivatives:

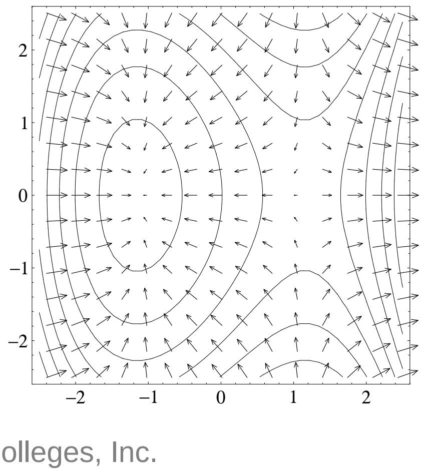
$$\text{grad } f = \nabla f = (f_x, f_y) = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right).$$

The gradient vector field

The partial derivatives are functions, so the gradient varies from point to point. Thus, gradients form a **vector field** in the same sense that a dynamical system does (see chapter 8). We draw the gradient vector $(f_x(x, y), f_y(x, y))$ as an arrow whose tail is at the point (x, y).

Example. $f(x, y) = x^3 - 4x - y^2$, $\quad$ grad $f = (3x^2 - 4, -2y)$

Contours and gradient vectors together



The vectors are rescaled for clarity

There is one thing you should notice about the previous figure: we drew the gradient vectors much shorter than they actually are. For instance, at the origin grad $f = (-4, 0)$, but the arrow *as drawn* is closer to $(-.25, 0)$. The purpose of rescaling is to keep the vectors out of each other's way, so the overall pattern easier to see.

The relation between gradients and contours

The example shows contours as well as gradients so we can see how the two are related. The result is very striking. Even though the vectors vary in length and direction, and the contours vary in direction and spacing, the two are related the same way they were for a *linear* function (page 542ff). First of all, each vector is perpendicular to the level curve that passes through its tail. Second, the vectors get longer where the spacing between level curves gets smaller. The similarity is no accident, of course; it is a consequence of local linearity. We can summarize and extend our observations in the following theorem. It is just a modification of the earlier theorem on the gradient of a linear function (page 542).

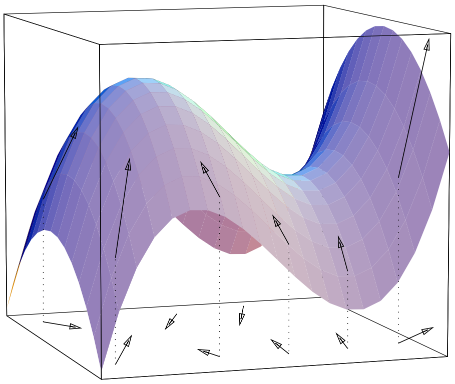
Theorem. The gradient vector field of the function $z = f(x, y)$ is perpendicular to its contour lines. At each point, the *direction* of the gradient is the direction in which z increases most rapidly; the *length* is equal to the maximum rate of increase.

A proof

To see why this theorem is true, just look in a microscope. The gradient and the contours become the gradient and the contours of the linear approximation at the point where the microscope is focused. But we already know the theorem is true for linear functions, so there is nothing more to prove.

The gradient points uphill

There is a direct connection between the gradient field of a function $z = f(x, y)$ and its graph. Since the gradient (which is a vector in the x, y-plane) points in the direction in which z increases most rapidly, it points in the direction in which the graph is tilted up. Thus, if we project the gradient onto the graph, as we do in the figure at the left, it points directly "uphill". Since f is not a linear function, both the steepness and the uphill direction vary from point to point. The gradient vector field also varies; in this way, it keeps track of the steepness of the graph and the direction of its tilt.



The Gradient of a Function of Three Variables

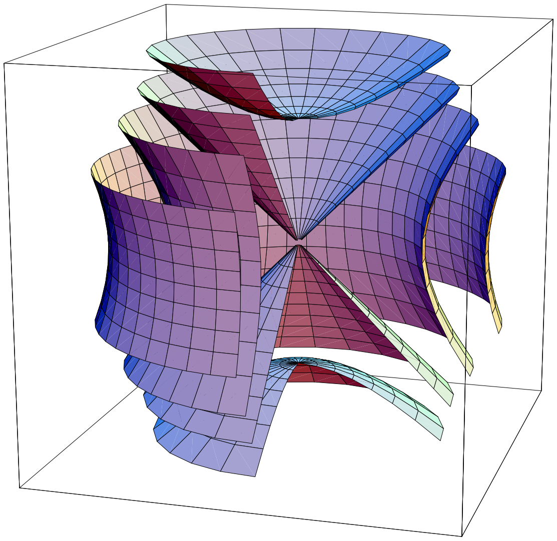
Let's take a brief glance at the gradient of a function $f(x, y, z)$. It has three components:

$$\operatorname{grad} f = \nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right) = (f_x, f_y, f_z),$$

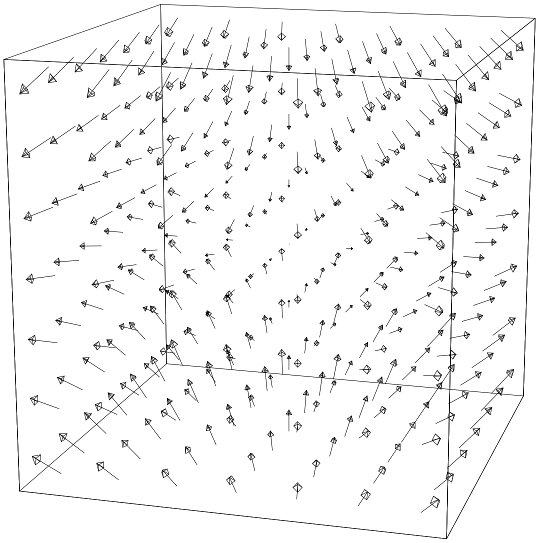
and it defines a vector field in x, y, z-space. At each point, the gradient of f is perpendicular to the level set through that point, and it points in the direction in which f increases most rapidly.

Visualizing a three-dimensional vector field

Example. In the two boxes below you can compare the gradient field of $f(x, y, z) = x^2 + y^2 - z^2$ with its level sets. At first glance, you may not find a clear pattern to the gradient vectors. After all, the picture is three-dimensional, and it is difficult to tell whether an arrow is near the front of the box or the back. However, there is a pattern: at the top and the bottom of the box, arrows point inward; closer to the middle of the box, they flare outward. The lowest values of the function occur along the z-axis, inside the shallow bowls that sit at the top and the bottom of the box. The highest values occur outside the "equatorial belt" formed by the outermost level set. This is where the x, y-plane meets the middle of the box. Notice also that the level sets are symmetric around the z-axis. The gradient field has the same symmetry, though this is harder to see.



$x^2 + y^2 - z^2 = c$



$\nabla f = (2x, 2y, -2z)$

Exercises

In many of these exercises it will be essential to have a computer program to make graphs and contour plots of functions of two variables.

1. a) Obtain the graph of $z = x^3 - 4x - y^2$ on a domain centered at the point $(x, y) = (1.5, -1)$, and magnify the graph until it looks like a plane.

b) Estimate, by eye, the slopes in the x-direction and the y-direction of the plane you found in part (a). [You should find numbers between +2 and +3 in both cases.]

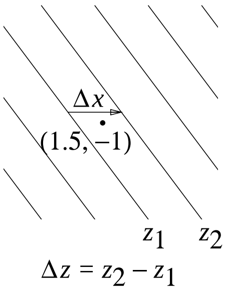
2. Obtain a contour plot of $z = x^3 - 4x - y^2$ on a domain centered at the point $(x, y) = (1.5, -1)$, and magnify the plot until the contours look straight, parallel, and equally spaced. Compare your results with the plots on page 535.

3. Continuation. The purpose of this exercise is to estimate the rate of change $\Delta z/\Delta x$ at $(x, y) = (1.5, -1)$, using the most highly magnified contour plot you constructed in the last exercise.

a) What is the horizontal spacing Δx between the two contours closest to the point $(1.5, -1)$? See the illustration at the right.

b) Find the z-levels z_1 and z_2 of those contours, and then compute $\Delta z = z_2 - z_1$.

c) Compare the value of $\Delta z/\Delta x$ you now obtain with the slope in the x-direction that you estimated in exercise 1.



4. Continuation. Repeat all the work of the last exercise, this time for $\Delta z/\Delta y$.

Linear functions

5. a) Find the z-intercept of the graph of the linear function given by the formula

$$z - 3 = 2(x - 4) - 3(y + 1).$$

b) Write the formula for this linear function in intercept form.

6. a) Write, in initial-value form, the formula for the linear function $z = L(x, y)$ for which $\Delta z/\Delta x = 3$, $\Delta z/\Delta y = -2$, and $L(1, 4) = 0$.

b) Write the intercept form of the same function. What is the z-intercept of the graph of L?

7. a) Suppose z is a linear function of x and y, and $\Delta z/\Delta x = -7$, $\Delta z/\Delta y = 12$. If (x, y) changes from $(35, 24)$ to $(33, 33)$, what is the total change in z?

b) Suppose $z = 29$ when $(x, y) = (35, 24)$. What value does z have when $(x, y) = (33, 33)$? What value does z have when $(x, y) = (0, 0)$?

c) Write the intercept form of the formula for z in terms of x and y.

8. Suppose z is a linear function of x and y for which we have the following information:

x	5	7	0		4	0	4	
y	9	1	4	6	7	0		-1
z	2		12	-7	2		20	20

a) Fill in the blanks in this table.

b) Write the formula for z in intercept form.

9. a) Sketch the graph of $z = x - 2y + 7$ on the domain $0 \le x \le 3$, $0 \le y \le 3$.

b) Determine the slope of this graph in the x-direction, and indicate on your sketch where this slope can be found.

c) Determine the slope of this graph in the y-direction, and indicate on your sketch where this slope can be found.

10. Continuation. Draw the gradient vector of $z = x - 2y + 7$ in the x, y-plane, and then lift it up so it sits on the graph you drew in the previous exercise. Does the gradient point directly "uphill"?

11. Sketch the graph of the linear function $z = L(x, y)$ for which

$$\frac{\Delta z}{\Delta x} = -1, \qquad \frac{\Delta z}{\Delta y} = .6, \qquad L(1, 1) = 8.$$

Be sure your graph shows clearly the slopes in the x-direction and the y-direction, and the z-intercept.

12. Continuation. Draw the gradient of the function L from the previous exercise, and lift it up so it sits on the graph of L you drew there. Does the gradient point directly uphill?

13. What is the equation of the linear function $z = L(x, y)$ whose graph (a) has a slope of -4 in the x-direction, (b) a slope of $+5$ in the y-direction, and (c) passes through the point $(x, y, z) = (2, -9, 0)$?

14. Suppose $z = L(x, y)$ is a linear function whose graph contains the three points

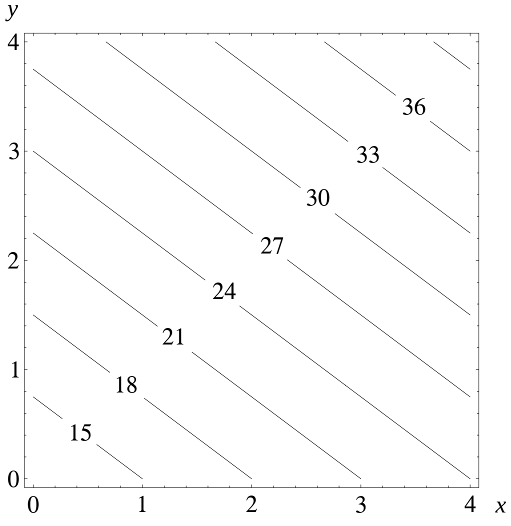
$$(1, 1, 2), \qquad (0, 5, 4), \qquad (-3, 0, 12).$$

a) Determine the partial rates of change $\Delta z/\Delta x$ and $\Delta z/\Delta y$.

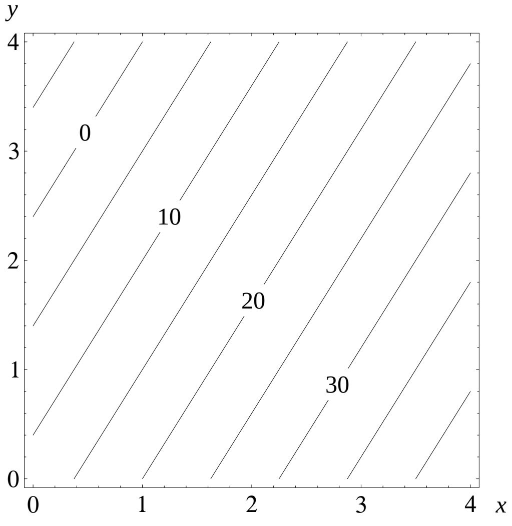
b) Where is the z-intercept of the graph?

c) If $(4, 1, c)$ is a point on the graph, what is the value of c?

d) Is the point $(2, 2, 4)$ on the graph? Explain your position.



contours of $L_1(x, y)$



contours of $L_2(x, y)$

15. The figure on the left above is the contour plot of a function $z = L_1(x, y)$.

a) What are the values of $L_1(1, 0)$, $L_1(2, 0)$, $L_1(3, 0)$, $L_1(0, 3)$, $L_1(2, 3)$, and $L_1(0, 0)$?

b) What are the partial rates $\Delta L_1/\Delta x$ and $\Delta L_1/\Delta y$?

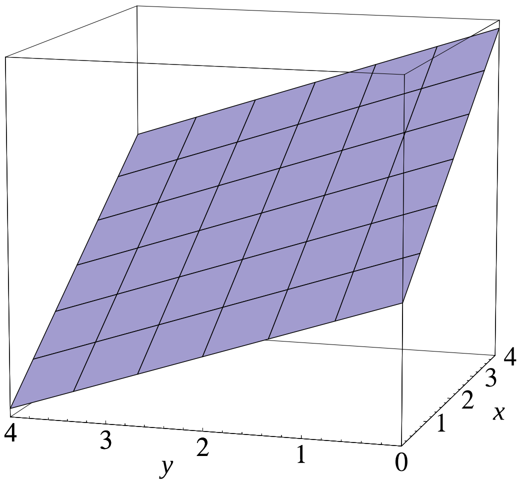
16. Continuation. Find the values of $L_1(3,2)$, $L_1(7,0)$, $L_1(7,7)$, $L_1(1.4,2.9)$, $L_1(-2,9)$, $L(-10,-100)$.

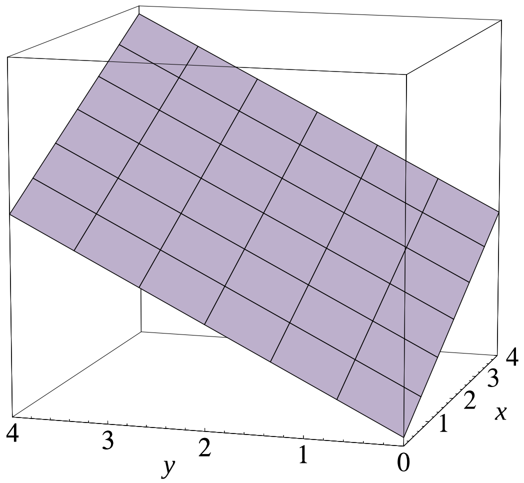
a) Find x so that $L_1(x,0)=0$. Find y so that $L_1(0,y)=0$.

17. Continuation. Write the intercept form of the formula for $L_1(x,y)$.

18. a) Find the partial rates $\Delta L_2/\Delta x$ and $\Delta L_2/\Delta y$ of the linear function L_2 whose contour plot is shown at the right on the previous page.

b) Obtain the intercept form of the formula for $L_2(x,y)$.





19. The figures above are the graphs of L_1 and L_2. Which is which? Explain your choice. (Note that both graphs are shown from the same viewpoint. The x-axis is on the right, and the y-axis is in the foreground. The z-axis has no scale on it.)

20. Determine the gradient vectors of L_1 and L_2.

a) Sketch the gradient vector of each function on the x,y-plane and on its own graph. Does the gradient point directly uphill in each case?

21. Suppose the gradient vector of the linear function $p=L(q,r)$ is grad $p=\nabla p=(5,-12)$. If $L(9,15)=17$, what is the value of $L(11,11)$?

22. a) What is the gradient vector of the function $w=2u+5v$?

b) At what point on the circle $u^2+v^2=1$ does w have its largest value? What is that value?

23. a) Write the formula of a linear function $z = L(x, y)$ whose gradient vector is grad $z = \nabla z = (-3, 4)$

b) Using your formula for L, calculate the total change in L when $\Delta x = 2$, $\Delta y = 1$.

24. a) Continuation; in particular, continue to use your formula for $z = L(x, y)$. What is the value of $L(0, 0)$?

b) What is the maximum value of z on the circle $x^2 + y^2 = 1$? At what point (x, y) does z achieve that value?

c) Determine the difference between the maximum and minimum values of L on the circle $x^2 + y^2 = 1$.

[Answer: The difference is 10, independent of the formula you use.]

25. a) What value does $z = 7x + 3y + 31$ have when $x = 5$ and $y = 2$?

b) If x increases by 2, how must y change so that the value of z doesn't change. (The change in y needed to keep z fixed when x changes is called the **trade-off**. See also chapter 3, page 174.)

The concept of a *trade-off*

c) What is the trade-off in y when x increases by α?

d) What is the trade-off in x when y increases by β?

26. a) Suppose $z = L(x, y)$ is a linear function for which $\Delta z/\Delta x = 5$ and $\Delta z/\Delta y = -2$. What is the trade-off in y when x increases by 50?

b) What is the trade-off in x when y increases by 1?

27. Suppose $z = L(x, y)$ is a linear function and suppose the trade-off in y when x increases by 1 is -4.

a) What is the trade-off in y when x is *decreased* by 3?

b) What is the trade-off in x when y is increased by 10? [Note that x and y are reversed here, in comparison to the earlier parts of this question.]

28. Suppose $z = L(x, y)$ is a linear function for which we know

$$L(3, 7) = -2, \qquad \frac{\Delta z}{\Delta x} = 2.$$

Suppose also that the trade-off in y when x increases by 10 is -4.

a) What is the value of $L(7, 7)$?

b) If $L(7, \beta) = -2$, what is the value of β?

c) What is the value of $\Delta z/\Delta y$?

d) Write the formula for $L(x, y)$ in intercept form.

29. a) Suppose the graph of the linear function $z = L(x, y)$ has a slope of -1.5 in the x-direction and -2.4 in the y-direction. What is the trade-off between x and y? That is, how much should y change when x is increased by the amount α?

b) Suppose the partial slopes become $+1.5$ and $+2.4$ in the x- and y-directions, respectively. How does that affect the trade-off? Explain.

30. a) Sketch in the x, y-plane the set of points where $z = 2x + 3y + 7$ has the value 34.

b) If $x = 10$ then what value must y have so that the point (x, y) is on the set in part (a)?

c) If x increases from 10 to 14, how must y change so that the point (x, y) stays on the set in part (a)? In other words, what is the trade-off?

The set in the last question is called a *trade-off line*. Do you see that it is just a contour line by another name?

31. a) Write, in intercept form, the formula for the linear function

$$w - 4 = 3(x - 2) - 7(y + 1) - 2(z - 5).$$

b) What is the gradient vector of the linear function in part (a)?

32. Suppose the gradient of the linear function $w = L(x, y, z)$ is $\nabla w = (1, -1, 4)$. If $L(3, 0, 5) = 10$, what is the value of $L(1, 2, 3,)$?

33. Describe the level sets of the function $w = f(x, y, z) = x + y + z$.

The microscope equation

34. Find the microscope equation for the function $f(x, y) = 3x^2 + 4y^2$ at the point $(x, y) = (2, -1)$.

35. a) Continuation. Use the microscope equation to estimate the values of $f(1.93, -1.05)$ and $f(2.07, -.99)$

b) Calculate the *exact* values of the quantities in part (a), and compare those values with the estimates. In particular, indicate how many digits of accuracy the estimates have.

36. Find the microscope equation for the function $f(x, y, z) = x^2 y \sin z$ at the point $(x, y, z) = (1, 1, \pi)$.

37. Suppose $f(87, 453) = 1254$ and

$$\frac{\partial f}{\partial x}(87, 453) = -3.4, \qquad \frac{\partial f}{\partial y}(87, 453) = 4.2.$$

Estimate the following values: $f(90, 453)$, $f(87, 450)$, $f(90, 450)$, and $f(100, 500)$. Explain how you got your estimates.

38. a) Continuation. Find an estimate for y to solve the equation $f(87, y) = 1250$.

b) Find an estimate for x to solve $f(x, 450) = 1275$.

39. Continuation: a trade-off. Go back to the starting values $x = 87$, $y = 453$, and $f(87, 453) = 1254$. If x increases from 87 to 88, how should y change to keep the value of the function fixed at 1254?

40. a) Suppose $Q(27.3, 31.9) = 15.7$ and $Q(27.9, 31.9) = 15.2$. Estimate the value of $\partial Q/\partial x(27.6, 31.9)$.

b) Estimate the value of $Q(27, 31.9)$.

41. Suppose $S(105, 93) = 10$, $S(110, 93) = 10.7$, $S(105, 95) = 9.3$. Estimate the value of $S(100, 100)$. Explain how you made your estimate.

42. Let P be the point $(x, y, z) = (173, -29, 553)$. Suppose $f(P) = 48$ and

$$\frac{\partial f}{\partial x}(P) = 7, \qquad \frac{\partial f}{\partial y}(P) = -2, \qquad \frac{\partial f}{\partial z}(P) = 5.$$

Estimate the value of $f(175, -30, 550)$, and explain what you did.

43. a) What is the equation of the tangent plane to the graph of $z = xy$ at the point $(x, y) = (2, -3)$?

b) Which has a higher z-intercept: the graph or the tangent plane?

44. a) Suppose the function $H(x, y)$ has the microscope equation

$$\Delta H \approx 2.53\,\Delta x - 1.19\,\Delta y$$

at the point $(x, y) = (35, 26)$. Sketch the gradient vector ∇H at that point.

b) Pick the point exactly one unit away from $(35, 26)$ at which you estimate H has the largest possible value.

[Answer: At $(35.324, 25.848)$, H is about 2.796 units larger than it is at $(35, 26)$.]

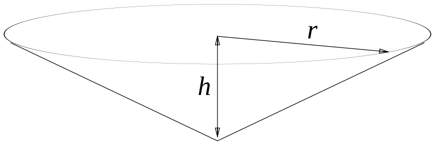
45. Write the microscope equation for the function $V(x, y) = x^2y$ at the point $(x, y) = (25, 10)$.

Refer to the discussion of error propagation in chapter 3.4

46. a) Continuation. Suppose a cardboard carton has a square base that is 25 inches on a side and a height of 10 inches. If there is an error of Δx inches in measuring the base and an error of Δy inches in measuring the height, how much error will there be in the calculated volume?

b) Why is this a continuation of the previous question?

47. Continuation. Which causes a larger percentage error in the calculated volume: a 1% error in the measurement of the length of the base, or a 1% error in the measurement of the height?



48. A large basin in the shape of a cone is to be used as a water reservoir. If the radius r is 186 meters and the depth h is 31 meters, how much water can the basin hold, in cubic meters?

49. a) Continuation. If there were a 3% error in the measurement of the radius, how much error would that lead to in the calculation of the capacity of the basin?

b) If there were a 5% error in the measurement of the depth, how much error would there in the calculated capacity of the basin?

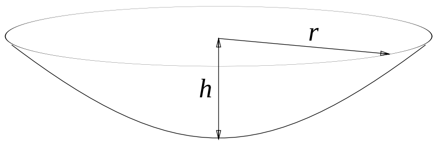
c) If *both* errors are present in the measurements, what is the total error in the calculated capacity of the basin?

50. Continuation. Suppose the measured radius of the basin ($r = 186$ meters) is assumed to be accurate to within 2%. The depth has been measured at 31 meters. Is it possible to make that measurement so accurate that the

calculated capacity is known to within 5%? How accurate does the depth measurement have to be?

51. Continuation. Suppose the accuracy of the radius measurement can only be guaranteed to be 3%. Is it still possible to measure the depth accurately enough to guarantee that the calculated capacity is accurate to within 5%? Explain.

52. Let's give the basin the more realistic shape of a parabolic bowl. If the radius r is still 186 meters and the depth h is still 31 meters, what is the capacity of the basin now? Is it larger or smaller than the conical basin of the same dimensions?



53. Continuation. Determine the error in the calculated capacity of the bowl if there were a 3% error in the measurement of the radius and, at the same time, a 5% error in the measurement of the depth.

54. Continuation. Is it possible for the calculated capacity to be accurate to within 5% when the measured radius is accurate to within 2%? How accurate does the measurement of the depth have to be to achieve this?

55. The energy of a certain pendulum whose position is x and velocity is v can be given by the formula

$$E(x,v) = 1 - \cos x + \tfrac{1}{2}v^2.$$

Suppose the position of the pendulum is known to be $x = \pi/2$ with a possible error of 5% and its velocity is $v = 2$ with possible error of 10%. What is the calculated value of the energy, and how accurately is that value known?

The conservation of energy leads to a trade-off

56. a) A frictionless pendulum conserves energy: as the pendulum moves, the value of E does not change. Suppose $x = \pi/2$ and $v = 2$, as in the previous exercise. When x decreases by $\pi/180$ (this is 1 degree), does v increase or decrease to conserve energy?

b) Approximately how much does v change when x decreases by $\pi/180$?

Linear approximations

57. a) Write the linear approximation to $f(x,y) = \sin x \cos y$ at the point $(x,y) = (0, \pi/2)$.

b) Write the equation of the tangent plane to the graph of $f(x, y)$ at the point $(x, y) = (0, \pi/2)$.

c) Where does the tangent plane in part (b) meet the x, y-plane?

58. Suppose $w(3, 4) = 2$ while

$$\frac{\partial w}{\partial x}(3,4) = -1, \qquad \frac{\partial w}{\partial y}(3,4) = 3.$$

Write the equation of the tangent plane of w at the point $(3, 4)$.

59. a) Write the equation of the tangent plane to the graph of the function $\varphi(x, y) = 3x^2 + 7xy - 2y^2 - 5x + 3y$ at an arbitrary point $(x, y) = (a, b)$.

b) At what point (a, b) is the tangent plane horizontal?

c) Magnify the graph of φ at the point you found in part (b) until it looks flat. Is it also horizontal?

The next few exercises concern the Lotka–Volterra differential equations and their linear approximations at an equilibrium point. Specifically, consider the *bounded growth* system from chapter 4.1 (pages 187–189):

$$\begin{aligned} R' &= .1\,R - .00001\,R^2 - .005\,RF, \\ F' &= .00004\,RF - .04\,F. \end{aligned}$$

60. Confirm that the system has an equilibrium point at $(R, F) = (1000, 18)$. Then obtain the phase portrait of the system near that point. What kind of equilibrium is there at $(1000, 18)$?

61. Obtain the linear approximations of the functions

$$\begin{aligned} g_1(R,F) &= .1\,R - .00001\,R^2 - .005\,RF, \\ g_2(R,F) &= .00004\,RF - .04\,F, \end{aligned}$$

at the point $(r, F) = (1000, 18)$. Call them $\ell_1(R, F)$ and $\ell_2(R, F)$, respectively.

62. Obtain the phase portrait of the **linear** dynamical system

$$\begin{aligned} R' &= \ell_1(R,F), \\ F' &= \ell_2(R,F). \end{aligned}$$

a) Does this system have an equilibrium at $(1000, 18)$? Compare this phase portrait to the phase portrait of the original *non*-linear system.

The gradient field

63. a) Make a sketch of the gradient vector field of $f(x, y) = x^2 - y^2$ on the domain $-2 \leq x \leq 2$, $-2 \leq y \leq 2$.

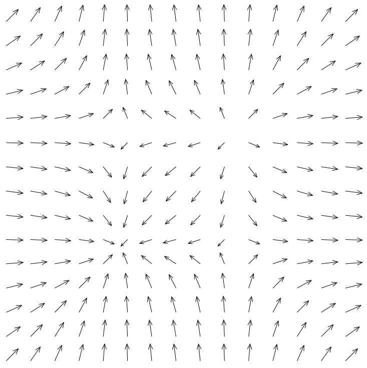
b) Mark on your sketch where the gradient field indicates the maximum and the minimum values of f are to be found in the domain.

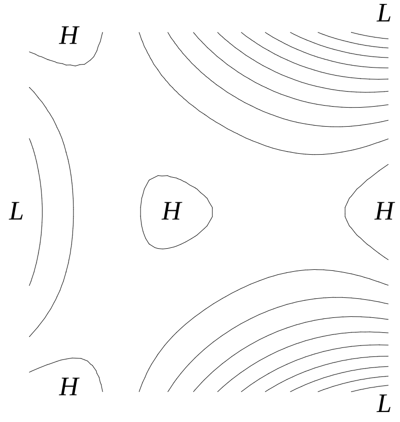
64. Repeat the previous exercise, using $f(x, y) = 2x + 4x^2 - x^4 - y^2$ and the domain $-2 \leq x \leq 2$, $-4 \leq y \leq 4$. (See exercises 5 and 6 in the previous section, page 528.)

65. Continuation. Add to your sketch in the previous exercise a contour plot of the function $f(x, y) = 2x + 4x^2 - x^4 - y^2$ and confirm that each gradient vector is perpend1 of 82icular to the contour that passes through its base. (Note: most vectors have no contour passing through their bases, so you have to infer the shape and position of such a contour from the contours that *are* drawn.)

66. Draw a plausible set of contour lines for the function whose gradient vector field is plotted on the left below.

67. Draw a plausible gradient vector field for the function whose contour plot is shown on the right below. H marks a local maximum and L a local minimum.





9.3 Optimization

The contexts for optimization

Optimization is the process of making the best choice from a range of possibilities. (“Optimum” is the Latin word for *best.*) We are all familiar with optimization in the economic arena: managers of an enterprise typically seek to to maximize profit or minimize cost by making conscious choices. It is perhaps more surprising to learn that we sometimes use the same language to describe physical processes. For instance, the atoms in a molecule are arranged so that their total energy is minimized. A light ray travels from one point to another along the path that takes the least time. Of course atoms and photons don’t make conscious choices. Nevertheless, the imagery of optimization is so vivid and useful that we try to invoke it whenever we can.

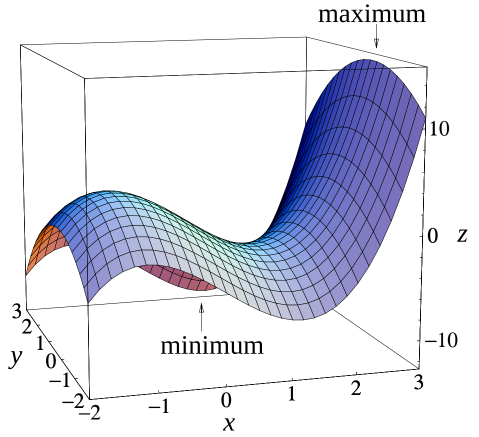
Constrained optimization

Usually, there is a restriction—called a **constraint**—on the choices that can be made to achieve to best possible outcome. For instance, consider a factory that makes tennis rackets. We can expect that the factory managers are instructed to minimize cost *while producing a given number of rackets.* This is their constraint. Production cost is a function of many quantities that the managers can control—the number of workers, the wage scale, and the cost of the raw materials are just a few. When the managers choose values for these quantities that minimize the cost function, they must be sure those values will also satisfy the constraint.

Mathematical optimization

In mathematical terms, optimization is the process of finding the minimum or maximum value of a function. The presence of constraints complicates this task, as you shall see.

Visual Inspection



The maximum value of a function is the highest point on its graph; the minimum value is the lowest. Shown at the right is the graph of $z = x^3 - 4x - y^2$ on the domain

$$-2 \leq x \leq 3 \qquad -2 \leq y \leq 3.$$

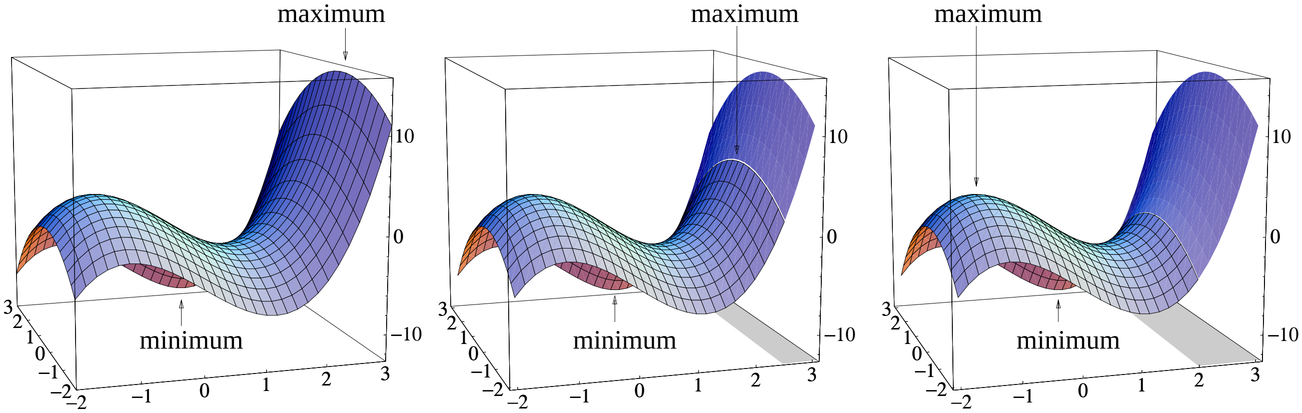
The maximum occurs where $(x, y) = (3, 0)$, and it has the value $z = 15$. The minimum is $z \approx -12.1$, when $(x, y) \approx (1.2, 3)$. Confirm this yourself by inspecting the graph and then calculating z.

Extremes, and local extremes

We'll use the term **extreme** to refer to either a maximum or a minimum. In the example we see several *local* extremes. These are points where the value of the function is larger or smaller than it is at any *nearby* point. There are local minima at both left-hand corners of the graph, at $(-2,-2)$ and $(-2,3)$. There is another along the front edge, near $(1.2,-2)$. There is a local maximum in the interior, near $(-1.2,0)$. To decide which local minimum is the *true* minimum, we must simply look. It is the same for local maxima.

The domain is a constraint

In our example, the domain of definition of the function acts as a *constraint*. If we change the domain, the positions of the extreme points can change. For example, suppose we change the position of the right-hand border in stages: first, $x \le 3$; second, $x \le 2.5$; third, $x \le 2$. (Since the graph itself doesn't change, we use a grid to show the part of the graph that satisfies the constraint in each case below.) At the start, the maximum is on the boundary. When we first move the boundary to the left (from $x = 3$ to $x = 2.5$), the maximum just moves with it. However, when we move the boundary farther (from $x = 2.5$ to $x = 2$), the maximum jumps from the boundary to an interior point (near $(-1.2,0)$). During these changes the minimum is not affected. It stays at the same place.



$x \le 3$ $x \le 2.5$ $x \le 2$

The maximum moves —it even jumps

Catastrophes

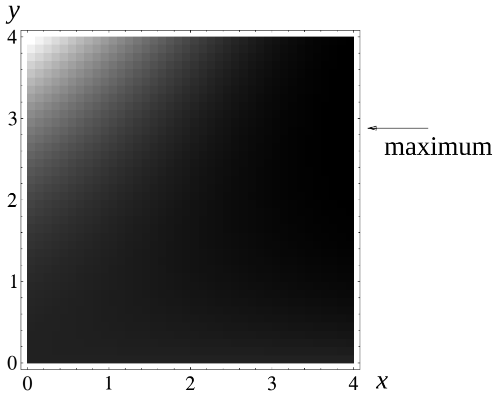
The sudden jump in the maximum is called a **catastrophe**. The figures explain what happens. We impose a constraint $x \le a$, and then we reduce the value of the parameter a. At first, the maximum is at the boundary point $(a,0)$. The position of this point changes smoothly with a, causing a gradual drop in the value of the maximum. Eventually, the maximum reaches the same value as the interior local maximum. (This happens when $a = 4/\sqrt{3}$;

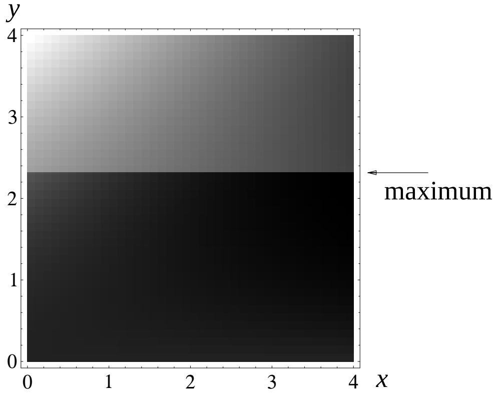
see the exercises.) If a continues to decrease, the local maximum at $(a, 0)$ then has a lower value than the local maximum at the interior point, so the true maximum jumps to the interior.

There are many variations on this pattern. Whenever a function depends on a parameter the positions of its extremes do, too. There are many ways for the position of an extreme to jump suddenly *while the parameter is changing gradually.* Any such jump is called a catastrophe.

Catastrophes make the task of optimization more interesting. If the maximum of a certain function gives the optimal solution to a problem, and that maximum jumps to a new position, then the optimal solution changes radically. For example, suppose the problem is to determine the minimum-energy configuration of atoms in a molecule. When the minimum jumps catastrophically, there is a new configuration of the same atoms, producing an *isomeric* form of the molecule.

The quest for an optimum does not have to involve mathematical tools. A good example is John Stuart Mill's philosophy of "the greatest good for the greatest number". In politics a catastrophe is called a revolution. Even scientific research pursues an optimum in raising the question: "What is the best way to explain certain phenomena?" The consensus in the scientific community can change catastrophically, in what is called a *paradigm shift* or an intellectual revolution. The geological theory of plate tectonics is a familiar example. Though proposed in the 1920s, it was dismissed until the 1960s, when it was suddenly and overwhelmingly accepted.





Density plots

We can also solve optimization problems by inspecting density plots. Suppose z is the yield from a process that is controlled by two inputs x and y, and

$$z = 3xy - 2y^2.$$

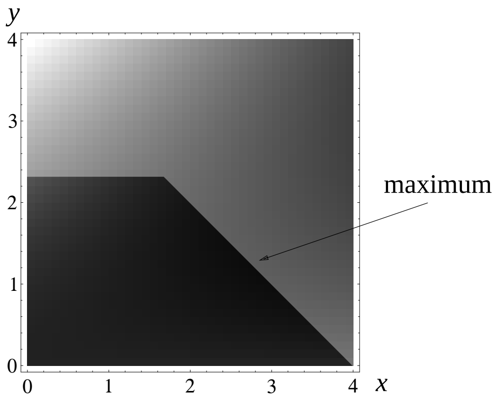
Initially we take $0 \leq x \leq 4$, $0 \leq y \leq 4$. The maximum yield is at the darkest spot in the upper density plot. It occurs on the right boundary, near $y = 3$.

The position of the maximum is subject to change if we have to impose further constraints. For instance, suppose the resource y is more limited than we first assumed, requiring us to set $y \leq 2.3$. With this added constraint we see that the maximum shifts to the corner $(x, y) = (4, 2.3)$. (The points shown in a lighter gray are the ones that have been removed from consideration by the new constraint.)

Besides limits on individual resources, we are often faced with a limit on *total* resources. In our case, let's suppose the limit has the form

$$x + y \leq 4.$$

That means all points above the line $x + y = 4$ must be removed from consideration. (They are shown in lighter gray.) This new constraint causes the maximum to shift yet again. It now appears near the point $(x, y) = (3, 1)$.

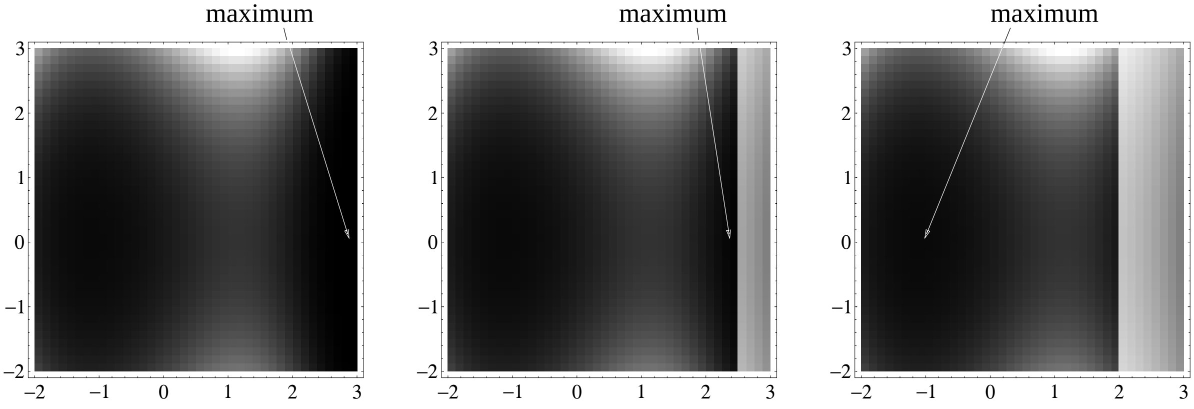


Constraints reduce optimality

As we add constraints that force the maximum to move, the density at each new maximum is less than it was at the previous one. Thus, the value of the maximum itself decreases. In other words, each added constraint makes the optimal solution slightly "less optimal" than it had been. This is only to be expected.

Extremes on the interior of a density plot

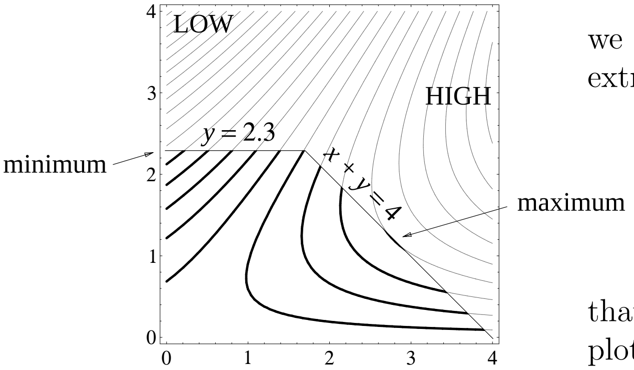
Of course the extremes may appear in the interior of the domain as well as on the boundary—and density plots can show this. Here are the density plots that correspond to the graphs of $z = x^3 - 4x - y^2$ that we saw on page 565. The maximum jumps to the interior when the value of a in the constraint $x \leq a$ drops below $a = 4/\sqrt{3}$. In all three plots the minimum appears as the bright spot at the top of the rectangle near where $x = 1$. (The exact location is $(x, y) = (2/\sqrt{3}, 3)$.) We can also see a local minimum at the bottom of the rectangle near $x = 1$. From the graph we know there are two more at the left corners of the rectangle, but they are harder to notice in the density plots.



$x \leq 3$ $x \leq 2.5$ $x \leq 2$

Contour plots

A density plot is useful for showing the general location of the highs and lows, but we can get a more precise picture by switching to a contour plot. Let's look at the contour plots of the same problems we just analyzed using density plots.



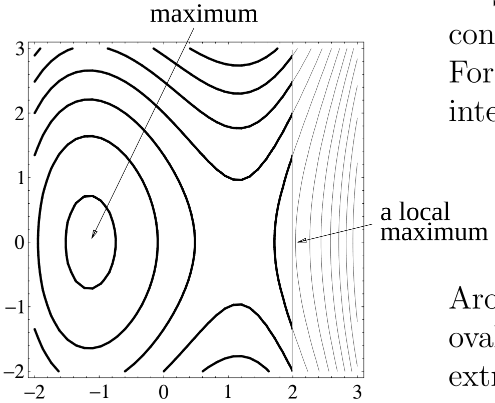
We'll start with the function $z = 3xy - 2y^2$ we first considered on page 566 and search for its extremes subject to the constraints

$$0 \le x,$$
$$0 \le y \le 2.3,$$
$$x + y \le 4,$$

that we introduced earlier. From the density plot on page 566 it was obvious that the values of z increase steadily from the upper left corner of the large square to the upper part of the right side. In the contour plot, though, we need some sort of labels to show us where the low and high values of z are to be found.

An extreme can occur where a contour is tangent to the boundary

To locate the extremes within the constrained region, we need to find contours that carry the lowest and the highest values of z. We can do this quite precisely. The contour that just passes through the upper left corner—and meets the region at that point alone—carries the lowest value of z. If you study the plot you can see that the contour carrying the *highest* value of z also touches the boundary at just a single point. It is the contour that is tangent to the line $x + y = 4$, and elsewhere lies outside the constrained region.



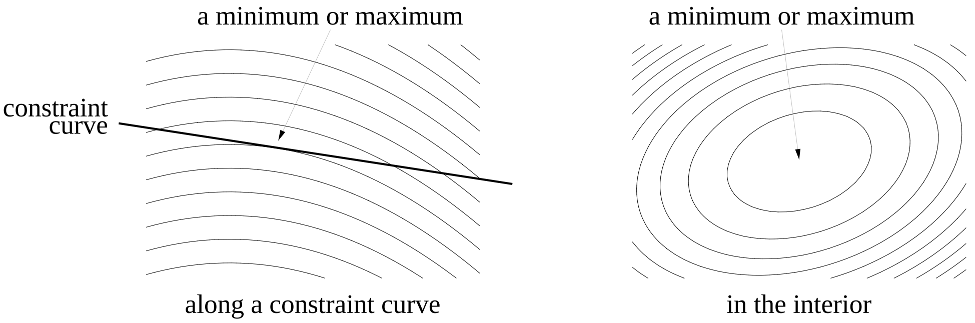
Suppose the extreme is in the *interior* of the constrained region, rather than on the boundary. For example, the function $z = x^3 - 4x - y^2$ has an interior maximum when

$$-2 \le x \le 2$$
$$-2 \le y \le 3.$$

Around the maximum there is a nest of concentric ovals. This pattern is characteristic for an interior extreme. Notice the local maximum where a contour is tangent to the boundary line $x = 2$.

Contour patterns at an extreme

These examples demonstrate that there are characteristic patterns that contour lines make near an extreme point. One pattern appears along a constraint curve; another pattern appears at interior points of a domain. We first met the pattern that appears at an interior extreme point when we discussed the 'standard' minimum $(x^2 + y^2)$ and maximum $(-x^2 - y^2)$ in section 1 (pages 519–522). **Caution**: the patterns you see here are "typical", but they do not *guarantee* the presence of an extreme point. The exercises give you a chance to explore some of the subtleties.

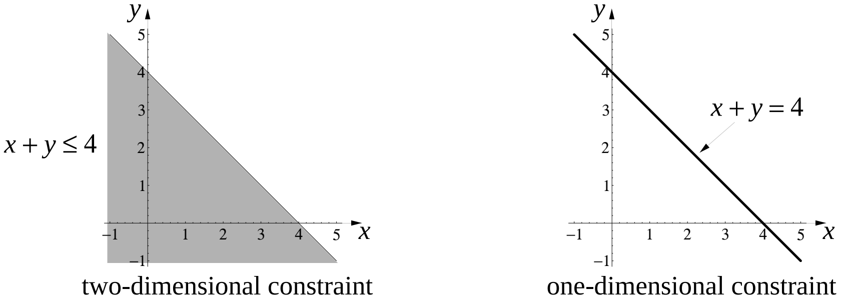


Patterns of contour lines at an extreme point

Dimension-reducing Constraints

How a constraint can reduce the dimension of a problem

Constraints appear frequently in optimization problems. Thus far, however, the constraints we considered have been described by *inequalities*, like $x+y \leq 4$. Initially, x and y give us the coordinates of a point in a two-dimensional plane. The effect of the constraint $x+y \leq 4$ is to restrict the points (x, y) to just a part of that plane—but it is still a *two-dimensional* part. Sometimes, though, the constraint is given by an *equality*, like $x + y = 4$. In that case, the points (x, y) are restricted to lie on a line—which is a one-dimensional set in the plane. The second constraint therefore reduces the dimension of the problem.



A standard form

There is a standard form for any constraint that reduces a two-dimensional optimization problem to a one-dimensional problem. The form is this:

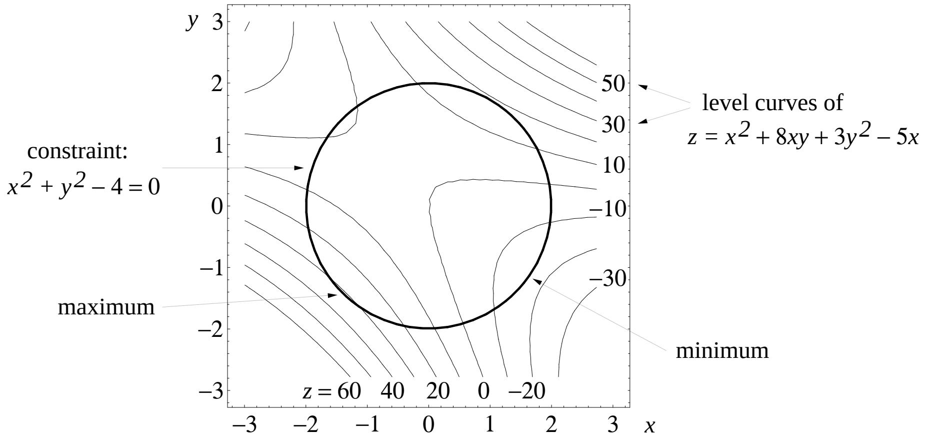
$$\text{constraint:} \quad g(x, y) = 0.$$

For instance, the constraint $x + y = 4$ can be written this way by setting $g(x, y) = x + y - 4$. We'll look at another example in a moment.

A comment on dimension

First, though, notice that $g(x, y) = 0$ is one of the contour lines of the function $g(x, y)$, namely, the contour at level zero. Since the contours of g are curves, we often call $g(x, y) = 0$ a **constraint curve**. This curve is *one*-dimensional, even though it might twist and turn in a two-dimensional plane. We say the curve is one-dimensional because it looks like a straight line under a microscope. Likewise, a curved surface is two-dimensional because it looks like a flat plane under a microscope.

Example. Find the extreme values of $f(x, y) = x^2 + 8xy + 3y^2 - 5x$ subject to the constraint $g(x, y) = x^2 + y^2 - 4 = 0$. The constraint curve is a circle of radius 2, and the level curves of $z = f(x, y)$ form a set of hyperbolas. We need to find the highest and the lowest z-levels on the constraint curve.



Finding the highest and lowest levels on the constraint curve

The z-levels are labelled around the right and the bottom edges of the figure. It is in the third quadrant, near the point $(-1.4, -1.5)$, that the constraint curve meets the highest z-level. At that point z is slightly more than 30. The constraint curve is evidently tangent to the contour there. The lowest z-level that the constraint curve meets is about $z = -16$, near the point $(1.6, -1.2)$. Picture in your mind how the contours between $z = -10$ and $z = -20$ fit together. Can you see that the constraint curve is tangent to contour at the minimum?

Locate points on the constraint circle with a single variable t

Example, continued. There is still more to say about the way the constraint $x^2 + y^2 - 4 = 0$ reduces the dimension of our problem. First of all, we can describe the coordinates of any point on the constraint circle by using the circular functions:

$$x = 2\cos t, \qquad y = 2\sin t.$$

We need the factor 2 because the circle has radius 2. These equations mean that x and y are now functions of a *single* variable, t. Next, consider the function

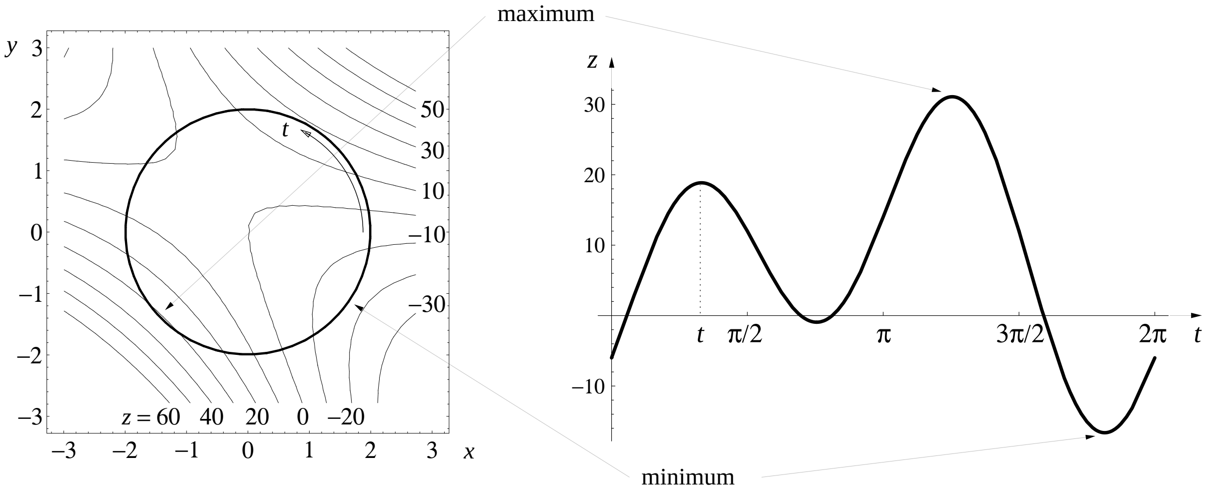
$$z = f(x, y) = x^2 + 8xy + 3y^2 - 5x$$

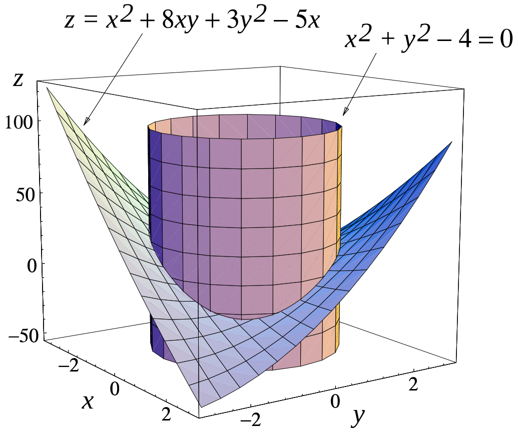
z becomes a function of t

that we seek to optimize *subject to the constraint.* But the constraint makes x and y functions of t. Thus, *when we take the constraint into account*, z itself becomes a function of t:

$$\begin{aligned} z &= f(2\cos t, 2\sin t) \\ &= (2\cos t)^2 + 8\,(2\cos t)(2\sin t) + 3(2\sin t)^2 - 5(2\cos t). \end{aligned}$$

The graph of *this* function is thus just an ordinary curve in the t, z-plane. It is shown on the right, below. A value of t determines a point on the circle, as shown in the contour plot on the left. (We have taken $t \approx \pi/3$.) The value of z at that point then determines the height of the graph on the right. As you can see, our chosen value of t puts z near a local maximum.

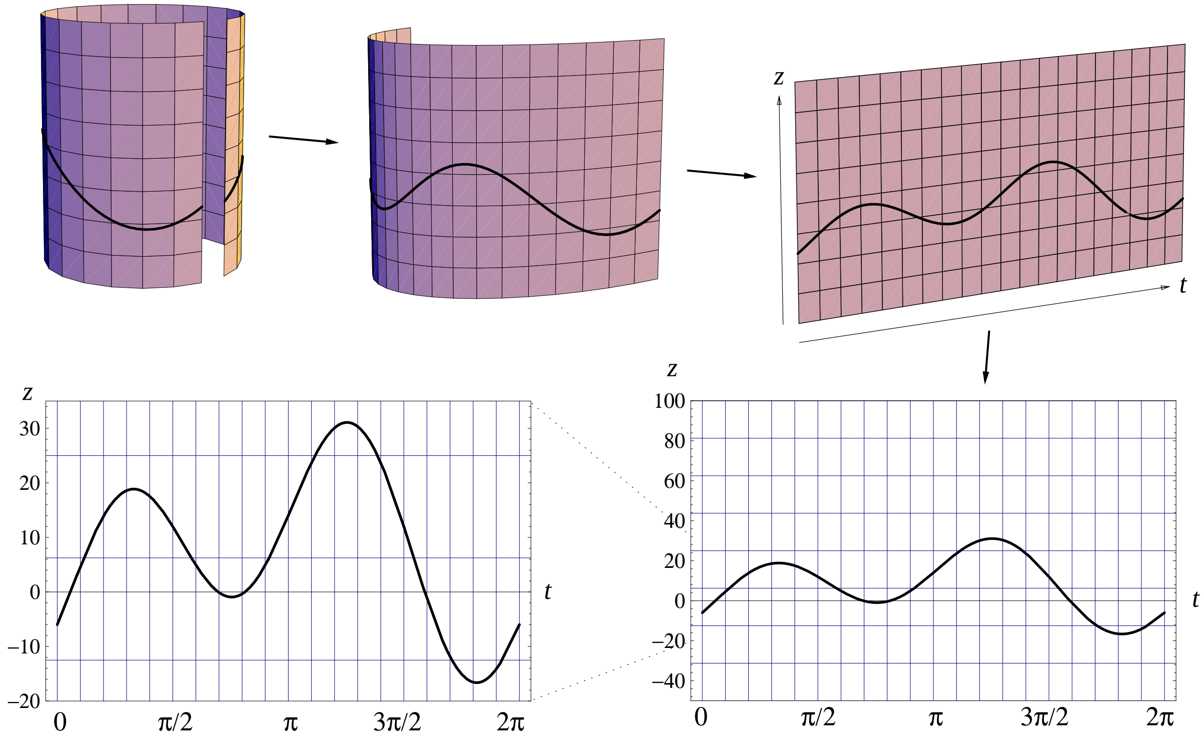




Example, continued. Let's look at the graph of $z = x^2 + 8xy + 3y^2 - 5x$. The constraint tells us that we should look *only* at the points on the graph that lie above the circle $x^2 + y^2 - 4 = 0$. These are the points where the graph intersects the cylinder that you see at the right. The intersection is a curve that goes up and down around the cylinder. At some point on the curve z has a maximum, and at some other point it has a minimum. (In fact, both the maximum and the minimum are visible in this view.)

Unwrap the cylinder

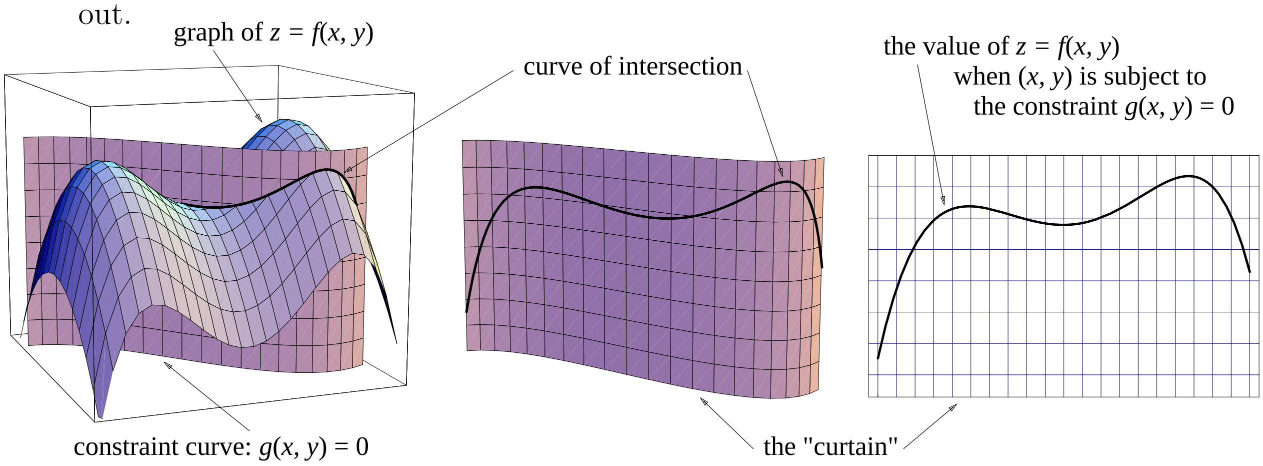
Since the intersection curve lies on the cylinder, we can get a better view of the curve if we slit open the cylinder and unwrap it. Follow the sequence clockwise from the upper left. We can use coordinates to describe the curve on the flattened cylinder. The t variable takes us around the cylinder, so it becomes the horizontal coordinate. The z variable measures vertical height. The z-range in the figure above is larger than we need: it goes from -50 to $+100$. In the bottom row on the left we have rescaled the z-axis so it runs from -20 to $+35$. Compare this graph with the one on the opposite page.



A general view of the dimension-reducing effect of a constraint

The example shows us how a constraint works to reduce the dimension of a problem in general. Suppose we want to maximize the value of the function $z = f(x, y)$, subject to the constraint $g(x, y) = 0$. Then:

- *Without* the constraint, we would look for the highest point on a *two*-dimensional surface in three-dimensional space.
- *With* the constraint, we would restrict our search to the vertical "curtain" that lies above the constraint curve. The graph intersects this curtain in a curve, so we end up looking for the highest point on a *one*-dimensional curve in a two-dimensional plane. We can think of this plane as the curtain after it has been unwrapped and straightened out.

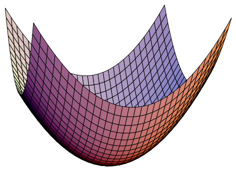

graph of z = f(x, y)
curve of intersection
the value of z = f(x, y) when (x, y) is subject to the constraint g(x, y) = 0
constraint curve: g(x, y) = 0
the "curtain"

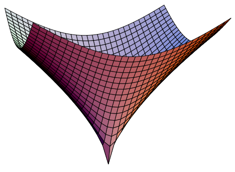
This is just a picture of the relation between the function and the constraint. We may still have to determine *analytically* the form that the function $f(x, y)$ takes when we impose the constraint $g(x, y) = 0$. You can find a number of possibilities in the exercises.

Extremes and Critical Points

Suppose that a function $f(x, y)$ has a maximum or a minimum at an *interior* point (a, b). Suppose also that the function is *locally linear* at (a, b), so we have a "bowl" rather than a "spike" or some other irregularity in the graph. Then we must have

$$\operatorname{grad} f(a, b) = \left(\frac{\partial f}{\partial x}(a, b), \frac{\partial f}{\partial y}(a, b) \right) = (0, 0).$$


bowl:
locally linear


spike:
not locally linear

A proof

Here is why. The gradient vector grad$f(a,b)$ tells us how the graph of $z = f(x,y)$ is tilted at the point (a,b). (We discuss the geometric meaning of the gradient on page 551.) At a maximum or a minimum, though, the graph is *not* tilted; it must be flat. Therefore, the gradient must be the zero vector.

Moving to an arbitrary number of input variables

The ideas here carry over to functions that have any number of input variables. First, we give a name to a point where the gradient is zero.

> **Definition**. A **critical point** of a locally linear function is one where the gradient vector is zero. Equivalently, all the first partial derivatives of the function are zero.

The observation we just made can now be restated as a theorem that connects extreme points and critical points.

> **Theorem**. If a locally linear function has a maximum or a minimum at an interior point of its domain, then that point must be a critical point.

A statement and its converse

The direction of the implication in this theorem is important. Here is the theorem, written in a very abbreviated form:

$$\textit{statement}: \qquad \text{extreme} \Longrightarrow \text{critical}.$$

When we reverse the direction of the implication, we get a new statement, abbreviated the same way:

$$\textit{converse}: \qquad \text{critical} \Longrightarrow \text{extreme}.$$

The converse of *this* theorem is not true

The converse says that a critical point must be an extreme point. But that is just not true. For example, an ordinary saddle point (a minimax) is a critical point, but it is not a minimum or a maximum.

Searching critical points for extremes

The theorem and the observation about its converse are both important in the *optimization process*—that is, the search for extremes. Together they offer us the following guidance:

- Search for the extremes of a function among its critical points.
- A critical point may be neither a maximum nor a minimum.

To see how we can find extremes by searching among the critical points of a function, we'll do a few examples. All the examples use the same basic idea. However, as the details get more complicated we bring in more powerful techniques.

Example 1. We'll start with the function

$$z = f(x, y) = x^3 - 4x - y^2$$

we have frequently used as a test case in this chapter.

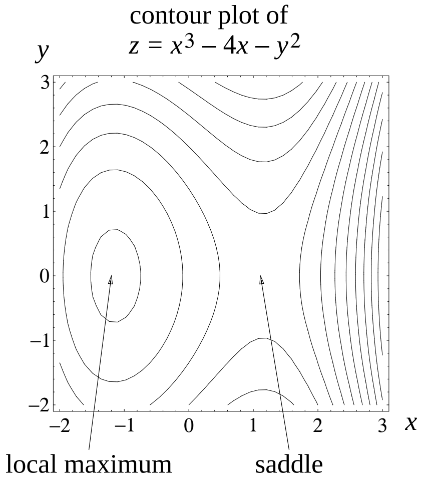
The critical points of f are the points that *simultaneously* satisfy the two equations

$$\frac{\partial f}{\partial x} = 3x^2 - 4 = 0,$$
$$\frac{\partial f}{\partial y} = -2y = 0.$$

It is clear that $y = 0$ and $x = \pm\sqrt{4/3} \approx \pm 1.1547$. The two critical points are therefore

$$\left(+\sqrt{4/3}, 0\right) \qquad \text{and} \qquad \left(-\sqrt{4/3}, 0\right).$$

If you check the contour plot you can see that $(-\sqrt{4/3}, 0)$ is a local maximum and $(+\sqrt{4/3}, 0)$ is a saddle point.



Example 2. Here is a somewhat more complicated function:

$$g(x, y) = 2x^2y - y^2 - 4x^2 + 3y.$$

The critical points are the solutions of the equations

$$\frac{\partial g}{\partial x} = 4xy - 8x = 0, \qquad \frac{\partial g}{\partial y} = 2x^2 - 2y + 3 = 0.$$

Algebraic methods will still work, even though both variables appear in both equations. For example, we can rewrite $\partial g/\partial y = 0$ as

$$2y = 2x^2 + 3 \qquad \text{or} \qquad y = x^2 + \tfrac{3}{2}.$$

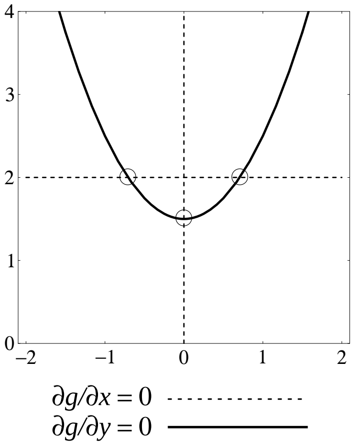
We can then substitute this expression for y into $\partial g/\partial x = 0$ and get

$$4x\left(x^2 + \tfrac{3}{2}\right) - 8x = 4x\left(x^2 + \tfrac{3}{2} - 2\right) = 4x\left(x^2 - \tfrac{1}{2}\right) = 0.$$

This implies $x = 0$ or $x = \pm\sqrt{1/2}$. For each x we can then find the corresponding y by from the equation $y = x^2 + \frac{3}{2}$.

Let's work through this a second time using a geometric approach. The equations $\partial g/\partial x = 0$ and $\partial g/\partial y = 0$ both define curves in the x, y-plane. The curve $\partial g/\partial y = 0$ is a parabola: $y = x^2 + \frac{3}{2}$. The equation $\partial g/\partial x = 0$ factors as

$$4x(y - 2) = 0 \qquad \text{or} \qquad x(y - 2) = 0.$$

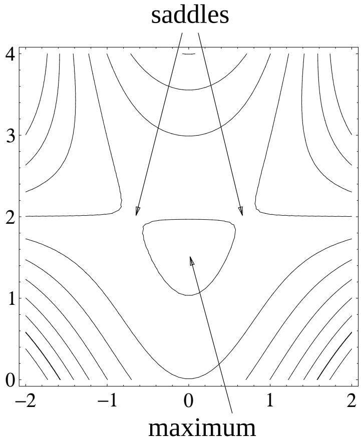


Now a product equals 0 precisely when one of its factors equals 0, so $\partial g/\partial x = 0$ implies that *either* $x = 0$ *or* $y - 2 = 0$. In other words, the "curve" $\partial g/\partial x = 0$ consists of two lines:

$$\begin{aligned} x &= 0, \quad \text{a vertical line,} \\ y &= 2, \quad \text{a horizontal line.} \end{aligned}$$

The two curves are shown in the figure at the right. They intersect in three points. (The place where the horizontal and vertical lines cross is *not* one of the intersection points.)

One of the immediate benefits of the geometric approach is to make it clear that the y-coordinate of a critical point is either 2 or $\frac{3}{2}$. The critical points are therefore



$$\left(-\sqrt{1/2}, 2\right), \quad \left(0, \tfrac{3}{2}\right), \quad \left(+\sqrt{1/2}, 2\right).$$

A glance at the contour plot of g makes it clear that the first and third of these are saddle points. The middle point is a local maximum. There are several ways to determine this. One is to look at the graph of g. Another is to look at a vertical slice of the graph through the line $x = 0$. Then $z = g(0, y) = -y^2 + 3y$. Here z has a maximum when $y = \frac{3}{2}$.

We first identified the local maximum of the function $z = x^3 - 4x - y^2$ on page 565. At the time, though, we could only estimate its position by eye. Now, however, we can specify its location exactly, because we have analytical tools for finding critical points. As the next example shows, these tools are useful even when we can't carry out the algebraic manipulations.

Example 3. Here is a function whose critical points we *can't* find using just algebraic computations:

$$z = h(x, y) = x^2y^2 - x^4 - y^4 - 2x^2 + 5xy + y$$

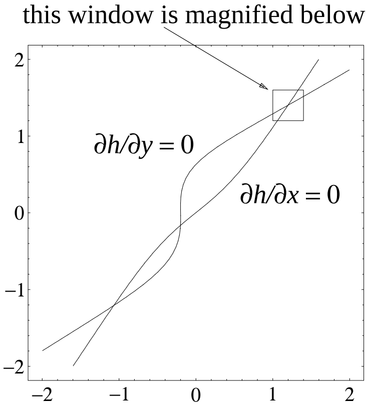
The equations for the critical points are

$$\frac{\partial h}{\partial x} = 2xy^2 - 4x^3 - 4x + 5y = 0,$$
$$\frac{\partial h}{\partial y} = 2x^2y - 4y^3 + 5x + 1 = 0.$$

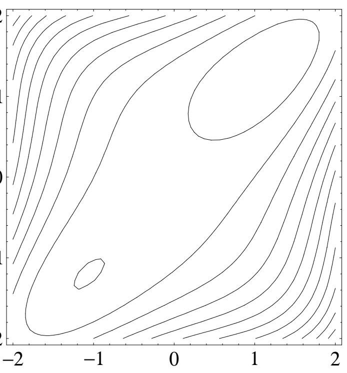
Even though we can't solve the equations algebraically, we can plot the curves they define by using the contour-plotting program of a computer. This is done at the right. The curves intersect in three points. (If you draw the plots on a large scale, you will find that these are still the only intersections.)

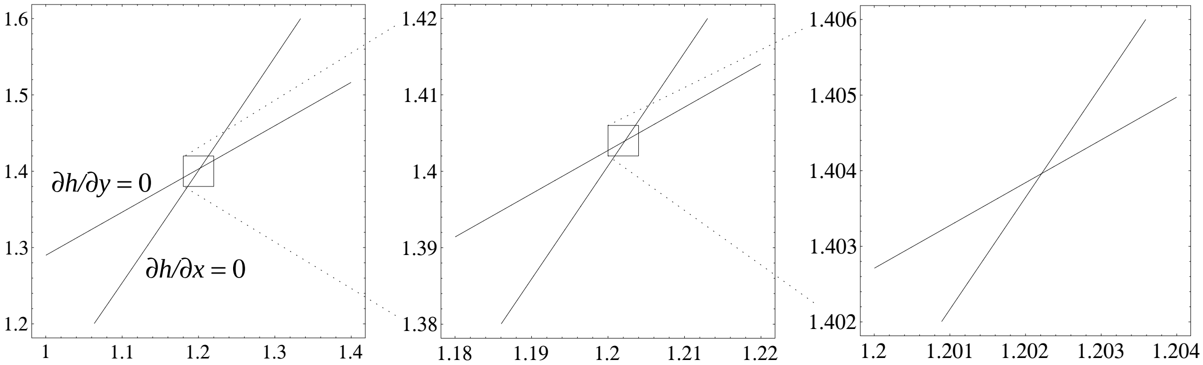
A contour plot of $z = h(x, y)$ itself reveals that the middle point is a saddle and the outer two are local maxima. Let's focus on the maximum in the upper right and determine its position more precisely.

We can always use a microscope. But if we magnify the contour plot, we just get a set of nested ovals. The maximum would lie somewhere inside the smallest—but we wouldn't know quite where. By contrast, the curves $\partial h/\partial x = 0$ and $\partial h/\partial y = 0$ give us a pair of "crosshairs" to focus on. Even with relatively little magnification we can see that the maximum is at $(1.202\ldots, 1.404\ldots)$.



contours of $z = h(x, y)$





The Method of Steepest Ascent

Walk uphill to get to a maximum

Here is yet another way to find the maximum of a function $z = f(x, y)$. Imagine that the graph of f is a landscape that you're standing on. You want to get to the highest point. To do that you just walk uphill. But the uphill direction on the landscape is given by the gradient vector field of f (see page 551). So you move to higher ground by following the gradient field.

Trajectories of the gradient dynamical system...

In fact, the gradient field defines a **dynamical system** of exactly the sort we studied in chapter 8. The differential equations are

$$\frac{dx}{dt} = \frac{\partial f}{\partial x}(x, y),$$
$$\frac{dy}{dt} = \frac{\partial f}{\partial y}(x, y).$$

...lead to the local maxima

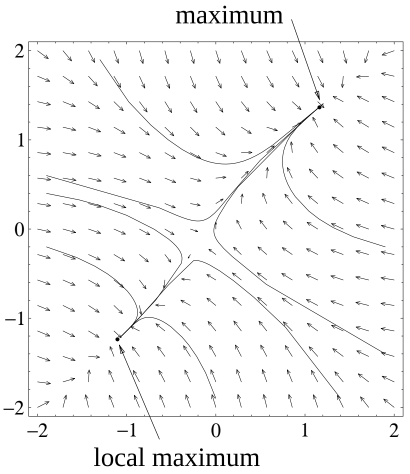
Because the gradient points uphill, the trajectories of this dynamical system also go uphill. Trajectories flow to the attractors of the system; these are the local maxima of the function. Furthermore, since the gradient points in the direction f increases *most rapidly*, the trajectories follow paths of **steepest ascent** to the maxima. This explains the name of the method.

Example 1. Let's see how the method of steepest ascent will find the local maxima of the function

$$z = h(x, y) = x^2y^2 - x^4 - y^4 - 2x^2 + 5xy + y$$

we considered in the previous example. The gradient field is

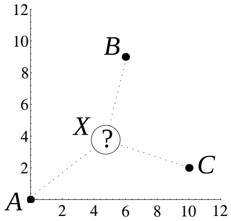
$$\frac{dx}{dt} = 2xy^2 - 4x^3 - 4x + 5y.$$
$$\frac{dy}{dt} = 2x^2y - 4y^3 + 5x + 1.$$



As you can see at the left, some of the trajectories flow to a local maximum near $(-1, -1)$, while other flow to the maximum whose position we determine on the opposite page. Each attractor has its own basin of attraction (as described in chapter 8). Therefore, the maximum found by the method of steepest ascent depends on the initial point of the trajectory.

The method of steepest *descent*

If we replace the gradient vectors by their negatives, the new field will point directly *downhill*—to the local minima. Using the trajectories of the negative gradient field to find the minima is thus called the method of steepest descent. In the following example we use this method to investigate an economic question.



Example 2. Manufacturing companies ship their products to regional warehouses from large distribution centers. Suppose a company has regional warehouses at A, B, and C, as shown on the map at the right. Where should it put its distribution center X so as to minimize the total cost of supplying the three regional warehouses?

This is a complicated problem that depends on many factors. For example, X probably should be put near major roads. The managers may also want to choose a location where labor costs are lower. Certainly, the total distance between the center and the three warehouses is important. Let's simply get a *first approximation* to a solution by concentrating on the last factor. We will find the position for X that minimizes the total straight-line distance (the distance "as the crow flies") from X to the three points A, B, and C. The map shows these distances as three dotted lines.

Minimize the total distance

To describe the various positions we have introduced a coordinate system in which

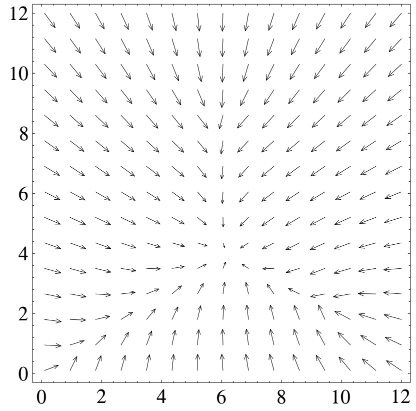
$$A: \ (0,0), \qquad B: \ (6,9), \qquad C: \ (10,2).$$

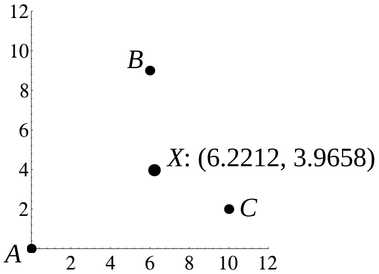
The coordinates here are arbitrary. That is, they don't represent miles, or kilometers, or any of the usual units of distance—but they are *proportional* to the usual units, so we can measure with them. If we let the unknown position of X be (x, y), then we seek to minimize the function

$$S(x,y) = \sqrt{x^2+y^2} + \sqrt{(x-6)^2+(y-9)^2} + \sqrt{(x-10)^2+(y-2)^2}.$$

According to the method of steepest descent, we want to find the attractor of this dynamical system:

$$\begin{aligned}
\frac{dx}{dt} &= -\frac{x}{\sqrt{x^2+y^2}} - \frac{x-6}{\sqrt{(x-6)^2+(y-9)^2}} - \frac{x-10}{\sqrt{(x-10)^2+(y-2)^2}}, \\
\frac{dy}{dt} &= -\frac{y}{\sqrt{x^2+y^2}} - \frac{y-9}{\sqrt{(x-6)^2+(y-9)^2}} - \frac{y-2}{\sqrt{(x-10)^2+(y-2)^2}}.
\end{aligned}$$





the optimum location for
the distributiuon center

The attractor is a global minimum

As you can see from the vector field of the dynamical system (above left), there is a single attractor, near the point $(6, 4)$. This implies the total distance function $S(x, y)$ has a single global minimum—which is what our intuition about the problem would lead us to expect.

To find the position of X more exactly, you can do the following. First, obtain a solution $(x(t), y(t))$ to the system of differential equations with an arbitrary initial condition—for example,

$$x(0) = 1, \qquad y(0) = 1.$$

The attractor is the limit point of a solution

Then, obtain the coordinates of the attractor by evaluating $(x(t), y(t))$ for larger and larger values of t, stopping when the values of $x(t)$ and $y(t)$ stabilize. You will find that

$$X = \lim_{t\to\infty} (x(t), y(t)) = (6.22120\ldots, 3.96577\ldots).$$

The method needs only a differential equation solver

The important point to note here is that it is not necessary to plot the vector field—or any other graphic aid, like a contour plot or graph. You simply need to solve a system of differential equations. For example, the values above were found by modifying the computer program SIRVALUE we introduced in chapter 2. In summary: *the method of steepest descent (or ascent) requires no graphical tools, but only a basic differential equation solver.*

Lagrange Multipliers

In searching for the interior extremes of a function $f(x, y)$, we have seen that it is helpful to solve the critical point equations:

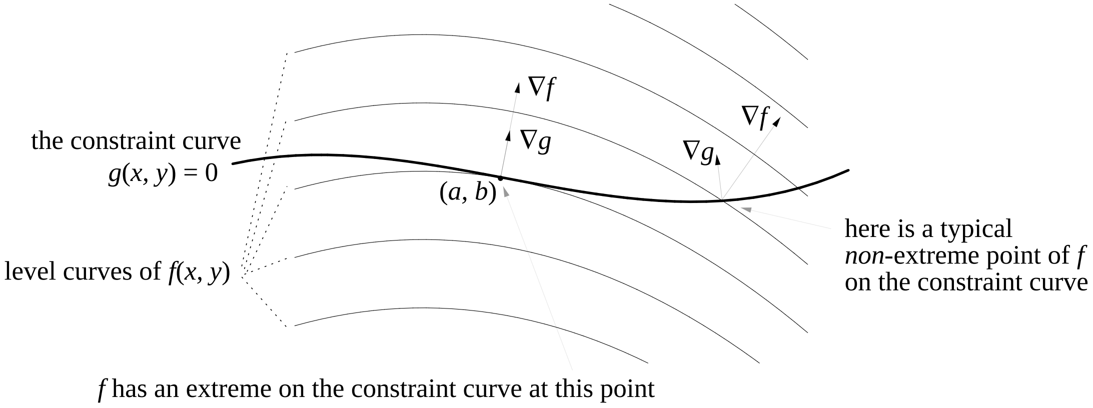
$$\frac{\partial f}{\partial x} = 0, \qquad \frac{\partial f}{\partial y} = 0.$$

Equations for a constrained extreme

There is a similar set of equations we can use in the search for the *constrained* extremes of f. These equations involve a new variable called a Lagrange multiplier.

∇f and ∇g must be parallel at a point where f has an extreme along the constraint curve $g = 0$

So, suppose the function $f(x, y)$ has an extreme on the constraint curve $g(x, y) = 0$ at the point (a, b). According to the following diagram, the gradient vectors ∇f and ∇g must be parallel at (a, b). (This means they are in the same direction or in opposite directions.) Here is why. We already know (see page 569) that the level curve of f that passes through the constrained maximum or minimum must be tangent to the constraint curve. Now, the gradient vector ∇f at any point is perpendicular to the level curve of f through that point—and the same is true for g. At a point where the level curves are tangent, the gradients ∇f and ∇g are perpendicular to the *same* curve, and must therefore be parallel.



The multiplier equation

Parallel vectors are multiples of each other. Specifically, at a point (a, b) where ∇f and ∇g are parallel, there must be a number λ for which

$$\nabla g(a, b) = \lambda \cdot \nabla f(a, b).$$

The multiplier λ is called a **Lagrange multiplier**. In the figure above, $\lambda \approx 1/2$. If ∇g and ∇f were in *opposite* directions, then λ would be negative.

Joseph Louis Lagrange (1736–1813) was a French mathematician and a younger contemporary of Leonhard Euler. Like Euler he played an important role in making calculus the primary analytical tool for study the physical world. He is particularly noted for his contributions to celestial mechanics, the field where Isaac Newton first applied the calculus.

If we write out the multiplier equation using the components of ∇f and ∇g, we get

$$\left(\frac{\partial g}{\partial x}(a,b), \frac{\partial g}{\partial y}(a,b)\right) = \lambda\left(\frac{\partial f}{\partial x}(a,b), \frac{\partial f}{\partial y}(a,b)\right) = \left(\lambda\,\frac{\partial f}{\partial x}(a,b), \lambda\,\frac{\partial f}{\partial y}(a,b)\right).$$

In a vector equation, the vectors are equal component by component. Thus,

$$\frac{\partial g}{\partial x}(a,b) = \lambda\,\frac{\partial f}{\partial x}(a,b)$$
$$\frac{\partial g}{\partial y}(a,b) = \lambda\,\frac{\partial f}{\partial y}(a,b)$$

How can we find a constrained maximum or minimum?

Let's return to the main question, which can be stated this way: How do we determine where the function $f(x,y)$ has a maximum or a minimum, subject to the constraint $g(x,y) = 0$? If we let (a,b) denote the point we seek, then we see that a and b satisfy three equations:

$$\text{the constraint equation}: \qquad g(a,b) = 0,$$
$$\text{the multiplier equations}: \quad \begin{cases} \dfrac{\partial g}{\partial x}(a,b) = \lambda\,\dfrac{\partial f}{\partial x}(a,b), \\[2ex] \dfrac{\partial g}{\partial y}(a,b) = \lambda\,\dfrac{\partial f}{\partial y}(a,b). \end{cases}$$

In fact, there are *three* unknowns in these equations: a, b, and λ. When we solve the three equations for the three unknowns, we will determine the location of the constrained extreme. (We'll also have a piece of information we can throw away: the value of λ.)

Example. Find the maximum of $f(x,y) = x^p y^{1-p}$ subject to the constraint $x + y = c$. There are two parameters in this problem: p and c. We assume that $0 < p < 1$ and $0 < c$. We introduce parameters to remind you that analytic methods (such as Lagrange multipliers) are especially valuable in solving problems that depend on parameters.

We let $g(x,y) = x + y - c$. Then

$$\nabla g = (1,1), \qquad \nabla f = \left(p x^{p-1} y^{1-p}, (1-p) x^p y^{-p}\right),$$

so the three equations we must solve are

$$\begin{aligned} x + y - c &= 0, \\ 1 &= \lambda p x^{p-1} y^{1-p}, \\ 1 &= \lambda (1-p) x^p y^{-p}. \end{aligned}$$

Since the second and third equations both equal 1, we can set them equal to each other:

$$\lambda p x^{p-1} y^{1-p} = \lambda (1-p) x^p y^{-p}.$$

We can cancel the two λs and combine the powers of x and y to get

$$p x^{-1} y = 1 - p \qquad \text{or} \qquad \frac{y}{x} = \frac{1-p}{p}.$$

According to the first of the three equations, $y = c - x$. If we substitution this expression in for y into the last equation, we get

$$\frac{c-x}{x} = \frac{1-p}{p} \qquad \text{or} \qquad p(c-x) = (1-p)x.$$

This equation reduces to $pc = x$, which gives us the x-coordinate of the maximum. To get the y-coordinate, we use $y = c - x = c - pc = c(1-p)$. To sum up, the maximum is at

$$(x, y) = (cp, c(1-p)) = c(p, 1-p).$$

Exercises

When searching for an extreme, be sure to zoom in on the graph or plot you are using as you narrow down the location of the point you seek.

1. Inspect the graph of $z = xy$ to find the maximum value of z subject to the constraints

$$x \geq 0, \qquad y \geq 0, \qquad 3x + 8y \leq 120.$$

2. Inspect the graph of $z = 5x + 2y$ to find the minimum value of z subject to the constraints

$$x \geq 0, \qquad y \geq 0, \qquad xy \geq 10.$$

3. Inspect a contour plot of $z = 3xy - y^2$ to find the maximum value of z subject to the constraints

$$x \geq 0, \qquad y \geq 0, \qquad x + y \leq 5.$$

4. Continuation. Add the constraint $x \leq a$ to the preceding three, where a is a parameter that takes values between 0 and 5. Find the maximum value of z subject to all four constraints. Describe how the position of the constraint depends on the value of the parameter a.

[Answer: The maximum is found at $(x, y) = (25/8, 15/8)$, as long as $a \geq 25/8$. When $a < 25/8$, the maximum is at $(x, y) = (a, 5 - a)$.]

5. Find the maximum and minimum values of $z = 12x - 5y$ when (x, y) is exactly 1 unit from the origin.

How is the gradient involved here?

6. Find the maximum and minimum value of $z = px + qy$ when (x, y) is exactly 1 unit from the origin.

a) At what point is the maximum achieved; at what point is the minimum achieved?

7. Use a graph to locate the maximum value of the function

$$z = 2xy - 5x^2 - 7y^2 + 2x + 3y.$$

There are no constraints.

8. Use a graph to find the maximum value of $z = 6x + 12y - x^3 - y^3$, subject to the constraints

$$x \geq 0, \qquad y \geq 0, \qquad x^2 + y^2 \leq 100.$$

9. a) Locate the position of the minimum of $x^4 - 2x^2 - \alpha x + y^2$ as a function of the parameter α.

b) The position of the minimum jumps catastrophically when α passes through a certain value. At what value of α does this happen, and what jump occurs in the minimum?

10. a) Locate the maximum of $x^3 + y^3 - 3x - 3y$ subject to

$$x \leq 3, \qquad y \leq 0, \qquad x + y \leq \beta.$$

The position of the maximum depends on the value of the parameter β, which you can assume lies between 0 and 5.

b) The position of the maximum jumps catastrophically when β crosses a certain threshold value β_0. What is β_0?

11. Find the maximum value of x^2y in the first quadrant, subject to the constraint $x + 5y = 10$.

12. Find the maximum and minimum values of $z = 3x + 4y$ subject to the single constraint $x^2 + 4xy + 5y^2 = 10$.

13. Find all the critical points of the following functions.

a) $3x^2 + 7xy + 2y^2 + 5x - 6y + 3$.

b) $\sin x \sin y$ on the domain $-4 \le x \le 4$, $-4 \le y \le 4$.

c) $\sin xy$ on the domain $-4 \le x \le 4$, $-4 \le y \le 4$.

d) $\exp(x^2 + y^2)$.

e) $x^3 + y^3 - 3x - 3y$.

f) $x^3 - 3xy^2 - x^2 - y^2$. [There are four critical points; three are saddles.]

14. a) Find the nine critical points of the function

$$C(x, y) = (x^2 + xy + y^2 - 1)(x^2 - xy + y^2 - 1).$$

Four are minima, four are saddles, and one is a maximum.

b) Mark the locations of the critical points on a suitable contour plot of $C(x, y)$.

15. Locate and classify the critical points of the energy integral of a pendulum:

$$E(x, v) = 1 - \cos x + \tfrac{1}{2}v^2.$$

a) Compare the *critical points* of E with the *equilibrium points* of the dynamical system associated with this energy integral.

16. Use the method of steepest descent to find the minimum of the function

$$z = p(x, y) = e^{2+y-x^2-y^2} \sin x.$$

The distribution problem

The next two exercises are modifications of the distribution problem on pages 579–580. In the example we assumed that deliveries from the center X to each of the regional warehouse A, B, and C happened equally often. These exercises assume that deliveries to some warehouses are more frequent than others.

17. Suppose that one truck makes deliveries to A 5 times each week, while a second truck is used to make 3 deliveries to B and 2 to C each week. It makes sense to locate X so that the *total weekly travel* is minimized, rather than just the total distance to the three warehouses. To get the total weekly travel, we should:

- multiply the distance from X to A by 5;
- multiply the distance from X to B by 3;
- multiply the distance from X to C by 2.

(Actually, the total *round-trip* distances are twice these values, but the proportions would remain the same, so we can use these numbers.) Thus, the function to minimize is

$$T(x,y) = 5\sqrt{x^2+y^2} + 3\sqrt{(x-6)^2+(y-9)^2} + 2\sqrt{(x-10)^2+(y-2)^2}.$$

a) Use the method of steepest descent to find the minimum of T.

b) Compare the location of the distribution center X as determined by T to its location determined by the function S of example 2 in the text. Would you *expect* the location to change? In what direction? Does the calculated change in position agree with your intuition?

18. Suppose that deliveries to A are twice as frequent as deliveries to either B or C. (For example, two trucks make the round-trip to A each day, but only one truck to B and one to C.) Where should the distribution center X be located in these circumstances? Explain how you got your answer.

[Answer: X should be at A. Does this surprise you?]

19. A company which has four offices around the country holds an annual meeting for its top executives. The location of each office, and the number of executives at that office, are given in the following table. (The coordinates x and y of the position are given in arbitrary units.)

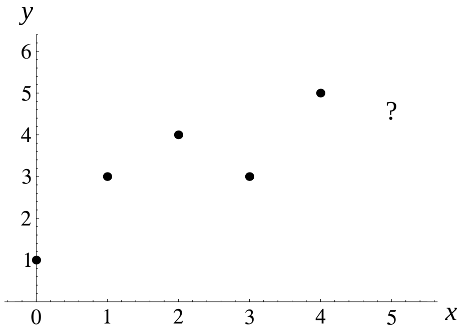
office	executives	x	y
A	32	200	300
B	17	1920	1100
C	20	2240	450
D	41	2875	1150

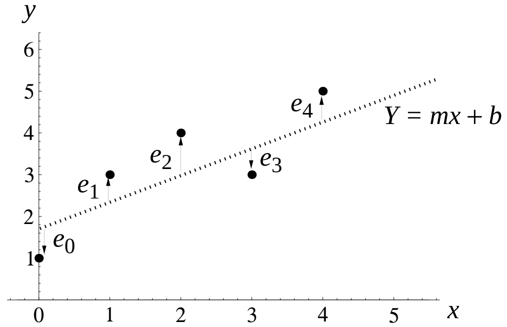
Where should the meeting be held if the location depends *solely* on the total travel cost for all the participants? Assume that the travel cost, per mile, is the same for every participant.

The best-fitting line

Suppose we've taken measurements of two quantities x and y, and obtained the results shown in the table and graph below. We assume that y depends on x according to some rule that we don't happen to know. In particular, we'd like to know what y is when $x = 5$. We have no data. Can we *predict* what y should be?

x	y
0	1
1	3
2	4
3	3
4	5
5	?





Here is a common approach to the question. We assume there is a simple underlying relation between y and x. However, the measurements that give us the data contain errors or "noise" of some sort that obscure the relationship. The simplest relation is a linear function, so we assume that there is a formula $Y = mx + b$ that describes the connection between x and y.

Which line should we choose? In other words, how should we choose m and b? Since the data points don't lie on a line, there is no perfect solution. For any choices, we must expect a difference e_j between the the

j-th data value y_j and the value $Y_j = m\,j + b$ predicted by the formula. These differences are the *errors* we assume are present.

A reasonable way to proceed is to **minimize** the total error. Even this involves choices. In the figure above, e_0 and e_3 are negative, so the ordinary total could be zero, or nearly so, even if the individual errors were large. We need a total that *ignores* the signs of the errors. Here is one:

$$\text{absolute error:} \qquad |e_0| + |e_1| + |e_2| + |e_3| + |e_4| = AE.$$

Here is another:

$$\text{squared error:} \qquad e_0^2 + e_1^2 + e_2^2 + e_3^2 + e_4^2 = SE.$$

In the following table we compare the data y_j with the calculated values Y_j and the resulting errors e_j.

x	y	Y	$e = y - Y$
0	1	b	$1 - b$
1	3	$m + b$	$3 - \ m - b$
2	4	$2m + b$	$4 - 2m - b$
3	3	$3m + b$	$3 - 3m - b$
4	5	$4m + b$	$5 - 4m - b$

The total errors are functions of m and b

To get the values of AE and SE, we take either the absolute values or the squares of the elements of the rightmost column, and then add. In particular, the table makes it clear that both total errors are functions of m and b. The absolute error is

$$AE(m, b) = |1 - b| + |3 - m - b| + |4 - 2m - b| + |3 - 3m - b| + |5 - 4m - b|.$$

20. Inspect a graph and a contour plot to determine the values of m and b which minimize $AE(m, b)$.

[Answer: Remarkable as is may seem, there is an entire line segment of solutions to this problem in the m, b-plane. One end of the line is near $(m, b) = (.67, 2.3)$, the other is near $(m, b) = (1, 1)$.]

21. a) Using a best-fitting line from the previous question, find the predicted value of y when $x = 5$.

b) Since there is a range of best-fitting lines, there should be a range of predicted values for y when $x = 5$. What is that range?

22. Write down the function $SE(m, b)$ that describes the squared error in the fit of a straight line to the data given above.

23. a) Use a graph and a contour plot to locate the minimum of the function $SE(m, b)$ from the previous exercise. Indicate how many digits of accuracy your answer has.

b) Use the method of steepest descent to locate the minimum. How many digits of accuracy does *this* method yield?

9.4 Chapter Summary

The Main Ideas

- The **graph** of a function of two variables is a two-dimensional surface in a three-dimensional space.
- A function of two variables can also be viewed using a **density plot**, a **terraced density plot**, or a **contour plot**. The latter is a set of **level curves** drawn on a flat plane.
- A **contour plot** of a function of three variables is a collection of **level surfaces** in three-dimensional space.
- The graph of a **linear function** is a flat plane, and its contour plot consists of straight, parallel, and equally-spaced lines.
- The **gradient** of a linear function is a vector whose components are the partial rates of change of the function.
- Under a **microscope**, the graph of a function of two variables becomes a flat plane. A contour plot turns into a set of straight, parallel, and equally-spaced lines.
- The multipliers in the **microscope equation** for a function are its **partial derivatives**:

$$\Delta z = \frac{\partial f}{\partial x_1}(a,b)\Delta x_1 + \cdots + \frac{\partial f}{\partial x_n}(a,b)\Delta x_n.$$

- The **gradient** of a function is a vector whose components are the partial derivatives of the function. Its magnitude and direction give the greatest **rate of increase** of the function at each point.
- **Optimization** is a process that involves finding the **maximum** or **minimum** value of a function. There may be **constraints** present that limit the scope of the search for an **extreme**.
- Extremes can be found at **critical points**, where all partial derivatives of a locally linear function are zero.
- The **method of steepest ascent** introduces the power of dynamical systems into the optimization process.

Expectations

- Using appropriate computer software, you should be able to make a **graph**, a **terraced density plot**, and a **contour plot** of a function of two variables.

- Using appropriate graphical representations of a function of two variables, you should be able to recognize the **maxima**, **minima**, and **saddle points**.

- You should be able to estimate the **partial rates of change** of a function of two variables at a point by zooming in on a contour plot.

- You should be able to recognize the various forms of a **linear function** of several variables and transform the representation of the function from one form to another.

- You should be able to describe the geometric meaning of the partial rates of change of a linear function of two variables.

- You should be able to find the **gradient** of a function of several variables at a point.

- You should know how the gradient of a function of two variables is related to its level curves.

- You should be able to write the **microscope equation** for a function of two variables at a point.

- You should be able to use the microscope equation for a function of two variables at a point to estimate values of the function at nearly points, to find the **trade-off** in one variable when the other changes by a fixed amount, and to estimate errors.

- You should be able to find the **linear aprroximation** to a function of two variables at a point.

- You should be able to find the equation of the **tangent plane** to the graph of a function of two variables at a point.

- You should be able to sketch the **gradient vector field** of a function of two variables in a specified domain.

- You should be able to sketch a plausible set of contour lines for a function whose gradient vecctor field is given; you should be able to sketch a plaausible gradient vector field for a ffunction whose contour plot is given.

- You should be able to find the critical points of a function of two variables, and you should be able to determine whether a critical point is an extreme by inspecting a graph or a contour plot.

- You should be able to find a local maximum of a function of two variables by the method of **steepest descent**.

- You should be able to find an extreme of a function of two variables subject to a constraint either by inspecting a graph or contour plot, or by the method of **Lagrange multipliers**.

Chapter 10

Series and Approximations

An important theme in this book is to give **constructive** definitions of mathematical objects. Thus, for instance, if you needed to evaluate

$$\int_0^1 e^{-x^2}\,dx,$$

you could set up a Riemann sum to evaluate this expression to any desired degree of accuracy. Similarly, if you wanted to evaluate a quantity like $e^{.3}$ from first principles, you could apply Euler's method to approximate the solution to the differential equation

$$y'(t) = y(t), \text{ with initial condition } y(0) = 1,$$

using small enough intervals to get a value for $y(.3)$ to the number of decimal places you needed. You might pause for a moment to think how you would get sin(5) to 7 decimal places—you wouldn't do it by drawing a unit circle and measuring the y-coordinate of the point where this circle is intersected by the line making an angle of 5 radians with the x-axis! Defining the sine function to be the solution to the second-order differential equation $y'' = -y$ with initial conditions $y = 0$ and $y' = 1$ when $t = 0$ is much better if we actually want to construct values of the function with more than two decimal accuracy.

Ordinary arithmetic lies at the heart of all calculations

What these examples illustrate is the fact that the only functions our brains or digital computers can evaluate directly are those involving the arithmetic operations of addition, subtraction, multiplication, and division. Anything else we or computers evaluate must ultimately be reducible to these four operations. But the only functions directly expressible in such terms are polynomials and rational functions (i.e., quotients of one polynomial by another). When you use your calculator to evaluate $\ln 2$, and the calculator shows .69314718056, it is really doing some additions, subtractions, multiplications, and divisions to compute this 11-digit *approximation* to $\ln 2$. There are no obvious connections to logarithms at all in what it does. One of the triumphs of calculus is the development of techniques for calculating highly accurate approximations of this sort quickly. In this chapter we will explore these techniques and their applications.

10.1 Approximation Near a Point or Over an Interval

Suppose we were interested in approximating the sine function—we might need to make a quick estimate and not have a calculator handy, or we might even be designing a calculator. In the next section we will examine a number of other contexts in which such approximations are helpful. Here is a third degree polynomial that is a good approximation in a sense which will be made clear shortly:

$$P(x) = x - \frac{x^3}{6}.$$

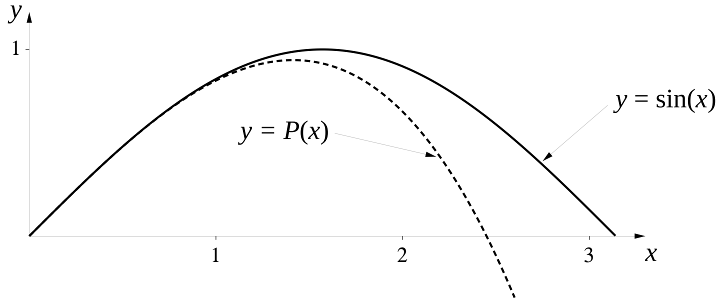
(You will see in section 2 where $P(x)$ comes from.)

If we compare the values of $\sin(x)$ and $P(x)$ over the interval $[0, 1]$ we get the following:

x	$\sin x$	$P(x)$	$\sin x - P(x)$
0.0	0.0	0.0	0.0
.2	.198669	.198667	.000002
.4	.389418	.389333	.000085
.6	.564642	.564000	.000642
.8	.717356	.714667	.002689
1.0	.841471	.833333	.008138

The fit is good, with the largest difference occurring at $x = 1.0$, where the difference is only slightly greater than .008.

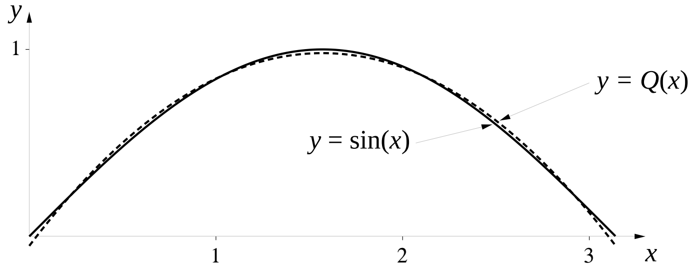
If we plot $\sin(x)$ and $P(x)$ together over the interval $[0, \pi]$ we see the ways in which $P(x)$ is both very good and not so good. Over the initial portion of the graph—out to around $x = 1$—the graphs of the two functions seem to coincide. As we move further from the origin, though, the graphs separate more and more. Thus if we were primarily interested in approximating $\sin(x)$ near the origin, $P(x)$ would be a reasonable choice. If we need to approximate $\sin(x)$ over the entire interval, $P(x)$ is less useful.



On the other hand, consider the second degree polynomial

$$Q(x) = -.4176977x^2 + 1.312236205x - .050465497$$

(You will see how to compute these coefficients in section 6.) When we graph $Q(x)$ and $\sin(x)$ together we get the following:



While $Q(x)$ does not fits the graph of $\sin(x)$ as well as $P(x)$ does near the origin, it is a good fit overall. In fact, $Q(x)$ exactly equals $\sin(x)$ at 4 values of x, and the greatest separation between the graphs of $Q(x)$ and $\sin(x)$ over the interval $[0, \pi]$ occurs at the endpoints, where the distance between the graphs is .0505 units.

What we have here, then, are two kinds of approximation of the sine function by polynomials: we have a polynomial $P(x)$ that behaves very much like the sine function near the origin, and we have another polynomial $Q(x)$ that keeps close to the sine function over the entire interval $[0, \pi]$. Which one is the "better" approximation depends on our needs. Each solves an important problem. Since finding approximations near a point has a neater

There's more than one way to make the "best fit" to a given curve solution—Taylor polynomials—we will start with this problem. We will turn to the problem of finding approximations over an interval in section 6.

10.2 Taylor Polynomials

The general setting. In chapter 3 we discovered that functions were locally linear at most points—when we zoomed in on them they looked more and more like straight lines. This fact was central to the development of much of the subsequent material. It turns out that this is only the initial manifestation of an even deeper phenomenon: Not only are functions locally linear, but, if we don't zoom in quite so far, they look locally like parabolas. From a little further back still they look locally like cubic polynomials, etc. Later in this section we will see how to use the computer to visualize these "local parabolizations", "local cubicizations", etc. Let's summarize the idea and then explore its significance:

> The functions of interest to calculus look locally like polynomials at most points of their domain. The higher the degree of the polynomial, the better typically will be the fit.

Comments The "at most points" qualification is because of exceptions like those we ran into when we explored locally linearity. The function $|x|$, for instance, was not locally linear at $x = 0$—it's not locally like any polynomial of higher degree at that point either. The issue of what "goodness of fit" means and how it is measured is a subtle one which we will develop over the course of this section. For the time being, your intuition is a reasonable guide—one fit to a curve is better than another near some point if it "shares more phosphor" with the curve when they are graphed on a computer screen centered at the given point.

The behavior of a function can often be inferred from the behavior of a local polynomialization

The fact that functions look locally like polynomials has profound implications conceptually and computationally. It means we can often determine the behavior of a function locally by examining the corresponding behavior of what we might call a "local polynomialization" instead. In particular, to find the values of a function near some point, or graph a function near some point, we can deal with the values or graph of a local polynomialization instead. Since we can actually evaluate polynomials directly, this can be a major simplification.

There is an extra feature to all this which makes the concept particularly attractive: not only are functions locally polynomial, it is easy to find the coefficients of the polynomials. Let's see how this works. Suppose we had some function $f(x)$ and we wanted to find the fifth degree polynomial that best fit this function at $x = 0$. Let's call this polynomial

We want the best fit at $x = 0$

$$P(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5.$$

To determine P, we need to find values for the six coefficients a_0, a_1, a_2, a_3, a_4, a_5.

Before we can do this, we need to define what we mean by the "best" fit to f at $x = 0$. Since we have six unknowns, we need six conditions. One obvious condition is that the graph of P should pass through the point $(0, f(0))$. But this is equivalent to requiring that $P(0) = f(0)$. Since $P(0) = a_0$, we thus must have $a_0 = f(0)$, and we have found one of the coefficients of $P(x)$. Let's summarize the argument so far:

The best fit should pass through the point $(0, f(0))$

> The graph of a polynomial passes through the point $(0, f(0))$ if and only if the polynomial is of the form
>
> $$f(0) + a_1x + a_2x^2 + \cdots .$$

But we're not interested in just any polynomial passing through the right point; it should be headed in the right direction as well. That is, we want the slope of P at $x = 0$ to be the same as the slope of f at this point—we want $P'(0) = f'(0)$. But

The best fit should have the right slope at $(0, f(0))$

$$P'(x) = a_1 + 2a_2x + 3a_3x^2 + 4a_4x^3 + 5a_5x^4,$$

so $P'(0) = a_1$. Our second condition therefore must be that $a_1 = f'(0)$. Again, we can summarize this as

> The graph of a polynomial passes through the point $(0, f(0))$ and has slope $f'(0)$ there if and only if it is of the form
>
> $$f(0) + f'(0)x + a_2x^2 + \cdots .$$

Note that at this point we have recovered the general form for the local linear approximation to f at $x = 0$: $L(x) = f(0) + f'(0)x$.

But there is no reason to stop with the first derivative. Similarly, we would want the way in which the slope of $P(x)$ is changing—we are now talking about $P''(0)$—to behave the way the slope of f is changing at $x = 0$, etc. Each higher derivative controls a more subtle feature of the shape of the graph. We now see how we could formulate reasonable additional conditions which would determine the remaining coefficients of $P(x)$:

> Say that $P(x)$ is the **best fit** to $f(x)$ at the point $x = 0$ if
> $P(0) = f(0),\ P'(0) = f'(0),\ P''(0) = f''(0), \ldots, P^{(5)}(0) = f^{(5)}(0)$.

The final criterion for best fit at $x = 0$

Since $P(x)$ is a fifth degree polynomial, all the derivatives of P beyond the fifth will be identically 0, so we can't control their values by altering the values of the a_k. What we are saying, then, is that we are using as our criterion for the best fit that all the derivatives of P as high as we can control them have the same values at $x = 0$ as the corresponding derivatives of f.

While this is a reasonable definition for something we might call the "best fit" at the point $x = 0$, it gives us no direct way to tell how good the fit really is. This is a serious shortcoming—if we want to approximate function values by polynomial values, for instance, we would like to know how many decimal places in the polynomial values are going to be correct. We will take up this question of goodness of fit later in this section; we'll be able to make measurements that allow us to to see how well the polynomial fits the function. First, though, we need to see how to determine the coefficients of the approximating polynomials and get some practice manipulating them.

Notation for higher derivatives

Note on Notation: We have used the notation $f^{(5)}(x)$ to denote the fifth derivative of $f(x)$ as a convenient shorthand for $f'''''(x)$, which is harder to read. We will use this throughout.

Finding the coefficients We first observe that the derivatives of P at $x = 0$ are easy to express in terms of $a_1, a_2, \ldots$. We have

$$\begin{aligned}
P'(x) &= a_1 + 2\,a_2 x + 3\,a_3 x^2 + 4\,a_4 x^3 + 5\,a_5 x^4,\\
P''(x) &= 2\,a_2 + 3\cdot 2\,a_3 x + 4\cdot 3\,a_4 x^2 + 5\cdot 4\,a_5 x^3,\\
P^{(3)}(x) &= 3\cdot 2\,a_3 + 4\cdot 3\cdot 2\,a_4 x + 5\cdot 4\cdot 3\,a_5 x^2,\\
P^{(4)}(x) &= 4\cdot 3\cdot 2\,a_4 + 5\cdot 4\cdot 3\cdot 2\,a_5 x,\\
P^{(5)}(x) &= 5\cdot 4\cdot 3\cdot 2\,a_5.
\end{aligned}$$

Thus $P''(0) = 2\,a_2$, $P^{(3)}(0) = 3 \cdot 2\,a_3$, $P^{(4)}(0) = 4 \cdot 3 \cdot 2\,a_4$, and $P^{(5)}(0) = 5 \cdot 4 \cdot 3 \cdot 2\,a_5$.

Factorial notation

We can simplify this a bit by introducing the **factorial** notation, in which we write $n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1$. This is called "$n$ factorial". Thus, for example, $7! = 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5040$. It turns out to be convenient to extend the factorial notation to 0 by defining $0! = 1$. (Notice, for instance, that this makes the formulas below work out right.) In the exercises you will see why this extension of the notation is not only convenient, but reasonable as well!

The desired rule for finding the coefficients

With this notation we can express compactly the equations above as $P^{(k)}(0) = k!\,a_k$ for $k = 0, 1, 2, \ldots 5$. Finally, since we want $P^{(k)}(0) = f^{(k)}(0)$, we can solve for the coefficients of $P(x)$:

$$a_k = \frac{f^{(k)}(0)}{k!} \quad \text{for } k = 0, 1, 2, 3, 4, 5.$$

We can now write down an explicit formula for the fifth degree polynomial which best fits $f(x)$ at $x = 0$ in the sense we've put forth:

$$P(x) = f(0) + f'(0)x + \frac{f^{(2)}(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 + \frac{f^{(4)}(0)}{4!}x^4 + \frac{f^{(5)}(0)}{5!}x^5.$$

We can express this more compactly using the Σ–notation we introduced in the discussion of Riemann sums in chapter 6:

$$P(x) = \sum_{k=0}^{5} \frac{f^{(k)}(0)}{k!}\,x^k.$$

We call this the **fifth degree Taylor polynomial for** $f(x)$. It is sometimes also called the fifth **order** Taylor polynomial.

It should be obvious to you that we can generalize what we've done above to get a best-fitting polynomial of any degree. Thus

General rule for the Taylor polynomial at $x = 0$

The **Taylor polynomial of degree n** approximating the function $f(x)$ at $x = 0$ is given by the formula

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!}\,x^k.$$

We also speak of the Taylor polynomial *centered at* $x = 0$.

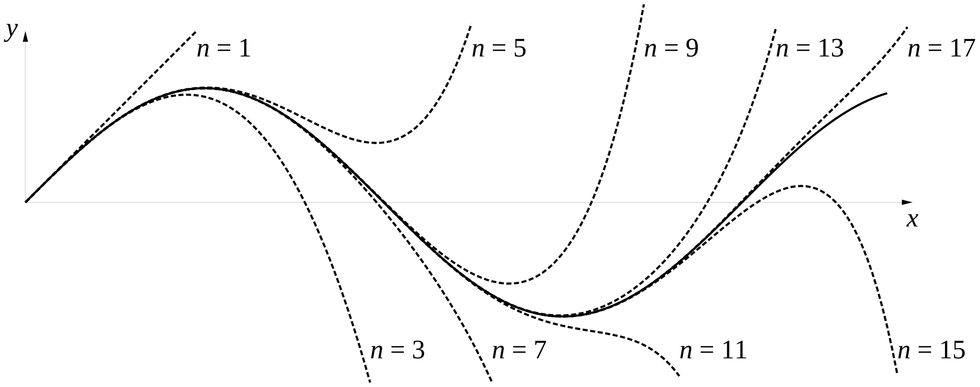
Example. Consider $f(x) = \sin(x)$. Then for $n = 7$ we have

$$\begin{aligned}
f(x) &= \sin(x), & f(0) &= 0,\\
f'(x) &= \cos(x), & f'(0) &= +1,\\
f^{(2)}(x) &= -\sin(x), & f^{(2)}(0) &= 0,\\
f^{(3)}(x) &= -\cos(x), & f^{(3)}(0) &= -1,\\
f^{(4)}(x) &= \sin(x), & f^{(4)}(0) &= 0,\\
f^{(5)}(x) &= \cos(x), & f^{(5)}(0) &= +1,\\
f^{(6)}(x) &= -\sin(x), & f^{(6)}(0) &= 0,\\
f^{(7)}(x) &= -\cos(x), & f^{(7)}(0) &= -1.
\end{aligned}$$

From this we can see that the pattern 0, +1, 0 , −1, ... will repeat forever. Substituting these values into the formula we get that for any odd integer n the n-th degree Taylor polynomial for $\sin(x)$ is

$$P_n(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \pm x^n n!.$$

Note that $P_3(x) = x - x^3/6$, which is the polynomial we met in section 1. We saw there that this polynomial seemed to fit the graph of the sine function only out to around $x = 1$. Now, though, we have a way to generate polynomial approximations of higher degrees, and we would expect to get better fits as the degree of the approximating polynomial is increased. To see how closely these polynomial approximations follow $\sin(x)$, here's the graph of $\sin(x)$ together with the Taylor polynomials of degrees $n = 1, 3, 5, \ldots, 17$ plotted over the interval $[0, 7.5]$:



The higher the degree of the polynomial, the better the fit

While each polynomial eventually wanders off to infinity, successive polynomials stay close to the sine function for longer and longer intervals—the Taylor polynomial of degree 17 is just beginning to diverge visibly by the time x reaches 2π. We might expect that if we kept going, we could find Taylor polynomials that were good fits out to $x = 100$, or $x = 1000$. This is indeed the case, although they would be long and cumbersome polynomials to work with. Fortunately, as you will see in the exercises, with a little cleverness we can use a Taylor polynomial of degree 9 to calculate $\sin(100)$ to 5 decimal place accuracy.

Approximating polynomials for other basic functions

Other Taylor Polynomials: In a similar fashion, we can get Taylor polynomials for other functions. You should use the general formula to verify the Taylor polynomials for the following basic functions. (The Taylor polynomial for $\sin(x)$ is included for convenient reference.)

$f(x)$	$P_n(x)$
$\sin(x)$	$x - \dfrac{x^3}{3!} + \dfrac{x^5}{5!} - \dfrac{x^7}{7!} + \cdots \pm \dfrac{x^n}{n!}$ (n odd)
$\cos(x)$	$1 - \dfrac{x^2}{2!} + \dfrac{x^4}{4!} - \dfrac{x^6}{6!} + \cdots \pm \dfrac{x^n}{n!}$ (n even)
e^x	$1 + x + \dfrac{x^2}{2!} + \dfrac{x^3}{3!} + \dfrac{x^4}{4!} + \cdots + \dfrac{x^n}{n!}$
$\ln(1-x)$	$-\left(x + \dfrac{x^2}{2} + \dfrac{x^3}{3} + \dfrac{x^4}{4} + \cdots + \dfrac{x^n}{n}\right)$
$\dfrac{1}{1-x}$	$1 + x + x^2 + x^3 + \cdots + x^n$

General rule for the Taylor polynomial at $x = a$

Taylor polynomials at points other than $x = 0$. Using exactly the same arguments we used to develop the best-fitting polynomial at $x = 0$, we can derive the more general formula for the best-fitting polynomial at any value of x. Thus, if we know the behavior of f and its derivatives at some point $x = a$, we would like to find a polynomial $P_n(x)$ which is a good approximation to $f(x)$ for values of x close to a.

Since the expression $x - a$ tells us how close x is to a, we use it (instead of the variable x itself) to construct the polynomials approximating f at $x = a$:

$$P_n(x) = b_0 + b_1(x-a) + b_2(x-a)^2 + b_3(x-a)^3 + \cdots + b_n(x-a)^n.$$

You should be able to apply the reasoning we used above to derive the following:

> The **Taylor polynomial of degree $\boldsymbol{n}$ centered at $\boldsymbol{x = a}$** approximating the function $f(x)$ is given by the formula
>
> $$\begin{aligned} P_n(x) &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^n(a)}{n!}(x-a)^n \\ &= \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k. \end{aligned}$$

Program: TAYLOR

```
Set up GRAPHICS
DEF fnfact(m)
     P = 1
     FOR r = 2 TO m
         P = P * r
     NEXT r
     fnfact = P
END DEF
DEF fnpoly(x)
     Sum = x
     Sign = -1
     FOR k = 3 TO 17 STEP 2
         Sum = Sum + Sign * x^k/fnfact(k)
         Sign = (-1) * Sign
     NEXT k
     fnpoly = Sum
END DEF
FOR x = 0 TO 3.14 STEP .01
     Plot the line from (x, fnpoly(x)) to (x + .01, fnpoly(x + .01))
NEXT x
```

A computer program for graphing Taylor polynomials Shown above is a program that evaluates the 17-th degree Taylor polynomial for $\sin(x)$ and graphs it over the interval $[0, 3.14]$. The first seven lines of the program constitute a subroutine for evaluating factorials. The syntax of such subroutines varies from one computer language to another, so be sure to use the format that's appropriate for you. You may even be using a language that already knows how to compute factorials, in which case you can omit the subroutine. The second set of 9 lines defines the function `poly` which evaluates the 17-th degree Taylor polynomial. Note the role of the variable `Sign`—it simply changes the sign back and forth from positive to negative as each new term is added to the sum. As usual, you will have to put in commands to set up the graphics and draw lines in the format your computer language uses. You can modify this program to graph other Taylor polynomials.

New Taylor Polynomials from Old

Given a function we want to approximate by Taylor polynomials, we could always go straight to the general formula for deriving such polynomials. On the other hand, it is often possible to avoid a lot of tedious calculation of derivatives by using a polynomial we've already calculated. It turns out that any manipulation on Taylor polynomials you might be tempted to try will probably work. Here are some examples to illustrate the kinds of manipulations that can be performed on Taylor polynomials.

Substitution in Taylor Polynomials. Suppose we wanted the Taylor polynomial for e^{x^2}. We know from what we've already done that for any value of u close to 0,

$$e^u \approx 1 + u + \frac{u^2}{2!} + \frac{u^3}{3!} + \frac{u^4}{4!} + \cdots + \frac{u^n}{n!}.$$

In this expression u can be anything, including another variable expression. For instance, if we set $u = x^2$, we get the Taylor polynomial

$$\begin{aligned} e^{x^2} &= e^u \\ &\approx 1 + u + \frac{u^2}{2!} + \frac{u^3}{3!} + \frac{u^4}{4!} + \cdots + \frac{u^n}{n!} \\ &= 1 + (x^2) + \frac{(x^2)^2}{2!} + \frac{(x^2)^3}{3!} + \frac{(x^2)^4}{4!} + \cdots + \frac{(x^2)^n}{n!} \\ &= 1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!} + \cdots + \frac{x^{2n}}{n!}. \end{aligned}$$

You should check to see that this is what you get if you apply the general formula for computing Taylor polynomials to the function e^{x^2}.

Similarly, suppose we wanted a Taylor polynomial for $1/(1+x^2)$. We could start with the approximation given earlier:

$$\frac{1}{1-u} \approx 1 + u + u^2 + u^3 + \cdots + u^n.$$

If we now replace u everywhere by $-x^2$, we get the desired expansion:

$$\begin{aligned}\frac{1}{1+x^2} = \frac{1}{1-(-x^2)} = \frac{1}{1-u} &\\ &\approx 1 + u + u^2 + u^3 + \cdots + u^n \\ &= 1 + (-x^2) + (-x^2)^2 + (-x^2)^3 + \cdots + (-x^2)^n \\ &= 1 - x^2 + x^4 - x^6 + \cdots \pm x^{2n}.\end{aligned}$$

Again, you should verify that if you start with $f(x) = 1/(1+x^2)$ and apply to f the general formula for deriving Taylor polynomials, you will get the preceding result. Which method is quicker?

Multiplying Taylor Polynomials. Suppose we wanted the 5-th degree Taylor polynomial for $e^{3x} \cdot \sin(2x)$. We can use substitution to write down polynomial approximations for e^{3x} and $\sin(2x)$, so we can get an approximation for their product by multiplying the two polynomials:

$$\begin{aligned}&e^{3x} \cdot \sin(2x) \\ &\quad\approx \left(1 + (3x) + \frac{(3x)^2}{2!} + \frac{(3x)^3}{3!} + \frac{(3x)^4}{4!} + \frac{(3x)^5}{5!}\right)\left((2x) - \frac{(2x)^3}{3!} + \frac{(2x)^5}{5!}\right) \\ &\quad\approx 2x + 6x^2 + \frac{23}{3}x^3 + 5x^4 - \frac{61}{60}x^5.\end{aligned}$$

Again, you should try calculating this polynomial directly from the general rule, both to see that you get the same result, and to appreciate how much more tedious the general formula is to use in this case.

In the same way, we can also divide Taylor polynomials, raise them to powers, and chain them by composition. The exercises provide examples of some of these operations.

Differentiating Taylor Polynomials. Suppose we know a Taylor polynomial for some function f. If g is the derivative of f, we can immediately get a Taylor polynomial for g (of degree one less) by differentiating the polynomial we know for f. You should review the definition of Taylor polynomial to see why this is so. For instance, suppose $f(x) = 1/(1-x)$ and $g(x) = 1/(1-x)^2$. Verify that $f'(x) = g(x)$. It then follows that

$$\frac{1}{(1-x)^2} = \frac{d}{dx}\left(\frac{1}{1-x}\right) \approx \frac{d}{dx}(1 + x + x^2 + \cdots + x^n)$$
$$= 1 + 2x + 3x^2 + \cdots + nx^{n-1}.$$

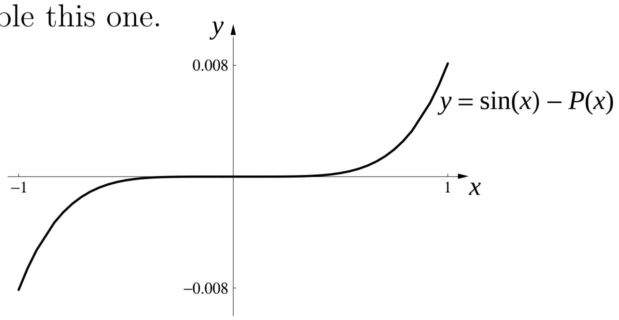
Integrating Taylor Polynomials. Again suppose we have functions $f(x)$ and $g(x)$ with $f'(x) = g(x)$, and suppose this time that we know a Taylor polynomial for g. We can then get a Taylor polynomial for f by antidifferentiating term by term. For instance, we find in chapter 11 that the derivative of $\arctan(x)$ is $1/(1+x^2)$, and we have seen above how to get a Taylor polynomial for $1/(1+x^2)$. Therefore we have

$$\arctan x = \int_0^x \frac{1}{1+t^2}dt \approx \int_0^x \left(1 - t^2 + t^4 - t^6 + \cdots \pm t^{2n}\right) dt$$
$$= t - \frac{1}{3}t^3 + \frac{1}{5}t^5 - \cdots \pm \frac{1}{2n+1}t^{2n+1}\Big|_0^x$$
$$= x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \cdots \pm \frac{1}{2n+1}x^{2n+1}.$$

Goodness of fit

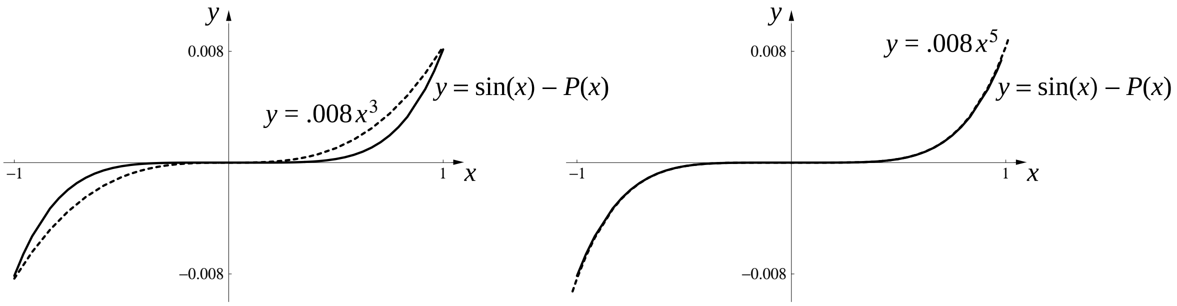
Graph the difference between a function and its Taylor polynomial

Let's turn to the question of *measuring* the fit between a function and one of its Taylor polynomials. The ideas here have a strong geometric flavor, so you should use a computer graphing utility to follow this discussion. Once again, consider the function $\sin(x)$ and its Taylor polynomial $P(x) = x - x^3/6$. According to the table in section 1, the difference $\sin(x) - P(x)$ got smaller as x got smaller. Stop now and graph the function $y = \sin(x) - P(x)$ near $x = 0$. This will show you exactly how $\sin(x) - P(x)$ depends on x. If you choose the interval $-1 \le x \le 1$ (and your graphing utility allows its vertical and horizontal scales to be set independently of each other), your graph should resemble this one.



The difference looks like a power of x

This graph looks very much like a cubic polynomial. If it really is a cubic, we can figure out its formula, because we know the value of $\sin(x) - P(x)$ is about .008 when $x = 1$. Therefore the cubic should be $y = .008\,x^3$ (because then $y = .008$ when $x = 1$). However, if you graph $y = .008\,x^3$ together with $y = \sin(x) - P(x)$, you should find a poor match (the left-hand figure, below.) Another possibility is that $\sin(x) - P(x)$ is more like a *fifth* degree polynomial. Plot $y = .008\,x^5$; it's so close that it "shares phosphor" with $\sin(x) - P(x)$ near $x = 0$.



Finding the multiplier

If $\sin(x) - P(X)$ were *exactly* a multiple of x^5, then $(\sin x - P(x))/x^5$ would be constant and would equal the value of the multiplier. What we actually find is this:

x	$\dfrac{\sin x - P(x)}{x^5}$
1.0	.0081377
0.5	.0082839
0.1	.0083313
0.05	.0083328
0.01	.0083333

suggesting $\displaystyle\lim_{x\to 0} \frac{\sin x - P(x)}{x^5} = .008333\ldots.$

How $P(x)$ fits $\sin(x)$

Thus, although the ratio is not constant, it appears to converge to a definite value—which we can take to be the value of the multipier:

$$\sin x - P(x) \approx .008333\,x^5 \quad \text{when} \quad x \approx 0.$$

We say that $\sin(x) - P(x)$ *has the same order of magnitude as* x^5 as $x \to 0$. So $\sin(x) - P(x)$ is about as small as x^5. Thus, if we know the size of x^5 we will be able to tell how close $\sin(x)$ and $P(x)$ are to each other.

Comparing two numbers

A rough way to measure how close two numbers are is to count the number of decimal places to which they agree. But there are pitfalls here; for instance, none of the decimals of 1.00001 and 0.99999 agree, even though the difference between the two numbers is only 0.00002. This suggests that a good way to compare two numbers is to look at their difference. Therefore, we say

$$A = B \text{ to } k \text{ decimal places} \quad \textit{means} \quad A - B = 0 \text{ to } k \text{ decimal places}$$

Now, a number equals 0 to k decimal places precisely when it *rounds off to* 0 (when we round it to k decimal places). Since X rounds to 0 to k decimal places if and only $|X| < .5 \times 10^{-k}$, we finally have a precise way to compare the size of two numbers:

$$\boxed{A = B \text{ to } k \text{ decimal places} \quad \textit{means} \quad |A - B| < .5 \times 10^{-k}.}$$

What the fit means computationally

Now we can say how close $P(x)$ is to $\sin(x)$. Since x is small, we can take this to mean $x = 0$ *to k decimal places*, or $|x| < .5 \times 10^{-k}$. But then,

$$|x^5 - 0| = |x - 0|^5 < (.5 \times 10^{-k})^5 < .5 \times 10^{-5k-1}$$

(since $.5^5 = .03125 < .5 \times 10^{-1}$). In other words, if $x = 0$ to k decimal places, then $x^5 = 0$ to $5k+1$ places. Since $\sin(x) - P(x)$ has the same order of magnitude as x^5 as $x \to 0$, $\sin(x) = P(x)$ to $5k+1$ places as well. In fact, because the multiplier in the relation

$$\sin x - P(x) \approx .008333\, x^5 \qquad (x \approx 0)$$

is .0083..., we gain two more decimal places of accuracy. (Do you see why?) Thus, finally, we see how reliable the polynomial $P(x) = x - x^3/6$ is for calculating values of $\sin(x)$:

$$\boxed{\begin{array}{l}\text{When } x = 0 \text{ to } k \text{ decimal places of accuracy, we can use } P(x)\\ \text{to calculate the first } 5k+3 \text{ decimal places of the value of } \sin(x).\end{array}}$$

Here are a few examples comparing $P(x)$ to the *exact* value of $\sin(x)$:

x	$P(x)$	$\sin(x)$
.0372	$.\underline{03719142}01920$	$.\underline{0371942}07856\ldots$
.0086	$.\underline{00859989}39907$	$.\underline{00859989}3991\ldots$
.0048	$.\underline{0047999815680}000$	$.\underline{0047999815680}212\ldots$

The underlined digits are guaranteed to be correct, based on the number of decimal places for which x agrees with 0. (Note that, according to our rule, $.0086 = 0$ to *one* decimal place, not two.)

Taylor's theorem

Order of magnitude

Taylor's theorem is the generalization of what we have just seen; it describes the goodness of fit between an arbitrary function and one of its Taylor polynomials. We'll state three versions of the theorem, gradually uncovering more information. To get started, we need a way to compare the order of magnitude of *any* two functions.

We say that $\varphi(x)$ has **the same order of magnitude** as $q(x)$ as $x \to a$, and we write $\varphi(x) = O(q(x))$ as $x \to a$, if there is a constant C for which

$$\lim_{x \to a} \frac{\varphi(x)}{q(x)} = C.$$

Now, when $\lim_{x \to a} \varphi(x)/q(x)$ is C, we have

$$\varphi(x) \approx Cq(x) \quad \text{when } x \approx a.$$

We'll frequently use this relation to express the idea that $\varphi(x)$ has the same order of magnitude as $q(x)$ as $x \to a$.

'Big oh' notation

The symbol O is an upper case "oh". When $\varphi(x) = O(q(x))$ as $x \to a$, we say *$\varphi(x)$ is 'big oh' of $q(x)$ as x approaches a.* Notice that the equal sign in $\varphi(x) = O(q(x))$ does *not* mean that $\varphi(x)$ and $O(q(x))$ are equal; $O(q(x))$ isn't even a function. Instead, the equal sign and the O together tell us that $\varphi(x)$ stands in a certain relation to $q(x)$.

Taylor's theorem, version 1. If $f(x)$ has derivatives up to order n at $x = a$, then

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R(x),$$

where $R(x) = O((x-a)^{n+1})$ as $x \to a$. The term $R(x)$ is called the **remainder**.

Informal language

This version of Taylor's theorem focusses on the general shape of the remainder function. Sometimes we just say the remainder has "order $n+1$", using this short phrase as an abbreviation for "the order of magnitude of the function $(x-a)^{n+1}$". In the same way, we say that *a function and its n-th degree Taylor polynomial at $x = a$ agree to order $n+1$ as $x \to a$.*

Notice that, if $\varphi(x) = O(x^3)$ as $x \to 0$, then it is also true that $\varphi(x) = O(x^2)$ (as $x \to 0$). This implies that we should take $\varphi(x) = O(x^n)$ to mean "φ has *at least* order n" (instead of simply "φ has order n"). In the same way, it would be more accurate (but somewhat more cumbersome) to say that $\varphi = O(q)$ means "φ has *at least* the order of magnitude of q".

Decimal places of accuracy

As we saw in our example, we can translate the order of agreement between the function and the polynomial into information about the number of decimal places of accuracy in the polynomial approximation. In particular, if $x - a = 0$ to k decimal places, then $(x - a)^n = 0$ to nk places, at least. Thus, as the order of magnitude n of the remainder increases, the fit increases, too. (You have already seen this illustrated with the sine function and its various Taylor polynomials, in the figure on page 600.)

A formula for the remainder

While the first version of Taylor's theorem tells us that $R(x)$ looks like $(x - a)^{n+1}$ in some general way, the next gives us a concrete formula. At least, it *looks* concrete. Notice, however, that $R(x)$ is expressed in terms of a number c_x (which depends upon x), but the formula doesn't tell us *how* c_x depends upon x. Therefore, if you want to use the formula to *compute* the value of $R(x)$, you can't. The theorem says only that c_x exists; it doesn't say how to find its value. Nevertheless, this version provides useful information, as you will see.

> **Taylor's theorem, version 2**. Suppose f has continuous derivatives up to order $n + 1$ for all x in some interval containing a. Then, for each x in that interval, there is a number c_x between a and x for which
>
> $$R(x) = \frac{f^{(n+1)}(c_x)}{(n+1)!}\,(x - a)^{n+1}.$$
>
> This is called **Lagrange's form of the remainder**.

Another formula for the remainder

We can use the Lagrange form as an aid to computation. To see how, return to the formula

$$R(x) \approx C(x - a)^{n+1} \quad (x \approx a)$$

that expresses $R(x) = O((x - a)^{n+1})$ as $x \to a$ (see page 608). The constant here is the limit

$$C = \lim_{x \to a} \frac{R(x)}{(x - a)^{n+1}}.$$

If we have a good estimate for the value of C, then $R(x) \approx C(x-a)^{n+1}$ gives us a good way to estimate $R(x)$. Of course, we could just evaluate the limit to determine C. In fact, that's what we did in the example; knowing $C \approx .008$ there gave us two more decimal places of accuracy in our polynomial approxmation to the sine function.

Determining C from f at $x = a$

But the Lagrange form of the remainder gives us another way to determine C:

$$\begin{aligned} C = \lim_{x \to a} \frac{R(x)}{(x-a)^{n+1}} &= \lim_{x \to a} \frac{f^{(n+1)}(c_x)}{(n+1)!} \\ &= \frac{f^{(n+1)}(\lim_{x \to a} c_x)}{(n+1)!} \\ &= \frac{f^{(n+1)}(a)}{(n+1)!}. \end{aligned}$$

In this argument, we are permitted to take the limit "inside" $f^{(n+1)}$ because $f^{(n+1)}$ is a continuous function. (That is one of the hypotheses of version 2.) Finally, since c_x lies between x and a, it follows that $c_x \to a$ as $x \to a$; in other words, $\lim_{x \to a} c_x = a$. Consequently, we get C *directly* from the function f itself, and we can therefore write

$$R(x) \approx \frac{f^{(n+1)}(a)}{(n+1)!}\,(x-a)^{n+1} \qquad (x \approx a).$$

An error bound

The third version of Taylor's theorem uses the Lagrange form of the remainder in a similar way to get an *error bound* for the polynomial approximation based on the size of $f^{(n+1)}(x)$.

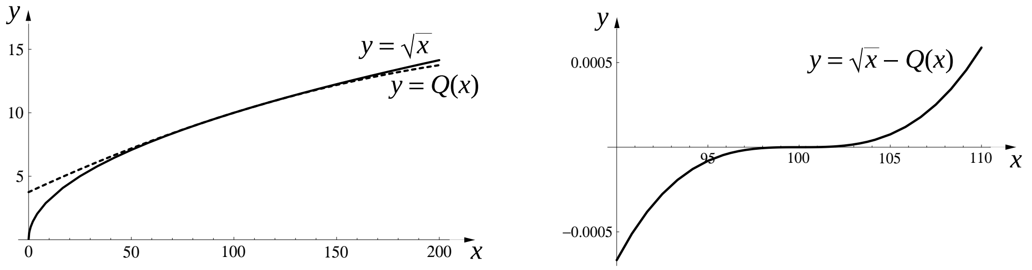
Taylor's theorem, version 3. Suppose that $|f^{(n+1)}(x)| \leq M$ for all x in some interval containing a. Then, for each x in that interval,

$$|R(x)| \leq \frac{M}{(n+1)!}\,|x-a|^{n+1}.$$

With this error bound, which is derived from knowledge of $f(x)$ near $x = a$, we can determine quite precisely how many decimal places of accuracy a Taylor polynomial approximation achieves. The following example illustrates the different versions of Taylor's theorem.

Example. Consider $\sqrt{x}$ near $x = 100$. The second degree Taylor polynomial for $\sqrt{x}$, centered at $x = 100$, is

$$Q(x) = 10 + \frac{(x-100)}{20} - \frac{(x-100)^2}{8000}.$$



Version 1: the remainder is $O((x-100)^3)$

Plot $y = Q(x)$ and $y = \sqrt{x}$ together; the result should look like the figure on the left, above. Then plot the remainder $y = \sqrt{x} - Q(x)$ near $x = 100$. This graph should suggest that $\sqrt{x} - Q(x) = O((x-100)^3)$ as $x \to 100$. In fact, this is what version 1 of Taylor's theorem asserts. Furthermore,

$$\lim_{x\to 100} \frac{\sqrt{x} - Q(x)}{(x-100)^3} \approx 6.25 \times 10^{-7};$$

check this yourself by constructing a table of values. Thus

$$\sqrt{x} - Q(x) \approx C(x-100)^3 \quad \text{where } C \approx 6.25 \times 10^{-7}.$$

Version 2: determining C in terms of $\sqrt{x}$ at $x = 100$

We can use the Lagrange form of the remainder (in version 2 of Taylor's theorem) to get the value of C another way—directly from the third derivative of $\sqrt{x}$ at $x = 100$:

$$C = \left.\frac{(x^{1/2})'''}{3!}\right|_{x=100} = \frac{\frac{1}{2}\cdot\frac{-1}{2}\cdot\frac{-3}{2}\cdot(100)^{-5/2}}{6} = \frac{1}{2^4\cdot 10^5} = 6.25 \times 10^{-7}.$$

This is the *exact* value, confirming the estimate obtained above.

Accuracy of the polynomial approximation

Let's see what the equation $\sqrt{x} - Q(x) \approx 6.25 \times 10^{-7}(x-100)^3$ tells us about the accuracy of the polynomial approximation. If we assume $|x-100| < .5 \times 10^{-k}$, then

$$\begin{aligned} |\sqrt{x} - Q(x)| &< 6.25 \times 10^{-7} \times (.5 \times 10^{-k})^3 \\ &= .78125 \times 10^{-(3k+7)} \ < \ .5 \times 10^{-(3k+6)}. \end{aligned}$$

Thus

$$x = 100 \text{ to } k \text{ decimal places} \Longrightarrow \sqrt{x} = Q(x) \text{ to } 3k+6 \text{ places.}$$

For example, if $x = 100.47$, then $k = 0$, so $Q(100.47) = \sqrt{100.47}$ to 6 decimal places. We find

$$Q(100.47) = \underline{10.023472}3875,$$

and the underlined digits should be correct. In fact,

$$\sqrt{100.47} = \underline{10.023472}4521\ldots.$$

Here is a second example. If $x = 102.98$, then we can take $k = -1$, so $Q(102.98) = \sqrt{102.98}$ to $3(-1) + 6 = 3$ decimal places. We find

$$Q(102.98) = \underline{10.147}88995, \qquad \sqrt{102.98} = \underline{10.147}906187\ldots.$$

Version 3: an explicit error bound

Let's see what additional light version 3 sheds on our investigation. Suppose we assume $x = 100$ to $k = 0$ decimal places. This means that x lies in the open interval $(99.5, 100.5)$. Version 3 requires that we have a bound on the size of the third derivative of $f(x) = \sqrt{x}$ over this interval. Now $f'''(x) = \frac{3}{8}x^{-5/2}$, and this is a decreasing function. (Check its graph; alternatively, note that its derivative is negative.) Its maximum value therefore occurs at the left endpoint of the (closed) interval $[99.5, 100.5]$:

$$|f'''(x)| \le f'''(99.5) = \tfrac{3}{8}\,(99.5)^{-5/2} < 3.8 \times 10^{-6}.$$

Therefore, from version 3 of Taylor's theorem,

$$|\sqrt{x} - Q(x)| < \frac{3.8 \times 10^{-6}}{3!}\,|x - 100|^3$$

Since $|x - 100| < .5$, $|x - 100|^3 < .125$, so

$$|\sqrt{x} - Q(x)| < \frac{3.8 \times 10^{-6} \times .125}{6} = .791667 \times 10^{-7} < .5 \times 10^{-6}.$$

This proves $\sqrt{x} = Q(x)$ to 6 decimal places—confirming what we found earlier.

Applications

Now you can do anything your calculator can!

Evaluating Functions. An obvious use of Taylor polynomials is to evaluate functions. In fact, whenever you ask a calculator or computer to evaluate a function—trigonometric, exponential, logarithmic—it is typically giving you the value of an appropriate polynomial (though not necessarily a Taylor polynomial).

Evaluating Integrals. The fundamental theorem of calculus gives us a quick way of evaluating a definite integral provided we can find an antiderivative for the function under the integral (cf. chapter 6.4). Unfortunately, many common functions, like e^{-x^2} or $(\sin x)/x$, don't have antiderivatives that can be expressed as finite algebraic combinations of the basic functions. Up until now, whenever we encountered such a function we had to rely on a Riemann sum to estimate the integral. But now we have Taylor polynomials, and it's easy to find an antiderivative for a polynomial! Thus, if we have an awkward definite integral to evaluate, it is reasonable to expect that we can estimate it by first getting a good polynomial approximation to the integrand, and then integrating this polynomial. As an example, consider the **error function**, erf(t), defined by

$$\operatorname{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-x^2}\, dx\,.$$

The error function

This is perhaps the most important integral in statistics. It is the basis of the so-called "normal distribution" and is widely used to decide how good certain statistical estimates are. It is important to have a way of obtaining fast, accurate approximations for erf(t). We have already seen that

$$e^{-x^2} \approx 1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \frac{x^8}{4!} - \cdots \pm \frac{x^{2n}}{n!}.$$

Now, if we antidifferentiate term by term:

$$\begin{aligned}\int e^{-x^2}\, dx &\approx \int \left(1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \frac{x^8}{4!} - \cdots \pm \frac{x^{2n}}{n!}\right) dx \\ &= \int 1\, dx - \int x^2\, dx + \int \frac{x^4}{2!}\, dx - \int \frac{x^6}{3!}\, dx + \cdots \pm \int \frac{x^{2n}}{n!}\, dx \\ &= x - \frac{x^3}{3} + \frac{x^5}{5 \cdot 2!} - \frac{x^7}{7 \cdot 3!} + \cdots \pm \frac{x^{2n+1}}{(2n+1) \cdot n!}.\end{aligned}$$

Thus,

$$\int_0^t e^{-x^2}\,dx \approx x - \frac{x^3}{3} + \frac{x^5}{5\cdot 2!} - \frac{x^7}{7\cdot 3!} + \cdots \pm \frac{x^{2n+1}}{(2n+1)\cdot n!}\Bigg|_0^t,$$

giving us, finally, an approximate formula for erf(t):

A formula for approximating the error function

$$\operatorname{erf}(t) \approx \frac{2}{\sqrt{\pi}}\left(t - \frac{t^3}{3} + \frac{t^5}{5\cdot 2!} - \frac{t^7}{\cdot 3!} + \cdots \pm \frac{t^{2n+1}}{(2n+1)\cdot n!}\right).$$

Thus if we needed to know, say, erf(1), we could quickly approximate it. For instance, letting $n = 6$, we have

$$\begin{aligned}\operatorname{erf}(1) &\approx \frac{2}{\sqrt{\pi}}\left(1 - \frac{1}{3} + \frac{1}{5\cdot 2!} - \frac{1}{7\cdot 3!} + \frac{1}{9\cdot 4!} - \frac{1}{11\cdot 5!} + \frac{1}{13\cdot 6!}\right)\\ &\approx \frac{2}{\sqrt{\pi}}\left(1 - \frac{1}{3} + \frac{1}{10} - \frac{1}{42} + \frac{1}{216} - \frac{1}{1320} + \frac{1}{9360}\right)\\ &\approx .746836\,\frac{2}{\sqrt{\pi}} \approx .842714,\end{aligned}$$

a value accurate to 4 decimals. If we had needed greater accuracy, we could simply have taken a larger value for n. For instance, if we take $n = 12$, we get the estimate .8427007929…, where all 10 decimals are accurate (i.e., they don't change as we take larger values n).

Evaluating Limits. Our final application of Taylor polynomials makes explicit use of the order of magnitude of the remainder. Consider the problem of evaluating a limit like

$$\lim_{x\to 0}\frac{1-\cos(x)}{x^2}.$$

Since both numerator and denominator approach 0 as $x \to 0$, it isn't clear what the quotient is doing. If we replace $\cos(x)$ by its third degree Taylor polynomial with remainder, though, we get

$$\cos(x) = 1 - \frac{1}{2!}x^2 + R(x),$$

and $R(x) = O(x^4)$ as $x \to 0$. Consequently, if $x \neq 0$ but $x \to 0$, then

$$\begin{aligned}\frac{1-\cos(x)}{x^2} &= \frac{1 - \left(1 - \frac{1}{2}x^2 + R(x)\right)}{x^2}\\ &= \frac{\frac{1}{2}x^2 - R(x)}{x^2} = \frac{1}{2} - \frac{R(x)}{x^2}.\end{aligned}$$

Since $R(x) = O(x^4)$, we know that there is some constant C for which $R(x)/x^4 \to C$ as $x \to 0$. Therefore,

$$\begin{aligned}\lim_{x\to 0}\frac{1-\cos(x)}{x^2} &= \frac{1}{2} - \lim_{x\to 0}\frac{R(x)}{x^2} = \frac{1}{2} - \lim_{x\to 0}\frac{x^2\cdot R(x)}{x^4}\\ &= \frac{1}{2} - \lim_{x\to 0} x^2 \cdot \lim_{x\to 0}\frac{R(x)}{x^4} = \frac{1}{2} - 0\cdot C = \frac{1}{2}.\end{aligned}$$

Extending the 'big oh' notation

There is a way to shorten these calculations—and to make them more transparent—by extending the way we read the 'big oh' notation. Specifically, we will read $O(q(x))$ as "some (unspecified) function that is the same order of magnitude as $q(x)$".

Then, instead of writing $\cos(x) = 1 - \frac{1}{2}x^2 + R(x)$, and then noting $R(x) = O(x^4)$ as $x \to 0$, we'll just write

$$\cos(x) = 1 - \tfrac{1}{2}x^2 + O(x^4) \qquad (x \to 0).$$

In this spirit,

$$\begin{aligned}\frac{1-\cos(x)}{x^2} &= \frac{1-\left(1-\frac{1}{2}x^2+O(x^4)\right)}{x^2}\\ &= \frac{\frac{1}{2}x^2 - O(x^4)}{x^2} = \tfrac{1}{2} + O(x^2) \qquad (x \to 0).\end{aligned}$$

We have used the fact that $\pm O(x^4)/x^2 = O(x^2)$. Finally, since $O(x^2) \to 0$ as $x \to 0$ (do you see why?), the limit of the last expression is just $1/2$ as $x \to 0$. Thus, once again we arrive at the result

$$\lim_{x\to 0}\frac{1-\cos(x)}{x^2} = \frac{1}{2}.$$

Exercises

1. Find a seventh degree Taylor polynomial centered at $x = 0$ for the indicated antiderivatives.

a) $\displaystyle\int \frac{\sin(x)}{x}\,dx.$

[Answer: $\displaystyle\int \frac{\sin(x)}{x}\,dx \approx x - \frac{x^3}{3\cdot 3!} + \frac{x^5}{5\cdot 5!} - \frac{x^7}{7\cdot 7!}$.]

b) $\displaystyle\int e^{x^2}\,dx.$

c) $\displaystyle\int \sin(x^2)\,dx.$

2. Plot the 7-th degree polynomial you found in part (a) above over the interval $[0, 5]$. Now plot the 9-th degree approximation on the same graph. When do the two polynomials begin to differ visibly?

3. Using the seventh degree Taylor approximation

$$E(t) \approx \int_0^t e^{-x^2}\,dx \; = t - \frac{t^3}{3} + \frac{t^5}{5 \cdot 2!} - \frac{t^7}{7 \cdot 3!},$$

calculate the values of $E(.3)$ and $E(-1)$. Give only the significant digits—that is, report only those decimals of your estimates that you think are fixed. (This means you will also need to calculate the ninth degree Taylor polynomial as well—do you see why?)

4. Calculate the values of $\sin(.4)$ and $\sin(\pi/12)$ using the seventh degree Taylor polynomial centered at $x = 0$

$$\sin(x) \approx x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!}.$$

Compare your answers with what a calculator gives you.

5. Find the third degree Taylor polynomial for $g(x) = x^3 - 3x$ at $x = 1$. Show that the Taylor polynomial is actually equal to $g(x)$—that is, the remainder is 0. What does this imply about the *fourth* degree Taylor polynomial for g at $x = 1$?

6. Find the seventh degree Taylor polynomial centered at $x = \pi$ for
(a) $\sin(x)$; (b) $\cos(x)$; (c) $\sin(3x)$.

7. In this problem you will compare computations using Taylor polynomials centered at $x = \pi$ with computations using Taylor polynomials centered at $x = 0$.

a) Calculate the value of $\sin(3)$ using a seventh degree Taylor polynomial centered at $x = 0$. How many decimal places of your estimate appear to be fixed?

b) Now calculate the value of sin(3) using a seventh degree Taylor polynomial centered at $x = \pi$. Now how many decimal places of your estimate appear to be fixed?

8. Write a program which evaluates a Taylor polynomial to print out sin(5°), sin(10°), sin(15°), . . . , sin(40°), sin(45°) accurate to 7 decimals. (Remember to convert to radians before evaluating the polynomial!)

9. **Why 0! = 1**. When you were first introduced to exponential notation in expressions like 2^n, n was restricted to being a positive integer, and 2^n was defined to be the product of 2 multiplied by itself n times. Before long, though, you were working with expressions like 2^{-3} and $2^{1/4}$. These new expressions weren't defined in terms of the original definition. For instance, to calculate 2^{-3} you wouldn't try to multiply 2 by itself -3 times—that would be nonsense! Instead, 2^{-m} is defined by looking at the key *properties* of exponentiation for positive exponents, and extending the definition to other exponents in a way that preserves these properties. In this case, there are two such properties, one for adding exponents and one for multiplying them:

$$\begin{array}{lll} \text{Property A:} & 2^m \cdot 2^n = 2^{m+n} & \text{for all positive } m \text{ and } n, \\ \text{Property M:} & (2^m)^n = 2^{mn} & \text{for all positive } m \text{ and } n. \end{array}$$

a) Show that to preserve property A we have to define $2^0 = 1$.

b) Show that we then have to define $2^{-3} = 1/2^3$ if we are to continue to preserve property A.

c) Show why $2^{1/4}$ must be $\sqrt[4]{2}$.

d) In the same way, you should convince yourself that a basic property of the factorial notation is that $(n+1)! = (n+1) \cdot n!$ for any positive integer n. Then show that to preserve this property, we have to define $0! = 1$.

e) Show that there is no way to define $(-1)!$ which preserves this property.

10. Use the general rule to derive the 5-th degree Taylor polynomial centered at $x = 0$ for the function

$$f(x) = (1 + x)^{\frac{1}{2}}.$$

Use this approximation to estimate $\sqrt{1.1}$. How accurate is this?

11. Use the general rule to derive the formula for the n-th degree Taylor polynomial centered at $x = 0$ for the function

$$f(x) = (1 + x)^c \text{ where } c \text{ is a constant.}$$

12. Use the result of the preceding problem to get the 6-th degree Taylor polynomial centered at $x = 0$ for $1/\sqrt[3]{1 + x^2}$.

[Answer: $1 - \dfrac{1}{3}x^2 + \dfrac{2}{9}x^4 - \dfrac{14}{81}x^6$.]

13. Use the result of the preceding problem to approximate

$$\int_0^1 \frac{1}{\sqrt[3]{1 + x^2}}\, dx.$$

14. Calculate the first 7 decimals of erf(.3). Be sure to show why you think all 7 decimals are correct. What degree Taylor polynomial did you need to produce these 7 decimals?

[Answer: $\mathrm{erf}(.3) = .3286267\ldots$.]

15. a) Apply the general formula for calculating Taylor polynomials centered at $x = 0$ to the tangent function to get the 5-th degree approximation.

[Answer: $\tan(x) \approx x + x^3/3 + 2x^5/15$.]

b) Recall that $\tan(x) = \sin(x)/\cos(x)$. Multiply the 5-th degree Taylor polynomial for $\tan(x)$ from part a) by the 4-th degree Taylor polynomial for $\cos(x)$ and show that you get the fifth degree polynomial for $\sin(x)$ (discarding higher degree terms).

16. Show that the n-th degree Taylor polynomial centered at $x = 0$ for $1/(1 - x)$ is $1 + x + x^2 + \cdots + x^n$.

17. Note that

$$\int \frac{1}{1 - x}\, dx = -\ln(1 - x).$$

Use this observation, together with the result of the previous problem, to get the n-th degree Taylor polynomial centered at $x = 0$ for $\ln(1 - x)$.

18. a) Find a formula for the n-th degree Taylor polynomial centered at $x = 1$ for $\ln(x)$.

b) Compare your answer to part (a) with the Taylor polynomial centered at $x = 0$ for $\ln(1 - x)$ you found in the previous problem. Are your results consistent?

19. a) The first degree Taylor polynomial for e^x at $x = 0$ is $1 + x$. Plot the remainder $R_1(x) = e^x - (1 + x)$ over the interval $-.1 \le x \le .1$. How does this graph demonstrate that $R_1(x) = O(x^2)$ as $x \to 0$?

b) There is a constant C_2 for which $R_1(x) \approx C_2x^2$ when $x \approx 0$. Why? Estimate the value of C_2.

20. This concerns the second degree Taylor polynomial for e^x at $x = 0$. Plot the remainder $R_2(x) = e^x - (1 + x + x^2/2)$ over the interval $-.1 \le x \le .1$. How does this graph demonstrate that $R_2(x) = O(x^3)$ as $x \to 0$?

a) There is a constant C_3 for which $R_2(x) \approx C_3x^3$ when $x \approx 0$. Why? Estimate the value of C_3.

21. Let $R_3(x) = e^x - P_3(x)$, where $P_3(x)$ is the third degree Taylor polynomial for e^x at $x = 0$. Show $R_3(x) = O(x^4)$ as $x \to 0$.

22. At first glance, Taylor's theorem says that

$$\sin(x) = x - \frac{1}{6}x^3 + O(x^4) \quad \text{as } x \to 0.$$

However, graphs and calculations done in the text (pages 605–607) make it clear that

$$\sin(x) = x - \frac{1}{6}x^3 + O(x^5) \quad \text{as } x \to 0.$$

Explain this. Is Taylor's theorem wrong here?

23. Using a suitable formula (that is, a Taylor polynomial with remainder) for each of the functions involved, find the indicated limit.

a) $\displaystyle\lim_{x\to 0} \frac{\sin(x)}{x}$ [Answer: 1]

b) $\displaystyle\lim_{x\to 0} \frac{e^x - (1 + x)}{x^2}$ [Answer: 1/2]

c) $\displaystyle\lim_{x\to 1} \frac{\ln x}{x - 1}$

d) $\displaystyle\lim_{x\to 0} \frac{x - \sin(x)}{x^3}$ [Answer: 1/6]

e) $\displaystyle\lim_{x\to 0}\frac{\sin(x^2)}{1-\cos(x)}$

24. Suppose $f(x) = 1 + x^2 + O(x^4)$ as $x \to 0$. Show that

$$(f(x))^2 = 1 + 2x^2 + O(x^4) \quad \text{as } x \to 0.$$

25. a) Using $\sin x = x - \frac{1}{6}x^3 + O(x^5)$ as $x \to 0$, show

$$(\sin x)^2 = x^2 - \tfrac{1}{3}x^4 + O(x^6) \quad \text{as } x \to 0.$$

b) Using $\cos x = 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4 + O(x^5)$ as $x \to 0$, show

$$(\cos x)^2 = 1 - x^2 + \tfrac{1}{3}x^4 + O(x^5) \quad \text{as } x \to 0.$$

c) Using the previous parts, show $(\sin x)^2 + (\cos x)^2 = 1 + O(x^5)$ as $x \to 0$. (Of course, you already know $(\sin x)^2 + (\cos x)^2 = 1$ *exactly.*)

26. a) Apply the general formula for calculating Taylor polynomials to the tangent function to get the 5-th degree approximation.
b) Recall that $\tan(x) = \sin(x)/\cos(x)$, so $\tan(x) \cdot \cos(x) = \sin(x)$. Multiply the fifth degree Taylor polynomial for $\tan(x)$ from part a) by the fifth degree Taylor polynomial for $\cos(x)$ and show that you get the fifth degree Taylor polynomial for $\sin(x)$ plus $O(x^6)$—that is, plus terms of order 6 and higher.

27. a) Using the formulas

$$\begin{aligned} e^u &= 1 + u + \tfrac{1}{2}u^2 + \tfrac{1}{6}u^3 + O(u^4) \qquad (u \to 0),\\ \sin x &= x - \tfrac{1}{6}x^3 + O(x^5) \qquad (x \to 0), \end{aligned}$$

show that $e^{\sin x} = 1 + x + \frac{1}{2}x^2 + O(x^4)$ as $x \to 0$.
b) Apply the general formula to obtain the third degree Taylor polynomial for $e^{\sin x}$ at $x = 0$, and compare your result with the formula in part (a).

28. Using $\displaystyle\frac{e^x - 1}{x} = 1 + \tfrac{1}{2}x + \tfrac{1}{6}x^2 + \tfrac{1}{24}x^3 + O(x^4)$ as $x \to 0$, show that

$$\frac{x}{e^x - 1} = 1 - \tfrac{1}{2}x + \tfrac{1}{12}x^2 + O(x^4) \qquad (x \to 0).$$

29. Show that the following are true as $x \to \infty$.

a) $x + 1/x = O(x)$.

b) $5x^7 - 12x^4 + 9 = O(x^7)$.

c) $\sqrt{1 + x^2} = O(x)$.

d) $\sqrt{1 + x^p} = O(x^{p/2})$.

30. a) Let $f(x) = \ln(x)$. Find the smallest bound M for which

$$|f^{(4)}(x)| \leq M \quad \text{when } |x - 1| \leq .5.$$

b) Let $P_3(x)$ be the degree 3 Taylor polynomial for $\ln(x)$ at $x = 1$, and let $R_3(x)$ be the remainder $R_3(x) = \ln(x) - P_3(x)$. Find a number K for which

$$|R(x)| \leq K\,|x - 1|^4$$

for all x satisfying $|x - 1| \leq .5$.

c) If you use $P_3(x)$ to approximate the value of $\ln(x)$ in the interval $.5 \leq x \leq 1.5$, how many digits of the approximation are correct?

d) Suppose we restrict the interval to $|x-1| \leq .1$. Repeat parts (a) and (b), getting *smaller* values for M and K. Now how many digits of the polynomial approximation $P_3(x)$ to $\ln(x)$ are correct, if $.9 \leq x \leq 1.1$?

"Little oh" notation. Similar to the "big oh" notation is another, called the "little oh": if

$$\lim_{x \to a} \frac{\phi(x)}{q(x)} = 0,$$

then we write $\phi(x) = o(q(x))$ and say ϕ is '*little oh*' of q as $x \to a$.

31. Suppose $\phi(x) = O(x^6)$ as $x \to 0$. Show the following.

a) $\phi(x) = O(x^5)$ as $x \to 0$.

b) $\phi(x) = o(x^5)$ as $x \to 0$.

c) It is false that $\phi(x) = O(x^7)$ as $x \to 0$. (One way you can do this is to give an explicit example of a function $\phi(x)$ for which $\phi(x) = O(x^6)$ but for which you can show $\phi(x) = O(x^7)$ is false.)

d) It is false that $\phi(x) = o(x^6)$ as $x \to 0$.

32. Sketch the graph $y = x\ln(x)$ over the interval $0 < x \leq 1$. Explain why your graph shows $\ln(x) = o(1/x)$ as $x \to 0$.

10.3 Taylor Series

In the previous section we have been talking about approximations to functions by their Taylor polynomials. Thus, for instance, we were able to write statements like

$$\sin(x) \approx x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!},$$

where the approximation was a good one for values of x not too far from 0. On the other hand, when we looked at Taylor polynomials of higher and higher degree, the approximations were good for larger and larger values of x. We are thus tempted to write

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \cdots ,$$

indicating that the sine function is equal to this "infinite degree" polynomial. This infinite sum is called the **Taylor series** centered at $x = 0$ for $\sin(x)$. But what would we even mean by such an infinite sum? We will explore this question in detail in section 5, but you should already have some intuition about what it means, for it can be interpreted in exactly the same way we interpret a more familiar statement like

You have seen infinite sums before

$$\begin{aligned}\frac{1}{3} &= .33333\ldots \\ &= \frac{3}{10} + \frac{3}{100} + \frac{3}{1000} + \frac{3}{10000} + \frac{3}{100000} + \cdots .\end{aligned}$$

Every decimal number is a sum of fractions whose denominators are powers of 10; 1/3 is a number whose decimal expansion happens to need an infinite number of terms to be completely precise. Of course, when a practical matter arises (for example, typing a number like 1/3 or π into a computer) just the beginning of the sum is used—the "tail" is dropped. We might write 1/3 as 0.33, or as 0.33333, or however many terms we need to get the accuracy we want. Put another way, we are saying that 1/3 is the *limit* of the finite sums of the right hand side of the equation.

Infinite degree polynomials are to be viewed like infinite decimals

Our new formulas for Taylor series are meant to be used exactly the same way: when a computation is involved, take only the beginning of the sum, and drop the tail. Just where you cut off the tail depends on the input value x and on the level of accuracy needed. Look at what happens when we we approximate the value of $\cos(\pi/3)$ by evaluating Taylor polynomials of increasingly higher degree:

$$\begin{aligned}
1 \quad &= 1.0000000 \\
1 - \frac{1}{2!}\left(\frac{\pi}{3}\right)^2 &\approx 0.4516887 \\
1 - \frac{1}{2!}\left(\frac{\pi}{3}\right)^2 + \frac{1}{4!}\left(\frac{\pi}{3}\right)^4 &\approx 0.5017962 \\
1 - \frac{1}{2!}\left(\frac{\pi}{3}\right)^2 + \frac{1}{4!}\left(\frac{\pi}{3}\right)^4 - \frac{1}{6!}\left(\frac{\pi}{3}\right)^6 &\approx 0.4999646 \\
1 - \frac{1}{2!}\left(\frac{\pi}{3}\right)^2 + \frac{1}{4!}\left(\frac{\pi}{3}\right)^4 - \frac{1}{6!}\left(\frac{\pi}{3}\right)^6 + \frac{1}{8!}\left(\frac{\pi}{3}\right)^8 &\approx 0.5000004 \\
1 - \frac{1}{2!}\left(\frac{\pi}{3}\right)^2 + \frac{1}{4!}\left(\frac{\pi}{3}\right)^4 - \frac{1}{6!}\left(\frac{\pi}{3}\right)^6 + \frac{1}{8!}\left(\frac{\pi}{3}\right)^8 - \frac{1}{10!}\left(\frac{\pi}{3}\right)^{10} &\approx 0.5000000
\end{aligned}$$

These sums were evaluated by setting $\pi = 3.141593$. As you can see, at the level of precision we are using, a sum that is six terms long gives the correct value. However, five, four, or even three terms may have been adequate for the needs at hand. The crucial fact is that these are all honest calculations using *only* the four operations of elementary arithmetic.

Note that if we had wanted to get the same 6 place accuracy for $\cos(x)$ for a larger value of x, we might need to go further out in the series. For instance $\cos(7\pi/3)$ is also equal to .5, but the tenth degree Taylor polynomial centered at $x = 0$ gives

$$1 - \frac{1}{2!}\left(\frac{7\pi}{3}\right)^2 + \frac{1}{4!}\left(\frac{7\pi}{3}\right)^4 - \frac{1}{6!}\left(\frac{7\pi}{3}\right)^6 + \frac{1}{8!}\left(\frac{7\pi}{3}\right)^8 - \frac{1}{10!}\left(\frac{7\pi}{3}\right)^{10} = -37.7302,$$

which is not even close to .5 . In fact, to get $\cos(7\pi/3)$ to 6 decimals, we need to use the Taylor polynomial centered at $x = 0$ of degree 30, while to get $\cos(19\pi/3)$ (also equal to .5) to 6 decimals we need the Taylor polynomial centered at $x = 0$ of degree 66!

The key fact, though, is that, for any value of x, if we go out in the series far enough (where what constitutes "far enough" will depend on x), we can approximate $\cos(x)$ to any number of decimal places desired. For any x, the value of $\cos(x)$ is the limit of the finite sums of the Taylor series, just as 1/3 is the limit of the finite sums of its infinite series representation.

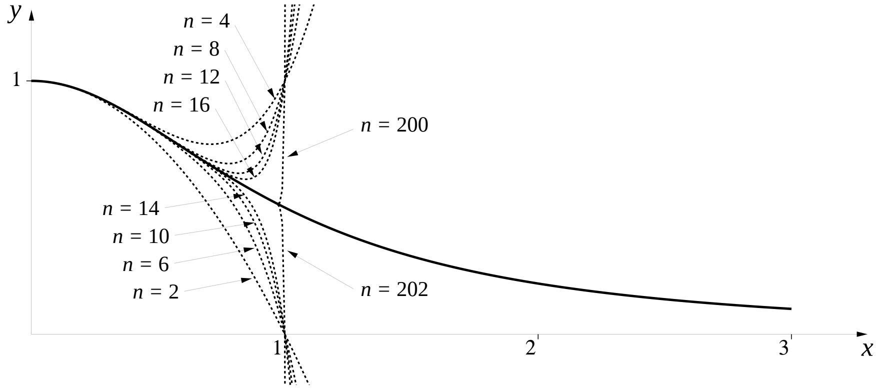
In general, given a function $f(x)$, its Taylor series centered at $x = 0$ will be

$$f(0) + f'(0)x + \frac{f^{(2)}(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 + \frac{f^{(4)}(0)}{4!}x^4 + \cdots = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!}x^k.$$

We have the following Taylor series centered at $x = 0$ for some common functions:

$f(x)$	Taylor series for $f(x)$
$\sin(x)$	$x - \dfrac{x^3}{3!} + \dfrac{x^5}{5!} - \dfrac{x^7}{7!} + \cdots$
$\cos(x)$	$1 - \dfrac{x^2}{2!} + \dfrac{x^4}{4!} - \dfrac{x^6}{6!} + \cdots$
e^x	$1 + x + \dfrac{x^2}{2!} + \dfrac{x^3}{3!} + \dfrac{x^4}{4!} + \cdots$
$\ln(1 - x)$	$-\left(x + \dfrac{x^2}{2} + \dfrac{x^3}{3} + \dfrac{x^4}{4} + \cdots\right)$
$\dfrac{1}{1 - x}$	$1 + x + x^2 + x^3 + \cdots$
$\dfrac{1}{1 + x^2}$	$1 - x^2 + x^4 - x^6 + \cdots$
$(1 + x)^c$	$1 + cx + \dfrac{c(c-1)}{2!}x^2 + \dfrac{c(c-1)(c-2)}{3!}x^3 + \cdots$

While it is true that $\cos(x)$ and e^x equal their Taylor series, just as $\sin(x)$ did, we have to be more careful with the last four functions. To see why this is, let's graph $1/(1 + x^2)$ and its Taylor polynomials $P_n(x) = 1 - x^2 + x^4 - x^6 + \cdots \pm x^n$ for $n = 2$, 4, 6, 8, 10, 12, 14, 16, 200, and 202. Since all the graphs are symmetric about the y-axis (why is this?), we draw only the graphs for positive x:



It appears that the graphs of the Taylor polynomials $P_n(x)$ approach the graph of $1/(1+x^2)$ very nicely *so long as* $x < 1$. If $x \geq 1$, though, it looks like there is no convergence, no matter how far out in the Taylor series we go. We can thus write

A Taylor series may not converge for all values of x

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \cdots \qquad \text{for } |x| < 1,$$

where the restriction on x is essential if we want to use the = sign. We say that the interval $-1 < x < 1$ is the **interval of convergence** for the Taylor series centered at $x = 0$ for $1/(1+x^2)$. Some Taylor series, like those for $\sin(x)$ and e^x, converge for all values of x—their interval of convergence is $(-\infty, \infty)$. Other Taylor series, like those for $1/(1+x^2)$ and $\ln(1-x)$, have finite intervals of convergence.

Brook Taylor (1685–1731) was an English mathematician who developed the series that bears his name in his book *Methodus incrementorum* (1715). He did not worry about questions of convergence, but used the series freely to attack many kinds of problems, including differential equations.

Remark On the one hand it is perhaps not too surprising that a function should equal its Taylor series—after all, with more and more coefficients to fiddle with, we can control more and more of the behavior of the associated polynomials. On the other hand, we are saying that a function like $\sin(x)$ or e^x has its behavior for all values of x completely determined by the value of the function and all its derivatives at a single point, so perhaps it is surprising after all!

Exercises

1. a) Suppose you wanted to use the Taylor series centered at $x = 0$ to calculate $\sin(100)$. How large does n have to be before the term $(100)^n/n!$ is less than 1?

b) If we wanted to calculate $\sin(100)$ directly using this Taylor series, we would have to go very far out before we began to approach a limit at all closely. Can you use your knowledge of the way the circular functions behave to calculate $\sin(100)$ much more rapidly (but still using the Taylor series centered at $x = 0$)? Do it.

c) Show that we can calculate the sine of any number by using a Taylor series centered at $x = 0$ *either* for $\sin(x)$ *or* for $\cos(x)$ to a suitable value 0f x between 0 and $\pi/4$.

2. a) Suppose we wanted to calculate $\ln 5$ to 7 decimal places. An obvious place to start is with the Taylor series centered at $x = 0$ for $\ln(1 - x)$:

$$-\left(x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \cdots\right)$$

with $x = -4$. What happens when you do this, and why? Try a few more values for x and see if you can make a conjecture about the interval of convergence for this Taylor series.

[Answer: The Taylor series converges for $-1 \leq x < 1$.]

b) Explain how you could use the fact that $\ln(1/A) = -\ln A$ for any real number $A > 0$ to evaluate $\ln x$ for $x > 2$. Use this to compute $\ln 5$ to 7 decimals. How far out in the Taylor series did you have to go?

c) If you wanted to calculate $\ln 1.5$, you could use the Taylor series for $\ln(1-x)$ with either $x = -1/2$, which would lead directly to $\ln 1.5$, or you could use the series with $x = 1/3$, which would produce $\ln(2/3) = -\ln 1.5$. Which method is faster, and why?

3. We can improve the speed of our calculations of the logarithm function slightly by the following series of observations:

a) Find the Taylor series centered at $u = 0$ for $\ln(1 + u)$.

[Answer: $u - u^2/2 + u^3/3 - u^4/4 + u^5/5 + \cdots$]

b) Find the Taylor series centered at $u = 0$ for

$$\ln\left(\frac{1-u}{1+u}\right).$$

(Remember that $\ln(A/B) = \ln A - \ln B$.)

c) Show that *any* $x > 0$ can be written in the form $(1 - u)/(1 + u)$ for some suitable $-1 < u < 1$.

d) Use the preceding to evaluate $\ln 5$ to 7 decimal places. How far out in the Taylor series did you have to go?

4. a) Evaluate arctan(.5) to 7 decimal places.

b) Try to use the Taylor series centered at $x = 0$ to evaluate arctan(2) directly—what happens? Remembering what the arctangent function means geometrically, can you figure out a way around this difficulty?

5. a) **Calculating π** The Taylor series for the arctangent function,

$$\arctan x = x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \cdots \pm \frac{1}{2n+1}x^{2n+1} + \cdots ,$$

lies behind many of the methods for getting lots of decimals of π rapidly. For instance, since $\tan\left(\frac{\pi}{4}\right) = 1$, we have $\frac{\pi}{4} = \arctan 1$. Use this to get a series expansion for π. How far out in the series do you have to go to evaluate π to 3 decimal places?

b) The reason the preceding approximations converged so slowly was that we were substituting $x = 1$ into the series, so we didn't get any help from the x^n terms in making the successive corrections get small rapidly. We would like to be able to do something with values of x between 0 and 1. We can do this by using the addition formula for the tangent function:

$$\tan(\alpha + \beta) = \frac{\tan\alpha + \tan\beta}{1 - \tan\alpha\tan\beta}.$$

Use this to show that

$$\frac{\pi}{4} = \arctan\left(\frac{1}{2}\right) + \arctan\left(\frac{1}{5}\right) + \arctan\left(\frac{1}{8}\right).$$

Now use the Taylor series for each of these three expressions to calculate π to 12 decimal places. How far out in the series do you have to go? Which series did you have to go the farthest out in before the 12th decimal stabilized? Why?

6. **Raising e to imaginary powers** One of the major mathematical developments of the last century was the extension of the ideas of calculus to **complex numbers**—i.e., numbers of the form $r + s\,i$, where r and s are real numbers, and i is a new symbol, defined by the property that $i \cdot i = -1$. Thus $i^3 = i^2\,i = -i$, $i^4 = i^2\,i^2 = (-1)(-1) = 1$, and so on. If we want to extend our standard functions to these new numbers, we proceed as we did in the previous section and look for the crucial *properties* of these functions

to see what they suggest. One of the key properties of e^x as we've now seen is that it possesses a Taylor series:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots .$$

But this property only involves operations of ordinary arithmetic, and so makes perfectly good sense even if x is a complex number

a) Show that if s is any real number, we must define $e^{i\,s}$ to be $\cos(s)+i\,\sin(s)$ if we want to preserve this property.

b) Show that $e^{\pi\,i} = -1$.

c) Show that if $r + s\,i$ is any complex number, we must have

$$e^{r+s\,i} = e^r(\cos s + i\,\sin s)$$

if we want complex exponentials to preserve all the right properties.

d) Find a complex number $r + si$ such that $e^{r+s\,i} = -5$.

7. **Hyperbolic trigonometric functions** The hyperbolic trigonometric functions are defined by the formulas

$$\cosh(x) = \frac{e^x + e^{-x}}{2}, \qquad \sinh(x) = \frac{e^x - e^{-x}}{2}.$$

(The names of these functions are usually pronounced "cosh" and "cinch.") In this problem you will explore some of the reasons for the adjectives *hyperbolic* and *trigonometric*.

a) Modify the Taylor series centered at $x = 0$ for e^x to find a Taylor series for $\cosh(x)$. Compare your results to the Taylor series centered at $x = 0$ for $\cos(x)$.

b) Now find the Taylor series centered at $x = 0$ for $\sinh(x)$. Compare your results to the Taylor series centered at $x = 0$ for $\sin(x)$.

c) Parts (a) and (b) of this problem should begin to explain the *trigonometric* part of the story. What about the *hyperbolic* part? Recall that the familiar trigonometric functions are called *circular functions* because, for any t, the point $(\cos t, \sin t)$ is on the unit circle with equation $x^2 + y^2 = 1$ (cf. chapter 7.2). Show that the point $(\cosh t, \sinh t)$ lies on the hyperbola with equation $x^2 - y^2 = 1$.

8. Consider the Taylor series centered at $x = 0$ for $(1 + x)^c$.

a) What does the series give if you let $c = 2$? Is this reasonable?

b) What do you get if you set $c = 3$?

c) Show that if you set $c = n$, where n is a positive integer, the Taylor series will terminate. This yields a general formula—the **binomial theorem**—that was discovered by the 12th century Persian poet and mathematician, Omar Khayyam (c. 1050–1130), and generalized by Newton to the form you have just obtained. Write out the first three and the last three terms of this formula.

d) Use an appropriate substitution for x and a suitable value for c to derive the Taylor series for $1/(1 - u)$. Does this agree with what we previously obtained?

e) Suppose we want to calculate $\sqrt{17}$. We might try letting $x = 16$ and $c = 1/2$ and using the Taylor series for $(1 + x)^c$. What happens when you try this?

f) We can still use the series to help us, though, if we are a little clever and write

$$\sqrt{17} = \sqrt{16+1} = \sqrt{16\left(1+\frac{1}{16}\right)} = \sqrt{16}\cdot\sqrt{1+\frac{1}{16}} = 4\cdot\sqrt{1+\frac{1}{16}}.$$

Now apply the series using $x = 1/16$ to evaluate $\sqrt{17}$ to 7 decimal place accuracy. How many terms does it take?

g) Use the same kind of trick to evaluate $\sqrt[3]{30}$.

Evaluating Taylor series rapidly Suppose we wanted to plot the Taylor polynomial of degree 11 associated with $\sin(x)$. For each value of x, then, we would have to evaluate

$$P_{11}(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \frac{x^{11}}{11!}.$$

Since the length of time it takes the computer to evaluate an expression like this is roughly proportional to the number of multiplications and divisions involved (additions and subtractions, by comparison, take a negligible amount of time), let's see how many of these operations are needed to evaluate $P_{11}(x)$. To calculate x^{11} requires 10 multiplications, while 11! requires

9 (if we are clever and don't bother to multiply by 1 at the end!), so the evaluation of the term $x^{11}/11!$ will require a total of 20 operations (counting the final division). Similarly, evaluating $x^9/9!$ requires 16 operations, $x^7/7!$ requires 12, on down to $x^3/3!$, which requires 4. Thus the total number of multiplications and divisions needed is

$$4 + 8 + 12 + 16 + 20 = 60.$$

This is not too bad, although if we were doing this for many different values of x, which would be the case if we wanted to graph $P_{11}(x)$, this would begin to add up. Suppose, though, that we wanted to graph something like $P_{51}(x)$ or $P_{101}(x)$. By the same analysis, evaluating $P_{51}(x)$ for a single value of x would require

$$4 + 8 + 12 + 16 + 20 + 24 + \cdots + 96 + 100 = 1300$$

multiplications and divisions, while evaluation of $P_{101}(x)$ would require 5100 operations. Thus it would take roughly 20 times as long to evaluate $P_{51}(x)$ as it takes to evaluate $P_{11}(x)$, while $P_{101}(x)$ would take about 85 times as long.

9. Show that, in general, the number of multiplications and divisions needed to evaluate $P_n(x)$ is roughly $n^2/2$.

We can be clever, though. Note that $P_{11}(x)$ can be written as

$$x\left(1-\frac{x^2}{2\cdot 3}\left(1-\frac{x^2}{4\cdot 5}\left(1-\frac{x^2}{6\cdot 7}\left(1-\frac{x^2}{8\cdot 9}\left(1-\frac{x^2}{10\cdot 11}\right)\right)\right)\right)\right).$$

10. How many multiplications and divisions are required to evaluate this expression?
[Answer: $3 + 4 + 4 + 4 + 4 + 1 = 20$.]

11. Thus this way of evaluating $P_{11}(x)$ is roughly three times as fast, a modest saving. How much faster is it if we use this method to evaluate $P_{51}(x)$?
[Answer: The old way takes roughly 13 times as long.]

12. Find a general formula for the number of multiplications and divisions needed to evaluate $P_n(x)$ using this way of grouping.

Finally, we can extend these ideas to reduce the number of operations even further, so that evaluating a polynomial of degree n requires only n multiplications, as follows. Suppose we start with a polynomial

$$p(x) = a_0 + a_1\,x + a_2\,x^2 + a_3\,x^3 + \cdots + a_{n-1}\,x^{n-1} + a_n\,x^n.$$

We can rewrite this as

$$a_0 + x\,(a_1 + x\,(a_2 + \ldots + x\,(a_{n-2} + x\,(a_{n-1} + a_n\,x))\ldots))\,.$$

You should check that with this representation it requires only n multiplications to evaluate $p(x)$ for a given x.

13. a) Write two computer programs to evaluate the 300th degree Taylor polynomial centered at $x = 0$ for e^x, with one of the programs being the obvious, standard way, and the second program being this method given above. Evaluate $e^1 = e$ using each program, and compare the length of time required.

b) Use these two programs to graph the 300th degree Taylor polynomial for e^x over the interval $[0, 2]$, and compare times.

10.4 Power Series and Differential Equations

So far, we have begun with functions we already know, in the sense of being able to calculate the value of the function and all its derivatives at at least one point. This in turn allowed us to write down the corresponding Taylor series. Often, though, we don't even have this much information about a function. In such cases it is frequently useful to assume that there is some infinite polynomial—called a **power series**—which represents the function, and then see if we can determine the coefficients of the polynomial.

This technique is especially useful in dealing with differential equations. To see why this is the case, think of the alternatives. If we can approximate the solution $y = y(x)$ to a certain differential equation to an acceptable degree of accuracy by, say, a 20-th degree polynomial, then the only storage space required is the insignificant space taken to keep track of the 21 coefficients. Whenever we want the value of the solution for a given value of x, we can then get a quick approximation by evaluating the polynomial at x. Other alternatives are much more costly in time or in space. We could use Euler's method to grind out the solution at x, but, as you've already discovered, this can be a slow and tedious process. Another option is to calculate lots of values and store them in a table in the computer's memory. This not only takes up a lot of memory space, but it also only gives values for a finite set of values of x, and is not much faster than evaluating a polynomial. Until 30 years ago, the table approach was the standard one—all scientists and mathematicians had a handbook of mathematical functions containing hundreds of pages of numbers giving the values of every function they might need.

To see how this can happen, let's first look at a familiar differential equation whose solutions we already know:

$$y' = y.$$

Of course, we know by now that the solutions are $y = ae^x$ for an arbitrary constant a (where $a = y(0)$). Suppose, though, that we didn't already know how to solve this differential equation. We might see if we can find a power series of the form

$$y = a_0 + a_1\,x + a_2\,x^2 + a_3\,x^3 + \cdots + a_n\,x^n + \cdots$$

that solves the differential equation. Can we, in fact, determine values for the coefficients $a_0, a_1, a_2, \cdots, a_n, \cdots$ that will make $y' = y$?

Using the rules for differentiation, we have

$$y' = a_1 + 2a_2\,x + 3a_3\,x^2 + 4a_4\,x^3 + \cdots + na_n\,x^{n-1} + \cdots .$$

Two polynomials are equal if and only if the coefficients of corresponding powers of x are equal. Therefore, if $y' = y$, it would have to be true that

$$\begin{aligned} a_1 &= a_0 \\ 2a_2 &= a_1 \\ 3a_3 &= a_2 \\ &\ \vdots \\ na_n &= a_{n-1} \\ &\ \vdots \end{aligned}$$

Therefore the values of a_1, a_2, a_3, ... are not arbitrary; indeed, each is determined by the preceding one. Equations like these—which deal with a sequence of quantities and relate each term to those earlier in the sequence—are called **recursion relations**. These recursion relations permit us to express every a_n in terms of a_0:

$$\begin{array}{llll} a_1 = a_0 & & & \\ a_2 = \dfrac{1}{2}\,a_1 & & & = \dfrac{1}{2}\,a_0 \\ a_3 = \dfrac{1}{3}\,a_2 & = \dfrac{1}{3}\cdot\dfrac{1}{2}\,a_0 & & = \dfrac{1}{3!}\,a_0 \\ a_4 = \dfrac{1}{4}\,a_3 & = \dfrac{1}{4}\cdot\dfrac{1}{3!}\,a_0 & & = \dfrac{1}{4!}\,a_0 \\ \ \vdots & \ \vdots & & \ \vdots \\ a_n = \dfrac{1}{n}\,a_{n-1} & = \dfrac{1}{n}\cdot\dfrac{1}{(n-1)!}\,a_0 & & = \dfrac{1}{n!}\,a_0 \\ \ \vdots & \ \vdots & & \ \vdots \end{array}$$

Notice that a_0 remains "free": there is no equation that determines its value. Thus, without additional information, a_0 is *arbitrary.* The series for y now becomes

$$y = a_0 + a_0\,x + \frac{1}{2!}\,a_0\,x^2 + \frac{1}{3!}\,a_0\,x^3 + \cdots + \frac{1}{n!}\,a_0\,x^n + \cdots$$

or

$$y = a_0\left[1 + x + \frac{1}{2!}\,x^2 + \frac{1}{3!}\,x^3 + \cdots + \frac{1}{n!}\,x^2 + \cdots\right].$$

But the series in square brackets is just the Taylor series for e^x—we have derived the Taylor series from the differential equation alone, without using any of the other properties of the exponential function. Thus, we again find that the solutions of the differential equation $y' = y$ are

$$y = a_0 e^x,$$

where a_0 is an arbitrary constant. Notice that $y(0) = a_0$, so the value of a_0 will be determined if the initial value of y is specified.

Note In general, once we have derived a power series expression for a function, that power series will also be the Taylor series for that function. Although the two series are the same, the term Taylor series is typically reserved for those settings where we were able to evaluate the derivatives through some other means, as in the preceding section.

Bessel's Equation

For a new example, let's look at a differential equation that arises in an enormous variety of physical problems (wave motion, optics, the conduction of electricity and of heat and fluids, and the stability of columns, to name a few):

$$x^2 \cdot y'' + x \cdot y' + (x^2 - p^2) \cdot y = 0\,.$$

This is called the **Bessel equation of order $\boldsymbol{p}$**. Here p is a parameter specified in advance, so we will really have a different set of solutions for each value of p. To determine a solution completely, we will also need to specify the initial values of $y(0)$ and $y'(0)$. The solutions of the Bessel equation of order p are called **Bessel functions of order $\boldsymbol{p}$**, and the solution for a given value of p (together with particular initial conditions which needn't concern us here) is written $J_p(x)$. In general, there is no formula for a Bessel function in terms of simpler functions (although it turns out that a few special cases like $J_{1/2}(x), J_{3/2}(x), \ldots$ can be expressed relatively simply). To evaluate such a function we could use Euler's method, or we could try to find a power series solution.

Friedrich Wilhelm Bessel (1784–1846) was a German astronomer who studied the functions that now bear his name in his efforts to analyze the perturbations of planetary motions, particularly those of Saturn.

Consider the Bessel equation with $p = 0$. We are thus trying to solve the differential equation

$$x^2 \cdot y'' + x \cdot y' + x^2 \cdot y = 0.$$

By dividing by x, we can simplify this a bit to

$$x \cdot y'' + y' + x \cdot y = 0.$$

Let's look for a power series expansion

$$y = b_0 + b_1\,x + b_2\,x^2 + b_3\,x^3 + \cdots .$$

We have

$$y' = b_1 + 2b_2x + 3b_3x^2 + 4b_4x^3 + \cdots + (n+1)b_{n+1}x^n + \cdots$$

and

$$y'' = 2b_2 + 6b_3x + 12b_4x^2 + 20b_5x^3 + \cdots + (n+2)(n+1)b_{n+2}x^n + \cdots ,$$

We can now use these expressions to calculate the series for the combination that occurs in the differential equation:

$$\begin{array}{rcccccc}
xy'' = & & 2b_2\,x & + & 6b_3\,x^2 & + \cdots \\
y' = & b_1\ + & 2b_2\,x & + & 3b_3\,x^2 & + \cdots \\
xy = & & b_0\,x & + & b_1\,x^2 & + \cdots \\
\hline
xy'' + y' + xy = & b_1\ + & (4b_2 + b_0)x & + & (9b_3 + b_1)x^2 & + \cdots
\end{array}$$

In general, the coefficient of x^n in the combination will be

Finding the coefficient of x^n

$$(n+1)n\,b_{n+1} + (n+1)\,b_{n+1} + b_{n-1} = (n+1)^2\,b_{n+1} + b_{n-1}.$$

If the power series y is to be a solution to the original differential equation, the infinite series for $xy'' + y' + xy$ must equal 0. This in turn means that every coefficient of that series must be 0. We thus get

$$\begin{aligned}
b_1 &= 0,\\
4b_2 + b_0 &= 0,\\
9b_3 + b_1 &= 0,\\
&\ \vdots\\
n^2b_n + b_{n-2} &= 0.\\
&\ \vdots
\end{aligned}$$

If we now solve these recursively as before, we see first off that since $b_1 = 0$, it must also be true that

$$b_k = 0 \qquad \text{for every odd } k.$$

For the even coefficients we have

$$\begin{aligned} b_2 &= -\frac{1}{2^2}b_0, \\ b_4 &= -\frac{1}{4^2}b_2 = \frac{1}{2^2 4^2}b_0, \\ b_6 &= -\frac{1}{6^2}b_4 = -\frac{1}{2^2 4^2 6^2}b_0, \end{aligned}$$

and, in general,

$$b_{2n} = \pm\frac{1}{2^2 4^2 6^2 \cdots (2n)^2}b_0 = \pm\frac{1}{2^{2n}(n!)^2}b_0.$$

Thus any function y satisfying the Bessel equation of order 0 must be of the form

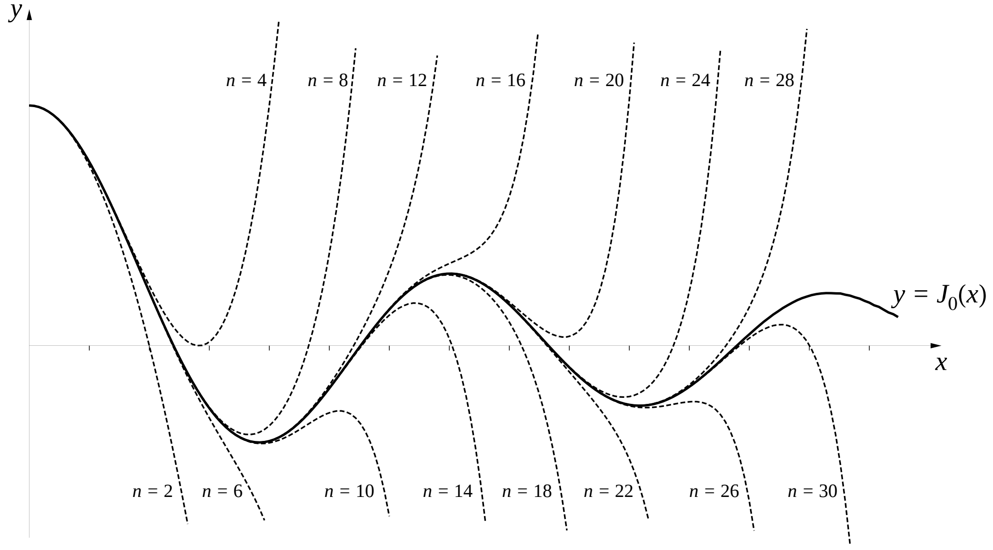
$$y = b_0\left(1 - \frac{x^2}{2^2} + \frac{x^4}{2^4(2!)^2} - \frac{x^6}{2^6(3!)^2} + \cdots\right).$$

In particular, if we impose the initial condition $y(0) = 1$ (which requires that $b_0 = 1$), we get the 0-th order Bessel function $J_0(x)$:

$$J_0(x) = 1 - \frac{x^2}{4} + \frac{x^4}{64} - \frac{x^6}{2304} + \frac{x^8}{147456} + \cdots.$$

The graph of the Bessel function J_0

Here is the graph of $J_0(x)$ together with the polynomial approximations of degree 2, 4, 6, ..., 30 over the interval $[0, 14]$:



The graph of J_0 is suggestive: it appears to be oscillatory, with decreasing amplitude. Both observations are correct: it can in fact be shown that J_0 has infinitely many zeroes, spaced roughly π units apart, and that $\lim_{x\to\infty} J_0(x) = 0$.

The S-I-R Model One More Time

In exactly the same way, we can find power series solutions when there are several interacting variables involved. Let's look at the example we've considered at a number of points in this text to see how this works. In the S-I-R model we basically wanted to solve the system of equations

The S-I-R model

$$\begin{aligned} S' &= -aSI, \\ I' &= aSI - bI, \\ R' &= bI, \end{aligned}$$

where a and b were parameters depending on the specific situation. Let's look for solutions of the form

$$\begin{aligned} S &= s_0 + s_1\,t + s_2\,t^2 + s_3\,t^3 + \cdots\,, \\ I &= i_0 + i_1\,t + i_2\,t^2 + i_3\,t^3 + \cdots\,, \\ R &= r_0 + r_1\,t + r_2\,t^2 + r_3\,t^3 + \cdots\,. \end{aligned}$$

If we put these series in the equation $S' = -a\,S\,I$, we get

$$\begin{aligned} s_1 + 2s_2t + 3s_3t^2 + \cdots &= -a(s_0 + s_1t + s_2t^2 + \cdots)(i_0 + i_1t + i_2t^2 + \cdots) \\ &= -a(s_0i_0 + (s_0i_1 + s_1i_0)t + (s_0i_2 + s_1i_1 + s_2i_0)t^2 + \cdots). \end{aligned}$$

As before, if the two sides of the differential equation are to be equal, the coefficients of corresponding powers of t must be equal:

Finding the coefficients of the power series for $S(t)$

$$\begin{aligned} s_1 &= -as_0i_0, \\ 2s_2 &= -a(s_0i_1 + s_1i_0), \\ 3s_3 &= -a(s_0i_2 + s_1i_1 + s_2i_0), \\ &\vdots \\ ns_n &= -a(s_0i_{n-1} + s_1i_{n-2} + \ldots + s_{n-2}i_1 + s_{n-1}i_0) \\ &\vdots \end{aligned}$$

Recursion again

While this looks messy, it has the crucial recursive feature—each s_k is expressed in terms of previous terms. That is, if we knew all the s and the i coefficients out through the coefficients of, say, t^6 in the series for S and I, we could immediately calculate s_7. We again have a *recursion relation.*

Finding the power series for $I(t)$

We could expand the equation $I' = aSI - bI$ in the same way, and get recursion relations for the coefficients i_k. In this model, though, there is a shortcut if we observe that since $S' = -aSI$, and since $I' = aSI - bI$, we have $I' = -S' - bI$. If we substitute the power series in this expression and equate coefficients, we get

$$ni_n = -ns_n - bi_{n-1},$$

which leads to

$$i_n = -s_n - \frac{b}{n} i_{n-1}$$

—so if we know s_n and i_{n-1}, we can calculate i_n.

We are now in a position to calculate the coefficients as far out as we like. For we will be given values for a and b when we are given the model. Moreover, since $s_0 = S(0) =$ the initial S-population, and $i_0 = I(0) =$ the initial I-population, we will also typically be given these values as well. But knowing s_0 and i_0, we can determine s_1 and then i_1. But then, knowing these values, we can determine s_2 and then i_2, and so on. Since the arithmetic is tedious, this is obviously a place for a computer. Here is a program that calculates the first 50 coefficients in the power series for $S(t)$ and $I(t)$:

Program: SIRSERIES

```
DIM S(0 to 50), I(0 to 50)
a = .00001
b = 1/14
S(0) = 45400
I(0) = 2100
FOR k = 1 TO 50
     Sum = 0
     FOR j = 0 TO k - 1
          Sum = Sum + S(j) * I(k - j - 1)
     NEXT j
     S(k) = -a * SUM/k
     I(k) = -S(k) - b * I(k - 1)/k
NEXT k
```

Comment: The opening command in this program introduces a new feature. It notifies the computer that the variables `S` and `I` are going to be **arrays**—strings of numbers—and that each array will consist of 51 elements. The element `S(k)` corresponds to what we have been calling s_k. The integer `k` is called the **index** of the term in the array. The indices in this program run from 0 to 50.

The effect of running this program is thus to create two 51-element arrays, `S` and `I`, containing the coefficients of the power series for S and I out to degree 50. If we just wanted to see these coefficients, we could have the computer list them. Here are the first 35 coefficients for S (read across the rows):

```
        45400         −953.4      −172.3611      −17.982061      −.86969127
 5.4479852e-2   1.5212707e-2   1.4463108e-3    4.3532884e-5   −7.9100481e-6
−1.4207959e-6  −1.0846994e-7  −6.512610e-10    9.304633e-10    1.256507e-10
 7.443310e-12  −2.191966e-13  −9.787285e-14  −1.053428e-14   −4.382620e-16
 4.230290e-17   9.548369e-18   8.321674e-19    1.760392e-20   −5.533369e-21
−8.770972e-22  −6.101928e-23   3.678170e-25    6.193375e-25    7.627253e-26
 4.011923e-27  −1.986216e-28  −6.318305e-29   −6.271724e-30   −2.150100e-31
```

Thus the power series for S begins

$$45400-953.4t-172.3611t^2-17.982061t^3-.86969127t^4+\cdots-2.15010\times10^{-31}t^{34}+\cdots$$

In the same fashion, we find that the power series for I begins

$$2100+803.4t+143.66824t^2+14.561389t^3+.60966648t^4+\cdots+2.021195\times10^{-31}t^{34}+\cdots$$

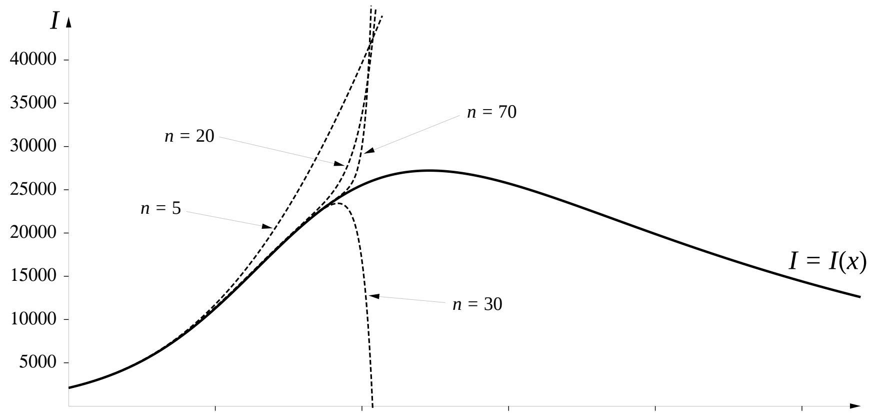
Extending SIRSERIES to graph S and I

If we now wanted to graph these polynomials over, say, $0 \le t \le 10$, we can do it by adding the following lines to SIRSERIES. We first define a couple of short subroutines `SUS` and `INF` to calculate the polynomial approximations for $S(t)$ and $I(t)$ using the coefficients we've derived in the first part of the program. (Note that these subroutines calculate polynomials in the straightforward, inefficient way. If you did the exercises in section 3 which developed techniques for evaluating polynomials rapidly, you might want to modify these subroutines to take advantage of the increased speed available.) Remember, too, that you will need to set up the graphics at the beginning of the program to be able to plot.

Extension to **SIRSERIES**

```
DEF SUS(x)
     Sum = S(0)
     FOR j = 1 TO 50
         Sum = Sum + S(j) * x^j
     NEXT j
     SUS = Sum
END DEF
DEF INF(x)
     Sum = I(0)
     FOR j = 1 TO 50
         Sum = Sum + I(j) * x^j
     NEXT j
     INF = Sum
END DEF
FOR x = 0 TO 10 STEP .01
     Plot the line from (x, SUS(x)) to (x + .01, SUS(x + .01))
     Plot the line from (x, INF(x)) to (x + .01, INF(x + .01))
NEXT x
```

Here is the graph of $I(t)$ over a 25-day period, together with the polynomial approximations of degree 5, 20, 30, and 70.

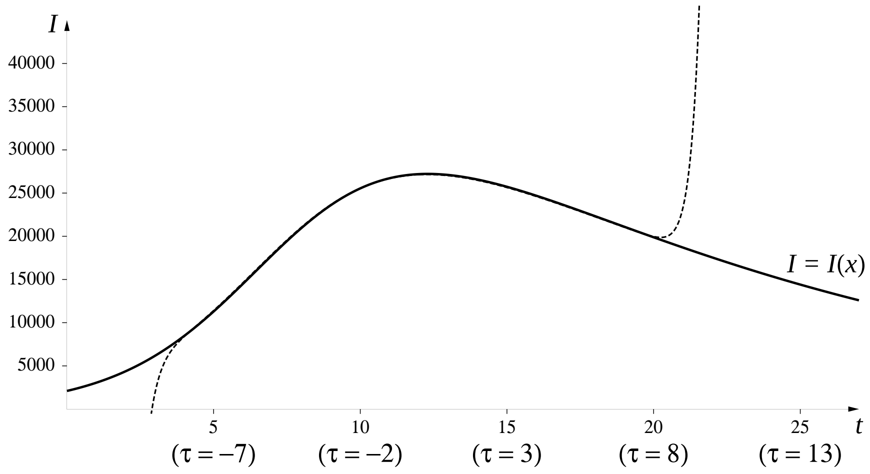


Note that these polynomials appear to converge to $I(t)$ only out to values of t around 10. If we needed polynomial approximations beyond that point, we could shift to a different point on the curve, find the values of I and S there by Euler's method, then repeat the above process. For instance, when

$t = 12$, we get by Euler's method that $S(12) = 7670$ and $I(12) = 27,136$. If we now shift our clock to measure time in terms of $\tau = t - 12$, we get the following polynomial of degree 30:

$$27136+143.0455\tau-282.0180\tau^2+23.5594\tau^3+.4548\tau^4+\cdots+1.2795\times 10^{-25}\tau^{30}$$

Here is what the graph of this polynomial looks like when plotted with the graph of I. On the horizontal axis we list the t-coordinates with the corresponding τ-coordinates underneath.



The interval of convergence seems to be approximately $4 < t < 20$. Thus if we combine this polynomial with the 30-th degree polynomial from the previous graph, we would have very accurate approximations for I over the entire interval $[0, 20]$.

Exercises

1. Find power series solutions of the form

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n + \cdots$$

for each of the following differential equations.

a) $y' = 2xy$.

b) $y' = 3x^2y$.

c) $y'' + xy = 0$.

d) $y'' + xy' + y = 0$.

2. a) Find power series solutions to the differential equation $y'' = -y$. Start with

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n + \cdots .$$

Notice that, in the recursion relations you obtain, the coefficients of the even terms are completely independent of the coefficients of the odd terms. This means you can get two separate power series, one with only even powers, with a_0 as arbitrary constant, and one with only odd powers, with a_1 as arbitrary constant.

b) The two power series you obtained in part a) are the Taylor series centered at $x = 0$ of two familiar functions; which ones? Verify that these functions do indeed satisfy the differential equation $y'' = -y$.

3. a) Find power series solutions to the differential equation $y'' = y$. As in the previous problem, the coefficients of the even terms depend only on a_0, and the coefficients of the odd terms depend only on a_1. Write down the two series, one with only even powers and a_0 as an arbitrary constant, and one with only odd powers, with a_1 as an arbitrary constant.

b) The two power series you obtained in part a) are the Taylor series centered at $x = 0$ of two *hyperbolic trigonometric functions* (see the exercises in section 3). Verify that these functions do indeed satisfy the differential equation $y'' = y$.

4. a) Find power series solutions to the differential equation $y' = xy$, starting with

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n + \cdots .$$

What recursion relations do you get? Is $a_1 = a_3 = a_5 = \cdots = 0$?

b) Verify that

$$y = e^{x^2/2}$$

satisfies the differential equation $y' = xy$. Find the Taylor series for this function and compare it with the series you obtained in a) using the recursion relations.

5. **The Bessel Equation.**

a) Take $p = 1$. The solution satisfying the initial condition $y' = 1/2$ when $x = 0$ is defined to be the first order Bessel function $J_1(x)$. (It will turn out that y has to be 0 when $x = 0$, so we don't have to specify the initial value of y; we have no choice in the matter.) Find the first five terms of the power series expansion for $J_1(x)$. What is the coefficient of x^{2n+1}?

b) Show by direct calculation from the series for J_0 and J_1 that

$$J_0' = -J_1.$$

c) To see, from another point of view, that $J_0' = -J_1$, take the equation

$$x \cdot J_0'' + J_0' + x \cdot J_0 = 0$$

and differentiate it. By doing some judicious cancelling and rearranging of terms, show that

$$x^2 \cdot (J_0')'' + x \cdot (J_0')' + (x^2 - 1)(J_0') = 0.$$

This demonstrates that J_0' is a solution of the Bessel equation with $p = 1$.

6. a) When we found the power series expansion for solutions to the 0-th order Bessel equation, we found that all the odd coefficients had to be 0. In particular, since b_1 is the value of y' when $x = 0$, we are saying that all solutions have to be flat at $x = 0$. This should bother you a bit. Why can't you have a solution, say, that satisfies $y = 1$ and $y' = 1$ when $x = 0$?

b) You might get more insight on what's happening by using Euler's method, starting just a little to the right of the origin and moving left. Use Euler's method to sketch solutions with the initial values

i.	$y = 2$	$y' = 1$	when	$x = 1$,
ii.	$y = 1.1$	$y' = 1$	when	$x = .1$,
iii.	$y = 1.01$	$y' = 1$	when	$x = .01$.

What seems to happen as you approach the y-axis?

7. **Legendre's differential equation**

$$(1 - x^2)y'' - 2xy' + \ell(\ell + 1)y = 0$$

arises in many physical problems—for example, in quantum mechanics, where its solutions are used to describe certain orbits of the electron in a hydrogen atom. In that context, the parameter ℓ is called the *angular momentum* of the electron; it must be either an integer or a "half-integer" (i.e., a number like 3/2). Quantum theory gets its name from the fact that numbers like the angular momentum of the electron in the hydrogen atom are "quantized", that is, they cannot have just any value, but must be a multiple of some "quantum"—in this case, the number 1/2.

a) Find power series solutions of Legendre's equation.

b) Quantization of angular momentum has an important consequence. Specifically, when ℓ is an integer it is possible for a series solution to stop—that is, to be a polynomial. For example, when $\ell = 1$ and $a_0 = 0$ the series solution is just $y = a_1 x$—all higher order coefficients turn out to be zero. Find polynomial solutions to Legendre's equation for $\ell = 0, 2, 3, 4$, and 5 (consider $a_0 = 0$ or $a_1 = 0$). These solutions are called, naturally enough, **Legendre polynomials**.

8. It turns out that the power series solutions to the S-I-R model have a finite interval of convergence. By plotting the power series solutions of different degrees against the solutions obtained by Euler's method, estimate the interval of convergence.

9. a) **Logistic Growth** Find the first five terms of the power series solution to the differential equation

$$y' = y(1 - y).$$

Note that this is just the logistic equation, where we have chosen our units of time and of quantities of the species being studied so that the carrying capacity is 1 and the intrinsic growth rate is 1.

b) Using the initial condition $y = .1$ when $x = 0$, plot this power series solution on the same graph as the solution obtained by Euler's method. How do they compare?

c) Do the same thing with initial conditions $y = 2$ when $x = 0$.

10.5 Convergence

We have written expressions such as

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \cdots ,$$

meaning that for any value of x the series on the right will converge to $\sin(x)$. There are a couple of issues here. The first is, what do we even mean when we say the series "converges", and how do we prove it converges to $\sin(x)$? If x is small, we can convince ourselves that the statement is true just by trying it. If x is large, though, say $x = 100^{100}$, it would be convenient to have a more general method for proving the stated convergence. Further, we have the example of the function $1/(1 + x^2)$ as a caution—it seemed to converge for small values of x ($|x| < 1$), but not for large values.

Convergence means essentially that "decimals stabilize"

Let's first clarify what we mean by convergence. It is, essentially, the intuitive notion of "decimals stabilizing" that we have been using all along. To make explicit what we've been doing, let's write a "generic" series

$$b_0 + b_1 + b_2 + \cdots = \sum_{m=0}^{\infty} b_m .$$

When we evaluated such a series, we looked at the **partial sums**

$$\begin{aligned}
S_1 &= b_0 + b_1 &&= \sum_{m=0}^{1} b_m, \\
S_2 &= b_0 + b_1 + b_2 &&= \sum_{m=0}^{2} b_m, \\
S_3 &= b_0 + b_1 + b_2 + b_3 &&= \sum_{m=0}^{3} b_m, \\
&\vdots && \vdots \\
S_n &= b_0 + b_1 + b_2 + \ldots + b_n &&= \sum_{m=0}^{n} b_m, \\
&\vdots && \vdots
\end{aligned}$$

Typically, when we calculated a number of these partial sums, we noticed that beyond a certain point they all seemed to agree on, say, the first 8 decimal places. If we kept on going, the partial sums would agree on the first 9 decimals, and, further on, on the first 10 decimals, etc. This is precisely what we mean by convergence:

The infinite series

$$b_0 + b_1 + b_2 + \cdots = \sum_{m=0}^{\infty} b_m$$

converges if, no matter how many decimal places are specified, it is always the case that the partial sums eventually agree to at least this many decimal places.

Put more formally, we say the series converges if, given any number D of decimal places, it is always possible to find an integer N_D such that if k and n are both greater than N_D, then S_k and S_n agree to at least D decimal places.

The number defined by these stabilizing decimals is called the **sum** of the series.

If a series does not converge, we say it **diverges**.

What it means for an infinite sum to converge

In other words, for me to prove to you that the Taylor series for $\sin(x)$ converges at $x = 100^{100}$, you would specify a certain number of decimal places, say 5000, and I would have to be able to prove to you that if you took partial sums with enough terms, they would all agree to at least 5000 decimals. Moreover, I would have to be able to show the same thing happens if you specify *any* number of decimal places you want agreement on.

How can this be done? It seems like an enormously daunting task to be able to do for any series. We'll tackle this challenge in stages. First we'll see what goes wrong with some series that don't converge—divergent series. Then we'll look at a particular convergent series—the **geometric series**—that's relatively easy to analyze. Finally, we will look at some more general rules that will guarantee convergence of series like those for the sine, cosine, and exponential functions.

Divergent Series

Suppose we have an infinite series

$$b_0 + b_1 + b_2 + \cdots = \sum_{m=0}^{\infty} b_m,$$

and consider two successive partial sums, say

$$S_{122} = b_0 + b_1 + b_2 + \ldots + b_{122} = \sum_{m=0}^{122} b_m$$

and

$$S_{123} = b_0 + b_1 + b_2 + \ldots + b_{122} + b_{123} = \sum_{m=0}^{123} b_m.$$

Note that these two sums are the same, except that the sum for S_{123} has one more term, b_{123}, added on. Now suppose that S_{122} and S_{123} agree to 19 decimal places. In section 2 we defined this to mean $|S_{123} - S_{122}| < .5 \times 10^{-19}$. But since $S_{123} - S_{122} = b_{123}$, this means that $|b_{123}| < .5 \times 10^{-19}$. To phrase this more generally,

> Two successive partial sums, S_n and S_{n+1}, agree out to k decimal places if and only if $|b_{n+1}| < .5 \times 10^{-k}$.

A necessary condition for convergence

But since our definition of convergence required that we be able to fix any specified number of decimals provided we took partial sums lengthy enough, it must be true that *if the series converges*, the individual terms b_k must become arbitrarily small if we go out far enough. Intuitively, you can think of the partial sums S_k as being a series of approximations to some quantity. The term b_{k+1} can be thought of as the "correction" which is added to S_k to produce the next approximation S_{k+1}. Clearly, if the approximations are eventually becoming good ones, the corrections made should become smaller and smaller. We thus have the following necessary condition for convergence:

> If the infinite series $b_0 + b_1 + b_2 + \cdots = \displaystyle\sum_{m=0}^{\infty} b_m$ converges,
>
> then $\displaystyle\lim_{k\to\infty} b_k = 0$.

Necessary and *sufficient* mean different things

Remark: It is important to recognize what this criterion does and does not say—it is a **necessary** condition for convergence (i.e., every convergent sequence has to satisfy the condition $\lim_{k\to\infty} b_k = 0$)—but it is not a **sufficient** condition for convergence (i.e., there are some divergent sequences that also have the property that $\lim_{k\to\infty} b_k = 0$). The criterion is usually used to detect some divergent series, and is more useful in the following form (which you should convince yourself is equivalent to the preceding):

If $\lim\limits_{k\to\infty} b_k \neq 0$, (either because the limit doesn't exist at all, or it equals something besides 0), then the infinite series

$$b_0 + b_1 + b_2 + \cdots = \sum_{m=0}^{\infty} b_m \qquad \text{diverges.}$$

Detecting divergent series

This criterion allows us to detect a number of divergent series right away. For instance, we saw earlier that the statement

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \cdots$$

appeared to be true only for $|x| < 1$. Using the remarks above, we can see why this series has to diverge for $|x| \geq 1$. If we write $1 - x^2 + x^4 - x^6 + \cdots$ as $b_0 + b_1 + b_2 + \cdots$, we see that $b_k = (-1)^k x^{2k}$. Clearly b_k does not go to 0 for $|x| \geq 1$—the successive "corrections" we make to each partial sum just become larger and larger, and the partial sums will alternate more and more wildly from a huge positive number to a huge negative number. Hence the series converges *at most* for $-1 < x < 1$. We will see in the next subsection how to prove that it really does converge for all x in this interval.

Using exactly the same kind of argument, we can show that the following series also diverge for $|x| > 1$:

$f(x)$	Taylor series for $f(x)$
$\ln(1-x)$	$-\left(x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \cdots\right)$
$\frac{1}{1-x}$	$1 + x + x^2 + x^3 + \cdots$
$(1+x)^c$	$1 + cx + \frac{c(c-1)}{2!}x^2 + \frac{c(c-1)(c-2)}{3!}x^3 + \cdots$
$\arctan x$	$x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots \pm \frac{x^{2n+1}}{2n+1}$

The details are left to the exercises. While these common series all happen to diverge for $|x| > 1$, it is easy to find other series that diverge for $|x| > 2$ or $|x| > 17$ or whatever—see the exercises for some examples.

The Harmonic Series

We stated earlier in this section that simply knowing that the individual terms b_k go to 0 for large values of k does not guarantee that the series

$$b_0 + b_1 + b_2 + \cdots$$

will converge. Essentially what can happen is that the b_k go to 0 slowly enough that they can still accumulate large values. The classic example of such a series is the **harmonic series**:

An important counterexample

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots = \sum_{i=1}^{\infty} \frac{1}{i}.$$

It turns out that this series just keeps getting larger as you add more terms. It is eventually larger than 1000, or 1 million, or 100^{100} or .... This fact is established in the exercises. A suggestive argument, though, can be quickly given by observing that the harmonic series is just what you would get if you substituted $x = 1$ into the power series

$$x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \cdots .$$

But this is just the Taylor series for $-\ln(1 - x)$, and if we substitute $x = 1$ into this we get $-\ln 0$, which isn't defined. Also, $\lim_{x \to 0} -\ln x = +\infty$.

The Geometric Series

A series occurring frequently in a wide range of contexts is the **geometric series**

$$G(x) = 1 + x + x^2 + x^3 + x^4 + \cdots ,$$

This is also a sequence we can analyze completely and rigorously in terms of its convergence. It will turn out that we can then reduce the analysis of the convergence of several other sequences to the behavior of this one.

By the analysis we performed above, if $|x| \geq 1$ the individual terms of the series clearly don't go to 0, and the series therefore diverges. What about the case where $|x| < 1$?

To avoid divergence, $|x|$ must be less than 1

The starting point is the partial sums. A typical partial sum looks like:

$$S_n = 1 + x + x^2 + x^3 + \cdots + x^n\,.$$

A simple expression for the partial sum S_n

This is a finite number; we must find out what happens to it as n grows without bound. Since S_n is finite, we can calculate with it. In particular,

$$xS_n = x + x^2 + x^3 + \cdots + x^n + x^{n+1}.$$

Subtracting the second expression from the first, we get

$$S_n - xS_n = 1 - x^{n+1},$$

and thus (if $x \neq 1$)

$$S_n = \frac{1 - x^{n+1}}{1 - x}.$$

(What is the value of S_n if $x = 1$?)

This is a handy, compact form for the partial sum. Let us see what value it has for various values of x. For example, if $x = 1/2$, then

$$\begin{array}{rcccccccc} n: & 1 & 2 & 3 & 4 & 5 & 6 & \cdots \rightarrow & \infty \\ S_n: & 1 & \frac{3}{2} & \frac{7}{4} & \frac{15}{8} & \frac{31}{16} & \frac{63}{32} & \cdots \rightarrow & 2 \end{array}$$

Finding the limit of S_n as $n \to \infty$

It appears that as $n \to \infty$, $S_n \to 2$. Can we see this algebraically?

$$\begin{aligned} S_n &= \frac{1 - (1/2)^{n+1}}{1 - \frac{1}{2}} \\ &= \frac{1 - (1/2)^{n+1}}{1/2} \\ &= 2 \cdot (1 - (1/2)^{n+1}) = 2 - (1/2)^n. \end{aligned}$$

As $n \to \infty$, $(1/2)^n \to 0$, so the values of S_n become closer and closer to 2. The series converges, and its sum is 2.

Summing another geometric series

Similarly, when $x = -1/2$, the partial sums are

$$S_n = \frac{1 - (-1/2)^{n+1}}{3/2} = \frac{2}{3}\left(1 \pm \frac{1}{2^{n+1}}\right).$$

The presence of the ± sign does not alter the outcome: since $(1/2^{n+1}) \to 0$, the partial sums converge to $2/3$. Therefore, we can say the series converges and its sum is $2/3$.

In exactly the same way, though, for any x satisfying $|x| < 1$ we have

$$S_n = \frac{1 - x^{n+1}}{1 - x} = \frac{1}{1 - x}(1 - x^{n+1}),$$

and as $n \to \infty$, $x^{n+1} \to 0$. Therefore, $S_n \to 1/(1 - x)$. Thus the series converges, and its sum is $1/(1 - x)$.

To summarize, we have thus proved that

Convergence and divergence of the geometric series

The geometric series

$$G(x) = 1 + x + x^2 + x^3 + x^4 + \cdots$$

converges for all x such that $|x| < 1$. In such cases the sum is

$$\frac{1}{1 - x}.$$

The series diverges for all other values of x.

As a final comment, note that the formula

$$S_n = \frac{1 - x^{n+1}}{1 - x}$$

is valid for all x except $x = 1$. Even though the partial sums aren't converging to any limit if $x > 1$, the formula can still be useful as a quick way for summing powers. Thus, for instance

$$1 + 3 + 9 + 27 + 81 + 243 = \frac{1 - 3^6}{1 - 3} = \frac{1 - 729}{-2} = \frac{-728}{-2} = 364,$$

and

$$1 - 5 + 25 - 125 + 625 - 3125 = \frac{1 - (-5)^6}{1 - (-5)} = \frac{1 - 15625}{6} = -264.$$

Alternating Series

A large class of common power series consists of the **alternating series**—series in which the terms are alternately positive and negative. The behavior of such series is particularly easy to analyze, as we shall see in this section. Here are some examples of alternating series we've already encountered :

Many common series are alternating, at least for some values of x

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots ,$$
$$[.15in] \cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots ,$$
$$[.15in] \frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \cdots , \qquad (\text{for } |x| < 1).$$

Other series may be alternating for some, but not all, values of x. For instance, here are two series that are alternating for negative values of x, but not for positive values:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots ,$$

$$\ln(1-x) = -(x + x^2 + x^3 + x^4 + \cdots) \qquad (\text{for } |x| < 1).$$

Convergence criterion for alternating series. Let us write a generic alternating series as

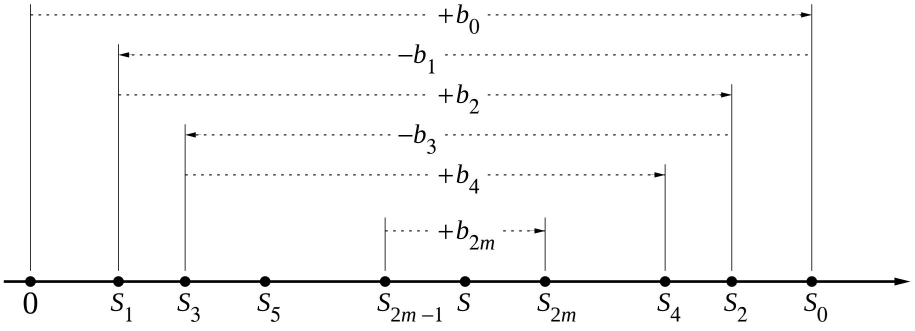
$$b_0 - b_1 + b_2 - b_3 + \cdots + (-1)^m b_m + \cdots ,$$

where the b_m are positive. It turns out that an alternating series converges if the terms b_m both consistently shrink in size and approach zero:

$$\boxed{\begin{array}{c} b_0 - b_1 + b_2 - b_3 + \cdots + (-1)^m b_m + \cdots \qquad \text{converges if} \\ 0 < b_{m+1} \le b_m \quad \text{for all } m \quad \text{and} \quad \lim_{m\to\infty} b_m = 0. \end{array}}$$

Alternating series are easy to test for convergence

It is this property that makes alternating series particularly easy to deal with. Recall that this is *not* a property of series in general, as we saw by the example of the harmonic series. The reason it is true for alternating series becomes clear if we view the behavior of the partial sums geometrically:



We mark the partial sums S_n on a number line. The first sum $S_0 = b_0$ lies to the right of the origin. To find S_1 we go to the left a distance b_1. Because $b_1 \leq b_0$, S_1 will lie between the origin and S_0. Next we go to the right a distance b_2, which brings us to S_2. Since $b_2 \leq b_1$, we will have $S_2 \leq S_0$. The next move is to the left a distance b_3, and we find $S_3 \geq S_1$. We continue going back and forth in this fashion, each step being less than or equal to the preceding one, since $b_{m+1} \leq b_m$. We thus get

$$0 \leq S_1 \leq S_3 \leq S_5 \leq \ldots \leq S_{2m-1} \leq \cdots \leq S_{2m} \leq \ldots \leq S_4 \leq S_2 \leq S_0.$$

The partial sums oscillate, with the exact sum trapped between consecutive partial sums

The partial sums oscillate back and forth, with all the odd sums on the left increasing and all the even sums on the right decreasing. Moreover, since $|S_n - S_{n-1}| = b_n$, and since $\lim_{n\to\infty} b_n = 0$, the difference between consecutive partial sums eventually becomes arbitrarily small—the oscillations take place within a smaller and smaller interval. Thus given any number of decimal places, we can always go far enough out in the series so that S_k and S_{k+1} agree to that many decimal places. But if n is any integer greater than k, then, since S_n lies between S_k and S_{k+1}, S_n will also agree to that many decimal places—those decimals will be fixed from k on out. The series therefore converges, as claimed—the sum is the unique number S that is greater than all the odd partial sums and less than all the even partial sums.

A simple estimate for the accuracy of the partial sums

For a convergent alternating series, we also have a particularly simple bound for the error when we approximate the sum S of the series by partial sums.

If $$S_n = b_0 - b_1 + b_2 - \cdots \pm b_n,$$

and if $0 < b_{m+1} \leq b_m$ for all m and $\lim\limits_{m\to\infty} b_m = 0,$

(so the series converges), then

$$|S - S_n| < b_{n+1}.$$

In words, the error in approximating S by S_n is less than the next term in the series.

Proof: Suppose n is odd. Then we have, as above, that $S_n < S < S_{n+1}$. Therefore $0 < S - S_n < S_{n+1} - S_n = b_{n+1}$. If n is even, a similar argument shows $0 < S_n - S < S_n - S_{n+1} = b_{n+1}$. In either case, we have $|S - S_n| < b_{n+1}$, as claimed.

Note further that we also know whether S_n is too large or too small, depending on whether n is even or odd.

Estimating the error in approximating $\cos(.7)$

Example. Let's apply the error estimate for an alternating series to analyze the error if we approximate cos(.7) with a Taylor series with three terms:

$$\cos(.7) \approx 1 - \frac{1}{2!}(.7)^2 + \frac{1}{4!}(.7)^4 = 0.765004166\ldots.$$

Since the last term in this partial sum was an addition, this approximation is too big. To get an estimate of how far off it might be, we look at the next term in the series:

$$\frac{1}{6!}(.7)^6 = .0001634\ldots.$$

We thus know that the correct value for cos(.7) is somewhere in the interval

$$.76484 = .76500 - .00016 \le \cos(.7) \le .76501,$$

so we know that cos(.7) begins .76 . . . and the third decimal is either a 4 or a 5. Moreover, we know that $\cos(.7) = .765$ rounded to 3 decimal places.

If we use the partial sum with four terms, we get

$$\cos(.7) \approx 1 - \frac{1}{2!}(.7)^2 + \frac{1}{4!}(.7)^4 - \frac{1}{6!}(.7)^6 = .764840765\ldots,$$

and the error would be less than

$$\frac{1}{8!}(.7)^8 = .0000014\ldots < .5 \times 10^{-5},$$

so we could now say that $\cos(.7) = .76484\ldots$.

How many terms are needed to obtain accuracy to 12 decimal places?

If we wanted to know in advance how far out in the series we would have to go to determine cos(.7) to, say, 12 decimals, we could do it by finding a value for n such that

$$b_n = \frac{1}{n!}(.7)^n \le .5 \times 10^{-12}.$$

With a little trial and error, we see that $b_{12} \approx .3 \times 10^{-10}$, while $b_{14} < 10^{-13}$. Thus if we take the value of the 12th degree approximation for cos(.7), we can be assured that our value will be accurate to 12 places.

We have met this capability of getting an error estimate in a single step before, in version 3 of Taylor's theorem. It is in contrast to the approximations made in dealing with general series, where we typically had to look at the pattern of stabilizing digits in the succession of improving estimates to get a sense of how good our approximation was, and even then we had no guarantee.

It may be possible to convert a given problem to one involving alternating series

Computing e. Because of the fact that we can find sharp bounds for the accuracy of an approximation with alternating series, it is often desirable to convert a given problem to this form where we can. For instance, suppose we wanted a good value for e. The obvious thing to do would be to take the Taylor series for e^x and substitute $x = 1$. If we take the first 11 terms of this series we get the approximation

$$e = e^1 \approx 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{10!} = 2.718281801146\ldots,$$

but we have no way of knowing how many of these digits are correct.

Suppose instead, that we evaluate e^{-1}:

$$e^{-1} \approx 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{1}{10!} = .367879464286\ldots.$$

Since $1/(11!) = .000000025\ldots$, we know this approximation is accurate to at least 7 decimals. If we take its reciprocal we get

$$1/.3678794624286\ldots = 2.718281657666\ldots,$$

which will then be accurate to 6 decimals (in the exercises you will show why the accuracy drops by 1 decimal place), so we can say $e = 2.718281\ldots$.

The Radius of Convergence

We have seen examples of power series that converge for all x (like the Taylor series for $\sin x$) and others that converge only for certain x (like the series for $\arctan x$). How can we determine the convergence of an arbitrary power series of the form

$$a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n + \cdots ?$$

We must suspect that this series *may not* converge for all values of x. For example, does the Taylor series

$$1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots$$

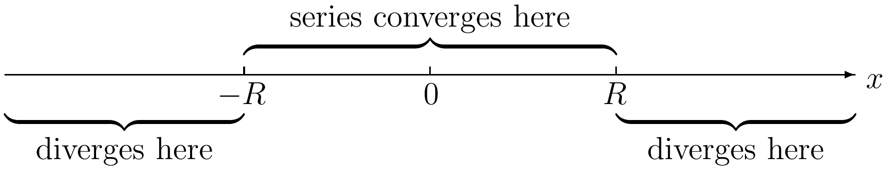
converge for all values of x, or only some? When it converges, does it converge to e^x or to something else? After all, this Taylor series is designed to look like e^x only near $x = 0$; it remains to be seen how well the function and its series match up far from $x = 0$.

The question of convergence has a definitive answer. It goes like this: if the power series

$$a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n + \cdots \ .$$

The answer to the convergence question

converges for a particular value of x, say $x = s$, then it automatically converges for any *smaller* value of x (meaning any x that is closer to the origin than s is; i.e, any x for which $|x| < |s|$). Likewise, if the series *diverges* for a particular value of x, then it also diverges for any value farther from the origin. In other words, the values of x where the series converges are not interspersed with the values where it diverges. On the contrary, within a certain distance R from the origin there is only convergence, while beyond that distance there is only divergence. The number R is called the **radius of convergence** of the series, and the range where it converges is called its **interval of convergence**.



The radius of convergence of a power series

An obvious example of the radius of convergence is given by the geometric series

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots$$

Radius of convergence of the geometric series

We know that this converges for $|x| < 1$ and diverges for $|x| > 1$. Thus the radius of convergence is $R = 1$ in this case.

It is possible for a power series to converge for all x; if that happens, we take R to be ∞. At the other extreme, the series may converge only for $x = 0$. (When $x = 0$ the series collapses to its constant term a_0, so it certainly converges *at least* when $x = 0$.) If the series converges only for $x = 0$, then we take R to be 0.

At $x = R$ the series may diverge or converge; different things happen for different series. The same is true when $x = -R$. The radius of convergence tells us *where* the switch from convergence to divergence happens. It does not tell us *what* happens at the place where the switch occurs. If we know that the series converges for $x = \pm R$, then we say that $[-R, R]$ is the interval of convergence. If the series converges when $x = R$ but *not* when $x = -R$, then the interval of convergence is $(-R, R]$, and so on.

The Ratio Test

There are several ways to determine the radius of convergence of a power series. One of the simplest and most useful is by means of the **ratio test**. Because the power series to which we apply this test need not include *consecutive* powers of x (think of the Taylor series for $\cos x$ or $\sin x$) we'll write a "generic" series as

$$b_0 + b_1 + b_2 + \cdots = \sum_{m=0}^{\infty} b_m$$

Here are three examples of the use of this notation.

1. The Taylor series for e^x is $\displaystyle\sum_{m=0}^{\infty} b_m$, where

$$b_0 = 1, \quad b_1 = x, \quad b_2 = \frac{x^2}{2!}, \cdots, b_m = \frac{x^m}{m!}.$$

2. The Taylor series for $\cos x$ is $\displaystyle\sum_{m=0}^{\infty} b_m$, where

$$b_0 = 1, \quad b_1 = \frac{-x^2}{2!}, \quad b_2 = \frac{x^4}{4!}, \cdots, b_m = (-1)^m \frac{x^{2m}}{(2m)!}.$$

3. We can even describe the series

$$17 + x + x^2 + x^4 + x^6 + x^8 + \cdots = 17 + x + \sum_{m=2}^{\infty} x^{2m-2}$$

in our generic notation, in spite of the presence of the first two terms "$17 + x$" which don't fit the pattern of later ones. We have $b_0 = 17$, $b_1 = x$, and then $b_m = x^{2m-2}$ for $m = 2, 3, 4, \ldots$.

Convergence is determined by the "tail" of the series

The question of convergence for a power series is unaffected by the "beginning" of the series; only the pattern in the "tail" matters. (Of course the *value* of the power series is affected by all of its terms.) So we can modify our generic notation to fit the circumstances at hand. No harm is done if we don't begin with b_0.

Using this notation we can state the ratio test (but we give no proof).

$$\boxed{\begin{array}{c}\textbf{Ratio Test: the series } \boldsymbol{b_0 + b_1 + b_2 + b_3 + \cdots + b_n + \cdots} \\ \textbf{converges if } \displaystyle\lim_{m\to\infty} \frac{|\boldsymbol{b_{m+1}}|}{|\boldsymbol{b_m}|} < \mathbf{1}.\end{array}}$$

The ratio test for the geometric series . . .

Let's see what the ratio test says about the geometric series:

$$1 + x + x^2 + x^3 + \cdots .$$

We have $b_m = x^m$, so the ratio we must consider is

$$\frac{|b_{m+1}|}{|b_m|} = \frac{|x^{m+1}|}{|x^m|} = \frac{|x|^{m+1}}{|x|^m} = |x|.$$

(Be sure you see why $|x^m| = |x|^m$.) Obviously, this ratio has the same value for all m, so the limit

$$\lim_{m\to\infty} |x| = |x|$$

exists and is less than 1 precisely when $|x| < 1$. Thus the geometric series converges for $|x| < 1$—which we already know is true. This means that the radius of convergence of the geometric series is $R = 1$.

. . . and for e^x

Look next at the Taylor series for e^x:

$$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots = \sum_{m=0}^{\infty} \frac{x^m}{m!}.$$

For negative x this is an alternating series, so by the criterion for convergence of alternating series we know it converges for all $x < 0$. The radius of convergence should then be ∞. We will use the ratio test to show that in fact this series converges for *all* x.

In this case

$$b_m = \frac{x^m}{m!},$$

so the relevant ratio is

$$\frac{|b_{m+1}|}{|b_m|} = \left|\frac{x^{m+1}}{(m+1)!}\right| \cdot \left|\frac{m!}{x^m}\right| = \frac{|x^{m+1}|}{|x^m|} \cdot \frac{m!}{(m+1)!} = |x| \cdot \frac{1}{m+1} = \frac{|x|}{m+1}.$$

Unlike the example with the geometric series, the value of this ratio depends on m. For any particular x, as m gets larger and larger the numerator stays the same and the denominator grows, so this ratio gets smaller and smaller. In other words,

$$\lim_{m\to\infty} \frac{|b_{m+1}|}{|b_m|} = \lim_{m\to\infty} \frac{|x|}{m+1} = 0.$$

Since this limit is less than 1 for any value of x, the series converges for *all* x, and thus the radius of convergence of the Taylor series for e^x is $R = \infty$, as we expected.

One of the uses of the theory developed so far is that it gives us a new way of specifying functions. For example, consider the power series

$$\sum_{m=0}^{\infty} (-1)^m \frac{2^m}{m^2+1}\, x^m = 1 - x + \frac{4}{5}x^2 - \frac{8}{10}x^3 + \frac{16}{17}x^4 + \cdots .$$

In this case $b_m = (-1)^m \dfrac{2^m}{m^2+1} x^m$, so to find the radius of convergence, we compute the ratio

$$\frac{|b_{m+1}|}{|b_m|} = \frac{2^{m+1}|x|^{m+1}}{(m+1)^2+1} \cdot \frac{m^2+1}{2^m|x|^m} = 2|x| \frac{m^2+1}{m^2+2m+2}.$$

Finding the limit of $|b_{m+1}|/|b_m|$ may require some algebra

To figure out what happens to this ratio as m grows large, it is helpful to rewrite the factor involving the m's as

$$\frac{m^2 \cdot (1+1/m^2)}{m^2 \cdot (1+2/m+2/m^2)} = \frac{1+1/m^2}{1+2/m+2/m^2}.$$

Now we can see that

$$\lim_{m\to\infty} \frac{|b_{m+1}|}{|b_m|} = 2|x| \cdot \frac{1}{1} = 2|x|.$$

The limit value is less than 1 precisely when $2|x| < 1$, or, equivalently, $|x| < 1/2$, so the radius of convergence of this series is $R = 1/2$. It follows that for $|x| < 1/2$, we have a new function $f(x)$ defined by the power series:

$$f(x) = \sum_{m=0}^{\infty} (-1)^m \frac{2^m}{m^2+1}\, x^m.$$

We can also discuss the radius of convergence of a power series

$$a_0 + a_1(x-a) + a_2(x-a)^2 + \cdots + a_m(x-a)^m + \cdots$$

centered at a location $x = a$ other than the origin. The radius of convergence of a power series of this form can be found by the **ratio test** in exactly the same way it was when $a = 0$.

Example. Let's apply the ratio test to the Taylor series centered at $a = 1$ for $\ln(x)$:

$$\ln(x) = \sum_{m=1}^{\infty} \frac{(-1)^{m-1}}{m}\,(x-1)^m.$$

We can start our series with b_1, so we can take $b_m = \dfrac{(-1)^{m-1}}{m}\,(x-1)^m$. Then the ratio we must consider is

$$\frac{|b_{m+1}|}{|b_m|} = \frac{|x-1|^{m+1}}{m+1} \cdot \frac{m}{|x-1|^m} = |x-1| \cdot \frac{m}{m+1} = |x-1| \cdot \frac{1}{1+1/m}.$$

Then

$$\lim_{m\to\infty} \frac{|b_{m+1}|}{|b_m|} = |x-1| \cdot 1 = |x-1|.$$

From this we conclude that this series converges for $|x-1| < 1$. This inequality is equivalent to

$$-1 < x - 1 < 1,$$

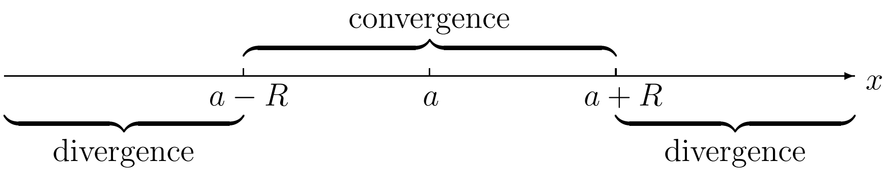
which is an interval of "radius" 1 about $a = 1$, so the radius of convergence is $R = 1$ in this case. We may also write the interval of convergence for this power series as

$$0 < x < 2.$$

More generally, using the ratio test we find that a power series centered at a converges in an interval of "radius" R (and width $2R$) around the point $x = a$ on the x-axis. Ignoring what happens at the endpoints, we say the **interval of convergence** is

$$a - R < x < a + R.$$

Here is a picture of what this looks like:



The convergence of a power series centered at a

Exercises

1. Find a formula for the sum of each of the following power series by performing suitable operations on the geometric series and the formula for *its* sum.

a) $1 - x^3 + x^6 - x^9 + \cdots$.

b) $x^2 + x^6 + x^{10} + x^{14} + \cdots$.

c) $1 - 2x + 3x^2 - 4x^3 + \cdots$.

d) $x + 2x^2 + 3x^3 + 4x^4 + \cdots$.

e) $x + \dfrac{x^2}{2} + \dfrac{x^3}{3} + \dfrac{x^4}{4} + \cdots$.

2. Determine the value of each of the following infinite sums. (Each os these sums is a geometric or related series evaluated at a particular value of x.)

a) $\dfrac{1}{4} + \dfrac{1}{16} + \dfrac{1}{64} + \dfrac{1}{256} + \cdots$.

b) $.02020202\ldots$.

c) $-\dfrac{5}{2} + \dfrac{5}{4} - \dfrac{5}{8} + \cdots$.

d) $\dfrac{1}{1} - \dfrac{2}{2} + \dfrac{3}{4} - \dfrac{4}{8} + \dfrac{5}{16} - \dfrac{6}{32} + \cdots$.

e) $\dfrac{1}{1 \cdot 10} + \dfrac{1}{2 \cdot 10^2} + \dfrac{1}{3 \cdot 10^3} + \dfrac{1}{4 \cdot 10^4} + \cdots$.

3. **The Multiplier Effect**. Economists know that the effect on a local economy of tourist spending is greater than the amount actually spent by the tourists. The *multiplier effect* quantifies this enlarged effect. In this problem you will see that calculating the multiplier effect involves summing a geometric series.

Suppose that, over the course of a year, tourists spend a total of A dollars in a small resort town. By the end of the year, the townspeople are therefore A dollars richer. Some of this money leaves the town—for example, to pay state and federal taxes or to pay off debts owed to "big city" banks. Some of it stays in town but gets put away as savings. Finally, a certain fraction of the original amount is spent in town, by the townspeople themselves. Suppose

3/5-ths is spent this way. The tourists and the townspeople *together* are therefore responsible for spending

$$S = A + \frac{3}{5}A \quad \text{dollars}$$

in the town that year. The second amount—$\frac{3}{5}A$ dollars—is *recirculated* money.

Since one dollar looks much like another, the recirculated money should be handled the same way as the original tourist dollars: some will leave the town, some will be saved, and the remaining 3/5-ths will get recirculated a *second* time. The twice-recirculated amount is

$$\frac{3}{5} \times \frac{3}{5}A \quad \text{dollars,}$$

and we must revise the calculation of the total amount spent in the town to

$$S = A + \frac{3}{5}A + \left(\frac{3}{5}\right)^2 A \quad \text{dollars.}$$

But the twice-recirculated dollars look like all the others , so 3/5-ths of them will get recirculated a *third* time. Revising the total dollars spent yet again, we get

$$S = A + \frac{3}{5}A + \left(\frac{3}{5}\right)^2 A + \left(\frac{3}{5}\right)^3 A \quad \text{dollars.}$$

This process never ends: no matter how many times a sum of money has been recirculated, 3/5-ths of it is recirculated once more. The total amount spend in the town is thus given by a *series.*

a) Write the series giving the total amount of money spent in the town and calculate its sum.

b) Your answer in a) is a certain multiple of A—what is the multiplier?

c) Suppose the recirculation rate is r instead of 3/5. Write the series giving the total amount spent and calculate its sum. What is the multiplier now?

d) Suppose the recirculation rate is 1/5; what is the multiplier in this case?

e) Does a lower recirculation rate produce a smaller multiplier effect?

4. Which of the following alternating series converge, which diverge? Why?

a) $\displaystyle\sum_{n=1}^{\infty}(-1)^n\frac{n}{n^2+1}$

b) $\displaystyle\sum_{n=1}^{\infty}(-1)^n\frac{1}{\sqrt{3n+2}}$

c) $\displaystyle\sum_{n=2}^{\infty}(-1)^n\frac{n}{\ln n}$

d) $\displaystyle\sum_{n=1}^{\infty}(-1)^n\frac{n}{5n-4}$

e) $\displaystyle\sum_{n=1}^{\infty}(-1)^n\frac{\arctan n}{n}$

f) $\displaystyle\sum_{n=1}^{\infty}(-1)^n\frac{(1.0001)^n}{n^{10}+1}$

g) $\displaystyle\sum_{n=1}^{\infty}(-1)^n\frac{1}{n^{1/n}}$

h) $\displaystyle\sum_{n=2}^{\infty}(-1)^n\frac{1}{\ln n}$

i) $\displaystyle\sum_{n=1}^{\infty}(-1)^n\frac{n!}{n^n}$

j) $\displaystyle\sum_{n=1}^{\infty}(-1)^n\frac{n!}{1\cdot 3\cdot 5\cdots(2n-1)}$

5. For each of the sums in the preceding problem that converges, use the alternating series criterion to determine how far out you have to go before the sum is determined to 6 decimal places. Give the sum for each of these series to this many places.

6. Find a value for n so that the nth degree Taylor series for e^x gives at least 10 place accuracy for all x in the interval $[-3,0]$.

7. We defined the harmonic series as the infinite sum

$$1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots=\sum_{i=1}^{\infty}\frac{1}{i}.$$

a) Use a calculator to find the partial sums

$$S_n=1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots\frac{1}{n}$$

for $n = 1, 2, 3, \ldots, 12$.

b) Use the following program to find the value of S_n for $n = 100$. Modify the program to find the values of S_n for $n = 500$, 1000, and 5000.

Program: HARMONIC

```
n = 100
sum = 0
FOR i = 1 TO n
     sum = sum + 1/i
NEXT i
PRINT  n, sum
```

c) Group the terms in the harmonic series as indicated by the parentheses:

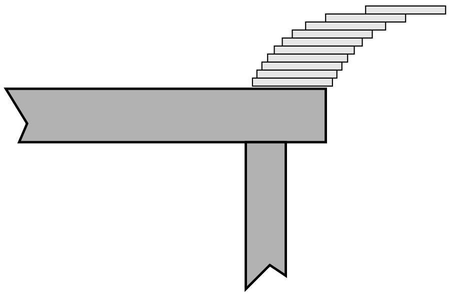
$$1+\left(\frac{1}{2}\right)+\left(\frac{1}{3}+\frac{1}{4}\right)+\left(\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}\right)+ \\ +\left(\frac{1}{9}+\cdots+\frac{1}{16}\right)+\left(\frac{1}{17}+\cdots+\frac{1}{32}\right)+\cdots.$$

Explain why each parenthetical grouping totals at least 1/2.

d) Following the pattern in part (c), if you add up the terms of the harmonic series forming S_n for $n = 2^k$, you can arrange the terms as $1 + k$ such groupings. Use this fact and the result of c) to explain why S_n exceeds $1 + k \cdot \frac{1}{2}$.

e) Use part (d) to explain why the harmonic series *diverges*.

f) You might try this problem if you've studied physics—enough to know how to locate the center of mass of a system. Suppose you had n cards and wanted to stack them on the edge of a table with the top of the pile leaning out over the edge. How far out could you get the pile to reach if you were careful? Let's choose our units so the length of each card is 1. Clearly if $n = 1$, the farthest reach you could would be $\frac{1}{2}$. If $n = 2$, you could clearly place the top card to extend half a unit beyond the bottom card. For the system to be stable, the center of mass of the two cards must be to the left of the edge of the table. Show that for this to happen, the bottom card can't extend more than 1/4 unit beyond the edge. Thus with $n =$ 2, the maximum extension of the pile is $\frac{1}{2}+\frac{1}{4}=\frac{3}{4}$. The picture at the right shows 10 cards stacked carefully.



Prove that if you have n cards, the stack can be built to extend a distance of

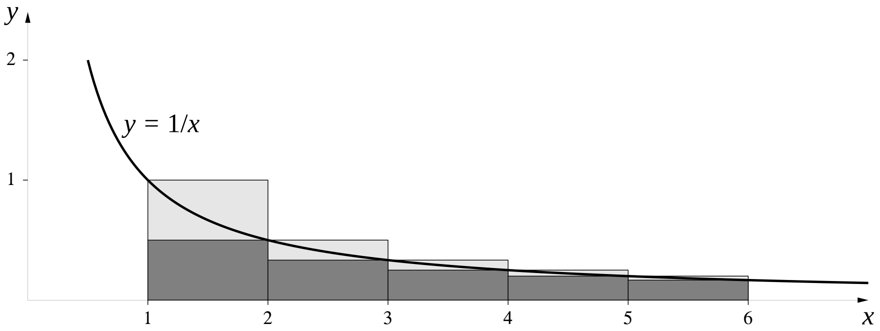
$$\frac{1}{2}+\frac{1}{4}+\frac{1}{6}+\frac{1}{8}+\cdots+\frac{1}{2n}=\frac{1}{2}\left(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots+\frac{1}{n}\right).$$

In the light of what we have just proved about the harmonic series, this shows that if you had enough cards, you could have the top card extending 100 units beyond the edge of the table!

8. **An estimate for the partial sums of the harmonic series**. You may notice in part (d) of the preceding problem that the sum $S_n = 1+1/2+1/3+\cdots+1/n$ grows in proportion to the exponent k of $n = 2^k$; i.e., the sum grows like the logarithm of n. We can make this more precise by comparing the value of S_n to the value of the integral

$$\int_1^n \frac{1}{x}\,dx = \ln(n).$$

a) Let's look at the case $n = 6$ to see what's going on. Consider the following picture:



Show that the lightly shaded region plus the dark region has area equal to S_5, which can be rewritten as $S_6 - \frac{1}{6}$. Show that the dark region alone has area $S_6 - 1$. Hence prove that

$$S_6 - 1 < \int_1^6 \frac{1}{x}\,dx < S_6 - \frac{1}{6},$$

and conclude that

$$\frac{1}{6} < S_6 - \ln(6) < 1.$$

b) Show more generally that $\displaystyle \frac{1}{n} < S_n - \ln(n) < 1.$

c) Use part (b) to get upper and lower bounds for the value of S_{10000}.

[Answer: $9.21044 < S_{10000} < 10.21035$.]

d) Use the result of part (b) to get an estimate for how many cards you would need in part (f) of the preceding problem to make the top of the pile extend 100 units beyond the edge of the table.

[Answer: It would take approximately 10^{87} cards—a number which is on the same order of magnitude as the number of atoms in the universe!]

Remarkably, partial sums of the harmonic series exceed $\ln(n)$ in a very regular way. It turns out that

$$\lim_{n\to\infty} \{S_n - \ln(n)\} = \gamma,$$

where $\gamma = .5772\ldots$ is called **Euler's constant.** (You have seen another constant named for Euler, the base $e = 2.7183\ldots$.) Although one can find the decimal expansion of γ to any desired degree of accuracy, no one knows whether γ is a rational number or not.

9. Show that the power series for $\arctan x$ and $(1+x)^c$ diverge for $|x| > 1$. Do the series converge or diverge when $|x| = 1$?

10. Find the radius of convergence of each of the following power series.

a) $1 + 2x + 3x^2 + 4x^3 + \cdots$.

b) $x + 2x^2 + 3x^3 + 4x^4 + \cdots$.

c) $1 + \frac{1}{1^2}x + \frac{1}{2^2}x^2 + \frac{1}{3^2}x^3 + \frac{1}{4^2}x^4 + \cdots$. [Answer: $R = 1$]

d) $x^3 + x^6 + x^9 + x^{12} + \cdots$.

e) $1 + (x+1) + (x+1)^2 + (x+1)^3 + \cdots$.

f) $17 + \frac{1}{3}x + \frac{1}{3^2}x^2 + \frac{1}{3^3}x^3 + \frac{1}{3^4}x^4 + \cdots$.

11. Write out the first five terms of each of the following power series, and determine the radius of convergence of each.

a) $\displaystyle\sum_{n=0}^{\infty} nx^n$. [Answer: $R = 1$]

b) $\displaystyle\sum_{n=0}^{\infty} \frac{n^2}{2^n}x^n$. [Answer: $R = 2$]

c) $\displaystyle\sum_{n=0}^{\infty} (n+5)^2\, x^n$. [Answer: $R = 1$]

d) $\displaystyle\sum_{n=0}^{\infty} \frac{99}{n^n} x^n.$ [Answer: $R = \infty$]

e) $\displaystyle\sum_{n=0}^{\infty} n!\, x^n.$ [Answer: $R = 0$]

12. Find the radius of convergence of the Taylor series for $\sin x$ and for $\cos x$. For which values of x can these series therefore represent the sine and cosine functions?

13. Find the radius of convergence of the Taylor series for $f(x) = 1/(1+x^2)$ at $x = 0$. (See the table of Taylor series in section 3.) What is the radius of convergence of this series? For which values of x can this series therefore represent the function f? Do these x values constitute the *entire* domain of definition of f?

14. In the text we used the alternating series for e^x, $x < 0$, to approximate e^{-1} accurate to 7 decimal places. The claim was made that in taking the reciprocal to obtain an estimate for e, the accuracy drops by one decimal place. In this problem you will see why this is true.

a) Consider first the more general situation where two functions are reciprocals, $g(x) = 1/f(x)$. Express $g'(x)$ in terms of $f(x)$ and $f'(x)$.

b) Use your answer in part a) to find an expression for the *relative error* in g, $\Delta g/g(x) \approx g'(x)\Delta x/g(x)$, in terms of $f(x)$ and $f'(x)$. How does this compare to the relative error in f?

c) Apply your results in part b) to the functions e^x and e^{-x} at $x = 1$. Since e is about 7 times as large as $1/e$, explain why the error in the estimate for e should be about 7 times as large as the error in the estimate for $1/e$.

10.6 Approximation Over Intervals

A powerful result in mathematical analysis is the **Stone–Weierstrass Theorem**, which states that given any continuous function $f(x)$ and any interval $[a, b]$, there exist polynomials that fit f over this interval to any level of accuracy we care to specify. In many cases, we can find such a polynomial simply by taking a Taylor polynomial of high enough degree. There are several ways in which this is not a completely satisfactory response, however. First, some functions (like the absolute value function) have corners or other places where they aren't differentiable, so we can't even build a Taylor series at such points. Second, we have seen several functions (like $1/(1 + x^2)$) that have a finite interval of convergence, so Taylor polynomials may not be good fits no matter how high a degree we try. Third, even for well-behaved functions like $\sin(x)$ or e^x, we may have to take a very high degree Taylor polynomial to get the same overall fit that a much lower degree polynomial could achieve.

In this section we will develop the general machinery for finding polynomial approximations to functions over given intervals. In chapter 12.4 we will see how this same approach can be adapted to approximating periodic functions by **trigonometric polynomials**.

Approximation by polynomials

Two possible criteria for best fit

Example. Let's return to the problem introduced at the beginning of this chapter: find the second degree polynomial which best fits the function $\sin(x)$ over the interval $[0, \pi]$. Just as we did with the Taylor polynomials, though, before we can start we need to agree on our criterion for the best fit. Here are two obvious candidates for such a criterion:

1. The second degree polynomial $Q(x)$ is the best fit to $\sin(x)$ over the interval $[0, \pi]$ if the *maximum* separation between $Q(x)$ and $\sin(x)$ is smaller than the maximum separation between $\sin(x)$ and any other second degree polynomial:

$$\max_{0 \le x \le \pi} |\sin(x) - Q(x)| \qquad \text{is the smallest possible.}$$

2. The second degree polynomial $Q(x)$ is the best fit to $\sin(x)$ over the interval $[0, \pi]$ if the *average* separation between $Q(x)$ and $\sin(x)$ is smaller

than the average separation between $\sin(x)$ and any other second degree polynomial:

$$\frac{1}{\pi}\int_0^\pi |\sin(x) - Q(x)|\,dx \qquad \text{is the smallest possible.}$$

Why we don't use either criterion

Unfortunately, even though their clear meanings make these two criteria very attractive, they turn out to be largely unusable—if we try to apply either criterion to a specific problem, including our current example, we are led into a maze of tedious and unwieldy calculations.

Instead, we use a criterion that, while slightly less obvious than either of the two we've already articulated, still clearly measures some sort of "best fit" and has the added virtue of behaving well mathematically. We accomplish this by modifying criterion 2 slightly. It turns out that the major difficulty with this criterion is the presence of absolute values. If, instead of considering the average separation between $Q(x)$ and $\sin(x)$, we consider the average of the *square* of the separation between $Q(x)$ and $\sin(x)$, we get a criterion we can work with. (Compare this with the discussion of the best-fitting line in the exercises for chapter 9.3.) Since this is a definition we will be using for the rest of this section, we frame it in terms of arbitrary functions g and h, and an arbitrary interval $[a, b]$:

Given two functions g and h defined over an interval $[a, b]$, we define the **mean square separation** between g and h over this interval to be

$$\frac{1}{(b-a)}\int_a^b (g(x) - h(x))^2\, dx.$$

Note: In this setting the word **mean** is synonymous with what we have called "average". It turns out that there is often more than one way to define the term "average"—the concepts of median and mode are two other natural ways of capturing "averageness", for instance—so we use the more technical term to avoid ambiguity.

The criterion we shall use

We can now rephrase our original problem as: find the second degree polynomial $Q(x)$ whose mean squared separation from $\sin(x)$ over the interval $[0, \pi]$ is as small as possible. In mathematical terms, we want to find coefficients a_0, a_1, and a_2 which such that the integral

$$\int_0^\pi \left(\sin(x) - (a_0 + a_1\,x + a_2\,x^2)\right)^2\,dx$$

is minimized. The solution $Q(x)$ is called the quadratic **least squares approximation** to $\sin(x)$ over $[0, \pi]$.

The key to solving this problem is to observe that a_0, a_1, and a_2 can take on any values we like and that this integral can thus be considered a function of these three variables. For instance, if we couldn't think of anything cleverer to do, we might simply try various combinations of a_0, a_1, and a_2 to see how small we could make the given integral. Therefore another way to phrase our problem is

A mathematical formulation of the problem

> Find values for a_0, a_1, and a_2 that minimize the function
>
> $$F(a_0, a_1, a_2) = \int_0^\pi \left(\sin(x) - (a_0 + a_1\,x + a_2\,x^2)\right)^2\,dx\,.$$

We know how to find points where functions take on their extreme values—we look for the places where the partial derivatives are 0. But how do we differentiate an expression involving an integral like this? It turns out that for all continuous functions, or even functions with only a finite number of breaks in them, we can simply interchange integration and differentiation. Thus, in our example,

$$\begin{aligned}\frac{\partial}{\partial a_0}F(a_0, a_1, a_2) &= \frac{\partial}{\partial a_0}\int_0^\pi \left(\sin(x) - (a_0 + a_1\,x + a_2\,x^2)\right)^2\,dx\\ &= \int_0^\pi \frac{\partial}{\partial a_0}\left(\sin(x) - (a_0 + a_1\,x + a_2\,x^2)\right)^2\,dx\\ &= \int_0^\pi 2\left(\sin(x) - (a_0 + a_1\,x + a_2\,x^2)\right)(-1)\,dx.\end{aligned}$$

Similarly we have

$$\frac{\partial}{\partial a_1}F(a_0, a_1, a_2) = \int_0^\pi 2\left(\sin(x) - (a_0 + a_1\,x + a_2\,x^2)\right)(-x)\,dx,$$

$$\frac{\partial}{\partial a_2}F(a_0, a_1, a_2) = \int_0^\pi 2\left(\sin(x) - (a_0 + a_1\,x + a_2\,x^2)\right)(-x^2)\,dx.$$

We now want to find values for a_0, a_1, and a_2 that make these partial derivatives simultaneously equal to 0. That is, we want

Setting the partials equal to zero gives equations for a_0, a_1, a_2

$$\int_0^\pi 2\left(\sin(x) - (a_0 + a_1\,x + a_2\,x^2)\right)(-1)\,dx = 0,$$

$$\int_0^\pi 2\left(\sin(x) - (a_0 + a_1\,x + a_2\,x^2)\right)(-x)\,dx = 0,$$

$$\int_0^\pi 2\left(\sin(x) - (a_0 + a_1\,x + a_2\,x^2)\right)(-x^2)\,dx = 0,$$

which can be rewritten as

$$\int_0^\pi \sin(x)\,dx = \int_0^\pi \left(a_0 + a_1\,x + a_2\,x^2\right)\,dx,$$

$$\int_0^\pi x\,\sin(x)\,dx = \int_0^\pi \left(a_0\,x + a_1\,x^2 + a_2\,x^3\right)\,dx,$$

$$\int_0^\pi x^2\,\sin(x)\,dx = \int_0^\pi \left(a_0\,x^2 + a_1\,x^3 + a_2\,x^4\right)\,dx.$$

All of these integrals can be evaluated relatively easily (see the exercises for a hint on evaluating the integrals on the left–hand side). When we do so, we are left with

Evaluating the integrals gives three linear equations

$$\begin{aligned} 2 &= \pi a_0 + \frac{\pi^2}{2}a_1 + \frac{\pi^3}{3}a_2, \\ \pi &= \frac{\pi^2}{2}a_0 + \frac{\pi^3}{3}a_1 + \frac{\pi^4}{4}a_2, \\ \pi^2 - 4 &= \frac{\pi^3}{3}a_0 + \frac{\pi^4}{4}a_1 + \frac{\pi^5}{5}a_2. \end{aligned}$$

But this is simply a set of three linear equations in the unknowns a_0, a_1, and a_2, and they can be solved in the usual ways. We could either replace each expression in π by a corresponding decimal approximation, or we could keep everything in terms of π. Let's do the latter; after a bit of tedious arithmetic we find

$$\begin{aligned}
a_0 &= \frac{12}{\pi} - \frac{120}{\pi^3} &&= -.050465\ldots,\\
a_1 &= \frac{-60}{\pi^2} + \frac{720}{\pi^4} &&= 1.312236\ldots,\\
a_2 &= \frac{60}{\pi^3} - \frac{720}{\pi^5} &&= -.417697\ldots,
\end{aligned}$$

and we have

$$Q(x) = -.050465 + 1.312236\,x - .417698\,x^2,$$

which is the equation given in section 1 at the beginning of the chapter.

The analysis we gave for this particular case can clearly be generalized to apply to any function over any interval. When we do this we get

How to find least squares polynomial approximations in general

Given a function g over an interval $[a, b]$, then the n-th degree polynomial

$$P(x) = c_0 + c_1\,x + c_2\,x^2 + \cdots + c_n\,x^n$$

whose mean square distance from g is a minimum has coefficients that are determined by the following $n+1$ equations in the $n+1$ unknowns c_0, c_1, c_2, ..., c_n:

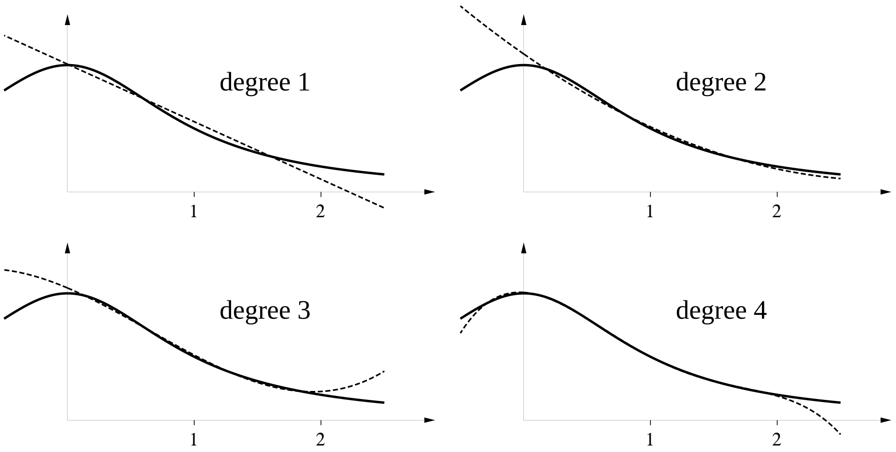
$$\begin{aligned}
\int_a^b g(x)\,dx &= c_0 \int_a^b dx + c_1 \int_a^b x\,dx + \cdots + c_n \int_a^b x^n\,dx,\\
\int_a^b x\,g(x)\,dx &= c_0 \int_a^b x\,dx + c_1 \int_a^b x^2\,dx + \cdots + c_n \int_a^b x^{n+1}\,dx,\\
&\vdots\\
\int_a^b x^n\,g(x)\,dx &= c_0 \int_a^b x^n\,dx + c_1 \int_a^b x^{n+1}\,dx + \cdots + c_n \int_a^b x^{2n}\,dx.
\end{aligned}$$

All the integrals on the right-hand side can be evaluated immediately. The integrals on the left-hand side will typically need to be evaluated numerically, although simple cases can be evaluated in closed form. Integration by parts is often useful in these cases. The exercises contain several problems using this technique to find approximating polynomials.

Solving the equations is a job for the computer

The real catch, though, is not in obtaining the equations—it is that solving systems of equations by hand is excruciatingly boring and subject to frequent arithmetic mistakes if there are more than two or three unknowns involved. Fortunately, there are now a number of computer packages available which do all of this for us. Here are a couple of examples, where the details are left to the exercises.

Example. Let's find polynomial approximation for $1/(1+x^2)$ over the interval $[0, 2]$. We saw earlier that the Taylor series for this function converges only for $|x| < 1$, so it will be no help. Yet with the above technique we can derive the following approximations of various degrees (see the exercises for details):

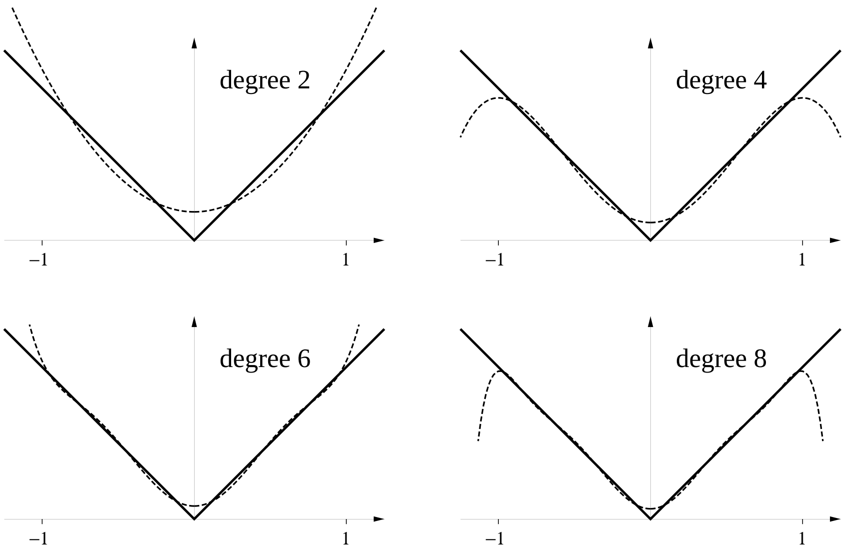


Here are the corresponding equations of the approximating polynomials:

degree	polynomial
1	$1.00722 - .453645\,x$
2	$1.08789 - .695660\,x + .121008\,x^2$
3	$1.04245 - .423017\,x - .219797\,x^2 + .113602\,x^3$
4	$1.00704 - .068906\,x - 1.01653\,x^2 + .733272\,x^3 - .154916\,x^4$

The technique works even when differentiability fails

Example. We can even use this new technique to find polynomial approximations for functions that aren't differentiable at some points. For instance, let's approximate the function $h(x) = |x|$ over the interval $[-1, 1]$. Since this function is symmetric about the y–axis, and we are approximating it over an interval that is symmetric about the y–axis, only even powers of x will appear. (See the exercises for details.) We get the following approximations of degrees 2, 4, 6, and 8:



Here are the corresponding polynomials:

degree	polynomial
2	$.1875 + .9375\,x^2$
4	$.117188 + 1.64062\,x^2 - .820312\,x^4$
6	$.085449 + 2.30713\,x^2 - 2.81982\,x^4 + 1.46631\,x^6$
8	$.067291 + 2.960821\,x^2 - 6.415132\,x^4 + 7.698173\,x^6 - 3.338498\,x^8$

The technique is useful for data functions

A Numerical Example. If we have some function which exists only as a set of data points—a numerical solution to a differential equation, perhaps, or the output of some laboratory instrument—it can often be quite useful to replace the function by an approximating polynomial. The polynomial takes up much less storage space and is easier to manipulate. To see how this works, let's return to the S-I-R model we've studied before

$$\begin{aligned} S' &= -.00001\,SI, \\ I' &= .00001\,SI - I/14, \\ R' &= I/14, \end{aligned}$$

with initial values $S(0) = 45400$ and $I(0) = 2100$.

Let's find an 8th degree polynomial $Q(t) = i_0 + i_1 t + i_2 t^2 + \cdots + i_8 t^8$ approximating I over the time interval $0 \le t \le 40$. We can do this by a minor modification of the Euler's method programs we've been using all along. Now, in addition to keeping track of the current values for S and I as we go along, we will also need to be calculating Riemann sums for the integrals

$$\int_0^{40} t^k I(t)\,dt \qquad \text{for } k = 0,\ 1,\ 2,\ \ldots,\ 8,$$

as we go through each iteration of Euler's method.

The importance of using the right-sized units

Since the numbers involved become enormous very quickly, we open ourselves to various sorts of computer roundoff error. We can avoid some of these difficulties by **rescaling** our equations—using units that keep the numbers involved more manageable. Thus, for instance, suppose we measure S, I, and R in units of 10,000 people, and suppose we measure time in "decadays", where 1 decaday = 10 days. When we do this, our original differential equations become

$$\begin{aligned} S' &= -SI, \\ I' &= SI - I/1.4, \\ R' &= I/1.4, \end{aligned}$$

with initial values $S(0) = 4.54$ and $I(0) = 0.21$. The integrals we want are now of the form

$$\int_0^{4} t^k I(t)\,dt \qquad \text{for } k = 0,\ 1,\ 2,\ \ldots,\ 8.$$

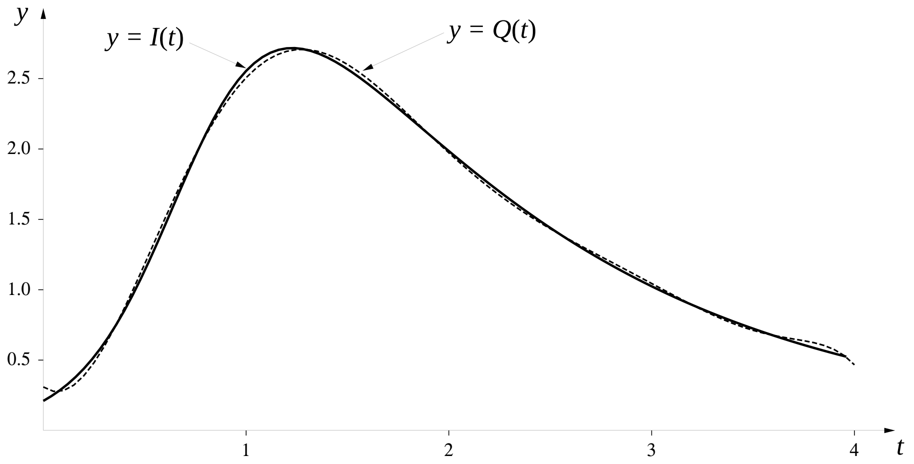
Using Simpson's rule helps reduce errors

The use of Simpson's rule (see chapter 11.3) will also reduce errors. It may be easiest to calculate the values of I first, using perhaps 2000 values, and store them in an array. Once you have this array of I values, it is relatively quick and easy to use Simpson's rule to calculate the 9 integrals needed. If you later decide you want to get a higher-degree polynomial approximation, you don't have to re-run the program.

Once we've evaluated these integrals, we set up and solve the corresponding system of 9 equations in the 9 unknown coefficients i_k. We get the following 8-th degree approximation

$$\begin{aligned} Q(t) = {} & .3090 - .9989\,t + 7.8518\,t^2 - 3.6233\,t^3 - 3.9248\,t^4 + 4.2162\,t^5 \\ & - 1.5750\,t^6 + .2692\,t^7 - .01772\,t^8\,. \end{aligned}$$

When we graph Q and I together over the interval $[0, 4]$ (decadays), we get



—a reasonably good fit.

A Caution. Numerical least-squares fitting of the sort performed in this last example fairly quickly pushes into regions where the cumulative effects of the inaccuracies of the original data, the inaccuracies of the estimates for the integrals, and the immense range in the magnitude of the numbers involved all combine to produce answers that are obviously wrong. Rescaling the equations and using good approximations for the integrals can help put off the point at which this begins to happen.

Exercises

1. To find polynomial approximations for $\sin(x)$ over the interval $[0, \pi]$, we needed to be able to evaluate integrals of the form

$$\int_0^\pi x^n \sin(x)\, dx.$$

The value of this integral clearly depends on the value of n, so denote it by I_n.

a) Evaluate I_0 and I_1. Suggestion: use integration by parts (Chapter 11.3) to evaluate I_1.

[Answer: $I_0 = 2$, and $I_1 = \pi$.]

b) Use integration by parts twice to prove the general **reduction formula**:

$$I_{n+2} = \pi^{n+2} - (n+2)(n+1)I_n \qquad \text{for all } n \geq 0.$$

c) Evaluate I_2, I_3, and I_4.

[Answer: $I_2 = \pi^2 - 4$, $I_3 = \pi^3 - 6\pi$, and $I_4 = \pi^4 - 12\pi^2 + 48$.]

d) If you have access to a computer package that will solve a system of equations, find the 4-th degree polynomial that best fits the sine function over the interval $[0, \pi]$. What is the maximum difference between this polynomial and the sine function over this interval?

[Answer: $.00131 + .98260\,x + .05447\,x^2 - .23379\,x^3 + .03721\,x^4$, with maximum difference occuring at the endpoints.]

2. To find polynomial approximations for $|x|$ over the interval $[-1, 1]$, we needed to be able to evaluate integrals of the form

$$\int_{-1}^{1} x^n |x|\, dx.$$

As before, let's denote this integral by I_n.

a) Show that

$$I_n = \begin{cases} \dfrac{2}{n+2} & \text{if } n \text{ is even,} \\ 0 & \text{if } n \text{ is odd.} \end{cases}$$

b) Derive the quadratic least squares approximation to $|x|$ over $[-1, 1]$.

c) If you have access to a computer package that will solve a system of equations, find the 10-th degree polynomial that best fits $|x|$ over the interval $[-1, 1]$. What is the maximum difference between this polynomial and $|x|$ over this interval?

3. To find polynomial approximations for $1/(1 + x^2)$ over the interval $[0, 2]$, we needed to be able to evaluate integrals of the form

$$\int_0^2 \frac{x^n}{1 + x^2}\, dx.$$

Call this integral I_n.

a) Evaluate I_0 and I_1.

[Answer: $I_0 = \arctan(1) = \pi/4$, and $I_1 = (\ln 2)/2 = .3465736$.]

b) Prove the general reduction formula:

$$I_{n+2} = \frac{2^{n+1}}{n+1} - I_n \qquad \text{for } n = 0,\ 1,\ 2,\ \ldots.$$

c) Evaluate I_2, I_3, and I_4.

[Answer: $I_2 = 2 - \dfrac{\pi}{4}, I_3 = 2 - \dfrac{\ln 2}{2}, I_4 = \dfrac{2}{3} + \dfrac{\pi}{4}$.]

d) If you have access to a computer package that will solve a system of equations, find the 4-th degree polynomial that best fits $1/(1+x^2)$ over the interval $[0, 2]$. What is the maximum difference between this polynomial and the function over this interval?

4. Set up the equations (including evaluating all the integrals) for finding the best fitting 6-th degree polynomial approximation to $\sin(x)$ over the interval $[-\pi, \pi]$.

5. In the *S*-*I*-*R* model, find the best fitting 8-th degree polynomial approximation to $S(t)$ over the interval $0 \le t \le 40$.

10.7 Chapter Summary

The Main Ideas

- **Taylor polynomials** approximate functions at a point. The Taylor polynomial $P(x)$ of degree n is the **best fit** to $f(x)$ at $x = a$; that is, P satisfies the following conditions: $P(a) = f(a)$, $P'(a) = f'(a)$, $P''(a) = f''(a)$, …, $P^{(n)}(a) = f^{(n)}(a)$.

- **Taylor's theorem** says that a function and its Taylor polynomial of degree n agree to order $n + 1$ near the point where the polynomial is centered. Different versions expand on this idea.

- If $P(x)$ is the Taylor polynomial approximating $f(x)$ at $x = a$, then $P(x)$ approximates $f(x)$ for values of x near a; $P'(x)$ approximates $f'(x)$; and $\int P(x)\,dx$ approximates $\int f(x)\,dx$.

- A **Taylor series** is an infinite sum whose partial sums are Taylor polynomials. Some functions *equal* their Taylor series; among these are the sine, cosine and exponential functions.

- A **power series** is an "infinite" polynomial

$$a_0 + a_1x + a_xx^2 + \cdots + a_nx^n + \cdots .$$

If the solution of a differential equation can be represented by a power series, the coefficients a_n can be determined by **recursion relations** obtained by substituting the power series into the differential equation.

- An infinite series **converges** if, no matter how many decimal places are specified, all the partial sums eventually agree to at least this many decimal places. The number defined by these stabilizing decimals is called the **sum** of the series. If a series does not converge, we say it **diverges**.

- If the series $\sum_{m=0}^{\infty} b_m$ converges, then $\lim_{m\to\infty} b_m = 0$. The important counter-example of the **harmonic series** $\sum_{m=1}^{\infty} 1/m$ shows that $\lim_{m\to\infty} b_m = 0$ is a necessary but not sufficient condition to guarantee convergence.

- The **geometric series** $\sum_{m=0}^{\infty} x^m$ converges for all x with $|x| < 1$ and diverges for all other x.

- An **alternating series** $\sum_{m=0}^{\infty}(-1)^m b_m$ converges if $0 < b_{m+1} \leq b_m$ for all m and $\lim_{m\to\infty} b_m = 0$. For a convergent alternating series, the error in approximating the sum by a partial sum is less than the next term in the series.

- A convergent power series converges on an **interval of convergence** of width $2R$; R is called the **radius of convergence**. The **ratio test** can be used to find the radius of convergence of a power series: $\sum_{m=0}^{\infty} b_m$ converges if $\lim_{m\to\infty} |b_{m+1}|/|b_m| < 1$.

- A polynomial $P(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ is the **best fitting** approximation to a function $f(x)$ on an interval $[a, b]$ if $a_0, a_1, \cdots, a_n$ are chosen so that the **mean squared separation** between P and f

  $$\frac{1}{b-a}\int_a^b (P(x) - f(x))^2\,dx$$

  is as small as possible. The polynomial P is also called the **least squares approximation** to f on $[a, b]$.

Expectations

- Given a differentiable function $f(x)$ at a point $x = a$, you should be able to write down any of the **Taylor polynomials** or the **Taylor series** for f at a.

- You should be able to use the program TAYLOR to graph Taylor polynomials.

- You should be able to obtain new Taylor polynomials by substitution, differentiation, anti-differentiation and multiplication.

- You should be able to use Taylor polynomials to find the value of a function to a specified degree of accuracy, to approximate integrals and to find limits.

- You should be able to determine the order of magnitude of the agreement between a function and one of its Taylor polynomials.

- You should be able to find the power series solution to a differential equation.

- You should be able to test a series for divergence; you should be able to check a series for convergence using either the **alternating series test** or the **ratio test**.

- You should be able to find the sum of a **geometric series** and its interval of convergence.

- You should be able to estimate the error in an approximation using partial sums of an alternating series.

- You should be able to find the **radius of convergence** of a series using the ratio test.

- You should be able to set up the equations to find the **least squares** polynomial approximation of a particular degree for a given function on a specified interval. Working by hand or, if necessary, using a computer package to solve a system of equations, you should be able to find the coefficients of the least squares approximation.

Chapter 11

Techniques of Integration

Chapter 6 introduced the integral. There it was defined numerically, as the limit of approximating Riemann sums. Evaluating integrals by applying this basic definition tends to take a long time if a high level of accuracy is desired. If one is going to evaluate integrals at all frequently, it is thus important to find **techniques of integration** for doing this efficiently. For instance, if we evaluate a function at the midpoints of the subintervals, we get much faster convergence than if we use either the right or left endpoints of the subintervals.

A powerful class of techniques is based on the observation made at the end of chapter 6, where we saw that the fundamental theorem of calculus gives us a second way to find an integral, using antiderivatives. While a Riemann sum will usually give us only an approximation to the value of an integral, an antiderivative will give us the exact value. The drawback is that antiderivatives often can't be expressed in **closed form**—that is, as a **formula** in terms of named functions. Even when antiderivatives can be so expressed, the formulas are often difficult to find. Nevertheless, such a formula can be so powerful, both computationally and analytically, that it is often worth the effort needed to find it. In this chapter we will explore several techniques for finding the antiderivative of a function given by a formula.

We will conclude the chapter by developing a numerical method—Simpson's rule—that gives a good estimate for the value of an integral with relatively little computation.

11.1 Antiderivatives

Definition

Recall that we say F is an **antiderivative** of f if $F' = f$. Here are some examples.

FUNCTION:	x^2	$1/y$	$\sin u$	$2\sin t\,\cos t$	2^z
	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$
ANTIDERIVATIVE:	$\dfrac{x^3}{3}$	$\ln y$	$-\cos u$	$\sin^2 t$	$\dfrac{2^z}{\ln 2}$

Undo a differentiation

Notice that you go up (↑) from the bottom row to the top by carrying out a differentiation. To go down (↓) you must "undo" that differentiation. The process of reversing, or undoing, a differentiation has come to be called **antidifferentiation**. You should differentiate each function on the bottom row to check that it is an antiderivative of the function above it.

A function has many antiderivatives

While a function can have only one derivative, it has many antiderivatives. For example, $1 - \cos u$ and $99 - \cos u$ are also antiderivatives of the function $\sin u$ because

$$(1 - \cos u)' = \sin u = (99 - \cos u)'.$$

In fact, every function $C - \cos u$ is an antiderivative of $\sin u$, for any constant C whatsoever. This observation is true in general. That is, if F is an antiderivative of a function f, then so is $F + C$, for any constant C. This follows from the addition rule for derivatives:

$$(F + C)' = F' + C' = F' + 0 = f.$$

A caution

It is tempting to claim the converse—that *every* antiderivative of f is equal to $F + C$, for some appropriately chosen value of C. In fact, you will often see this statement written. The statement is true, though, only for continuous functions. If the function f has breaks in its domain, then there will be more antiderivatives than those of the form $F + C$ for a *single* constant C—over each piece of the domain of f, F can be modified by a *different* constant and still yield an antiderivative for f. Exercises 18, 19, and 20 at the end of this section explore this for a couple of cases. If f is continuous, though, $F + C$ will cover all the possibilities, and we sometimes say that $F + C$ is *the* antiderivative of f. For the sake of keeping a compact notation, we will even write this when the domain of f consists of more than

What the '$+\,C$' term really means

one interval. You should understand, though, that in such cases, over each piece F can be modified by a different constant

Antiderivatives of basic functions

For future reference we collect a list of basic functions whose antiderivatives we already know. Remember that each antiderivative in the table can have an arbitrary constant added to it.

function	antiderivative
x^p	$\dfrac{x^{p+1}}{p+1}, \quad p \neq -1$
$1/x$	$\ln x$
$\sin x$	$-\cos x$
$\cos x$	$\sin x$
e^x	e^x
b^x	$\dfrac{b^x}{\ln b}$

All of these antiderivatives are easily verified and could have been derived with at most a little trial and error fiddling to get the right constant. You should notice one incongruity: the function $1/x$ is defined for all $x \neq 0$, but its listed antiderivative, $\ln x$, is only defined for $x > 0$. In exercise 18 (page 697) you will see how to find antiderivatives for $1/x$ over its entire domain.

Two of our basic functions—$\ln x$ and $\tan x$—do not appear in the left column of the table. This happens because there is no simple multiple of a basic function whose derivative is equal to either $\ln x$ or $\tan x$. It turns out that these functions *do* have antiderivatives, though, that can be expressed as more complicated combinations of basic functions. In fact, by differentiating $x \ln x - x$ you should be able to verify that it is an antiderivative of $\ln x$. Likewise, $-\ln(\cos x)$ is an antiderivative of $\tan x$. It would take a long time to stumble on these antiderivatives by inspection or by trial and error. It is the purpose of later sections to develop techniques which will enable us to discover antiderivatives like these quickly and efficiently. In particular, the antiderivative of $\ln x$ is derived in chapter 11.3 on page 712, while the antiderivative of $\tan x$ is derived in chapter 11.5 on page 744.

There are a couple of other functions that don't appear in the above table whose antiderivatives are needed frequently enough that they should become part of your repertoire of elementary functions that you recognize immediately:

function	antiderivative
$\dfrac{1}{\sqrt{1-x^2}}$	$\arcsin x$
$\dfrac{1}{1+x^2}$	$\arctan x$

The antiderivatives are inverse trigonometric functions, which we've had no need for until now. We introduce them immediately below. They are examples of functions that occur more often for their antiderivative properties than for themselves. Note that the derivatives of the inverse trigonometric functions have no obvious reference to trigonometric relations. In fact, they often occur in settings where there are no triangles or periodic functions in sight. Let's see how the derivatives of these inverse functions are derived.

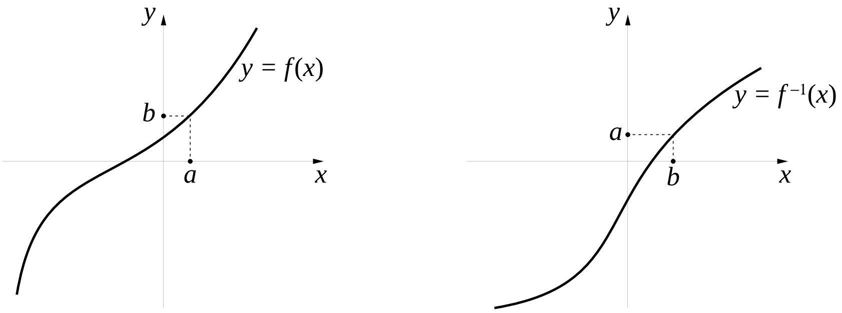
Inverse Functions

We discussed inverse functions in chapter 4.4. Here's a quick summary of the main points made there. Two functions f and g are **inverses** if

$$\begin{aligned} f(g(a)) &= a, \\ \text{and} \quad g(f(b)) &= b, \end{aligned}$$

for every a in the domain of g and every b in the domain of f. It follows that the *range* of f is the same as the *domain* of g, and vice versa. We write $g = f^{-1}$ to indicate that g is the inverse of f, and $f = g^{-1}$ to indicate that f is the inverse of g.

The graphs of $y = f(x)$ and $y = g(x)$ are mirror reflections about the line $y = x$. As the figure below shows, the mirror image of a point with coordinates (a, b) is the point with coordinates (b, a).



This connection between the graphs is a direct translation of the definitions into graphical language, since

$$\begin{aligned}(a,b) \text{ is on the graph of } y = g(x) &\iff g(a) = b \qquad \text{(definition of graph)}\\ &\iff f(b) = a \qquad \text{(definition of inverse)}\\ &\iff (b,a) \text{ is on the graph of } y = f(x).\end{aligned}$$

Because of the connection between the graphs, it follows immediately that if the graph of f is locally linear at the point (b,a) with slope m, then the graph of g will be locally linear at the point (a,b) with slope $1/m$. Algebraically, this is expressed as

$$g'(a) = 1/f'(b),$$

where $a = f(b)$ and $b = g(a)$. We get same result by differentiating the expression $f(g(x)) = x$, using the chain rule:

$$1 = x' = (f(g(x))' = f'(g(x))g'(x),$$

and therefore

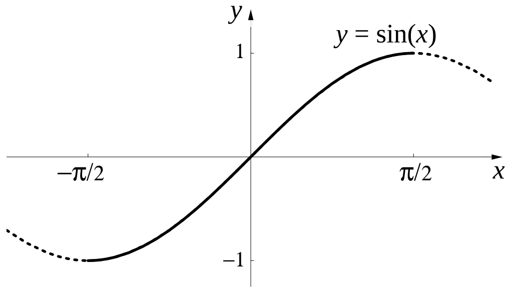
$$g'(x) = \frac{1}{f'(g(x))}$$

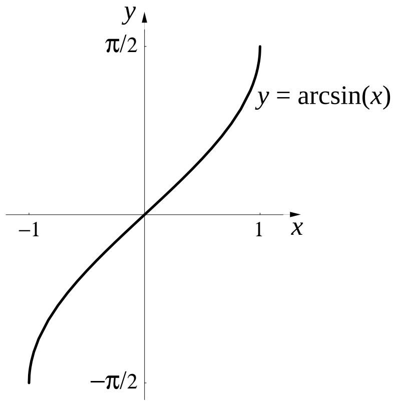
for any value of x for which $g(x)$ is defined.

Inverse trigonometric functions

The arcsine function In the discussion in chapter 4 we saw that a function has an inverse only when it is **one-to-one**, so if we want an inverse, we often have to restrict the domain of a function to a region where it is one-to-one. This is certainly the case with the sine function, which takes the same value infinitely many times. The standard choice of domain on which the sine function is one-to-one is $[-\pi/2, \pi/2]$. Over this interval the sine function increases from -1 to 1. We can then define an inverse function, which we call the **arcsine function**, written $\arcsin x$, whose domain is the interval $[-1, 1]$, and whose range is $[-\pi/2, \pi/2]$. Since the sine function is strictly increasing on its domain, the arcsine function will be strictly increasing on its domain as well—do you see why this has to be?

Another notation for the arcsine function is $\sin^{-1} x$; this form commonly appears on one of the buttons on a calculator.





To find the derivative of the arcsine function, let $f(x) = \sin x$, and let $g(x) = \arcsin x$. Then by the remarks above, we have

$$g'(x) = \frac{1}{f'(g(x))} = \frac{1}{\cos\left(g(x)\right)} = \frac{1}{\cos(\arcsin x)}.$$

The function $\cos(\arcsin x)$ can be expressed in another form, which we will obtain two ways, one algebraic, the other geometric. Both perspectives are useful.

The algebraic approach. Recall that for any input t, $\sin^2 t + \cos^2 t = 1$. This can be solved for $\cos t$ as $\cos t = \pm\sqrt{1 - \sin^2 t}$. That is, the cosine of anything is the square root of 1 minus the square of the sine of that input, with a possible minus sign needed out front, depending on the context. Since the output of the arcsine function lies in the range $[-\pi/2, \pi/2]$, and the cosine function is positive (or 0) for numbers in this interval, it follows that $\cos(\arcsin x) \geq 0$ for any value of x in the domain of the arcsine function. Therefore,

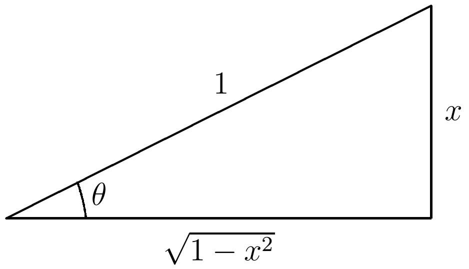
$$\cos(\arcsin x) = \sqrt{1 - \sin^2(\arcsin x)} = \sqrt{1 - (\sin(\arcsin x))^2} = \sqrt{1 - x^2},$$

since $\sin(\arcsin x) = x$ by definition of inverse functions. It follows that

$$(\arcsin x)' = \frac{1}{\cos\left(\arcsin x\right)} = \frac{1}{\sqrt{1 - x^2}},$$

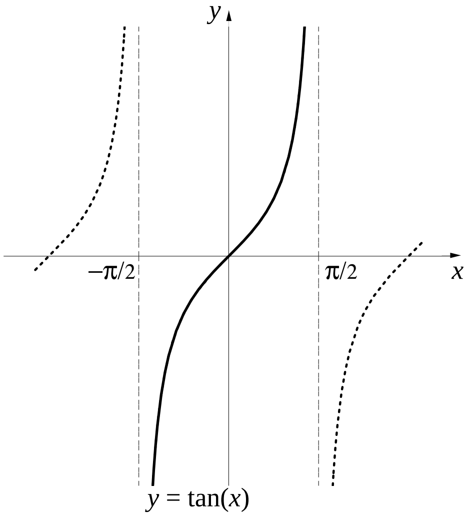
as we indicated above on page 684.

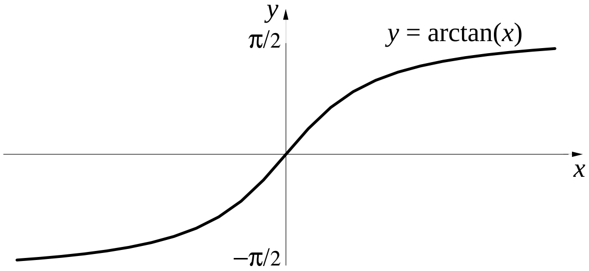
The geometric approach Introduce a new variable $\theta = \arcsin x$, so that $x = \sin\theta$. We can represent these relationships in the following picture:



Notice that we have labelled the side opposite the angle θ as x and the hypotenuse as 1. This ensures that $\sin\theta = x$. By the Pythagorean theorem, the remaining side must then be $\sqrt{1-x^2}$. From this picture it is then obvious that $\cos(\arcsin x) = \cos\theta = \sqrt{1-x^2}/1 = \sqrt{1-x^2}$, as before.

The arctangent function To get an inverse for the tangent function, we again need to limit its domain. Here the standard choice is to restrict it to the interval $(-\pi/2, \pi/2)$ (*not* including the endpoints this time, since $\tan(-\pi/2)$ and $\tan(\pi/2)$ aren't defined). Over this domain the tangent function increases from $-\infty$ to $+\infty$. We can then define the inverse of the tangent, called the **arctangent function**, written $\arctan x$, whose domain is the interval $(-\infty, \infty)$, and whose range is $(-\pi/2, \pi/2)$. Again, both functions are increasing over their domains.





To find the derivative, we proceed as we did with $\arcsin x$, letting $f(x) = \tan x$, and $g(x) = \arctan x$. This time we get

$$g'(x) = \frac{1}{f'(g(x))} = \frac{1}{\sec^2(g(x))} = \frac{1}{\sec^2(\arctan x)}.$$

This expression also has a different form; we obtain it from another trigonometric identity, as we did before, deriving the desired result algebraically and geometrically.

Algebraic We start as before with the identity $\sin^2 t + \cos^2 t = 1$. Dividing through by $\cos^2 t$, we get the equivalent identity $\tan^2 t + 1 = \sec^2 t$. That is, the square of the secant of any input is just 1 plus the square of the tangent applied to the same input. In particular,

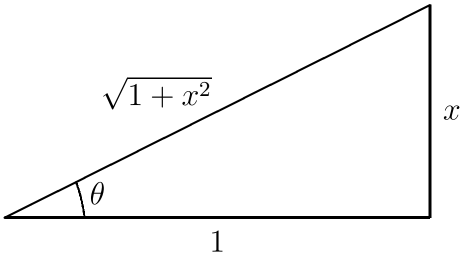
$$\sec^2(\arctan x) = \tan^2(\arctan x) + 1 = (\tan(\arctan x))^2 + 1 = x^2 + 1;$$

since $\tan(\arctan x) = x$. It follows that

$$(\arctan x)' = \frac{1}{\sec^2(\arctan x)} = \frac{1}{1+x^2},$$

as we indicated above on page 684.

Geometric Let $\theta = \arctan x$, so $x = \tan\theta$. Again we can draw and label a triangle reflecting these relationships:



Notice that this time we have labelled the side opposite the angle θ as x and the adjacent side as 1 to ensure that $\tan\theta = x$. Again by the Pythagorean theorem, the hypotenuse must be $1+x^2$. From this picture it is then obvious that $\sec(\arctan x) = \sec\theta = (\sqrt{1+x^2})/1 = \sqrt{1+x^2}$, so $\sec^2(\arctan x) = 1 + x^2$ again.

Notation

According to the fundamental theorem of calculus (see chapter 6.4), every accumulation function

$$A(x) = \int_a^x f(t)\,dt$$

is an antiderivative of the function f, no matter at what point $t = a$ the accumulation begins—so long as the function is defined over the entire interval from $t = a$ to $t = x$. (This is an important caution when dealing with functions like $1/x$, for instance, for which the integral from, say, -1 to 1 makes no sense.) In other words, the expression

$$\int_a^x f(t)\,dt$$

represents an antiderivative of f. The influence of the fundamental theorem is so pervasive that this expression—with the "limits of integration" a and x omitted—is used to denote an antiderivative:

Write an antiderivative as an integral

> **Notation**: The antiderivative of f is $\int f(x)\,dx$.

With this new notation the antiderivatives we have listed so far can be written in the following form.

$$\begin{aligned}
\int x^p\,dx &= \frac{1}{p+1}\,x^{p+1} + C \qquad (\,p \neq -1) \\
\int \frac{1}{x}\,dx &= \ln x + C \\
\int \sin x\,dx &= -\cos x + C \\
\int \cos x\,dx &= \sin x + C \\
\int e^x\,dx &= e^x + C \\
\int b^x\,dx &= \frac{1}{\ln b\, b^x} + C \\
\int \frac{1}{\sqrt{1-x^2}}\,dx &= \arcsin x + C \\
\int \frac{1}{1+x^2}\,dx &= \arctan x + C
\end{aligned}$$

The basic antiderivatives again

The integration sign $\int$ now has two distinct meanings. Originally, it was used to describe the *number*

The *definite* integral is a number. . .

$$\int_a^b f(x)\,dx,$$

which was always calculated as the limit of a sequence of Riemann sums. Because this integral has a definite numerical value, it is called the **definite integral**. In its new meaning, the integration sign is used to describe the antiderivative

$$\int f(x)\,dx,$$

. . . while the *indefinite* integral is a function

which is a *function*, not a number. To contrast the new use of $\int$ with the old, and to remind us that the new expression is a variable quantity, it is called the **indefinite integral**. The function that appears in either a definite or an indefinite integral is called the **integrand**. The terms "antiderivative" and "indefinite integral" are completely synonymous. We will tend to use the former term in general discussions, using the latter term when focusing on the process of finding the antiderivative.

Because an indefinite integral represents an antiderivative, the process of finding an antiderivative is sometimes called **integration**. Thus the term *integration*, as well as the symbol for it, has two distinct meanings.

Using Antiderivatives

According to the fundamental theorem, we can use an *indefinite* integral to find the value of a *definite* integral—and this largely explains the importance of antiderivatives. In the language of indefinite integrals, the statement of the fundamental theorem in the box on page 410 takes the following form.

$$\boxed{\int_a^b f(x)\,dx = F(b) - F(a), \text{ where } F(x) = \int f(x)\,dx.}$$

Example 1 Find $\displaystyle\int_1^4 x^2\,dx$. We have that

$$\int x^2\,dx = \tfrac{1}{3}x^3 + C;$$

it follows that

$$\int_1^4 x^2\,dx = \tfrac{1}{3}4^3 + C - \left(\tfrac{1}{3}1^3 + C\right) = \frac{64}{3} + C - \frac{1}{3} - C = 21.$$

Example 2 Find $\int_0^{\pi/2} \cos t\,dt$. This time the indefinite integral we need is

$$\int \cos t\,dt = \sin t + C.$$

The value of the definite integral is therefore

$$\int_0^{\pi/2} \cos t\,dt = \sin \pi/2 + C - (\sin 0 + C) = 1 + C - 0 - C = 1.$$

In each example the two appearances of C cancel each other. Thus C does not appear in the final result. This implies that it does not matter which value of C we choose to do the calculation. Usually, we just take $C = 0$.

The calculation doesn't depend on C, so take $C = 0$

Notation: Because expressions like $F(b) - F(a)$ occur often when we are using indefinite integrals, we use the abbreviation

$$F(b) - F(a) = F(x)\Big|_a^b .$$

Thus,

$$\int_1^4 x^2\,dx = \frac{x^3}{3}\Big|_1^4 = \frac{64}{3} - \frac{1}{3} = 21.$$

There are clear advantages to using antiderivatives to evaluate definite integrals: we get exact values and we avoid many lengthy calculations. The difficulty is that the method works only if we can find a formula for the antiderivative.

There are several reasons why we might not find the formula we need. For instance, the antiderivative we want may be a function we have never seen before. The function $\arctan x$ is an example.

We may not recognize the antiderivative...

Even if we have a broad acquaintance with functions, we may still not be able to find the formula for a given antiderivative. The reason is simple: for most functions we can write down, the antiderivative is just not among the basic functions of calculus. For example, none of the basic functions, in any form or combination, equals

...and there may even be no formula for it

$$\int e^{x^2} dx \qquad \text{or} \qquad \int \frac{\sin x}{x}\,dx.$$

This does not mean that e^{x^2}, for example, has no antiderivative. On the contrary, the accumulation function

$$\int_0^x e^{t^2} dt$$

is an antiderivative of e^{x^2}. It can be evaluated, graphed, and analyzed like any other function. What we lack is a *formula* for this antiderivative in terms of the basic functions of calculus.

Finding Antiderivatives

In the rest of this chapter we will be deriving a number of statements involving antiderivatives. It is important to remember what such statements mean.

The following statements are completely equivalent:

$$\int f(x)\,dx = F(x) + C \qquad \text{and} \qquad F' = f.$$

We differentiate to verify a statement about antiderivatives

In other words, a statement about antiderivatives can be verified by looking at a statement about derivatives. If it is claimed that the antiderivative of f is F, you check the statement by seeing if it is true that $F' = f$. This was how we verified the elementary antiderivatives we've considered so far. Another way of expressing these relationships is

$$\left(\int f(x)\,dx\right)' = f(x) \quad \text{and} \quad \int F'(x)\,dx = F(x) + C$$

for any functions f and F.

This duality between statements about derivatives and statements about antiderivatives also holds when applied to more general statements. Many of the most useful techniques for finding antiderivatives are based on converting the general rules for taking derivatives of sums, products, and chains into equivalent antiderivative form. We'll start with the simplest combinations, which involve a **function multiplied by a constant** and a **sum of two functions**.

The constant multiple and addition rules

derivative form	antiderivative form
$(k \cdot F)' = k \cdot F'$	$\displaystyle\int k \cdot f\,dx = k \cdot \int f\,dx$
$(F + G)' = F' + G'$	$\displaystyle\int (f+g)\,dx = \int f\,dx + \int g\,dx$

Let's verify these rules. Set $\int f = F$ and $\int g = G$. Then the first rule is claiming that $\int (k \cdot f) = k \cdot F$—the antiderivative of a constant times a function equals the constant times the antiderivative of the function. To verify this, we have to show that when we take the derivative of the right-hand side, we get the function under the integral on the left-hand side. But $(k \cdot F)' = k \cdot F'$ (by the derivative rule), which is just $k \cdot f$, which is what we had to show.

Similarly, to show that $\int (f + g) = F + G$—the antiderivative of the sum of two functions is the sum of their separate antiderivatives—we differentiate the right-hand side and find $(F + G)' = F' + G'$ (by the derivative rule for sums) which is just $f + g$, so the rule is true.

Example 3 This example illustrates the use of both the addition and the constant multiple rules.

$$\begin{aligned}\int (7e^x + \cos x)\,dx &= \int 7e^x\,dx + \int \cos x\,dx \\ &= 7\int e^x\,dx + \int \cos x\,dx \\ &= 7e^x + \sin x + C.\end{aligned}$$

To verify this answer, you should take the derivative of the right-hand side to see that it equals the integrand on the left-hand side.

In the following sections we will develop the antidifferentiation rules that correspond to the product rule and to the chain rule. They are called *integration by parts* and *integration by substitution*, respectively.

Trial and error

While antiderivatives can be hard to find, they are easy to check. This makes "trial and error" a good strategy. In other words, if you don't immediately see what the antiderivative should be, but the function doesn't look too bad, try guessing. When you differentiate your guess, what you see may lead you to a better guess.

Example 4 Find

$$F(x) = \int \cos(3x)\, dx.$$

Since the derivative of $\sin u$ is $\cos u$, it is reasonable to try $\sin(3x)$ as an antiderivative for $\cos(3x)$. Therefore:

$$\begin{aligned} \text{FIRST GUESS: } & F(x) = \sin(3x), \\ \text{CHECK: } & F'(x) = \cos(3x) \cdot 3 \neq \cos(3x). \end{aligned}$$

We wanted $F' = \cos(3x)$ but we got $F' = 3\cos(3x)$. The chain rule gave us an extra—and unwanted—factor of 3. We can compensate for that factor by multiplying our first guess by 1/3. Then

$$\begin{aligned} \text{SECOND GUESS: } & F(x) = \tfrac{1}{3}\sin(3x), \\ \text{CHECK: } & F'(x) = \tfrac{1}{3}\cos(3x) \cdot 3 = \cos(3x). \end{aligned}$$

Thus $\displaystyle\int \cos(3x)\, dx = \tfrac{1}{3}\sin(3x)$.

Tables of Integrals

Because indefinite integrals are difficult to calculate, reference manuals in mathematics and science often include tables of integrals. There are sometimes many hundreds of individual formulas, organized by the type of function being integrated. A modest selection of such formulas can be found at the back of this book. You should take some time to learn how these tables are arranged and get some practice using them. You should also check some of the more unlikely looking formulas by differentiating them to see that they really are the antiderivatives they are claimed to be.

Computers are having a major impact on integration techniques. We saw in the last chapter that any continuous function—even one given by the output from some laboratory recording device or as the result of a numerical technique like Euler's method—can be approximated by a polynomial (usually using lots of computation!), and the antiderivative of a polynomial is easy to find.

Integration can now be done quickly and efficiently by computer software

Moreover, computer software packages which can find any existing formula for a definite integral are becoming widespread and will probably have a profound impact on the importance of integration techniques over the next several years. Just as hand-held calculators have rendered obsolete many traditional arts—like using logarithms for performing multiplications or knowing how to interpolate in trig tables—there is likely to be a decreased importance placed on humans' being adept at some of the more esoteric integration techniques. While some will continue to derive pleasure from becoming proficient in these skills, for most users it will generally be much faster, and more accurate, to use an appropriate software package. Nevertheless, for those going on in mathematics and the physical sciences, it will still to be useful to be able to perform some of the simpler integrations by hand reasonably rapidly. The subsequent sections of this chapter develop the most commonly needed techniques for doing this.

Exercises

1. What is the inverse of the function $y = 1/x$? Sketch the graph of the function and its inverse.

2. What is the inverse of the function $y = 1/x^3$? Sketch the graph of the the function and its inverse. Do the same for $y = 3x - 2$.

3. a) Let $\theta = \arctan x$. Then $\tan\theta = x$. Refer to the picture on page 688 showing the relationship between x and θ. Use this drawing to show that $\arctan(x) + \arctan(1/x)$ is constant—that is, its value doesn't depend on x. What is the value of the constant?

b) Use part (a), and the derivative of $\arctan x$, to find the derivative of $\arctan(1/x)$.

c) Use the chain rule to verify your answer to part (b).

4. The logarithm for the base b is defined as the inverse to the exponential function with base b:

$$y = \log_b x \qquad \text{if} \qquad x = b^y.$$

Using only the fact that $dx/dy = \ln b \cdot b^y$, deduce the formula

$$\frac{dy}{dx} = \frac{1}{\ln b} \cdot \frac{1}{x}.$$

Note: this is purely an algebra problem; you don't need to invoke any differentiation rules.

5. a) Define $\arccos x$, the inverse of the cosine function. Be sure to limit the domain of the cosine function to an interval on which it is one-to-one.

b) Sketch the graph $y = \arccos x$. How did you limit the range of y?

c) Determine dy/dx.

6. a) If $\theta = \arcsin x$, refer to the picture on page 687 reflecting the relation between θ and x. Using this picture, proceed as in problem 3 to to show that the sum $\arcsin x + \arccos x$ is constant. What is the value of the constant?

b) Use part (a), and the derivative of $\arcsin x$, to determine the derivative of $\arccos x$. Does this result agree with what you got in the last exercise?

7. Find a formula for $\displaystyle\int \frac{dx}{\sqrt{1-x^2}}$.

8. Verify that the antiderivatives given in the table on page 689 are correct.

9. Find an antiderivative of each of the following functions. Don't hesitate to use the "trial and error" method of Example 4 above.

$$\begin{array}{llll}
3 & 5t & -5t & 3-5t \\
7x^4 & \dfrac{1}{y^3} & e^{2z} & u+\dfrac{1}{u} \\
(1+w^3)^2 & \cos(5v) & x^9+5x^7-2x^5 & \sin t\,\cos t
\end{array}$$

10. Find a formula for each of the following indefinite integrals.

a) $\displaystyle\int 3x\,dx$

b) $\displaystyle\int 3u\,du$

c) $\displaystyle\int e^z\,dz$

d) $\displaystyle\int 5t^4\,dt$

e) $\displaystyle\int 7y+\frac{1}{y}\,dy$

f) $\displaystyle\int 7y-\frac{4}{y^2}\,dy$

g) $\displaystyle\int 5\sin w - 2\cos w\,dw$

h) $\displaystyle\int dx$

i) $\displaystyle\int e^{z+2}\,dz$

j) $\displaystyle\int \cos(4x)\,dx$

k) $\displaystyle\int \frac{5}{1+r^2}\,dr$

l) $\displaystyle\int \frac{1}{\sqrt{1-4s^2}}\,ds$

11. a) Find an antiderivative $F(x)$ of $f(x) = 7$ for which $F(0) = 12$.

b) Find an antiderivative $G(x)$ of $f(x) = 7$ for which $G(3) = 1$.

c) Do $F(x)$ and $G(x)$ differ by a constant? If so, what is the value of that constant?

12. a) Find an antiderivative $F(t)$ of $f(t) = t + \cos t$ for which $F(0) = 3$.

b) Find an antiderivative $G(t)$ of $f(t) = t + \cos t$ for which $G(\pi/2) = -5$.

c) Do $F(t)$ and $G(t)$ differ by a constant? If so, what is the value of that constant?

13. Find an antiderivative of the function $a + by$ when a and b are fixed constants.

14. a) Verify that $(1 + x^3)^{10}$ is an antiderivative of $30x^2(1 + x^3)^9$.

b) Find an antiderivative of $x^2(1 + x^3)^9$.

c) Find an antiderivative of $x^2 + x^2(1 + x^3)^9$.

15. a) Verify that $x \ln x$ is an antiderivative of $1 + \ln x$.

b) Find an antiderivative of $\ln x$. [Do you see how you can use part (a) to find this antiderivative?]

16. Recall that $F(y) = \ln(y)$ is an antiderivative of $1/y$ for $y > 0$. According to the text, *every* antiderivative of $1/y$ over this domain must be of the form $\ln(y) + C$ for an appropriate value of C.

a) Verify that $G(y) = \ln(2y)$ is also an antiderivative of $1/y$.

b) Find C so that $\ln(2y) = \ln(y) + C$.

17. Verify that $-\cos^2 t$ is an antiderivative of $2 \sin t \, \cos t$. Since you already know $\sin^2 t$ is an antiderivative, you should be able to show

$$-\cos^2 t = \sin^2 t + C$$

for an appropriate value of C. What is C?

18. Since the function $\ln x$ is defined only when $x > 0$, the equation

$$\int \frac{1}{x}\, dx = \ln x + C$$

applies only when $x > 0$. However, the integrand $1/x$ is defined when $x < 0$ as well. Therefore, it makes sense to ask what the integral (i.e., antiderivative) of $1/x$ is when $x < 0$.

a) When $x < 0$ then $-x > 0$ so $\ln(-x)$ is defined. In these circumstances, show that $\ln(-x)$ is an antiderivative of $1/x$.

b) Now put these two "pieces" of antiderivative together by defining the function

$$F(x) = \begin{cases} \ln(-x) & \text{if } x < 0, \\ \ln(x) & \text{if } x > 0 \end{cases}$$

Sketch together the graphs of the functions $F(x)$ and $1/x$ in such a way that it is clear that $F(x)$ is an antiderivative of $1/x$.

c) Explain why $F(x) = \ln|x|$. For this reason a table of integrals often contains the entry

$$\int \frac{1}{x}\,dx = \ln|x| + C.$$

d) Every function $\ln|x| + C$ is an antiderivative of $1/x$, but there are even more. As you will see, this can happen because the domain of $1/x$ is broken into two parts. Let

$$G(x) = \begin{cases} \ln(-x) & \text{if } x < 0, \\ \ln(x) + 1 & \text{if } x > 0. \end{cases}$$

Sketch together the graphs of the functions $G(x)$ and $1/x$ in such a way that it is clear that $G(x)$ is an antiderivative of $1/x$.

e) Explain why there is no value of C for which

$$\ln|x| + C = G(x).$$

This shows that the functions $\ln|x| + C$ do not exhaust the set of antiderivatives of $1/x$.

f) Construct two more antiderivatives of $1/x$ and sketch their graphs. What is the general form of the new antiderivatives you have constructed? (A suggestion: you should be able to use two separate constants C_1 and C_2 to describe the general form.)

19. On page 683 of the text there is an antiderivative for the tangent function:

$$\int \tan x\,dx = -\ln(\cos x).$$

However, this is not defined when x makes $\cos x$ either zero or negative.

a) How many separate intervals does the domain of $\tan x$ break down into?

b) For what values of x is $\cos x$ equal to zero, and for what values is it negative?

c) Modify the antiderivative $-\ln(\cos x)$ so that it *is* defined when $\cos x$ is negative. (How is this problem with the logarithm function treated in the previous question?)

d) In a typical table of integrals you will find the statement

$$\int \tan x\,dx = -\ln|\cos x| + C.$$

Explain why this does not cover all the possibilities.

e) Give a more precise expression for $\int \tan x\,dx$, modelled on the way you answered part (f) of problem 18. How many different constants will you need?

f) Find a function G that is an antiderivative for $\tan x$ and that also satisfies the following conditions:

$$G(0) = 5, \qquad G(\pi) = -23, \qquad G(17\pi) = 197.$$

20. In the table on page 689 the antiderivative of x^p is given as

$$\frac{1}{p+1}x^{p+1} + C.$$

For some values of p this is correct, with only a single constant C needed. For other values of p, though, the domain of x^p will consist of more than one piece, and $\frac{1}{p+1}x^{p+1}$ can be modified by a different constant over each piece. For what values of p does this happen?

21. Find $F'(x)$ for the following functions. In parts (a), (b), and (d) do the problems two ways: by finding an antiderivative, and by using the fundamental theorem to get the answer without evaluating an antiderivative. Check that the answers agree.

a) $F(x) = \displaystyle\int_0^x (t^2 + t^3)\,dt.$

b) $F(x) = \displaystyle\int_1^x \frac{1}{u}\,du.$

c) $F(x) = \displaystyle\int_1^x \frac{v}{1+v^3}\,dv.$

d) $F(x) = \displaystyle\int_0^{x^2} \cos t\,dt.$

e) $F(x) = \displaystyle\int_1^{x^2} \frac{v}{1+v^3}\,dv.$ [Hint: let $u = x^2$ and use the chain rule.]

Comment: It may seem that parts (c) and (e) are more difficult than the others. However, there is a way to apply the fundamental theorem of calculus here to get answers to parts (c) and (e) quickly and with little effort.

22. Consider the two functions

$$F(x) = \sqrt{1+x^2} - 1 \quad \text{and} \quad G(x) = \int_0^x \frac{t}{\sqrt{1+t^2}}\,dt.$$

a) Show that F and G both satisfy the initial value problem

$$y' = \frac{x}{\sqrt{1+x^2}}, \qquad y(0) = 0.$$

b) Since an initial value problem typically has a *unique* solution, F and G should be equal. Assuming this, determine the exact value of the following definite integrals.

$$\int_0^1 \frac{t}{\sqrt{1+t^2}}\,dt, \qquad \int_0^2 \frac{t}{\sqrt{1+t^2}}\,dt, \qquad \int_0^5 \frac{t}{\sqrt{1+t^2}}\,dt.$$

23. The connection between integration and differentiation that is provided by the fundamental theorem of calculus makes it possible to determine an integral by solving a differential equation. For example, the accumulation function

$$A(x) = \int_0^x e^{-t^2}\,dt$$

is the solution to the initial value problem

$$y' = e^{-x^2}, \qquad y(0) = 0.$$

Therefore, $A(x)$ can be found by solving the differential equation. As you have seen, Euler's method is a useful way to solve differential equations.

a) Use either a program (e.g., PLOT) or a differential equation solver on a computer to get a graphical solution $A(x)$ to the initial value problem above.

b) Sketch the graph of $y = A(x)$ over the domain $0 \le x \le 4$.

c) Your graph should increase from left to right. How can you tell this even before you see the computer output?

d) Your graph should level off as x increases. Determine $A(5)$, $A(10)$, $A(30)$. (Approximations provided by the computer are adequate here.)

e) Estimate $\lim_{x\to\infty} A(x)$. [The *exact* value is $\sqrt{\pi}/2$.]

f) Determine $\int_0^1 e^{-t^2} dt$ and $\int_0^2 e^{-t^2} dt$.

g) Determine $\int_1^2 e^{-t^2} dt$.

24. Find the area under the curve $y = x^3 + x$ for x between 1 and 4. (See chapter 6.3.)

25. Find the area under the curve $y = e^{3x}$ for x between 0 and $\ln 3$.

26. The **average value** of the function $f(x)$ on the interval $a \le x \le b$ is the integral

$$\frac{1}{b-a}\int_a^b f(x)\,dx.$$

(See the discussion of average value in chapter 6, pages 397–399.)

a) Find the average value of each of the functions $y = x$, x^2, x^3, and x^4 on the interval $0 \le x \le 1$.

b) Explain, using the graphs of $y = x$ and $y = x^2$, why the average value of x^2 is less than the average value of x on the interval $[0, 1]$.

27. a) What are the maximum, minimum, and average values of the function

$$f(x) = x + 2e^{-x}$$

on the interval $[0, 3]$?

b) Sketch the graph of $y = f(x)$ on the interval $[0, 3]$. Draw the line $y = \mu$, where μ is the average value you found in part (a).

c) For which x does the graph of $y = f(x)$ lie above the line $y = \mu$, and for which x does it lie below the line? The region between the graph and the line has two parts—one is above the line (and below the graph) and the other is below the line (and above the graph). Shade these two regions and compare their areas: which is larger? [The two are equal.]

11.2 Integration by Substitution

In the preceding section we converted a couple of general rules for differentiation—the rule for the derivative of a constant times a function and the rule for the derivative of the sum of two functions—into equivalent rules in integral form. In this section we will develop the integral form of the chain rule and see some of the ways this can be used to find antiderivatives.

Suppose we have functions F and G, with corresponding derivatives f and g. Then the chain rule says

$$(F(G(x)))' = F'(G(x))\,G'(x) = f(G(x))\,g(x).$$

If we now take the indefinite integral of these equations, we get

$$F(G(x)) + C = \int (F(G(x)))'\,dx = \int f(G(x))\,g(x)\,dx,$$

where C can be any constant.

The integral form of the chain rule

Reversing these equalities to get a statement about integrals, we obtain:

$$\boxed{\int \boldsymbol{f(G(x))\,g(x)\,dx = F(G(x)) + C.}}$$

This somewhat unpromising expression turns out to be surprisingly useful. Here's how: Suppose we want to find an indefinite integral, and see in the integrand a *pair* of functions G and g, where $G' = g$ and where $g(x)$ can be factored out of the integrand. We then find a function f so that the integrand can be written in the form $f(G(x))g(x)$. Now we only have to find an antiderivative for f. Once we have such an antiderivative, call it F, then the solution to our original problem will be $F(G(x))$. Thus, while our original antiderivative problem is not yet solved, it has been reduced to a

Reduction methods transform problems into equivalent, simpler, problems different, simpler antiderivative problem that, when all goes well, will be easier to evaluate. Such **reduction methods** are typical of many integration techniques. We will see other examples in the remainder of this chapter.

Example 1 Suppose we try to find a formula for the integral

$$\int 3x^2\left(1+x^3\right)^7 dx.$$

One way would be to multiply out the expression $\left(1+x^3\right)^7$, making the integrand a polynomial with many separate terms of different degrees. (The highest degree would be 23; do you see why?) We could then carry out the integration "term-by-term," using the rules for sums and constant multiples of powers of x that were given in the previous section. But this is tedious—even excruciating.

Instead, notice that the expression $3x^2$ is the derivative of $1+x^3$. If we let $G(x)=1+x^3$ and $g(x)=3x^2$, we can then write the integrand as

$$3x^2\left(1+x^3\right)^7=(G(x))^7 g(x).$$

Now $(G(x))^7$ is clearly just $f(G(x))$ where f is the function which raises its input to the 7th power—$f(x)=x^7$ for any input x. But we recognize f as an elementary function whose antiderivative we can write down immediately as $F(x)=\frac{1}{8}x^8$. Thus the solution to our original problem will be

$$\int 3x^2\left(1+x^3\right)^7 dx=F(G(x))+C=\frac{1}{8}\left(1+x^3\right)^8+C.$$

As with any integration problem, we can check our answer by taking the derivative of the right-hand side to see if it agrees with the integrand on the left. You should do this whenever you aren't quite sure of your technique (or your arithmetic!).

Note that the term $3x^2$ that appeared in the integrand above was essential for the procedure to work. The integral

$$\int\left(1+x^3\right)^7 dx$$

cannot be found by substitution, even though it appears to have a simpler form. (Of course, the integral *can* be found by multiplying out the integrand.)

A compact notation for expressing the integral form of the chain rule

Using differential notation So far the symbol dx (the **differential** of x) under the integral sign has simply been an appendage, tagging along to suggest the Δx portion of the Riemann sums approximating definite integrals. It turns out we can take advantage of this notation to use the integral form of the chain rule more compactly.

Instead of naming the functions G and g as above, we introduce a new variable $u = G(x)$. Then

$$\frac{du}{dx} = G'(x) = g(x),$$

and it is suggestive to multiply out this "quotient" to get

$$du = g(x)\,dx.$$

While this is reminiscent of the microscope equation we met in chapter 3, and 18th century mathematicians took this equation seriously as a relation between two "infinitesimally small" quantities dx and du, we will view it only as a convenient mnemonic device. To see how this simplifies computations, reconsider the previous example. If we let $u = 1 + x^3$, then $du = 3x^2\,dx$, so we can write

$$3x^2(1+x^3)^7\,dx = (\underbrace{1+x^3}_{u})^7\,\underbrace{3x^2\,dx}_{du} = u^7\,du.$$

It follows that

$$\begin{aligned}\int 3x^2\left(1+x^3\right)^7 dx &= \int u^7\,du\\ &= \tfrac{1}{8}u^8 + C\\ &= \tfrac{1}{8}\left(1+x^3\right)^8 + C,\end{aligned}$$

as before. We have arrived at the same answer without having to introduce the cumbersome language of all the auxiliary functions—we simply **substituted** the variable u for a certain expression in x (which we called $G(x)$ before), and replaced $G'(x)dx$ by du. For this reason this technique is called **integration by substitution**. You should always be clear, though, that integration by substitution is just the integral form of the chain rule, a relationship that becomes clear whenever you check the answer substitution gives you.

Example 2 Can we use the method of substitution to find

$$\int \frac{e^{5x}}{6+e^{5x}}\,dx?$$

The numerator is almost the derivative of the denominator. This suggests we let $G(x) = 6 + e^{5x}$, giving $g(x) = G'(x) = 5e^{5x}$. Since we need to be able to factor $g(x)$ out of the integrand, we multiply numerator and denominator by 5 to get

$$\begin{aligned}\int \frac{e^{5x}}{6+e^{5x}}\,dx &= \int \frac{1}{5}\cdot\frac{1}{6+e^{5x}}5e^{5x}\,dx\\ &= \int \frac{1}{5}\cdot\frac{1}{G(x)}g(x)\,dx\\ &= \frac{1}{5}\int f(G(x))g(x)\,dx,\end{aligned}$$

where f is just the reciprocal function—$f(x) = 1/x$. But an antiderivative for f is just $\ln x$, so the desired antiderivative is just

$$\frac{1}{5}F(G(x)) + C = \frac{1}{5}\ln(6+e^{5x}) + C.$$

As usual, you should check this answer by differentiating to see that you really do get the original function.

Now let's see how this works using differential notation. If we set $u = 6 + e^{5x}$, then

$$\frac{du}{dx} = 5e^{5x} \qquad \text{and} \qquad du = 5e^{5x}\,dx.$$

Again we insert a factor of 5 in the numerator and an identical one in the denominator to balance it. Substitutions for u and du then yield the following:

$$\begin{aligned}\int \frac{1}{5}\cdot\frac{5e^{5x}}{6+e^{5x}}\,dx &= \frac{1}{5}\int \frac{5e^{5x}\,dx}{6+e^{5x}}\\ &= \frac{1}{5}\int \frac{du}{u}\\ &= \frac{1}{5}\ln(u) + C\\ &= \frac{1}{5}\ln(6+e^{5x}) + C,\end{aligned}$$

as before.

The basic structure

The two examples above have the same structure. In both, a certain function of x is selected and called u; part of the integrand, namely $u'\,dx$, becomes du, the rest becomes one of the basic functions of u. Specifically:

integrand	u	du	function of u
$3x^2\,(1+x^3)^7\,dx$	$1+x^3$	$3x^2\,dx$	u^7
$\dfrac{e^{5x}}{6+e^{5x}}\,dx$	$6+e^{5x}$	$5e^{5x}\,dx$	$\dfrac{1}{5}\cdot\dfrac{1}{u}$

Note that you may have to do a bit of algebraic reshaping of the integrand to cast it in the proper form. For example, we had to insert a factor of 5 to the numerator of the second example to make the numerator be the derivative of the denominator. There is no set routine to be followed to find an antiderivative most efficiently, or even any way to know whether a particular method will work until you try it. Success comes with experience and a certain amount of intelligent fiddling until something works out.

Example 3 The method of substitution is useful in simple problems, too. Consider

$$\int \cos(3t)\,dt.$$

If we set $u = 3t$, then $du = 3\,dt$ and

$$\begin{aligned}\int \cos(3t)\,dt &= \int \cos(u)\cdot\tfrac{1}{3}\,du\\ &= \tfrac{1}{3}\int \cos(u)\,du\\ &= \tfrac{1}{3}\sin(u)+C\\ &= \tfrac{1}{3}\sin(3t)+C.\end{aligned}$$

Substitution in Definite Integrals

Until now we have been using the technique of substitution to find antiderivatives—that is, to evaluate *indefinite integrals*. Refer back to the integral form of the chain rule given in the box on page 702, and see what happens when we use this equation to evaluate a *definite integral*. Suppose we want to evaluate

$$\int_a^b f(G(x))\,g(x)\,dx.$$

We know that $F(G(x))$ is an antiderivative, so by the fundamental theorem we have

$$\int_a^b f(G(x))\, g(x)\, dx = F(G(x))\big|_a^b = F(G(b)) - F(G(a)).$$

Now suppose we make the substitution $u = G(x)$ and $du = g(x)\, dx$. Then as x goes from a to b, u will go from $G(a)$ to $G(b)$. Moreover, we have

$$\int_{G(a)}^{G(b)} f(u)\, du = F(u)\big|_{G(a)}^{G(b)} = F(G(b)) - F(G(a)),$$

so the two definite integrals have the same value. In other words,

If we make the substitution $u = G(x)$, then

$$\boldsymbol{\int_a^b f(G(x))\, g(x)\, dx = F(G(b)) - F(G(a)) = \int_{G(a)}^{G(b)} f(u)\, du.}$$

This means that to evaluate a definite integral by substitution, we can do everything in terms of u. We don't ever need to find an antiderivative for the original integrand in terms of x or use the original limits of integration.

Example 4 Consider the definite integral

$$\int_0^{\pi/2} \frac{\cos x\, dx}{1 + \sin x}.$$

We can evaluate this integral by making the substitution $u = 1 + \sin x$, and $du = \cos x\, dx$. Moreover, as x goes from 0 to $\pi/2$, u goes from 1 to 2. Therefore

$$\int_0^{\pi/2} \frac{\cos x\, dx}{1 + \sin x} = \int_1^2 \frac{du}{u} = \ln u\big|_1^2 = \ln 2.$$

Check that this is the same answer you would have gotten if you had expressed the antiderivative $\ln u$ in terms of x and evaluated the result at the limits on the original integral.

Example 5 Evaluate

$$\int_0^1 6x^2(1 + x^3)^4\, dx.$$

With the substitution $u = 1 + x^3$ and $du = 3x^2\,dx$, as x goes from 0 to 1, u goes from 1 to 2. Our integral thus becomes

$$\int_0^1 6x^2(1+x^3)^4\,dx = \int_0^2 2u^4\,du = \frac{2}{5}u^5\bigg|_1^2 = \frac{64}{5}.$$

Exercises

1. Evaluate the following using substitution. Do parts (a) through (e) in two ways: i. by writing the integrand in the form $f(G(x))g(x)$ (or y or t or whatever the variable is) for appropriate functions f, G, and g, with $G' = g$, and then finding $F = \int f$; and ii. using differential notation. Do the remaining parts in the way you feel most confident.

a) $\displaystyle\int 2y(y^2+1)^{50}\,dy$

b) $\displaystyle\int \sin(5z)\,dz$

c) $\displaystyle\int \frac{e^{\sqrt{x}}}{\sqrt{x}}\,dx$

d) $\displaystyle\int (5t+7)^{50}\,dt$

e) $\displaystyle\int 3u^2\,\sqrt[3]{u^3+8}\,du$

f) $\displaystyle\int \frac{1}{2v+1}\,dv$

g) $\displaystyle\int \tan x\,dx$

h) $\displaystyle\int \tan^2(x)\sec^2(x)\,dx$

i) $\displaystyle\int \sec(x/2)\tan(x/2)\,dx$

j) $\displaystyle\int \sin(w)\,\sqrt{\cos(w)}\,dw$

k) $\displaystyle\int \frac{\sin(\sqrt{s})}{\sqrt{s}}\,ds$

l) $\displaystyle\int \sqrt{3-x}\,dx$

m) $\displaystyle\int \frac{dr}{r\ln r}$

n) $\displaystyle\int e^x\sin(1+e^x)\,dx$

o) $\displaystyle\int \frac{y}{1+y^2}\,dy$

p) $\displaystyle\int \frac{w}{\sqrt{1-w^2}}\,dw$

q) $\displaystyle\int \frac{1}{1+4y^2}\,dy$

r) $\displaystyle\int \frac{1}{\sqrt{1-9w^2}}\,dw$

2. Use integration by substitution to find the numerical value of the following. In four of these you should get your answer in two ways: i) by finding an antiderivative for the given integrand, and ii) by using the observation in the box on page 707, comparing the results. You should also check your results for three of the problems by finding numerical estimates for the integrals using RIEMANN.

a) $\displaystyle\int_0^1 \frac{e^s}{e^s+1}\,ds$

b) $\displaystyle\int_0^{\ln e} \frac{e^s}{e^s+1}\,ds$

c) $\displaystyle\int_1^3 \frac{1}{2x+1}\,dx$

d) $\displaystyle\int_{-3}^{-1} \frac{1}{2x+1}\,dx$

e) $\displaystyle\int_0^1 \frac{t}{\sqrt{1+t^2}}\,dt$

f) $\displaystyle\int_0^1 \frac{\sin(\pi\sqrt{t})}{\sqrt{t}}\,dt$

g) $\displaystyle\int_0^2 \frac{1}{1+(x^2/4)}\,dx$

h) $\displaystyle\int_0^{\frac{1}{3}} \frac{1}{\sqrt{1-9y^2}}\,dy$

3. This question concerns the integral $I = \displaystyle\int \sin x \, \cos x \, dx$.

a) Find I by using the substitution $u = \sin x$.

b) Find I by using the substitution $u = \cos x$.

c) Compare your answers to (a) and (b). Are they the same? If not, how do they differ? Since both answers are antiderivatives of $\sin x \, \cos x$, they should differ only by a constant. Is that true here? If so, what is the constant?

d) Now calculate the value of the *definite* integral

$$\int_0^{\pi/2} \sin x \, \cos x \, dx$$

twice, using the two *indefinite* integrals you found in (a) and (b). Do the two values agree, or disagree? Is your result consistent with what you expect?

4. a) Find all functions $y = F(x)$ that satisfy the differential equation

$$\frac{dy}{dx} = x^2\left(1+x^3\right)^{13}.$$

b) From among the functions $F(x)$ you found in part (a), select the one that satisfies $F(0) = 4$.

c) From among the functions $F(x)$ you found in part (a), select the one that satisfies $F(-1) = 4$.

5. Find a function $y = G(t)$ that solves the initial value problem

$$\frac{dy}{dt} = te^{-t^2} \qquad y(0) = 3.$$

6. a) What is the average value of the function $f(x) = x/\sqrt{1+x^2}$ on the interval $[0, 2]$?

b) Show that the average value of the $f(x)$ on the interval $[-2, 2]$ is 0. Sketch a graph of $y = f(x)$ on this interval, and explain how the graph also shows that the average is 0.

7. a) Sketch the graph of the function $y = xe^{-x^2}$ on the interval $[0, 5]$.

b) Find the area between the graph of $y = xe^{-x^2}$ and the x-axis for $0 \le x \le 5$.

c) Find the area between the graph of $y = xe^{-x^2}$ and the x-axis for $0 \le x \le b$. Express your answer in terms of the quantity b, and denote it $A(b)$. Is $A(5)$ the same number you found in part(b)? What are the values of $A(10)$, $A(100)$, $A(1000)$?

d) It is possible to argue that the area between the graph of $y = xe^{-x^2}$ and the *entire* positive x-axis is 1/2. Can you develop such an argument?

8. a) Use a computer graphing utility to establish that

$$\sin^2 x = \frac{1 - \cos(2x)}{2}.$$

Sketch these graphs.

b) Find a formula for $\int \sin^2 x \, dx$. (Suggestion: replace $\sin^2 x$ by the expression involving $\cos(2x)$, above, and integrate by substitution.)

c) What is the average value of $\sin^2 x$ on the interval $[0, \pi]$? What is its average value on any interval of the form $[0, k\pi]$, where k is a whole number?

d) Explain your results in part (c) in terms of the graph of $\sin^2 x$ you drew in part (a).

e) Here's a differential equations proof of the identity in part (a). Let $f(x) = \sin^2 x$, and let $g(x) = (1 - \cos(2x))/2$. Show that both of these functions satisfy the initial value problem

$$y'' = 2 - 4y \quad \text{with} \quad y(0) = 0 \quad \text{and} \quad y'(0) = 0.$$

Hence conclude the two functions must be the same.

11.3 Integration by Parts

An integral form of the product rule

As in the previous section, suppose we have functions F and G, with corresponding derivatives f and g. If we use the product rule to differentiate $F(x) \cdot G(x)$, we get:

$$(F \cdot G)' = F \cdot G' + F' \cdot G = F \cdot g + f \cdot G.$$

We can turn this into a statement about indefinite integrals:

$$\int (F \cdot g + f \cdot G)\, dx = \int (F \cdot G)'\, dx = F(x) \cdot G(x) + C.$$

Unfortunately, in this form the statement is not especially useful; it applies only when the integrand has two terms of the special form $f \cdot g' + f' \cdot g$. However, if we rewrite the statement in the form

$$\boxed{\int \boldsymbol{F \cdot g\, dx = F \cdot G} - \int \boldsymbol{f \cdot G\, dx}}$$

it becomes very useful.

Example 1 We will use the formula in the box to find

$$\int x \cdot \cos x\, dx.$$

If we label the parts of this integrand as follows:

$$F(x) = x \qquad\qquad g(x) = \cos x,$$

then we have

$$f(x) = 1 \qquad \text{and} \qquad G(x) = \sin x.$$

According to the formula,

$$\begin{aligned}\int x \cdot \cos x\, dx &= x \cdot \sin x - \int \sin x\, dx\\ &= x \cdot \sin x + \cos x + C.\end{aligned}$$

Integrate only part of the integrand

The integrand is first broken into two parts—in this case, x and $\cos x$. One part is differentiated while the other part is integrated. (The part we integrated is $g(x) = \cos x$, and we got $G(x) = \sin x$.) For this reason, the rule described in the box is called **integration by parts**.

As with integration by substitution, integration by parts exchanges one integration task for another: Instead of finding an antiderivative for $F \cdot g$, we must find one for $f \cdot G$. The idea is to "trade-in" one integration problem for a more readily solvable one.

Example 2 Use integration by parts to find

$$\int \ln(x)\,dx.$$

At first glance we can't integrate by parts, because there aren't two parts! But note that we can write

$$\ln(x) = \ln(x) \cdot 1,$$

and then set

$$F(x) = \ln(x), \qquad g(x) = 1.$$

This implies

$$f(x) = \frac{1}{x} \qquad \text{and} \qquad G(x) = x,$$

and the integration by parts formula now gives us

$$\begin{aligned}\int \ln(x)\,dx &= x \cdot \ln(x) - \int x \cdot \frac{1}{x}\,dx \\ &= x \cdot \ln(x) - \int 1\,dx \\ &= x \cdot \ln(x) - x + C.\end{aligned}$$

The ingredients of a successful integration by parts

We thus see that integration by parts—like integration by substitution—is an art rather than a set routine. If integration by parts is to work, several things must happen. First, you need to see that the method might actually apply. (In Example 2 this wasn't obvious.) Next, you need to identify the parts of the integrand that will be differentiated and integrated, respectively. The wrong choices can lead you away from a solution, rather than towards one. (See example 3 below for a cautionary tale.) Finally, you need to be able to carry out the integration of the new integral $\int f \cdot G\,dx$. As you work you may have to reshape the integrand algebraically. Technique comes with practice, and luck is useful, too.

Example 3 Use integration by parts to find

$$\int t \cdot e^t \, dt.$$

Set

$$F(t) = e^t, \qquad g(t) = t;$$

then

$$f(t) = e^t, \qquad G(t) = \frac{t^2}{2}.$$

The integration by parts formula then gives

$$\int t \cdot e^t \, dt = \frac{t^2}{2} e^t - \int \frac{t^2}{2} e^t \, dt.$$

While this is a true statement, we are not better off—the new integral is *not* simpler than the original. A solution is eluding us here. You will have a chance to do this problem properly in the exercises.

What went wrong?

Exercises

1. Use integration by parts to find a formula for each of the following integrals.

a) $\int x \sin x \, dx$

b) $\int te^t \, dt$

c) $\int we^{-w} \, dw$

d) $\int x \ln x \, dx$

e) $\int \arcsin x \, dx$

f) $\int \arctan x \, dx$

g) $\int x^2 e^{-x} \, dx$

h) $\int u^2 \cos u \, du$

i) $\int x \sec^2 x \, dx$

j) $\int e^{2x}(x + e^x) \, dx$

(Suggestion for part (g): Apply integration by parts twice. After the first application you should have an integral that can itself be evaluated using integration by parts.)

2. Use integration by parts to obtain a formula for

$$\int (\ln x)^2\,dx.$$

Choose $f = \ln x$ and also $g' = \ln x$. To continue you need to find g, the antiderivative of $\ln x$, but this has already been obtained in the text.

3. a) Find $\displaystyle\int x^2 e^x\,dx$.

b) Find $\displaystyle\int x^3 e^x\,dx$. (Reduce this to part (a)).

c) Find $\displaystyle\int x^4 e^x\,dx$.

d) What is the general pattern here? Find a formula for $\displaystyle\int x^n e^x\,dx$, where n is any positive integer.

e) Find $\displaystyle\int e^x\left(5x^2 - 3x + 7\right)\,dx$.

4. a) Draw the graph of $y = \arctan x$ over the interval $0 \le x \le 1$. You could have gotten the same graph by thinking of x as a function of y—write down this relationship and the corresponding y interval.

b) Evaluate

$$\int_0^1 \arctan x\,dx$$

and show on your graph the area this corresponds to.

c) Evaluate

$$\int_0^{\pi/4} \tan y\,dy$$

and show on your graph the area this corresponds to.

d) If we add the results of part (b) and part (c), what do you get? From the geometry of the picture, what should you have gotten?

5. Repeat the analysis of the preceding problem by calculating the value of

$$\int_0^2 x^3\,dx + \int_0^8 y^{1/3}\,dy,$$

and seeing if it agrees with what you would predict by looking at the graphs.

6. Generalize the preceding two problems to the case where f and g are any two functions that are inverses of each other whose graphs pass through the origin.

7. a) What is the average value of the function $\ln x$ on the interval $[1, e]$?

b) What is the average value of $\ln x$ on $[1, b]$? Express this in terms of b. Discuss the following claim: The average value of $\ln x$ on $[1, b]$ is approximately $\ln(b) - 1$ when b is large.

8. a) Sketch the graph of $f(x) = xe^{-x}$ on the interval $[0, 4]$.

b) What is the area between the graph of $y = f(x)$ and the x-axis for $0 \leq x \leq 4$?

c) What is the area between the graph of $y = f(x)$ and the x-axis for $0 \leq x \leq b$? Express your answer in terms of b, and denote it $A(b)$. What is $A(100)$?

9. Find three solutions $y = f(t)$ to the differential equation

$$\frac{dy}{dt} = 5 - 2\ln t.$$

10. a) Find the solution $y = \varphi(t)$ to the initial value problem

$$\frac{dy}{dt} = te^{-t^2/2}, \qquad y(0) = 2.$$

b) The function $\varphi(t)$ increases as t increases. Show this first by sketching the graph of $y = \varphi(t)$. Show it also by referring to the differential equation that $\varphi(t)$ satisfies. (What is true about the derivative of an increasing function?)

c) Does the value of $\varphi(t)$ increase without bound as $t \to \infty$? If not, what value does $\varphi(t)$ approach?

11. a) **The differential form of Integration by Parts** If u and v are expressions in x, then the product rule can be written as

$$\frac{d}{dx}(u \cdot v) = \frac{du}{dx} \cdot v + u \cdot \frac{dv}{dx}.$$

Explain carefully how this leads to the following statement of integration by parts, and why it is equivalent to the form in the text:

$$\int u\,dv = uv - \int v\,du.$$

b) Solve a couple of the preceding problems using this notation.

Sine and cosine integrals

The purpose of the remaining exercises is to establish integral formulas that we will use to analyze Fourier polynomials and the power spectrum in chapter 12. In the first three, α is a constant:

$$\begin{aligned}
\int \sin^2 \alpha x\,dx &= \frac{x}{2} - \frac{1}{4\alpha}\sin 2\alpha x + C,\\
\int \cos^2 \alpha x\,dx &= \frac{x}{2} + \frac{1}{4\alpha}\sin 2\alpha x + C,\\
\int \sin \alpha x \cos \alpha x\,dx &= -\frac{1}{4\alpha}\cos 2\alpha x + C.
\end{aligned}$$

In the remaining four, α and β are *different* constants:

$$\begin{aligned}
\int \sin \alpha x \sin \beta x\,dx &= \frac{1}{\beta^2 - \alpha^2}\left(\alpha \cos \alpha x \sin \beta x - \beta \sin \alpha x \cos \beta x\right) + C,\\
\int \cos \alpha x \cos \beta x\,dx &= \frac{1}{\beta^2 - \alpha^2}\left(\beta \cos \alpha x \sin \beta x - \alpha \sin \alpha x \cos \beta x\right) + C,\\
\int \sin \alpha x \cos \beta x\,dx &= \frac{1}{\beta^2 - \alpha^2}\left(\beta \sin \alpha x \sin \beta x + \alpha \cos \alpha x \cos \beta x\right) + C,\\
\int \cos \alpha x \sin \beta x\,dx &= \frac{1}{\beta^2 - \alpha^2}\left(-\alpha \sin \alpha x \sin \beta x - \beta \cos \alpha x \cos \beta x\right) + C.
\end{aligned}$$

12. a) In the later exercises we shall make frequent use of the following "trigonometric identities":

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

Using a graphing package on a computer, graph together the functions

$$2\sin\alpha x\cos\alpha x \qquad \text{and} \qquad \sin 2\alpha x$$

to show that they seem to be identical. (That is, show that they "share phosphor.") Then do the same for the pairs of functions in the other two identities.

b) We can give a different argument for the identities above using the ideas we have developed in studying initial value problems. To prove the first identity, for instance, let $f(x) = 2\sin\alpha x\cos\alpha x$, and let $g(x) = \sin 2\alpha x$. Show that both functions satisfy

$$y'' = -4\alpha^2 y, \quad \text{with} \quad y(0) = 0 \quad \text{and} \quad y'(0) = 2\alpha.$$

Hence conclude the two functions must be the same.

c) Find an initial value problem that is satisfied by both $f(x) = \cos^2\alpha x - \sin^2\alpha x$ and by $g(x) = \cos 2\alpha x$.

13. **Evaluating** $\int \sin^2 x\,dx$

a) Using integration by parts, show that

$$\int \sin^2 x\,dx = -\sin x\cos x + \int \cos^2 x\,dx.$$

b) Using the identity $\sin^2\alpha x + \cos^2\alpha x = 1$, show that the new integral can be written as

$$x - \int \sin^2 x\,dx.$$

c) Combining (a) and (b) algebraically, show that

$$2\int \sin^2 x\,dx = -\sin x\cos x + x + C.$$

d) Using algebra and a trigonometric identity, conclude that

$$\int \sin^2 x\,dx = \frac{x}{2} - \frac{1}{4}\sin 2x + C.$$

14. Modify the argument of the preceding exercise to show

$$\int \sin^2 \alpha x \, dx = \frac{x}{2} - \frac{1}{4\alpha} \sin 2\alpha x + C.$$

and

$$\int \cos^2 \alpha x \, dx = \frac{x}{2} + \frac{1}{4\alpha} \sin 2\alpha x + C.$$

15. **Evaluating** $\int \sin^2 \alpha x \, dx$

Determine this integral anew, without using integration by parts, by carrying out the following steps.

a) From the trigonometric identities on page 716, deduce that

$$2 \sin^2 \alpha x = 1 - \cos 2\alpha x.$$

b) Using the formula in (a), conclude that

$$\int \sin^2 \alpha x \, dx = \frac{x}{2} - \frac{1}{4\alpha} \sin 2\alpha x + C.$$

16. **Evaluating** $\int \cos^2 \alpha x \, dx$

Using only algebra and the identity $\sin^2 \alpha x + \cos^2 \alpha x = 1$, show that the previous exercise implies

$$\int \cos^2 \alpha x \, dx = \frac{x}{2} + \frac{1}{4\alpha} \sin 2\alpha x + C.$$

17. **Evaluating** $\int \sin \alpha x \cos \alpha x \, dx$

a) Using the identity $\sin 2\alpha x = 2 \sin \alpha x \cos \alpha x$, deduce the following formula

$$\int \sin \alpha x \cos \alpha x \, dx = -\frac{1}{4\alpha} \cos 2\alpha x + C.$$

b) Using integration by substitution, obtain the alternative formula

$$\int \sin \alpha x \cos \alpha x \, dx = \frac{1}{2\alpha} \sin^2 \alpha x + C.$$

c) Show that your results in (a) and (b) are compatible. (For example, use exercise 15 (a).)

18. **Evaluating** $\int \sin \alpha x \sin \beta x \, dx$

a) Use integration by parts to show that

$$\int \sin \alpha x \sin \beta x \, dx = -\frac{1}{\alpha} \cos \alpha x \sin \beta x + \frac{\beta}{\alpha} \int \cos \alpha x \cos \beta x \, dx.$$

b) Using integration by parts again, show that the new integral in part (a) can be written as

$$\int \cos \alpha x \cos \beta x \, dx = \frac{1}{\alpha} \sin \alpha x \cos \beta x + \frac{\beta}{\alpha} \int \sin \alpha x \sin \beta x \, dx.$$

c) Let $J = \int \sin \alpha x \sin \beta x \, dx$; show that combining (a) and (b) gives

$$J = -\frac{1}{\alpha} \cos \alpha x \sin \beta x + \frac{\beta}{\alpha^2} \sin \alpha x \cos \beta x + \frac{\beta^2}{\alpha^2} J.$$

d) Solve (c) for J to find

$$\int \sin \alpha x \sin \beta x \, dx = \frac{1}{\beta^2 - \alpha^2}(\alpha \cos \alpha x \sin \beta x - \beta \sin \alpha x \cos \beta x) + C.$$

19. Imitate the methods of the preceding exercise to deduce

$$\int \cos \alpha x \cos \beta x \, dx = \frac{1}{\beta^2 - \alpha^2}(\beta \cos \alpha x \sin \beta x - \alpha \sin \alpha x \cos \beta x) + C$$

and

$$\int \sin \alpha x \cos \beta x \, dx = \frac{1}{\beta^2 - \alpha^2}(\beta \sin \alpha x \sin \beta x + \alpha \cos \alpha x \cos \beta x) + C.$$

20. **Evaluating** $\int \cos \alpha x \sin \beta x \, dx$

This integral is the same as one in the preceding exercise, if you exchange the factors α and β. Do that, and obtain the formula

$$\int \cos \alpha x \sin \beta x \, dx = 1\alpha^2 - \beta^2(\alpha \sin \alpha x \sin \beta x + \beta \cos \alpha x \cos \beta x) + C.$$

21. Determine the following.

a) $\displaystyle\int_0^{2\pi} \sin^2 x \, dx.$

b) $\displaystyle\int_0^{n\pi} \cos^2 x \, dx$, n a positive integer.

c) $\displaystyle\int_0^{2\pi} \sin x \sin \beta x \, dx$, $\beta \neq 1$.

d) $\displaystyle\int_0^{\pi} \sin x \sin \beta x \, dx$, $\beta \neq 1$.

11.4 Separation of Variables and Partial Fractions

One of the principal uses of integration techniques is to find closed form solutions to differential equations. If you look back at the methods we have developed so far in this chapter, they are all applicable to differential equations of the form $y' = f(t)$ for some function f—that is, the rate at which y changes is a function of the independent variable only. In such cases we only need to find an antiderivative F for f, choose the constant C to satisfy the initial value, and we have our solution. As we saw in the early chapters, though, the behavior of y' often depends on the values of y rather than on t—think of the S-I-R model or the various predator-prey problems. In this section we will see how our earlier techniques can be adapted to apply to problems of this sort as well.

The Differential Equation $y' = y$

A new method for solving $y' = y$

As you know, the exponential functions $y = Ce^t$ are the solutions to the differential equation $dy/dt = y$. Let's put aside this knowledge for a moment and rediscover these solutions using a new method. The method involves the connection between inverse functions and their derivatives. With it we will be able to explore a variety of problems that had been beyond our reach.

Find the inverse function instead

The idea behind the new method is quite simple: Instead of thinking of y as a function of t, convert to thinking of t as a function of y, thereby looking for the **inverse function**. We know that the derivative of the inverse function is the reciprocal of the derivative of the original function, so we can rewrite the given differential equation by using its reciprocal:

The new differential equation...

$$\frac{dy}{dt} = y \qquad \text{becomes} \qquad \frac{dt}{dy} = \frac{1}{y}.$$

Then solve the new differential equation $dt/dy = 1/y$. While this may not look very different, it has the property that the rate of change of the *dependent variable*—now t—is expressed as a function of the *independent* variable—now y. But this is just the form we have been considering in the earlier sections of this chapter. A solution to the new equation is a function $t = g(y)$ whose derivative is $1/y$. This is one of the basic antiderivatives listed in the table on page 689:

...and its solution

$$t = g(y) = \ln y + k,$$

where k is an arbitrary constant.

The inverse ...

The solution to the original differential equation $dy/dt = y$ is the inverse of $t = \ln y + k$. We find it by solving this equation for y:

$$\begin{aligned} t - k &= \ln y \\ e^{t-k} &= y \\ e^t \cdot e^{-k} &= y \end{aligned}$$

...solves the original problem

Thus $y(t) = Ce^t$, where we have replaced the constant e^{-k} by C.

Indefinite integrals

The differential equation expressed using differentials

The language of indefinite integrals and differentials again provides a convenient mnemonic for this new method. First, we use the original differential equation to relate the differentials dy and dt:

$$dy = \frac{dy}{dt}\,dt = y\,dt.$$

This use of differentials is introduced on page 704. At this point the equation makes sense for either t or y being the independent variable. Now if we try to integrate this equation with respect to t, we get

$$y = \int dy = \int y(t)\,dt.$$

We can't find the last integral, because we don't know what y is as a function of t. Remember, an indefinite integral is an antiderivative, so an expression of the form

$$\int f\,dt$$

represents a function $F(t)$ whose derivative is $f(t)$. If f is *not* given by a formula in t, there is no way to get a formula for $F(t)$.

Suppose, though, that we divide both sides of the differential equation $dy = y\,dt$ by y, and integrate with respect to y. The equation takes the form

$$\frac{dy}{y} = dt.$$

Now, if we introduce indefinite integrals, we have

$$\int \frac{dy}{y} = \int dt.$$

The variables are now *separated*

This time the variables y and t have been separated from each other, and we *can* find the integrals. In fact,

$$\ln y = \int \frac{dy}{y} = \int dt = t + b,$$

where b is an arbitrary constant. (We could just as easily have added the constant to the left side instead—do you see why we don't have to add a constant to both sides?) To complete the work we solve for y:

The solution once again

$$y = e^{t+b} = e^t \cdot e^b = Ce^t,$$

where $C = e^b$.

Summary

The first time we went through the method, we replaced

$$\frac{dy}{dt} = y \qquad \text{by} \qquad \frac{dt}{dy} = \frac{1}{y}.$$

These differential equations express the same relation between y and t. Each is just the reciprocal of the other. In the first, y depends on t; in the second, though, t depends on y. The second time we went through the method, using indefinite integrals, we replaced

$$dy = y\,dt \qquad \text{by} \qquad dt = \frac{dy}{y}.$$

This change was also algebraic, and it had the same effect: the dependent variable changed from y to t. More important, in the new differential equation (using the differentials themselves!) the variables are separated. That allows us to do the integration. We get a solution in the form $t = g(y)$. The solution to the original problem is the *inverse* $y = f(t)$ of $t = g(y)$.

To integrate, separate the variables

Separation of Variables

With the method of **separation of variables**, introduced in the previous pages, we can obtain formulas for solutions to a number of differential equations that were previously accessible only by Euler's method. Recall that one of the clear advantages of a *formula* is that it allows us to see how the parameters in the problem affect the solution. We'll look at two problems. First we'll show how the method can explain the rather baffling formula for supergrowth that we gave in chapter 4. Then, using the method of **partial fractions**, to be discussed next, we'll give a formula for logistic growth.

Supergrowth

In chapter 4.2 we modelled the growth of a population Q by the initial value problem

$$\frac{dQ}{dt} = kQ^{1.2}, \qquad Q(0) = A.$$

To get a formula for the solution, transform the differential equation in the following way:

Separate the variables

$$\frac{dQ}{dt} = k\,Q^{1.2} \quad \rightsquigarrow \quad dQ = k\,Q^{1.2}\,dt \quad \rightsquigarrow \quad \frac{dQ}{Q^{1.2}} = k\,dt.$$

Now integrate:

$$\int \frac{dQ}{Q^{1.2}} = \int k\,dt.$$

Because the variables have been separated, the integrals can be found:

$$\int \frac{dQ}{Q^{1.2}} = \int Q^{-1.2}\,dQ = \frac{1}{-.2}Q^{-.2}$$
$$\int k\,dt = kt + C$$

Solve for Q

Therefore, $\frac{1}{-.2}Q^{-.2} = kt + C$. Now we must solve this equation for Q. Here is one possible approach. First, we can write

$$Q^{-.2} = -.2(kt + C) \;=\; C_1 - .2kt.$$

To simplify the expression, we have replaced $-.2C$ by a *new* constant C_1. Since

$$(Q^{-.2})^{-5} = Q^{-.2\times -5} = Q^1 = Q,$$

we'll raise both sides of the previous equation to the power -5:

$$Q(t) = (Q^{-.2})^{-5} = (C_1 - .2kt)^{-5}.$$

Bring in the initial condition

The last step is to incorporate the initial condition $Q(0) = A$. According to the new formula for $Q(t)$,

$$Q(0) = (C_1 - .2k \cdot 0)^{-5} = C_1^{-5} = A.$$

Solving $A = C_1^{-5}$ for C_1, we get

$$C_1 = A^{-1/5} = \frac{1}{\sqrt[5]{A}}.$$

We are now done:

$$Q(t) = \left(\frac{1}{\sqrt[5]{A}} - .2kt\right)^{-5}.$$

Interpreting the formula

This is the formula that appears on page 215. It shows how the parameters k and A affect the solution. In particular, we called this *supergrowth* because the model predicts that the population Q becomes infinite when

$$\frac{1}{\sqrt[5]{A}} - .2kt = 0; \quad \text{that is, when} \quad t = \frac{1}{.2k\sqrt[5]{A}}.$$

Partial Fractions

Using separation of variables with a partial fractions decomposition (to be described below), we will obtain a formula for the solution to the logistic equation (chapter 4.1). The method of **partial fractions** is a useful tool for solving many integration problems.

Logistic growth

Consider this initial value problem associated with the logistic differential equation:

$$\frac{dP}{dt} = kP\left(1 - \frac{P}{C}\right), \qquad P(0) = A.$$

We will find a formula for the solution that incorporates the growth parameter k and the carrying capacity C.

Step 1: separate the variables

The first step is to transform the equation into one where the variables are separated:

$$\frac{dP}{dt} = kP\left(1 - \frac{P}{C}\right) \quad \rightsquigarrow \quad \frac{dP}{P(1 - P/C)} = k\,dt.$$

Integrating the new equation we get

$$\int \frac{dP}{P(1 - P/C)} = \int k\,dt.$$

The denominator has an unfamiliar form

We are stuck now, because the integral on the left doesn't appear in our table of integrals (page 689). If the denominator had *only* P or *only* $1 - P/C$, we could use the natural logarithm. The difficulty is that the denominator is the product of both terms.

There is a way out of the difficulty. We will use algebra to transform the integrand into a form we can work with. The first step is to simplify the denominator a bit:

$$\frac{1}{P(1 - P/C)} = \frac{1}{P(C/C - P/C)} = \frac{1}{P(C - P)/C} = \frac{C}{P(C - P)}.$$

(This wasn't essential; it just makes later steps easier to write.) The next step will be the crucial one. To understand why we take it, consider the rule for adding two fractions:

$$\frac{\alpha}{(x + a)} + \frac{\beta}{(x + b)} = \frac{\alpha(x + b) + \beta(x + a)}{(x + a)(x + b)}$$

Step 2: write the integrand as a sum of simple fractions

The denominator is a product—very much like the product $P(C-P)$ in our integrand! Perhaps we can write *that* as a sum of two simpler fractions:

$$\frac{C}{P(C-P)} = \frac{\alpha}{P} + \frac{\beta}{C-P}.$$

What values should α and β have? According to the rule for adding fractions,

$$\frac{\alpha}{P} + \frac{\beta}{C-P} = \frac{\alpha(C-P)+\beta P}{P(C-P)},$$

and this should equal the original integrand:

$$\frac{\alpha(C-P)+\beta P}{P(C-P)} = \frac{C}{P(C-P)}.$$

Determining α and β

Since the denominators are equal, the numerators must also be equal:

$$\alpha(C-P)+\beta P = C.$$

In fact, they must be equal *as polynomials in the variable* P. If we rewrite the last equation, collecting terms that involve the same power of P, we get

$$(\beta-\alpha)P + \alpha C = 0\cdot P + 1\cdot C.$$

Since two polynomials are equal precisely when their coefficients are equal, it follows that

$$\beta-\alpha = 0 \qquad \alpha = 1.$$

Thus $\alpha=\beta=1$, and we have

The partial fractions decomposition

$$\frac{C}{P(C-P)} = \frac{1}{P} + \frac{1}{C-P}.$$

The simpler expressions on the right are called **partial fractions**. Their denominators are the different *parts* of the denominator of the integrand. The equation that expresses the integrand as a sum of partial fractions is called a **partial fractions decomposition**.

We can now return to the integral equation we are trying to solve:

$$\int \frac{C}{P(C-P)}\,dP = \int k\,dt.$$

The right hand side equals $kt+b$, where b is the usual constant of integration. Thanks to the partial fractions decomposition, the left hand side can be written

Step 3: evaluate the integrals

$$\int \frac{C}{P(C-P)}\,dP = \int \frac{1}{P}\,dP + \int \frac{1}{C-P}\,dP.$$

The first integral on the right is straightforward:

$$\int \frac{1}{P}\,dP = \ln P.$$

The second can be solved by using the substitution $C - P = u$, with $dP = -du$:

$$\int \frac{1}{C-P}\,dP = \int \frac{-du}{u} = -\ln u = -\ln(C-P).$$

Putting everything together we find

$$\ln P - \ln(C-P) = \ln\left(\frac{P}{C-P}\right) = kt + b.$$

As we have seen, separation of variables usually leaves us with an inverse function to find. This problem is no different. We must solve the last equation for P. The first step is to exponentiate both sides:

Step 4: solve for P

$$\frac{P}{C-P} = e^{kt+b} = e^b \cdot e^{kt} = Be^{kt}.$$

To simplify the expression a bit, we have replaced e^b by $B = e^b$. Multiplying both sides by $C - P$ gives

$$P = Be^{kt}(C-P) = CBe^{kt} - PBe^{kt}.$$

Now bring the last term over to the left, and then factor out P:

$$P + PBe^{kt} = \left(1 + Be^{kt}\right)P = CBe^{kt}.$$

The final step is to divide by the coefficient $1 + Be^{kt}$:

The formula for $P(t)$

$$P(t) = \frac{CBe^{kt}}{1 + Be^{kt}}.$$

Lastly, we must see how the initial condition $P(0) = A$ affects the solution. We could substitute $t = 0$, $P = A$ into the last formula, but that produces an algebraic mess. We want to know how A affects the constant B,

Step 5: incorporate the initial condition and we can see that directly by making our substitutions into the equation above:

$$\frac{P}{C-P} = Be^{kt} \quad \rightsquigarrow \quad \frac{A}{C-A} = Be^0 = B.$$

Now replace B by $A/(C-A)$ in our formula for $P(t)$. This yields

$$P(t) = \frac{CAe^{kt}/(C-A)}{1+Ae^{kt}/(C-A)} = \frac{CAe^{kt}}{C-A+Ae^{kt}}$$

If we write $-A+Ae^{kt} = A(e^{kt}-1)$, we get one of the standard forms of the solution to the logistic equation:

The complete solution

$$P(t) = \frac{CAe^{kt}}{C+A(e^{kt}-1)}.$$

Remark. The method of partial fractions can be used to evaluate integrals of the form

$$\int \frac{dx}{(x+a_1)(x+a_2)\cdots(x+a_n)} \qquad \text{or} \qquad \int \frac{P(x)}{Q(x)}\,dx,$$

where $P(x)$ and $Q(x)$ are arbitrary polynomials. Tables of integrals and calculus references describe how the method works in these cases. As one example, though, let's compute the antiderivative of the cosecant function, since we will need it in the next section.

Example—The antiderivative of the cosecant We first use a trigonometric identity to transform the integral slightly:

$$\begin{aligned}\int \csc x\,dx &= \int \frac{1}{\sin x}\,dx \\ &= \int \frac{\sin x}{\sin^2 x}\,dx \\ &= \int \frac{\sin x\,dx}{1-\cos^2 x}.\end{aligned}$$

If we now make the substitution $u = \cos x$, with $du = -\sin x\,dx$, this becomes

$$\begin{aligned}\int \csc x\,dx &= \int \frac{-du}{1-u^2}\\ &= \int \frac{-1}{2}\left(\frac{1}{1+u}+\frac{1}{1-u}\right)du\\ &= -\frac{1}{2}\int\frac{du}{1+u}-\frac{1}{2}\int\frac{du}{1-u}\\ &= -\frac{1}{2}\ln(1+u)+\frac{1}{2}\ln(1-u)+C\\ &= \frac{1}{2}\ln\frac{1-u}{1+u}+C\\ &= \frac{1}{2}\ln\frac{1-\cos x}{1+\cos x}+C.\end{aligned}$$

We can simplify this slightly by multiplying both numerator and denominator by $1+\cos x$ to get

The final form of the antiderivative of the cosecant

$$\begin{aligned}\int \csc x\,dx = \frac{1}{2}\ln\frac{1-\cos^2 x}{(1+\cos x)^2}+C = \ln\left|\frac{\sin x}{1+\cos x}\right|+C\\ = -\ln|\csc x+\cot x|+C.\end{aligned}$$

Note that in the next to last line we used the general fact about logarithms that $n\ln A = \ln(A^n)$ for any value of n and any $A > 0$. Note also that since the domain of the secant function consists of infinitely many separate intervals. the "$+\ C$" at the end of the antiderivative needs to be interpreted as potentially a different value of C over each interval.

In the same fashion we can obtain the antiderivative for the secant function:

The antiderivative of the secant

$$\int \sec x\,dx = \ln|\sec x+\tan x|+C.$$

Exercises

Separation of variables

1. Use the method of separation of variables to find a formula for the solution of the differential equation $dy/dt = y + 5$. Your formula should contain an arbitrary constant to reflect the fact that many functions solve the differential equation.

2. Use the method of separation of variables to find formulas for the solutions to the following differential equations. In each case your formula should be expressed in terms of the input variable that is indicated (e.g., in part (a) it is t).

a) $dy/dt = 1/y$.

b) $dz/dx = 3/(z-2)$.

c) $dy/dx = x/y$

d) $dy/dx = y/x$

e) $du/dv = u/(u-1)$

f) $dv/dt = -\sqrt{v}$

3. **A cooling liquid**. According to Newton's law of cooling (see chapter 4.1), in a room where the ambient temperature is C, the temperature Q of a hot object will change according to the differential equation

$$\frac{dQ}{dt} = -k(Q - C).$$

The constant k gives the rate at which the object cools.

a) Find a formula for the solution to this equation using the method of separation of variables. Your formula should contain an arbitrary constant.

b) Suppose C is 20°C and k is .1° per minute per °C. If time t is measured in minutes, and $Q(0) = 90$°C, what will Q be after 20 minutes?

c) How long does it take for the temperature to drop to 30°C?

4. a) Suppose a cold drink at 36°F is sitting in the open air on a summer day when the temperature is 90°F. If the drink warms up at a rate of .2°F per minute per °F of temperature difference, write a differential equation to model what will happen to the temperature of the drink over time.

b) Obtain a formula for the temperature of the drink as a function of the number of minutes t that have passed since its temperature was 36°F.

c) What will the temperature of the drink be after 5 minutes; after 10 minutes?

d) How long will it take for the drink to reach 55°F?

5. **A leaking tank**. In chapter 4.2 we used the differential equation

$$\frac{dV}{dt} = -k\sqrt{V}$$

to model the volume $V(t)$ of water in a leaking tank after t hours (see page 222).

a) Use the method of separation of variables to show that

$$V(t) = \frac{k^2}{4}(C - t)^2$$

is a solution to the differential equation, for any value of the constant C.

b) Explain why the function

$$V(t) = \begin{cases} \dfrac{k^2}{4}(C-t)^2 & \text{if } 0 \le t \le C, \\ 0 & \text{if } C < t. \end{cases}$$

is *also* a solution to the differential equation. Why is *this* solution more relevant to the leaking tank problem than the solution in part (a)?

6. **A falling body with air resistance**. We have used the differential equation

$$\frac{dv}{dt} = -g - bv$$

to model the motion of a body falling under the influence of gravity (g) and air resistance (bv). Here v is the velocity of the body at time t. (See pages 224–225.)

a) Solve the differential equation by separating variables, and obtain

$$v(t) = \frac{1}{b}\left(Ce^{-bt} - g\right),$$

where C is an arbitrary constant.

b) Now impose the initial condition $v(0) = 0$ (so the body starts it fall from rest) to determine the value of C. What is the formula for $v(t)$ now?

c) Exercise 21 on page 224 gives the solution to the initial value problem as

$$v(t) = \frac{g}{b}\left(2^{-bt/.69} - 1\right).$$

Reconcile this expression with the one you obtained in part (b) of this exercise.

d) The distance $x(t)$ that the body has fallen by time t is given by the integral

$$x(t) = \int_0^t v(t)\,dt, \quad \text{because} \quad \frac{dx}{dt} = v \quad \text{and} \quad x(0) = 0.$$

Use your formula for $v(t)$ from part (b) to find $x(t)$.

7. a) **Supergrowth**. We have analyzed the differential equation

$$\frac{dQ}{dt} = kQ^p$$

when $p = 1.2$ (and, of course, when $p = 1$). Find a formula for the solution $Q(t)$ when $p = 2$. Your formula should contain an arbitrary constant C.

b) Add the initial condition $Q(0) = A$. This will fix the value of the constant C. What is the formula for $Q(t)$ when the initial condition is incorporated?

c) Your formula in part (b) should demonstrate that Q becomes infinite at some finite time $t = \tau$. When is τ? Your answer should be expressed in terms of the growth constant k and the initial population size A.

d) Suppose the values of k and A are known only imprecisely, and they could be in error by as much as 5%. That makes the value of τ uncertain. Which error causes the greater uncertainty: the error in k or the error in A? (See the discussion of error analysis for the supergrowth model on pages 216–217.)

8. **General supergrowth**. Find the solution to the initial value problem

$$\frac{dQ}{dt} = kQ^p, \qquad Q(0) = A$$

for *any* value of the power p. For which values of p does Q blow up to ∞ at a finite time $t = \tau$? What is τ?

Partial fractions

9. Use the method of partial fractions to determine the values of α, β, and γ in the following equations.

a) $\dfrac{1}{(x-1)(x+2)} = \dfrac{\alpha}{x-1} + \dfrac{\beta}{x+2}$

b) $\dfrac{x}{(x-1)(x+2)} = \dfrac{\alpha}{x-1} + \dfrac{\beta}{x+2}$

c) $\dfrac{1}{x(x^2-1)} = \dfrac{\alpha}{x} + \dfrac{\beta}{x-1} + \dfrac{\gamma}{x+1}$

d) $\dfrac{x}{2x^2+3x+1} = \dfrac{\alpha}{2x+1} + \dfrac{\beta}{x+1}$

e) $\dfrac{1}{x(x^2+1)} = \dfrac{\alpha}{x} + \dfrac{\beta x + \gamma}{x^2+1}$

[Note that x^2+1 can't be factored.]

10. Find a formula for each of these indefinite integrals.

a) $\displaystyle\int \frac{3\,dx}{(x-1)(x+2)}$ d) $\displaystyle\int \frac{x\,dx}{1-x^2}$

b) $\displaystyle\int \frac{5x+3}{(x-1)(x+2)}\,dx$ e) $\displaystyle\int \frac{1-u}{u^2-4}\,du$

c) $\displaystyle\int \frac{dt}{t(t^2-1)}$ f) $\displaystyle\int \frac{x^2+2x+1}{x(x^2+1)}\,dx$

11. Determine

a) $\displaystyle\int_2^3 \frac{3\,dx}{(x-1)(x+2)}$ c) $\displaystyle\int_0^{\pi/4} \frac{x\,dx}{1-x^2}$

b) $\displaystyle\int_2^4 \frac{dt}{t(t^2-1)}$ d) $\displaystyle\int_1^{\sqrt{3}} \frac{x^2+2x+1}{x(x^2+1)}\,dx$

12. Mirror the derivation of $\displaystyle\int \csc x\,dx$ to find $\displaystyle\int \sec x\,dx$.

13. Consider the particular logistic growth model defined by

$$\frac{dP}{dt} = .2P\left(1-\frac{P}{10}\right)\ \text{lbs/hr}, \qquad P(0) = .5\ \text{lbs}$$

(Compare this with the fermentation problems, pages 195–197.)

a) Obtain the formula for the solution to this initial value problem.

b) How large will P be after 3 hours; after 10 hours?

c) When will P reach one-half the carrying capacity–that is, for which t is $P = 5$ lbs?

14. Derive the formula for $\displaystyle\int \sec x\,dx$ given on page pagerefsecant, using methods similar to those used to find an antiderivative for the cosecant function.

11.5 Trigonometric Integrals

The preceding sections have covered the main integration techniques and concepts likely to be needed by most users of calculus. These techniques, together with the numerical methods discussed in chapter 11.6, should be part of the basic tool kit of every practitioner of calculus. For those going on in physics or mathematics, there are additional methods, largely involving trigonometric functions in various ways, that are sometimes useful. The purpose of this section is to develop the most commonly used of these techniques.

Recall that there are only a few simple antiderivatives we can write down immediately by inspection. All non-numerical integration techniques consist of finding transformations that will reduce some new class of integration problems to a class we already know how to solve. Once we have a new class of solvable problems, then we look for other classes of problems that can be reduced to this new class, and so on. The techniques we will be developing in this section involve ways of making such transformations through the use of basic trigonometric identities, typically in conjunction with integration by parts or by substitution. Before we proceed with the integration techniques, it will be helpful to list the trigonometric identities used.

Review of trigonometric identities

The most frequently used identity is

$$\sin^2 x + \cos^2 x = 1,$$

and the equivalent form obtained by dividing through by $\cos^2 x$:

$$\tan^2 x + 1 = \sec^2 x,$$

and by $\sin^2 x$:

$$1 + \cot^2 x = \csc^2 x.$$

The only other identities you will need have already been encountered:

$$\sin 2x = 2\sin x \cos x \quad \text{and} \quad \cos 2x = \cos^2 x - \sin^2 x,$$

plus the two other forms of the second of these identities,

$$\cos^2 x = \frac{1}{2}(1 + \cos 2x) \quad \text{and} \quad \sin^2 x = \frac{1}{2}(1 - \cos 2x).$$

Inverse Substitution

The method of substitution outlined in chapter 11.2 worked by taking a complicated integrand and breaking it down into simpler components, reducing the problem of finding an antiderivative for something in the form $f(G(x))g(x)$ to the problem of finding an antiderivative for f. In some cases, though, we go in the opposite direction: we have an integral $\int f(x)\,dx$ we want to find but can't evaluate directly. Instead, we can find a function $G(u)$ with derivative $g(u)$ such that we can find an antiderivative for $f(G(u))g(u)$. Since we know this integral is $F(G(u))$, we can now figure out what the desired function F must be. As with the earlier substitution techniques, this **inverse substitution** is conveniently expressed using differential notation.

Success sometimes comes by making things more complicated

Example 1 Suppose we want to evaluate

$$\int \sqrt{4-x^2}\,dx.$$

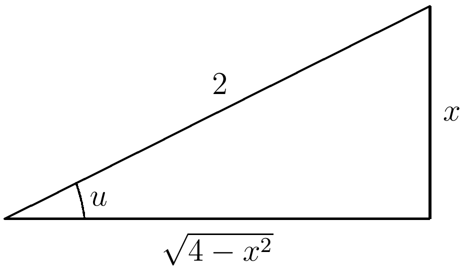
If we substitute $x = 2\sin u$, so that $dx = 2\cos u\,du$, look what happens:

$$\begin{aligned}\int \sqrt{4-x^2}\,dx &= \int \sqrt{4-(2\sin u)^2}\,2\cos u\,du\\ &= \int \sqrt{4-4\sin^2 u}\,2\cos u\,du\\ &= \int 2\sqrt{1-\sin^2 u}\,2\cos u\,du\\ &= \int 2\cos u\,2\cos u\,du\\ &= 4\int \cos^2 u\,du.\end{aligned}$$

But this is just an antiderivative we have already found in the exercises in chapter 11.3, namely

$$\begin{aligned}\int \cos^2 u\,du &= \frac{u}{2}+\frac{1}{4}\sin 2u + C\\ &= \frac{u}{2}+\frac{1}{4}\cdot 2\sin u\cos u + C\\ &= \frac{1}{2}(u+\sin u\cos u)+C.\end{aligned}$$

To find the desired antiderivative for the original function of x, we now replace u by its expression in terms of x by inverting the relationship: If $x = 2\sin u$, then $\sin u = x/2$, and $u = \arcsin(x/2)$. As we found in chapter 11.1, drawing a picture expressing the relationship between x and u makes it easy to visualize the other trigonometric functions:



From the picture we see that

$$\cos u = \frac{\sqrt{4-x^2}}{2} \qquad \text{and} \qquad \tan u = \frac{x}{\sqrt{4-x^2}}.$$

We can now find an expression for the desired antiderivative in terms of x:

$$\begin{aligned}\int \sqrt{4-x^2}\,dx &= 2(u + \sin u \cos u) + C \\ &= 2\left(\arcsin\frac{x}{2} + \frac{x}{2}\frac{\sqrt{4-x^2}}{2}\right) + C \\ &= 2\arcsin\frac{x}{2} + \frac{x\sqrt{4-x^2}}{2} + C.\end{aligned}$$

As usual, you should check this result by differentiating the right-hand side to see that you do obtain the integrand on the left.

Some useful substitutions

Similar substitutions allow us to evaluate other integrals involving square roots of quadratic expressions. Here is a summary of useful substitutions. In each case, a is a positive real number.

To transform	$a^2 - x^2$	let	$x = a\,\sin u$;
To transform	$a^2 + x^2$	let	$x = a\,\tan u$;
To transform	$x^2 - a^2$	let	$x = a\,\sec u$.

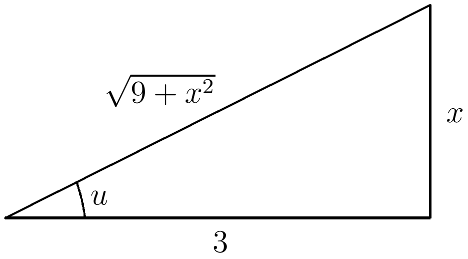
Example 2 Integrate

$$\int \frac{dx}{\sqrt{x^2+9}}.$$

If we set $x = 3\tan u$, then $dx = 3\,\sec^2 u\,du$, and the integral becomes

$$\begin{aligned}\int \frac{dx}{\sqrt{x^2+9}} &= \int \frac{3\,\sec^2 u\,du}{\sqrt{9\,\sec^2 u}}\\ &= \int \sec u\,du\\ &= \ln|\sec u + \tan u| + C,\end{aligned}$$

(as we saw in chapter 11.4). To express this in terms of x, we again draw a picture showing the relation between u and x:



From this picture we see that $\sec u = \sqrt{9+x^2}/3$. Therefore

$$\begin{aligned}\int \frac{dx}{\sqrt{x^2+9}} &= \ln\left|\frac{\sqrt{9+x^2}}{3} + \frac{x}{3}\right| + C\\ &= \ln\left|\frac{\sqrt{9+x^2}+x}{3}\right| + C = \ln\left|\sqrt{9+x^2}+x\right| + C',\end{aligned}$$

where $C' = C - \ln 3$ is a new constant. As usual, you should differentiate to check that this really is the claimed antiderivative

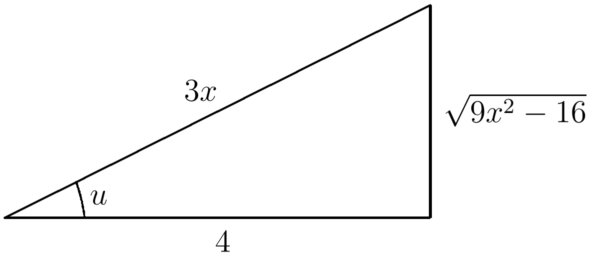
Example 3 Evaluate

$$\int \frac{dx}{\sqrt{9x^2-16}}.$$

We first write $\sqrt{9x^2-16}$ as $\sqrt{9}\sqrt{x^2-(16/9)} = 3\sqrt{x^2-(16/9)}$. Using the substitution $x = (4/3)\sec u$, with $dx = (4/3)\sec u \tan u\,du$ gives

$$\begin{aligned}\int \frac{dx}{\sqrt{9x^2-16}} &= \int \frac{(4/3)\sec u\,\tan u\,du}{3\sqrt{(16/9)\sec^2 u-(16/9)}}\\ &= \int \frac{(4/3)\sec u\,\tan u\,du}{3\cdot(4/3)\tan u} = \frac{1}{3}\int \sec u\,du\\ &= \frac{1}{3}\ln|\sec u+\tan u|+C.\end{aligned}$$

Again we need a picture to relate x and u:



Thus $\tan u = \sqrt{9x^2-16}/4$, which gives

$$\begin{aligned}\int \frac{dx}{\sqrt{9x^2-16}} &= \frac{1}{3}\ln|\sec u+\tan u|+C\\ &= \frac{1}{3}\ln\left|\frac{3x}{4}+\frac{\sqrt{9x^2-16}}{4}\right|+C\\ &= \frac{1}{3}\ln\left|3x+\sqrt{9x^2-16}\right|+C',\end{aligned}$$

where $C' = C-(\ln 4)/3$.

As usual, you should differentiate this final expression to confirm that it really is the desired antiderivative.

Inverse Substitution and Definite Integrals

We saw on page 707 in chapter 11.2 how to use substitution to evaluate a definite integral. When we transformed an integral originally expressed in terms of a variable x into one expressed in terms of a variable u, the two integrals had the same numerical value. The same can be done with the inverse substitution technique we have just been considering. Let's see how this works. Suppose we start with a function $f(x)$ to be integrated over an interval $[a, b]$. If we only knew an antiderivative F for f, we could easily write

$$\int_a^b f(x)\,dx = F(b) - F(a),$$

as usual. In the examples we've just been considering, we found the antiderivative for f by making a substitution $x = G(u)$ for some function G and then finding an antiderivative for $f(G(u))g(u)$, where $G' = g$. This antiderivative we know is $F(G(u))$, where F is the function we are trying to find. We were able to obtain F by replacing u by its expression in x. To do this we needed to find the inverse function G^{-1} for G, so that $x = G(u)$ was equivalent to $u = G^{-1}(x)$, and $F(x) = F(G(G^{-1}(x)))$. It is this last step we can eliminate in calculating definite integrals.

If we want x to go from a to b, what must u do? What interval of u values will get transformed to this interval of x values by G. The value of u such that $G(u) = a$ is $A = G^{-1}(a)$. Similarly the u value that gets transformed to b is $B = G^{-1}(b)$. Thus under the substitution $x = G(u)$, as u goes from A to B, x will go from a to b. Now look at the corresponding definite integral:

$$\begin{aligned}\int_A^B f(G(u))\,g(u)\,du &= F(G(u))\Big|_A^B \\ &= F(G(B)) - F(G(A)) \\ &= F(G(G^{-1}(b))) - F(G(G^{-1}(a))) \\ &= F(b) - F(a),\end{aligned}$$

which is just the desired value of the original definite integral. To summarize,

If we make the substitution $\boldsymbol{x = G(u)}$, then

$$\boldsymbol{\int_a^b f(x)\,dx = F(b) - F(a) = \int_A^B f(G(u))\,g(u)\,du,}$$

where $\boldsymbol{A = G^{-1}(a)}$ and $\boldsymbol{B = G^{-1}(b)}$.

Let's look back at a couple of the preceding examples to see how this works.

Example 4 Evaluate

$$\int_{-2}^{2} \sqrt{4 - x^2}\,dx.$$

In Example 1 we found an antiderivative for this function by making the substitution $x = 2\sin u = G(u)$. The inverse function is then $G^{-1}(x) = \arcsin(x/2)$. To get x to go from -2 to 2, u must go from $G^{-1}(-2) = \arcsin(-1) = -\pi/2 = A$ to $G^{-1}(2) = \arcsin 1 = \pi/2 = B$. Then to evaluate the integral from $x = -2$ to $x = 2$, we only need to evaluate the antiderivative we found for the u integral between $u = -\pi/2$ and $u = \pi/2$.

$$\int_{-2}^{2} \sqrt{4 - x^2}\, dx = 2(u + \sin u\, \cos u)\Big|_{-\pi/2}^{\pi/2} = \pi - (-\pi) = 2\pi.$$

(Note that this is just half the area of a circle of radius 2. How could we have foreseen this result from the form of the problem?)

Example 5 Suppose we wanted the integral

$$\int_0^3 \frac{dx}{\sqrt{x^2 + 9}}.$$

In Example 2 we found an antiderivative by letting $x = 3\tan u$. Here $G^{-1}(x) = \arctan(x/3)$. For x to go from 0 to 3, u must go from 0 to $\arctan 1 = \pi/4$. Using the u-antiderivative we found in Example 2, we have

$$\begin{aligned}\int_0^3 \frac{dx}{\sqrt{x^2 + 9}} &= \ln|\sec u + \tan u|\Big|_0^{\pi/4} \\ &= \ln|\sqrt{2} + 1| - \ln|1 + 0| = \ln(\sqrt{2} + 1).\end{aligned}$$

Completing The Square

Integrands involving terms of the form $Ax^2 + Bx + C$ can always be put in the form $A(u^2 \pm b^2)$ for a suitable variable u and constant b. The technique for doing this is the standard method of **completing the square**:

$$\begin{aligned}Ax^2 + Bx + C &= A\left(x^2 + \frac{B}{A}x\right) + C \\ &= A\left(x^2 + \frac{B}{A}x + \frac{B^2}{4A^2}\right) + C - \frac{B^2}{4A} \\ &= A\left(x + \frac{B}{2A}\right)^2 + \frac{4AC - B^2}{4A}.\end{aligned}$$

The substitutions

$$u = x + \frac{B}{2A} \qquad \text{and} \qquad b = \frac{\sqrt{|4AC - B^2|}}{2A}$$

then transform the problem to a form where we can use the techniques already developed. The following examples should make this clear.

Example 6 Consider the integral

$$\int \frac{dx}{x^2 + 4x + 5}.$$

This may not immediately remind us of anything we've seen before. But if we rewrite it in the form

$$\int \frac{dx}{(x^2 + 4x + 4) + 1} = \int \frac{dx}{(x + 2)^2 + 1},$$

it now begins to resemble something involving an arctangent. In fact, if we make the substitution $u = x + 2$, so $du = dx$, we can write

$$\begin{aligned}\int \frac{dx}{x^2 + 4x + 5} &= \int \frac{du}{u^2 + 1} \\ &= \arctan u + C \\ &= \arctan(x + 2) + C.\end{aligned}$$

Example 7 The technique of completing the square even works for expressions we could have factored directly, if we had noticed:

$$\begin{aligned}\int \frac{dx}{x^2 + 4x + 3} &= \int \frac{dx}{(x + 2)^2 - 1} \\ &= \int \frac{dx}{(x + 2 - 1)(x + 2 + 1)} \\ &= \int \frac{dx}{(x + 1)(x + 3)} \\ &= \frac{1}{2}\int \frac{dx}{x + 1} - \frac{1}{2}\int \frac{dx}{x + 3} \\ &= \frac{1}{2} \ln \left| \frac{x + 1}{x + 3} \right| + C.\end{aligned}$$

Example 8 Evaluate

$$\int \frac{dx}{\sqrt{6x - x^2}}.$$

Note that $6x - x^2 = -(x^2 - 6x) = -(x-3)^2 + 9 = 9 - (x-3)^2$. If we now substitute $x - 3 = 3u$, with $dx = 3du$, we get

$$\begin{aligned}
\int \frac{dx}{\sqrt{6x - x^2}} &= \int \frac{dx}{\sqrt{9 - (x-3)^2}} \\
&= \int \frac{3\,du}{\sqrt{9 - 9u^2}} = \int \frac{du}{\sqrt{1 - u^2}} \\
&= \arcsin u + C \\
&= \arcsin \frac{x-3}{3} + C.
\end{aligned}$$

Trigonometric Polynomials

A **trigonometric polynomial** is any sum of constant multiples of products of trigonometric functions. The preceding techniques have shown some cases where such trigonometric polynomials can arise, even though the original problem had no apparent reference to trigonometric functions. There are many different ways of breaking trigonometric polynomials down into special cases which can then be integrated. We will develop one way which has the virtue of using few special cases, so that it can be used fairly automatically. It also introduces a powerful tool—that of **reduction formula**—which can be used to generate mathematical results interesting in their own right. One example is the striking representation of π derived in chapter 12.1. Other examples are developed in the exercises at the end of this section.

Since every trigonometric function is expressible in terms of sines and cosines, any trigonometric polynomial can be written as a sum of terms of the form $c \sin^m x \cos^n x$ where c is a constant and m and n are integers—positive, negative, or 0. For instance, $5 \sec^2 x \tan^5 x$ can be rewritten as $5 \sin^5 x \cos^{-7} x$. To find antiderivatives for trigonometric polynomials, it therefore suffices to be able to evaluate integrals of the form

$$\int \sin^m x \cos^n x\, dx.$$

We will see how to find antiderivatives for functions of this sort by breaking the problem into a series of special cases:

Category I	either $m \geq 0$ or $n \geq 0$ (or both)
Case 1	$m = 1$ or $n = 1$
Case 2	$m = 0$ or $n = 0$
Category II	m and n both negative

Category I: Either $m \geq 0$ or $n \geq 0$ (or both)

Assume for the sake of explicitness that $m \geq 0$. We can then use the identity $\sin^2 x = 1 - \cos^2 x$ to replace $\sin^m x$ entirely by cosine terms if m is even, or to replace all but one of the sine terms by cosines if m is odd. A similar replacement can be made if $n \geq 0$.

Example 9

$$\begin{aligned}\sin^4 x \, \cos^6 x &= (1 - \cos^2 x)^2 \cdot \cos^6 x \\ &= (1 - 2\,\cos^2 x + \cos^4 x) \cdot \cos^6 x \\ &= \cos^6 x - 2\,\cos^8 x + \cos^{10} x.\end{aligned}$$

(Note that in this example we could just as well have expressed $\cos^6 x$ entirely in terms of $\sin x$.)

Example 10

$$\begin{aligned}\sin^3 x \, \cos^{-8} x &= \sin x \cdot (1 - \cos^2 x) \cdot \cos^{-8} x \\ &= \sin x \, \cos^{-8} x - \sin x \, \cos^{-6} x.\end{aligned}$$

Example 11

$$\begin{aligned}\sin^{-7} x \, \cos^7 x &= \sin^{-7} x \cdot (1 - \sin^2 x)^3 \cdot \cos x \\ &= \sin^{-7} x \, \cos x - 3\,\sin^{-5} x \, \cos x + 3\,\sin^{-3} x \, \cos x \\ &\quad - \sin^{-1} x \, \cos x.\end{aligned}$$

We can thus reduce any problem in Category I to one of two special cases:

Case 1	$m = 1$ or $n = 1$
Case 2	$m = 0$ or $n = 0$

We will now see how to find antiderivatives for these cases.

Case 1: $m = 1$ or $n = 1$ Since the two possibilities are analogous, we will consider the case with $n = 1$. Then m can be any real number at all, not necessarily an integer. We make the substitution $u = \sin x$, so that $du = \cos x\,dx$, and

$$\int \sin^m x \cos x\,dx = \int u^m\,du = \begin{cases} \dfrac{1}{m+1}u^{m+1} + C & \text{if } m \neq -1, \\ \ln|u| + C & \text{if } m = -1. \end{cases}$$

Replacing u by its expression in x we have the antiderivative:

$$\boxed{\int \sin^m x \cos x\,dx = \begin{cases} \dfrac{1}{m+1}\sin^{m+1} x + C & \text{if } m \neq -1, \\ \ln|\sin x| + C & \text{if } m = -1. \end{cases}}$$

The antiderivative of the cotangent

Remark: The instance $m = -1$ in this case is worth singling out, as it gives us an antiderivative for $\cot x$:

$$\int \cot x\,dx = \int \frac{\cos x}{\sin x}\,dx = \ln|\sin x| + C.$$

Integrals where $m = 1$ are handled in a completely analogous fashion. You should check that

$$\boxed{\int \cos^n x \sin x\,dx = \begin{cases} \dfrac{-1}{n+1}\cos^{n+1} x + C & \text{if } n \neq -1, \\ -\ln|\cos x| + C & \text{if } n = -1. \end{cases}}$$

The antiderivative of the tangent

Remark: Notice that $n = -1$ gives us an antiderivative for $\tan x$:

$$\int \tan x\,dx = \int \frac{\sin x}{\cos x}\,dx = -\ln|\cos x| + C.$$

Case II: $m = 0$ or $n = 0$ Again the two possibilities are analogous, so we will look at instances where $n = 0$. There are a number of clever ways for dealing with antiderivatives of functions of this form, many of them depending on special subcases according to whether m is even or odd, positive or negative, etc. We will develop a single method which deals with all cases in the same way.

Think of $\sin^n x$ as $\sin^{n-1} x \cdot \sin x$ and use integration by parts with

$$F(x) = \sin^{n-1} x \qquad \text{and} \qquad g(x) = \sin x;$$

then

$$f(x) = (n-1)\sin^{n-2} x\, \cos x \qquad \text{and} \qquad G(x) = -\cos x.$$

Therefore

$$\int \sin^n x\,dx = -\sin^{n-1} x\, \cos x + (n-1)\int \sin^{n-2} x\, \cos^2 x\,dx.$$

Now since $\cos^2 x = 1 - \sin^2 x$, we can rewrite the integral on the right-hand side as

$$\int \sin^{n-2} x\, \cos^2 x\,dx. = \int \sin^{n-2} x\,dx - \int \sin^n x\,dx$$

—an expression involving the original integral we are trying to evaluate! If we now substitute this expression in our original equation and bring all the terms involving $\sin^n x$ over to the left-hand side, we have

$$n\int \sin^n x\,dx = -\sin^{n-1} x\, \cos x + (n-1)\int \sin^{n-2} x\,dx,$$

so that

$$\boxed{\int \sin^n x\,dx = \frac{-1}{n}\sin^{n-1} x\, \cos x + \frac{n-1}{n}\int \sin^{n-2} x\,dx.}$$

We thus have a **reduction formula** which reduces the problem of finding an antiderivative for $\sin^n x$ to the problem of finding an antiderivative for $\sin^{n-2} x$. This in turn can be reduced to finding an antiderivative for $\sin^{n-4} x$, and so on, until we get down to having to find an antiderivative for $\sin x$ (if n is odd), or for 1 (if n is even).

A reduction formula

Example 12

$$\begin{aligned}\int \sin^5 x\,dx &= \frac{-1}{5}\sin^4 x\cos x + \frac{4}{5}\int \sin^3 x\,dx\\ &= \frac{-1}{5}\sin^4 x\cos x + \frac{4}{5}\left(\frac{-1}{3}\sin^2 x\cos x + \frac{2}{3}\int \sin x\,dx\right)\\ &= \frac{-1}{5}\sin^4 x\cos x - \frac{4}{15}\sin^2 x\cos x - \frac{8}{15}\cos x + C.\end{aligned}$$

Check this answer by taking the derivative of the right-hand side. To show that this derivative really is equal to the integrand on the left, you will need to express all the cosines in terms of sines.

Example 13 If we let $n = 2$, we quickly get the antiderivative for $\sin^2 x$ that we've needed at several points already:

$$\begin{aligned}\int \sin^2 x\,dx &= \frac{-1}{2}\sin x\cos x + \frac{1}{2}\int 1\,dx\\ &= \frac{x}{2} - \frac{1}{2}\sin x\cos x.\end{aligned}$$

In its current form, the reduction formula works best for $n > 0$

The reduction formula as stated is most convenient for $n > 0$, although it is true for any number $n \neq 0$. For if $n < 0$, though, we want to *increase* the exponent, replacing a problem of finding an antiderivative for $\sin^n x$ by a problem where the exponent is less negative. We can do this by rearranging the formula as

$$\int \sin^{n-2} x\,dx = \frac{n}{n-1}\int \sin^n x\,dx + \frac{1}{n-1}\sin^{n-1} x\cos x.$$

The reduction formula for negative exponents

Since we are interested in negative exponents, call $n-2$ by a new name, $-k$. But if $n-2 = -k$, then $n = -k+2$, and we can rewrite our formula as

$$\boxed{\int \sin^{-k} x\,dx = -\frac{1}{k-1}\sin^{-(k-1)} x\cos x + \frac{k-2}{k-1}\int \sin^{-(k-2)} x\,dx.}$$

With this formula we can reduce the problem of finding an antiderivative for $\sin^{-k} x$ to the problem of finding an antiderivative for $\sin^{-k+2} x$. This in turn can be reduced to finding an antiderivative for $\sin^{-k+4} x$, and so on, until we get up to having to find an antiderivative for $\sin^{-1} x$ (if k is odd), or for 1 (if k is even). All we need, then, is an antiderivative for $\sin^{-1} x$. But $\sin^{-1} x = \csc x$, and in chapter 11.4 (page 728) we found that

$$\int \csc x\,dx = \int \sin^{-1} x\,dx = -\ln|\csc x + \cot x| + C.$$

We can now handle antiderivatives for any negative integer exponent of the sine function.

Example 14 We can check this formula by trying $k = 2$, which will give us the antiderivative of $\csc^2 x$:

$$\begin{aligned}\int \csc^2 x &= \int \sin^{-2} x\,dx \\ &= -\frac{1}{1}\sin^{-1} x\ \cos x + \frac{0}{1}\int \sin^0 x\,dx \\ &= -\sin^{-1} x\ \cos x + C = -\cot x + C,\end{aligned}$$

as it should.

Example 15

$$\begin{aligned}\int \sin^{-3} x\,dx &= -\frac{1}{2}\sin^{-2} x\ \cos x + \frac{1}{2}\int \sin^{-1} x\,dx \\ &= \frac{1}{2}\left(-\sin^{-2} x \cos x - \ln|\csc x + \cot x|\right) + C.\end{aligned}$$

In the exercises you are asked to derive the following reduction formulas for the cosine function:

$$\boxed{\begin{aligned}\int \cos^m x\,dx &= \frac{1}{m}\cos^{m-1} x\ \sin x + \frac{m-1}{m}\int \cos^{m-2} x\,dx, \\ \int \cos^{-m} x\,dx &= \frac{1}{m-1}\cos^{-m+1} x\ \sin x + \frac{m-2}{m-1}\int \cos^{-m+2} x\,dx.\end{aligned}}$$

Category II: Both $m < 0$ and $n < 0$

If we divide the identity $\cos^2 x + \sin^2 x = 1$ by $\sin^2 x\ \cos^2 x$, we get the identity

$$\sin^{-2} x + \cos^{-2} x = \sin^{-2} x\ \cos^{-2} x.$$

We will now use this identity to express anything of the form $\cos^{-r} x\ \sin^{-s} x$ (where $r > 0$ and $s > 0$) as a sum of terms of the form $\sin^{-h} x$, or $\cos^{-i} x$, or $\sin x\ \cos^{-j} x$, or $\sin^{-k} x\ \cos x$. Since we learned how to find antiderivatives for expressions like these in the previous cases, we will then be done.

The trick in transforming $cos^{-r} x\ \sin^{-s} x$ to the desired form is to multiply by $(\cos x\ \cos^{-1} x)$ or $(\sin x\ \sin^{-1} x)$ as needed so that both the sine and the cosine terms appear to *even* negative exponents. Then simply keep using the identity above until there's nothing left to use it on. The following three examples should make clear how the reduction then works.

Example 16 (r and s both even already)

$$\begin{aligned}
\sin^{-4}x\,\cos^{-6}x &= \left(\sin^{-2}x\,\cos^{-2}x\right)^2\,\cos^{-2}x\\
&= \left(\sin^{-2}x+\cos^{-2}x\right)^2\,\cos^{-2}x\\
&= (\sin^{-4}x+2\sin^{-2}x\,\cos^{-2}x+\cos^{-4}x)\,\cos^{-2}x\\
&= (\sin^{-4}x+2(\sin^{-2}x+\cos^{-2}x)+\cos^{-4}x)\,\cos^{-2}x\\
&= \sin^{-4}x\,\cos^{-2}x+2\sin^{-2}x\,\cos^{-2}x+2\cos^{-4}x\\
&\qquad+\cos^{-6}x\\
&= \sin^{-2}x\,(\sin^{-2}x+\cos^{-2}x)+2(\sin^{-2}x+\cos^{-2}x)\\
&\qquad+2\cos^{-4}x+\cos^{-6}x\\
&= \sin^{-4}x+\sin^{-2}x\,\cos^{-2}x+2\sin^{-2}x+2\cos^{-2}x\\
&\qquad+2\cos^{-4}x+\cos^{-6}x\\
&= \sin^{-4}x+(\sin^{-2}x+\cos^{-2}x)+2\sin^{-2}x\\
&\qquad+2\cos^{-2}x+2\cos^{-4}x+\cos^{-6}x\\
&= \sin^{-4}x+3\sin^{-2}x+3\cos^{-2}x+2\cos^{-4}x+\cos^{-6}x.
\end{aligned}$$

While this process is tedious, it requires little thought—you simply replace $\sin^{-2}x\,\cos^{-2}x$ with $\sin^{-2}x+\cos^{-2}x$ at every opportunity until there is no negative-exponent sine term multiplying any negative-exponent cosine term. We will use this result to demonstrate how to deal with cases where either r or s (or both) is odd.

Example 17 (r even and s odd)

$$\begin{aligned}
\sin^{-4}x\,\cos^{-5}x &= \sin^{-4}x\,\cos^{-6}x\,\cos x\\
&= \sin^{-4}x\,\cos x+3\sin^{-2}x\,\cos x+3\cos^{-1}x\\
&\qquad+2\cos^{-3}x+\cos^{-5}x
\end{aligned}$$

Example 18 (both r and s odd)

$$\begin{aligned}
\sin^{-3}x\,\cos^{-5}x &= \sin x\,\sin^{-4}x\,\cos^{-6}x\,\cos x\\
&= \sin^{-3}x\,\cos x+3\sin^{-1}x\,\cos x+3\sin x\,\cos^{-1}x\\
&\qquad+2\sin x\,\cos^{-3}x+\sin x\,\cos^{-5}x
\end{aligned}$$

Exercises

1. Find the following antiderivatives (a is a positive constant):

a) $\displaystyle\int \frac{dx}{\sqrt{1-4x^2}}$

b) $\displaystyle\int \frac{dx}{\sqrt{1-4x^2}}$

c) $\displaystyle\int \frac{dx}{\sqrt{4+x^2}}$

d) $\displaystyle\int \frac{x\,dx}{\sqrt{4+x^2}}$

e) $\displaystyle\int \frac{dx}{(a^2-x^2)^{3/2}}$

f) $\displaystyle\int \frac{dx}{4+x^2}$

g) $\displaystyle\int \frac{x\,dx}{4+x^2}$

h) $\displaystyle\int \frac{dx}{x\sqrt{4+x^2}}$

i) $\displaystyle\int \frac{x\,dx}{\sqrt{x^2-a^2}}$

j) $\displaystyle\int \frac{dx}{(a^2+x^2)^2}$

2. Evaluate the following integrals:

a) $\displaystyle\int_1^{-1} \frac{dx}{4-x^2}$

b) $\displaystyle\int_1^{2} \sqrt{x^2-1}\,dx$

c) $\displaystyle\int_0^{\pi/3} x\,\sec^2 x\,dx$

d) $\displaystyle\int_0^{1} \frac{dx}{(2-x^2)^{3/2}}$

e) $\displaystyle\int_0^{\infty} \frac{dx}{9+x^2}$

f) $\displaystyle\int_a^{2a} x^3\sqrt{x^2-a^2}\,dx$

3. Sketch the ellipse $\dfrac{x^2}{a^2}+\dfrac{y^2}{b^2}=1$, labelling the coordinates of the points where it crosses the x-axis and the y-axis. Prove that the area of this ellipse is πab.

4. Find the following antiderivatives:

a) $\displaystyle\int \frac{dx}{\sqrt{x^2-2x-8}}$

b) $\displaystyle\int \frac{dx}{x^2+6x+10}$

c) $\displaystyle\int \frac{dx}{\sqrt{x^2+6x+8}}$

d) $\displaystyle\int \frac{x\,dx}{x^2+4x+5}$

e) $\displaystyle\int \frac{x\,dx}{\sqrt{5+4x-x^2}}$

f) $\displaystyle\int \frac{(2x+7)\,dx}{4x^2+4x+5}$

g) $\displaystyle\int \frac{(4x-3)\,dx}{\sqrt{-x^2-2x}}$

h) $\displaystyle\int \frac{dx}{(a^2-x^2-2x)^2}$

5. a) If $x = a \sec u$, where a is a constant, draw a right triangle containing an angle u with lengths of sides specified to reflect this relation between x, a, and u.

b) In terms of x and a, what is $\sin u$? What is $\cos u$? What is $\tan u$?

6. Evaluate the following:

a) $\displaystyle\int \frac{dx}{\sin x\ \cos x}$

b) $\displaystyle\int \cos^3 x\ \sin^{-4} x\, dx$

c) $\displaystyle\int \csc^4 x\ \cot^2 x\, dx$

d) $\displaystyle\int \sin 3x\ \cot 3x\, dx$

e) $\displaystyle\int_0^{\pi/2} \sin^n x\ \cos^3 x\, dx$

f) $\displaystyle\int \tan^5 x\, dx$

g) $\displaystyle\int \frac{\sin^3 5x\, dx}{\sqrt[3]{\cos 5x}}$

h) $\displaystyle\int \frac{\cos^3(\ln x)\, dx}{x}$

i) $\displaystyle\int \sec^4 x\ \ln(\tan x)\, dx$

j) $\displaystyle\int_0^{a/2} \frac{dx}{(a^2 - x^2)^{3/2}}$

7. Use the analysis of Example 17 (page 748) to find an antiderivative for $\sin^{-4} x\ \cos^{-5} x$.

Reduction formulas

8. Derive the reduction formulas for the cosine function given on page 747.

9. a) By writing $\tan^n x = \tan^{n-2} x(\sec^2 x - 1)$, get a reduction formula which expresses $\displaystyle\int \tan^n x\, dx$ in terms of $\displaystyle\int \tan^{n-2} x\, dx$.

b) Use this evaluation formula to find $\displaystyle\int \tan^6 x\, dx$.

c) Show that

$$\int_0^{\pi/4} \tan^n x\, dx = \begin{cases} \dfrac{1}{n-1} - \dfrac{1}{n-3} + \cdots \pm \dfrac{1}{3} \mp 1 \pm \pi/4 & \text{if } n \text{ is even,} \\ \dfrac{1}{n-1} - \dfrac{1}{n-3} + \cdots \pm \dfrac{1}{4} \mp \dfrac{1}{2} \pm \dfrac{1}{2} \ln 2 & \text{if } n \text{ is odd.} \end{cases}$$

d) Give a clear argument why $\displaystyle\lim_{n\to\infty}\int_0^{\pi/4}\tan^n x\,dx = 0.$

e) Prove that

$$\lim_{k\to\infty}\left(1-\frac{1}{3}+\frac{1}{5}-\cdots\pm\frac{1}{2k+1}\right)=\frac{\pi}{4},$$

and

$$\lim_{k\to\infty}\left(1-\frac{1}{2}+\frac{1}{3}-\cdots\pm\frac{1}{k}\right)=\ln 2.$$

10. a) By writing $\sec^n x$ as $\sec^{n-2} x\,\sec^2 x$ and using integration by parts, get a reduction formula which expresses $\displaystyle\int\sec^n x\,dx$ in terms of $\displaystyle\int\sec^{n-2} x\,dx$.

b) Since $\sec x = \cos^{-1} x$, the formula you got in part (a) could also have been obtained from the reduction formula for cosines on page 747. Try it and see if the formulas are in fact the same.

11. a) Find a reduction formula that expresses

$$\int x^n\,e^x\,dx \quad \text{in terms of} \quad \int x^{n-1}\,e^x\,dx.$$

b) Using the results of part (a), show that

$$\frac{1}{n!}\int_0^t x^n\,e^x\,dx = e^t\left(\frac{t^n}{n!}-\frac{t^{n-1}}{(n-1)!}+\frac{t^{n-2}}{(n-2)!}-\cdots\pm\frac{t^2}{2!}\mp t\pm 1\right)\mp 1.$$

c) Explain why, for a fixed value of t,

$$\lim_{n\to\infty}\frac{1}{n!}\int_0^t x^n\,e^x\,dx = 0.$$

d) Prove that

$$\lim_{n\to\infty}\left(1-t+\frac{t^2}{2!}-\frac{t^3}{3!}+\frac{t^4}{4!}-\cdots\pm\frac{t^n}{n!}\right)=e^{-t}.$$

12. a) Find a reduction formula expressing

$$\int\frac{dx}{(1+x^2)^n} \quad \text{in terms of} \quad \int\frac{dx}{(1+x^2)^{n-1}};$$

you can do this using integration by parts, or you can use a trigonometric substitution.

b) What is the exact value of $\displaystyle\int_0^1 \frac{dx}{(1+x^2)^5}$?

13. Our approach to integrating trigonometric polynomials was to express everything in the form $\sin^m x \cos^n x$. We can just as readily express everything in the form $\sec^j x \tan^k x$ (j and k integers—positive, negative, or 0), and develop our technique by dealing with various cases of this. See if you can work out the details, trying to parallel the approach developed in the text using sines and cosines as our basic functions.

11.6 Simpson's Rule

A return to numerical methods

This chapter has concentrated on formulas for antiderivatives, because a formula conveys compactly a lot of information. However, you must not lose sight of the fact that most antiderivatives cannot be found by such analytic methods. The integrand may be a data function, for instance, and thus have no formula. And even when the integrand is given by a formula, there may be no formula for the antiderivative itself. One possibility in such cases is to **approximate** such a function by a function—such as a polynomial—for which we can readily find an antiderivative. In chapter 10 we saw some methods for doing this. In chapter 12 we introduce Fourier series, providing another family of approximating functions for which antiderivatives can be readily obtained. Another approach is to find a desired definite integral using approximating rectangles, as we did in chapter 6.

Efficient numerical integration

In any case, numerical methods are inescapable, but accurate results require many calculations. This takes time—even on a modern high-speed computer. A numerical method is said to be **efficient** if it gets accurate results quickly, that is, with relatively few calculations. In chapter 6 we saw that *midpoint* Riemann sums are much more efficient than left or right *endpoint* Riemann sums. We will look at these and other methods in detail in this section. The most efficient method we will develop is called Simpson's rule.

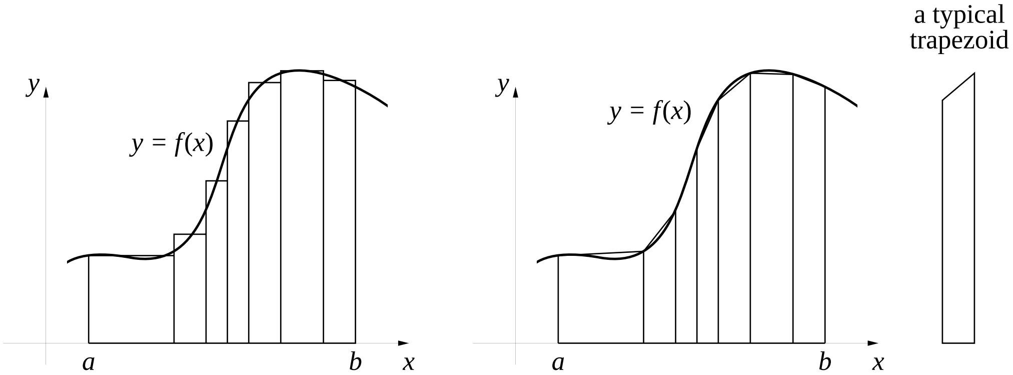
The Trapezoid Rule

We interpret the integral

$$\int_a^b f(x)\,dx$$

as the area under the graph $y = f(x)$ between $x = a$ and $x = b$. We interpret a Riemann sum as the total area of a collection of rectangles that approximate the area under the graph. The tops of the rectangles are level, and they represent the graph of step function. Clearly, we get a better approximation to the graph by using slanted lines. They form the tops of a sequence of **trapezoids** that approximate the area under the graph.

Replace rectangles by trapezoids

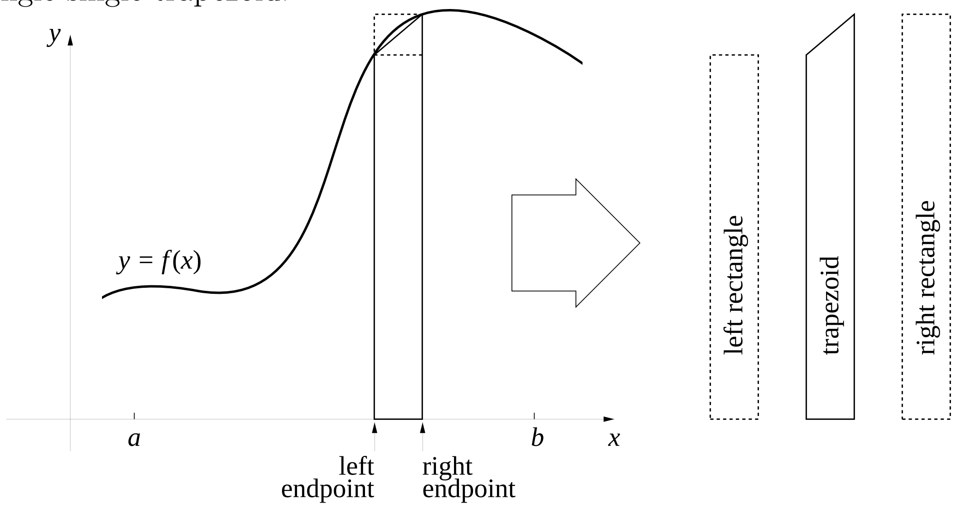


rectangular approximation

trapezoidal approximation

Let's figure out the areas of these trapezoids. They are related in a simple way to the rectangles that we would construct at the right and left endpoints to calculate Riemann sums. To see the relation, let's take a closer look at a single single trapezoid.

Each trapezoid is sandwiched between two rectangles. . .



. . . whose average area equals the area of the trapezoid

It is sandwiched between two rectangles, one taller and one shorter. In our picture the height of the taller rectangle is f(right endpoint). We will call it the *right rectangle*. The height of the shorter rectangle is f(left endpoint). We will call it the *left rectangle*. For other trapezoids the left rectangle may be the taller one. In any case, the trapezoid is exactly half-way between the two rectangles in size, and thus its area is the *average* of the areas of the rectangles:

$$\text{area trapezoid} = \tfrac{1}{2}\left(\text{area left rectangle} + \text{area right rectangle}\right).$$

The trapezoidal approximation is the average of left and right Riemann sums

If we sum over the areas of all of the trapezoids, the areas of the right rectangles sum to the right Riemann sum, and similarly for the left rectangles. In follows that

$$\text{trapezoidal approximation} = \tfrac{1}{2}(\text{left Riemann sum} + \text{right Riemann sum}).$$

The pictures make it clear that the trapezoidal approximation should be significantly better than either a left or a right Riemann sum. To test this numerically, let's get numerical estimates for the integral

$$\int_1^3 \frac{1}{x}\,dx = \ln 3 = 1.098612288668\ldots.$$

Comparing approximations

(The relation between the trapezoid approximation and the left and right Riemann sums holds for any choice Δx_k of subintervals. However, we will use equal subintervals to make the calculations simpler.) Here is how our four main estimates compare when we use 100 subintervals.

$n = 100$	approximation	error
right	1.09197525	6.63×10^{-3}
left	1.10530858	-6.69×10^{-3}
midpoint	1.09859747	1.48×10^{-5}
trapezoidal	1.09864191	-2.90×10^{-5}

The figures in this table are calculated to 8 decimal places, and the column marked *error* is the difference

$$1.09861229 - \text{approximation},$$

so that the error is negative if the approximation is too large. The left Riemann sum is too large, for instance.

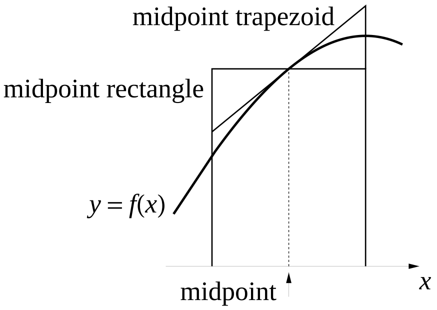
Before we comment on the differences between the estimates, let's gather more data. Here are the calculations for 1000 subintervals. Note that the midpoint and trapezoidal approximations are more than 1000 times better than the left or right Riemann sums!

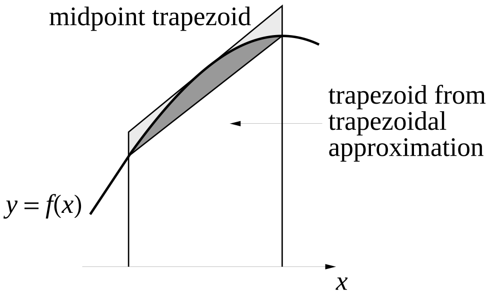
$n = 1000$	approximation	error
right	1.09794591	6.663×10^{-4}
left	1.09927925	-6.669×10^{-4}
midpoint	1.09861214	1.4×10^{-7}
trapezoidal	1.09861258	-2.9×10^{-7}

A surprise: the midpoint approximation is even better than the trapezoidal

We expected the trapezoidal approximation to be better than either the right or left Riemann sum. The surprising observation is that the midpoint Riemann sum is even better! In fact, it appears that the midpoint Riemann sum has only *half* the error of the trapezoidal approximation.

The figures below explain geometrically why the midpoint approximation is better than the trapezoidal. The first step is shown on the left. Take a midpoint rectangle (whose height is f(midpoint)), and rotate the top edge around the midpoint until it is tangent to the graph of $y = f(x)$. Call this a **midpoint trapezoid**. Notice that the trapezoid has the same area as the rectangle.





Shading depicts the errors

The second step is to compare the midpoint trapezoid to the one used in the trapezoidal approximation. This is done on the right. The error coming from the midpoint trapezoid is shaded light gray, while the error from the trapezoidal approximation is dark gray. The midpoint trapezoid is the better approximation. Since the midpoint rectangle has the same area as the midpoint trapezoid, we now see why the midpoint Riemann sum is more accurate than the trapezoidal approximation. This picture also explains why the errors of the two approximations have different signs, which we noticed first in the tables.

Simpson's Rule

Combine good approximations...

Our goal is a calculation scheme for integrals that gives an error is as small as possible. The trapezoidal and the midpoint approximations are both good—but we can combine them to get something even better. Here is why. The tables and the figure above indicate that the *errors* in the two approximations have opposite signs, and the midpoint error is only about half the size of the trapezoidal error (in absolute value). Thus, if we form the sum

$$2 \times \text{midpoint approximation} + \text{trapezoidal approximation},$$

...so that most of the error cancels

then most of the error will cancel. Now this sum is approximately three times the value of the integral, because each term in it approximates the integral itself. Therefore, if we divide by three, then

$$\tfrac{2}{3} \times \text{midpoint approximation} + \tfrac{1}{3} \times \text{trapezoidal approximation}$$

should be a superb approximation to the integral.

Let's try this approximation on our test integral

$$\int_1^3 \frac{1}{x}\,dx$$

with $n = 100$ subintervals. Using the numbers from the table on page 754, we obtain

$$\tfrac{2}{3} \times 1.09859747 + \tfrac{1}{3} \times 1.09864191 = 1.098612283,$$

which gives the value of the integral accurate to 8 decimal places. This method of approximating integrals is called **Simpson's rule**.

Using Riemann sums to carry out Simpson's rule

We can use the program RIEMANN to do the calculation. First calculate left and right Riemann sums, and take their average. That is the trapezoidal approximation. Then calculate the midpoint Riemann sum. Since

$$\text{trapezoid} = \tfrac{1}{2} \times \text{left} + \tfrac{1}{2} \times \text{right},$$

Simpson's rule reduces to this combination of left, right, and midpoint sums:

$$\begin{aligned}&\tfrac{2}{3} \times \text{midpoint} + \tfrac{1}{3} \times \text{trapezoidal}\\&\quad= \tfrac{2}{3}\text{midpoint} + \tfrac{1}{3}\left(\tfrac{1}{2} \times \text{left} + \tfrac{1}{2} \times \text{right}\right)\\&\quad= \tfrac{2}{3} \times \text{midpoint} + \tfrac{1}{6} \times \text{left} + \tfrac{1}{6} \times \text{right}\\&\quad= \tfrac{1}{6}\left(4 \times \text{midpoint} + \text{left} + \text{right}\right)\end{aligned}$$

Simpson's rule:

$$\int f(x)\,dx \approx \frac{1}{6}\left(\text{left sum} + \text{right sum} + 4 \times \text{midpoint sum}\right)$$

The accuracy of Simpson's rule

You can get even more accuracy if you keep track of more digits in the left, right, and midpoint Riemann sums. For example, if you estimate

$$\int_1^3 \frac{1}{x}\,dx = \ln 3 = 1.098\,612\,288\,668\ldots$$

to 14 decimal places, you will get the following.

left:	1.105 308 583 647 79
right:	1.091 975 250 314 45
midpoint:	1.098 597 475 005 31

When combined these give the estimate 1.098 612 288 997 3, which differs from the true value by less than 3.3×10^{-10}. In other words, the calculation is actually correct to 9 decimal places.

An error bound for Simpson's rule

It is possible to get a bound on the error produced by using Simpson's rule to estimate the value of

$$\int_a^b f(x)\,dx.$$

(See the discussion of error bounds in chapter 6.3.) Specifically,

$$\left|\int_a^b f(x)\,dx - \text{Simpson's rule}\right| \leq \frac{M(b-a)^5}{2880\,n^4},$$

where n is the number of subintervals used in the Riemann sums and M is a bound on the size of the fourth derivative of f:

$$|f^{(4)}(x)| \leq M \qquad \text{for all } a \leq x \leq b.$$

The crucial factor n^{-4}

The most important factor in the error bound is the n^4 that appears in the denominator. In our example $n = 100 = 10^2$, and this leads to the factor $1/n^4 = 10^{-8}$ in the error bound. As we saw, the actual error was less than 10^{-9}. Essentially, n is the number of computations we do, and error bound tells how many decimal places of accuracy we can count on. According to the error bound, a ten-fold increase in the number of computations produces four more decimal places of accuracy. *That* is why Simpson's method is efficient.

Exercises

Because Simpson's rule is so efficient, we can use it to get accurate values of some of the fundamental constants of mathematics. For example, since

$$4 \cdot \int_0^1 \frac{dx}{1+x^2} = 4\arctan(x)\,\bigg|_0^1 = 4\arctan(1) = 4 \cdot \frac{\pi}{4} = \pi,$$

we can estimate the value of π by using Simpson's rule to approximate this integral.

1. Evaluate the expression above (including the factor of 4) use Simpson's rule with $n = 2$, 4, 8, and 16. How accurate is each of these estimates of π; that is, how many decimal places of each estimate agree with the true value of π?

2. a) Over the interval $0 \le x \le 1$ it is true that

$$|f^{(4)}(x)| \le 96 \qquad \text{when} \qquad f(x) = \frac{4}{1+x^2}.$$

(You don't need to show this, but how might you do it?) Use this bound to show that $n = 256 = 2^8$ will guarantee that you can find the first 10 decimals of π by using the method of the previous question.

b) Show that if $n = 128 = 2^7$ then the error bound for Simpson's rule does *not* guarantee that you can find the first 10 decimals of π by the same method.

c) Run Simpson's rule with $n = 2^7$ to estimate π. How many decimal places *are* correct? Does this surprise you? In fact the error bound is too timid: it says that the error is no larger than the bound it gives, but the actual error may be much smaller. From your work in part (a), which power of 2 is sufficient to get 10 decimal places accuracy?

3. a) In chapter 6.3 (page 391) a left Riemann sum for

$$\int_0^1 e^{-x^2}\,dx$$

with 1000 equal subdivisions gave 3 decimal places accuracy. How many subdivisions n are needed to get that much accuracy using Simpson's rule? Let n be a power of 2. Start with $n = 1$ and increase n until three digits stabilize.

b) If you use Simpson's rule with $n = 1000$ to estimate this integral, how many digits stabilize?

4. On page 374 a midpoint Riemann sum with $n = 10000$ shows that

$$\int_1^3 \sqrt{1+x^3}\,dx = 6.229959\ldots.$$

How many subdivisions n are needed to get this much accuracy using Simpson's rule? Start with $n = 1$ and keep doubling it until seven digits stabilize.

11.7 Improper Integrals

The Lifetime of Light Bulbs

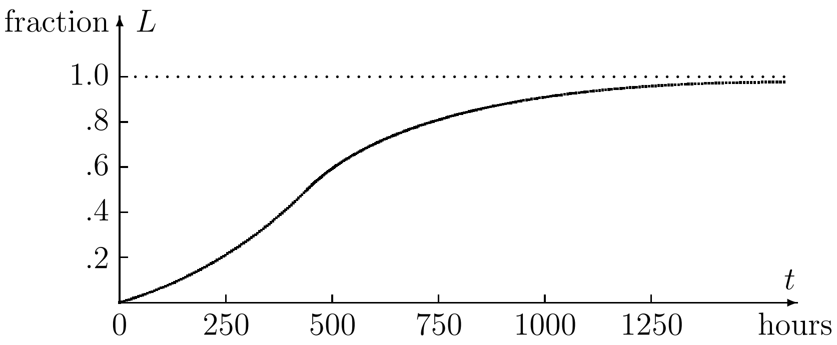
The lifetime of a light bulb is unpredictable

Ordinary light bulbs are supposed to burn about 700 hours, but of course some last longer while others burn out more quickly. It is impossible to know, in advance, the lifetime of a particular bulb you might buy, but it is possible to describe what happens to a large batch of bulbs.

Suppose we take a batch of 1000 light bulbs, start them burning at the same time, and note how long it takes each one to burn out. Let

$$L(t) = \text{fraction of bulbs that burn out before } t \text{ hours}$$

Then $L(t)$ might have a graph that looks like this:



In this example, $L(400) \approx .5$, so about half the bulbs burned out before 400 hours. Furthermore, all but a few have burned out by 1250 hours.

The burnout rate

Manufacturers are very concerned about the way the lifetime of light bulbs varies. They study the output of their factories on a regular basis. It is more common, though, for them to talk about the *rate* r at which bulbs burn out. The rate varies over time, too. In fact, in terms of L, r is just the derivative

$$r(t) = L'(t) \quad \text{bulbs per hour.}$$

However, if we *start* with the rate r, then we get L as the integral

$$L(t) = \int_0^t r(s)\,ds.$$

Lifetime is the integral of burnout rate...

This is yet another consequence of the fundamental theorem of calculus. The integral expression is quite handy. For example, the fraction of bulbs that burn out between $t = a$ hours and $t = b$ hours is

$$L(b) - L(a) = \int_a^b r(s)\,ds.$$

We can even use the integral to say that all the bulbs burn out eventually:

$$L(t) = \int_0^t r(s)\,ds = 1 \qquad \text{when } t \text{ is sufficiently large.}$$

...but there is no upper limit to the lifetime

In practice r is the average burnout rate for many batches of light bulbs, so we can't identify the precise moment when L becomes 1. All we can really say is

$$L(\infty) = \int_0^\infty r(s)\,ds.$$

This is called an **improper integral**, because it cannot be calculated directly: its "domain of integration" is infinite. By definition, its value is obtained as a limit of ordinary integrals:

An integral is *improper* if the domain of integration is infinite

$$\int_0^\infty r(s)\,ds = \lim_{b\to\infty}\int_0^b r(s)\,ds.$$

The **normal density function** of probability theory provides us with another example of an improper integral. In a simple form, the function itself is

$$f(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right).$$

(Recall, $\exp(x) = e^x$.) If x is any *normally distributed quantity* whose average value is 0, then the probability that a randomly chosen value of x lies between the numbers a and b is

$$\int_a^b f(x)\,dx.$$

Since the probability that x lies *somewhere* on the x-axis is 1, we have

The normal probability distribution involves an improper integral

$$\int_{-\infty}^\infty f(x)\,dx = 1.$$

This is an improper integral, and its value is defined by the limit

$$\int_{-\infty}^\infty f(x)\,dx = \lim_{b\to\infty}\int_{-b}^b f(x)\,dx.$$

In the exercises you will have a chance to evaluate this integral.

Evaluating Improper Integrals

An integral with an infinite domain of integration is only one kind of improper integral. A second kind has a finite domain of integration, but the integrand becomes infinite on that domain. For example,

An integral is also improper if its integrand becomes infinite

$$\int_0^1 \frac{dx}{x} \qquad \text{and} \qquad \int_0^1 \ln x\,dx$$

are both improper in this sense. In both cases, the integrand becomes infinite as $x \to 0$. Because the difficulty lies at the endpoint 0, we define

$$\int_0^1 \frac{dx}{x} = \lim_{a\to 0} \int_a^1 \frac{dx}{x}.$$

More generally,

$$\int_a^b f(x)\,dx$$

is an improper integral if $f(x)$ becomes infinite at some point c in the interval $[a,b]$. In that case we define

$$\int_a^b f(x)\,dx = \lim_{q\to 0}\left(\int_a^{c-q} f(x)\,dx + \int_{c+q}^b f(x)\,dx\right).$$

In effect, we avoid the bad spot but "creep up" on it in the limit.

Antiderivatives help find improper integrals

Indefinite integrals—that is, antiderivatives—can be a great help in evaluating improper integrals. Here are some examples.

Example 1. We can evaluate

$$\int_0^\infty e^{-x}\,dx$$

by noting first that $\displaystyle\int e^{-x}\,dx = -e^{-x}$. Therefore

$$\int_0^b e^{-x}\,dx = -e^{-x}\Big|_0^b = -e^{-b} - \left(-e^{-0}\right) = 1 - e^{-b}$$

and

$$\int_0^\infty e^{-x}\,dx = \lim_{b\to\infty}\int_0^b e^{-x}\,dx = \lim_{b\to\infty}\left(1-e^{-b}\right) = 1.$$

Example 2. To evaluate $\displaystyle\int_0^1 \ln x\,dx$, we use the indefinite integral

$$\int \ln x\,dx = x\,\ln x - x.$$

Thus

$$\int_a^1 \ln x\,dx = x\,\ln x - x\Big|_a^1 = -1 - (a\ln a - a) = a - 1 - a\ln a.$$

By direct calculation (using a graphing package, for instance) we can find

$$\lim_{a\to 0}\ a\ln a = 0;$$

therefore

$$\int_0^1 \ln x\,dx = \lim_{a\to 0}\int_a^1 \ln x\,dx = \lim_{a\to 0}\ a - 1 - a\ln a = -1.$$

You should not assume that an improper integral always has a finite value, though. Consider the next example.

Example 3. $\displaystyle\int_0^1 \frac{dx}{x} = \lim_{a\to 0}\int_a^1 \frac{dx}{x} = \lim_{a\to 0}\ln(x)\Big|_a^1 = \lim_{a\to 0}(\ln(1) - \ln(a)) = \infty.$

This is forced because $\lim_{a\to 0}\ \ln(a) = -\infty$, which you can see from the graph of the logarithm function.

Exercises

1. Find the value of each of the following improper integrals. (The value may be ∞.)

a) $\displaystyle\int_{-\infty}^0 e^x\,dx$

b) $\displaystyle\int_1^\infty \frac{du}{u}$

c) $\displaystyle\int_0^1 \frac{dy}{y^2}$

d) $\displaystyle\int_0^{\pi/2} \tan x\,dx$

e) $\displaystyle\int_0^\infty xe^{-x}\,dx$

f) $\displaystyle\int_1^\infty \frac{du}{u^2}$

g) $\displaystyle\int_0^\infty \frac{x}{1+x^2}\,dx$

h) $\displaystyle\int_1^3 \frac{x}{x^2-1}\,dx$

2. Use the reduction formula for $\displaystyle\int \frac{dx}{(1+x^2)^n}$ you found on page 752 to find the exact value of $\displaystyle\int_0^\infty \frac{dx}{(1+x^2)^{10}}$

The normal density function

The next two questions concern the improper integral

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx$$

of the normal density function defined on page 761. The goal is to determine the value of this integral.

3. First, use RIEMANN to estimate the value of

$$\frac{1}{\sqrt{2\pi}} \int_{-b}^{b} e^{-x^2/2}\, dx$$

when b has the different values 1, 10, 100, and 1000. On the basis of these results, estimate

$$\lim_{b \to \infty} \frac{1}{\sqrt{2\pi}} \int_{-b}^{b} e^{-x^2/2}\, dx.$$

This gives one estimate of the value of the improper integral.

4. a) To construct a second estimate, begin by sketching the graph of the normal density function

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

on an interval centered at the origin. Use the graph to argue that

$$\frac{1}{\sqrt{2\pi}} \int_{-b}^{b} e^{-x^2/2}\, dx = 2\left(\frac{1}{\sqrt{2\pi}} \int_{0}^{b} e^{-x^2/2}\, dx\right) = \sqrt{2/\pi} \int_{0}^{b} e^{-x^2/2}\, dx$$

and therefore

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2/\pi} \int_{0}^{\infty} e^{-x^2/2}\, dx.$$

b) Now consider the accumulation function

$$F(t) = \sqrt{2/\pi} \int_{0}^{t} e^{-x^2/2}\, dx.$$

We want to find $F(\infty) = \lim_{t \to \infty} F(t)$. According to the fundamental theorem of calculus, $y = F(t)$ satisfies the initial value problem

$$\frac{dy}{dt} = \sqrt{2/\pi} \cdot e^{-t^2/2}, \qquad y(0) = 0.$$

Use a differential equation solver (e.g., PLOT) to graph the solution $y = F(t)$ to this problem. From the graph determine

$$F(\infty) = \lim_{t \to \infty} F(t).$$

c) Does your results in part (b) and question 2 agree? Do they agree with the value the text claims for the improper integral. (Remember, the value is the probability that a randomly chosen number will lie *somewhere* on the number line between $-\infty$ and $+\infty$.)

The gamma function

The **factorial function** is defined for a positive integer n by the formula

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 3 \cdot 2 \cdot 1.$$

For example, $1! = 1$, $2! = 2$, $3! = 6$, $4! = 24$, and $10! = 3628800$. The factorial function is used often in diverse mathematical contexts, but its use is sometimes limited by the fact that it is defined only for positive integers. How might the function be defined on an expanded domain, so that we could deal with expressions like $tfrac12!$, for example? The *gamma function* answers this question.

The **gamma function** $\Gamma(x)$ is defined by the improper integral

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} \, dt.$$

Notice that t is the active variable in this integral. While the integration is being performed, x is treated as a constant; for the integral to converge, we need $x > 0$.

5. Show that $\Gamma(1) = 1$.

6. Using integration by parts, show that $\Gamma(x+1) = x \cdot \Gamma(x)$. You may use the fact that

$$\frac{t^p}{e^t} \to 0 \text{ as } t \to +\infty.$$

The property $\Gamma(x+1) = x \cdot \Gamma(x)$ makes the gamma function like the factorial function, because

$$(n+1)! = (n+1) \cdot \underbrace{n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 3 \cdot 2 \cdot 1}_{= n!} = (n+1) \cdot n!.$$

Notice there is a slight difference, though. We explore this now.

7. Using the property $\Gamma(x+1) = x \cdot \Gamma(x)$, calculate $\Gamma(2)$, $\Gamma(3)$, $\Gamma(4)$, $\Gamma(5)$, and $\Gamma(6)$. On the basis of this evidence, fill in the blank:

For a positive integer n, $\qquad \Gamma(n) =$ _____!

Using this relation, give a computable meaning to the expression $\frac{1}{2}!$.

8. Estimate the value of $\Gamma(1/2)$. (Exercises 2 and 3 above offer two ways to estimate the value of an improper integral.)

9. In fact, $\Gamma(1/2) = \sqrt{\pi}$ exactly. You can show this by employing several of the techniques developed in this chapter. Start with

$$\Gamma(1/2) = \int_0^\infty e^{-t} t^{-\frac{1}{2}}\, dt.$$

a) Make the substitution $u = (2t)^{\frac{1}{2}}$ and show that the integral becomes

$$\Gamma(1/2) = \sqrt{2} \int_0^\infty e^{-u^2/2} du.$$

b) From exercise 4 you know

$$\sqrt{2/\pi} \int_0^\infty e^{-u^2/2}\, du = 1.$$

(Check this.) Now, using some algebra, show $\Gamma(1/2) = \sqrt{\pi}$.

c) Compare your estimate for $\Gamma(1/2)$ from exercise 7 with the exact value $\sqrt{\pi}$.

10. a) Determine the exact values of $\Gamma(3/2)$ and $\Gamma(5/2)$.

b) In exercise 6 you gave a meaning to the expression $\frac{1}{2}!$; can you now give it an exact value?

11.8 Chapter Summary

The Main Ideas

- A function F is an **antiderivative** of f if $F' = f$. *Every* antiderivative of f is equal to $F + C$ for some appropriately chosen constant C. We write $\int f(x)\,dx = F(x) + C$.

- Differentiation rules for combinations of functions yield corresponding anti-differentiation rules. Among these are the **constant multiple** and **addition** rules. The chain rule for differentiation corresponds to **integration by substitution**. The product rule for differentiation corresponds to **integration by parts**.

- The derivative of a function and of its inverse are reciprocals. When $y = f(t)$ and $t = g(y)$ are inverses:

$$\frac{dt}{dy} = \frac{1}{dy/dt}.$$

- In some cases, the method of **separation of variables** can be used to find a *formula* for the solution of a differential equation.

- A numerical method for estimating an integral is **efficient** if it gets accurate results with relatively few calculations. The **trapezoidal approximation** is the average of a left and a right *endpoint* Riemann sum and is more efficient than either. *Midpoint* Riemann sums are even more efficient than trapezoidal approximations.

- The most efficient method developed in this chapter is **Simpson's rule**. Simpson's rule approximates an integral by

$$\int_a^b f(x)\,dx \approx \frac{1}{6}\left(\text{left sum} + \text{right sum} + 4 \times \text{midpoint sum}\right).$$

- An **improper integral** is one that cannot be calculated directly. The problem may be that its "domain of integration" is infinite or that the integrand becomes infinite on that domain. Its value is obtained as a limit of ordinary integrals.

Expectations

- You should be able to find antiderivatives of basic functions.
- You should be able to find antiderivatives of combinations of functions using the **constant multiple** and **addition** rules, as well as the **method of substitution** and **integration by parts**.
- You should be able to rewrite an integrand given as a quotient using the method of **partial fractions**.
- You should be able to express the derivative of an invertible function in terms of the derivative of its inverse. In particular, you should be able to differentiate the **arctangent**, **arcsine** and **arccosine** functions.
- You should be able to solve a differential equation using the method of **separation of variables**.
- You should be able to adapt the program RIEMANN to approximate integrals using the **trapezoid rule** and **Simpson's rule**.
- You should be able to find the value of an **improper integral** as the limit of ordinary integrals.

Chapter 12

Case Studies

To enable you to further explore the ways the concepts of calculus are used as analytical tools in scientific and mathematical investigations, this chapter presents four extended case studies. The four can be studied separately, although the first two and the last two are loosely linked

Stirling's Formula As an example of the way many of the ideas—Taylor series, numerical integration, reduction formulas, limits—developed in the earlier chapters of this book can be used in a tightly-reasoned argument to produce some powerful mathematical insights, in the first section we derive a famous formula approximating $n!$. This formula is then applied to the binomial probability distribution.

The Poisson Distribution Chapter 12.2 continues the probability theme by developing the Poisson distribution and using it to study the frequency of radioactive decay events.

The Power Spectrum Chapter 12.3 builds on the study of periodicity begun in chapter 7. We develop the Fourier transform, a basic tool in the sciences for detecting the relative strength of periodic components in a noisy data set.

Fourier Series Chapter 12.4 expands on some of the ideas in chapter 11. Here we develop tools for approximating functions over intervals using sums of sine and cosine terms. This is an extensively used method in a wide range of disciplines, from thermodynamics to music synthesis.

12.1 Stirling's Formula

Factorials in probability

Given a positive integer n, we define $n!$—pronounced n *factorial*—by the rule $n! = 1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n$. This is a convenient concept which occurs in a number of settings, particularly combinatorial and probabilistic ones. For instance, the probability of getting exactly n heads out of $2n$ tosses of a coin turns out to be

$$\frac{(2n)!}{2^{2n}(n!)^2}.$$

$n!$ is difficult to calculate

Unfortunately, evaluating $n!$ for values of n at all large is cumbersome at best. Although though many calculators will compute factorials, few of them can handle numbers as large as $1000!$. Even when we can evaluate $n!$, we are often as interested in the asymptotic behavior of a certain expression as much as in its exact value for specific n. For instance, using methods we develop below, it turns out that the above expression for the probability of n heads in $2n$ tosses is very close to $1/\sqrt{\pi n}$, with the approximation being more accurate the larger n is. In fact, for $n \geq 8$, the approximation is good to two places; for $n \geq 25$, the approximation gives three-place accuracy.

In his book *Methodus differentialis* (1730), the British mathematician James Stirling published the following approximation, now know as **Stirling's formula**, for the factorial operator:

$$n! \sim \sqrt{2\pi}\, n^{n+\frac{1}{2}}\, e^{-n}.$$

While the right-hand side may look much more complicated than the left, think which one you would rather evaluate for, say, $n = 100$. To see how good this approximation is, here are some comparisons:

		Stirling's
n	$n!$	approximation
2	2	1.9190
10	3,628,800	3,598,695.6
50	3.0414×10^{64}	3.0363×10^{64}
100	9.3326×10^{157}	9.3248×10^{157}
1000	4.02387×10^{2567}	4.02354×10^{2567}
10000	$2.84626 \times 10^{35659}$	$2.84624 \times 10^{35659}$

As an example of the way elementary ideas in calculus can be used to derive powerful and subtle results, we will outline a derivation of Stirling's approximation for $n!$. You should write up your own summary of this proof, filling in the gaps in the text below. We will work in two stages. In the first stage, we will show that

$$n! \sim c\, n^{n+\frac{1}{2}}\, e^{-n}.$$

for some constant c. In the second stage we will show that this constant is actually $\sqrt{2\pi}$.

Stage One: Deriving the General Form

We first observe that

$$\ln(n!) = \ln 1 + \ln 2 + \ldots + \ln n.$$

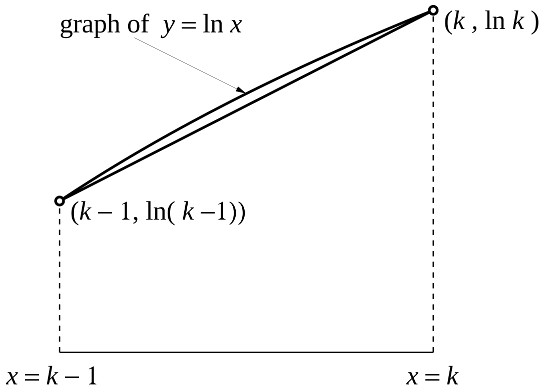
It turns out to be easier to prove things about this logarithmic form. In fact, we will deal most easily with

$$A_n = \ln 1 + \ln 2 + \ldots + \ln(n-1) + \frac{1}{2}\ln n.$$

Thus $\ln(n!) = A_n + \frac{1}{2}\ln n$. Even though $\ln 1 = 0$, it will be useful to retain the term in the expression for A_n.

We will find upper and lower bounds for A_n (and hence for $\ln(n!)$) by approximating the area under the curve $y = \ln x$ by certain inscribed and circumscribed trapezoids. We will then use these bounds to predict the asymptotic behavior of A_n for large values of n.

The upper bound: Note that if we inscribe a trapezoid under the graph of $y = \ln x$ between $x = k-1$ and $x = k$, its area will be $\frac{1}{2}(\ln(k-1) + \ln k)$. (How do we know that the straight line connecting the points $(k-1, \ln(k-1))$ and $(k, \ln k)$ will lie under the graph of $y = \ln x$?) The sum of the areas of all such trapezoids from $x = 1$ to $x = n$ is clearly less than the area under the curve $y = \ln x$ over the interval $[1, n]$.

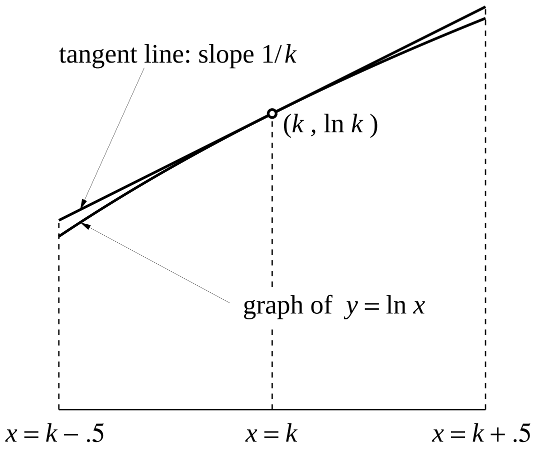


We therefore have the inequality

$$\frac{1}{2}(\ln 1 + \ln 2) + \frac{1}{2}(\ln 2 + \ln 3) + \cdots + \frac{1}{2}(\ln(n-1) + \ln n) < \int_1^n \ln x\, dx,$$

which is equivalent to $A_n < \int_1^n \ln x\, dx$.

The lower bound: On the other hand if we draw the tangent line to $y = \ln x$ at $x = k$ and form the trapezoid between $x = k - .5$ and $x = k + .5$, its area will just be $\ln k$ and will be greater than the area under the curve over the same interval. (We've used the fact—which you should check—that the area of a trapezoid equals the distance between the parallel sides times the distance between the midpoints of the other two sides.)



Adding up all such trapezoids, we get the inequality

$$\int_{\frac{3}{2}}^{n} \ln x\, dx < A_n.$$

Since we know that $\int \ln x\, dx = x \ln x - x$, we can evaluate these upper and lower bounds to conclude

$$n \ln n - n - \frac{3}{2} \ln \frac{3}{2} + \frac{3}{2} < A_n < n \ln n - n + 1,$$

which in turn yields

$$\left(n + \frac{1}{2}\right) \ln n - n + \frac{3}{2}\left(1 - \ln \frac{3}{2}\right) < \ln n! < \left(n + \frac{1}{2}\right) \ln n - n + 1.$$

Pause for a moment to observe that the difference

$$D_n = \left(n + \frac{1}{2}\right) \ln n - n + 1 - \ln n!$$

between the expressions on the two sides of the rightmost inequality is just the accumulated error from approximating the area under $y = \ln x$ by the inscribed trapezoids. Since the error over each interval is always positive, D_n must therefore get larger as n increases, We will need this fact shortly.

Returning to our inequalities, they can finally be rewritten as

$$\frac{3}{2}\left(1 - \ln \frac{3}{2}\right) < \ln n! - \left(n + \frac{1}{2}\right) \ln n + n < 1.$$

Evaluating the constants, we thus have that for any value of n,

$$.8918 < \ln n! - \left(n + \frac{1}{2}\right) \ln n + n < 1.$$

If we exponentiate, this becomes

$$2.395 < \frac{n!}{n^{n+\frac{1}{2}} e^{-n}} < 2.719,$$

or

$$2.395\, n^{n+\frac{1}{2}} e^{-n} < n! < 2.719\, n^{n+\frac{1}{2}} e^{-n}.$$

These estimates are often adequate

Notice that these bounds are already quite strong and would be adequate for many estimates. Moreover, they are true for any value of n. If we are only interested in large values of n, we can do a little better. Let

$$\delta_n = \ln n! - \left(n + \frac{1}{2}\right) \ln n + n.$$

Then $1 - \delta_n = (n + \frac{1}{2}) \ln n - n + 1 - \ln n!$ is just the expression we called D_n a moment ago and said had to be increasing as n increases. But if $D_n = 1 - \delta_n$ is increasing, it must be true that δ_n itself is decreasing as n gets larger. We thus must have $1 > \delta_1 > \delta_2 > \ldots > \delta_n \ldots > .8918$. There must therefore be some constant $d \geq .8918$ such that $\lim_{n\to\infty} \delta_n = d$. Define the constant c by $c = e^d$. Then

$$\lim_{n\to\infty} \frac{n!}{n^{n+\frac{1}{2}} e^{-n}} = c,$$

which is what we mean when we write

$$n! \sim c\, n^{n+\frac{1}{2}}\, e^{-n}.$$

This completes stage 1. In stage 2 we will see that $c = \sqrt{2\pi}$.

Stage Two: Evaluating c

We will do this using an interesting result of a 17th century English mathematician, John Wallis, who showed that

Wallis's formula

$$\lim_{n\to\infty} \frac{2}{1} \times \frac{2}{3} \times \frac{4}{3} \times \frac{4}{5} \times \frac{6}{5} \times \frac{6}{7} \times \cdots \times \frac{2n}{2n-1} \times \frac{2n}{2n+1} = \frac{\pi}{2}.$$

Suppose for the moment that we had proved Wallis's formula. We can express it in terms of factorials by noting that we can rewrite the product of the first n even numbers—$2 \times 4 \times 6 \times \ldots \times (2n)$—by factoring a 2 out of each term, leaving us

$$2 \times 4 \times 6 \times \ldots \times (2n) = 2^n\,(n!).$$

Similarly, we can take the product of the first n odd integers—$1 \times 3 \times 5 \times 7 \times \ldots \times (2n-1)$—and insert the missing even terms to get

$$\begin{aligned} 1 \times 3 \times 5 \times 7 \times \ldots \times (2n-1) &= \frac{1 \times 2 \times 3 \times 4 \times \ldots \times (2n-1) \times (2n)}{2 \times 4 \times 6 \times \ldots \times (2n)} \\ &= \frac{(2n)!}{2^n\,(n!)}. \end{aligned}$$

We can thus rewrite Wallis's formula as

$$\lim_{n\to\infty} \frac{(2^n\,n!)^4}{((2n)!)^2\,(2n+1)} = \frac{\pi}{2}.$$

If we now replace all the factorials by their corresponding expressions using Stirling's approximation, we get

$$\lim_{n\to\infty} \frac{2^{4n}c^4 n^{4n+2}e^{-4n}}{c^2(2n)^{4n+1}e^{-4n}(2n+1)} = \frac{\pi}{2},$$

which, after a great deal of cancelation, reduces to

$$\lim_{n\to\infty} \frac{c^2 n}{2(2n+1)} = \frac{\pi}{2}.$$

Now since

$$\lim_{n\to\infty} \frac{n}{2n+1} = \frac{1}{2},$$

this reduces to

$$\frac{c^2}{4} = \frac{\pi}{2},$$

so

$$c^2 = 2\pi,$$

and

$$c = \sqrt{2\pi},$$

as desired.

Deriving Wallis's formula

One way to derive Wallis's formula involves the integrals

$$I_k = \int_0^{\pi/2} \sin^k x \, dx.$$

Note that $I_0 > I_1 > I_2 > I_3 > \ldots$. Moreover, you should verify that

$$I_0 = \frac{\pi}{2} \qquad \text{and} \qquad I_1 = 1.$$

Using the reduction formula derived in chapter 11.5 for antiderivatives of $\sin^n x$, we have a similar reduction formula for the I_k:

$$\begin{aligned} I_k &= \int_0^{\pi/2} \sin^k x \, dx \\ &= \frac{-1}{k} \sin^{k-1} x \, \cos x \Big|_0^{\pi/2} + \frac{k-1}{k} \int_0^{\pi/2} \sin^{k-2} x \, dx \\ &= \frac{k-1}{k} I_{k-2}. \end{aligned}$$

This in turn leads to

$$I_k = \begin{cases} \dfrac{2n-1}{2n} \cdot \dfrac{2n-3}{2n-2} \cdots \dfrac{1}{2} \cdot \dfrac{\pi}{2} & \text{if } k = 2n \text{ is even,} \\ \dfrac{2n}{2n+1} \cdot \dfrac{2n-2}{2n-1} \cdots \dfrac{2}{3} \cdot 1 & \text{if } k = 2n+1 \text{ is odd.} \end{cases}$$

Further, note that

$$I_{2n+2}/I_{2n} = \frac{2n+1}{2n+2},$$

which has the limit 1 for large n. Since $I_{2n} > I_{2n+1} > I_{2n+2}$, it follows that I_{2n+1}/I_{2n} approaches 1 for large n. But this gives us

$$\begin{aligned} 1 &= \lim_{n\to\infty} I_{2n+1}/I_{2n} \\ &= \lim_{n\to\infty} \left(\frac{2n}{2n+1} \cdot \frac{2n-2}{2n-1} \cdots \frac{2}{3} \cdot 1 \right) \div \left(\frac{2n-1}{2n} \cdot \frac{2n-3}{2n-2} \cdots \frac{1}{2} \cdot \frac{\pi}{2} \right) \\ &= \lim_{n\to\infty} \frac{2n}{2n+1} \cdot \frac{2n}{2n-1} \cdot \frac{2n-2}{2n-1} \cdots \frac{2n-2}{2n-3} \cdots \frac{2}{3} \cdot \frac{2}{1} \cdot \frac{2}{\pi} \end{aligned}$$

If we multiply both sides of this equation by $\pi/2$, we get Wallis's formula.

Further refinements

Some refinements

Using even more careful methods of analysis, it is possible to improve on Stirling's approximation and derive approximations like

$$n! \sim \sqrt{2\pi}\, n^{n+\frac{1}{2}}\, e^{-n+\frac{1}{12n}-\frac{1}{360n^3}+\frac{1}{1260n^5}-\cdots}.$$

If we use this expression to approximate 1000!, for instance, our result is accurate for the first 24 digits.

While this approximation and Stirling's original one are good in the sense that they give more and more accurate digits the larger n gets—so that the *ratio* of $n!$ to either approximation goes to 1 as n gets large—they are bad in the sense that the *difference* between $n!$ and either approximation becomes infinite as n gets large.

The Binomial Distribution

One of the most frequently encountered concepts in probability theory is the **binomial probability distribution**. Suppose we repeat a certain experiment—flipping a penny, rolling a single die, mating a pair of fruit flies, feeding cholesterol to a lab rat—over and over. Suppose further that there is some outcome we are looking for—getting heads, rolling a 2, getting a red-eyed offspring, developing liver cancer in the rat—in each experiment. If p is the probability p of obtaining the looked-for outcome in any one experiment, denote by $P(n,k,p)$ the probability of the outcome happening exactly k times in n experiments. It turns out that

$$P(k,n,p) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}.$$

Example 1 How likely is it to get four 2's if we roll twelve dice? The probability of getting a 2 by throwing one die is $\dfrac{1}{6}$. Therefore the answer to the question is

$$P(12,4,\tfrac{1}{6}) = \frac{12!}{4!\,8!}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^8 = .0888281$$

—we should get exactly four 2's slightly less frequently than once out of every 11 times we roll twelve dice.

Example 2 What is the probability of getting exactly 47 heads if we flip 100 pennies? Since the probability of getting heads on a single toss of a penny is $\frac{1}{2}$,

$$P(100, 47, \tfrac{1}{2}) = \frac{100!}{47!\,53!}\left(\frac{1}{2}\right)^{100} = .0665905,$$

—on the average, if we flip 100 pennies, we should get 47 heads about once out of every 15 times.

The second example demonstrates the fact that calculating binomial probabilities can get very messy very quickly. Several of the exercises are designed to show how Stirling's formula can give us quick estimates that are easy to calculate and work with.

Exercises

1. Go through the derivation in this section and find several passages that seem to you to go a bit fast or skip over details. Rewrite these sections to make them clearer and more complete.

2. Confirm the values given in the table on page 770 for the approximations of 100! and 1000! that Stirling's formula produces.

3. **Rate of growth of n!** Factorials get very large very rapidly. The purpose of this exercise is to develop a sense of just how rapidly n! grows by comparing it to exponential functions.

Let N be some integer > 1, and consider the sequence $a_1, a_2, a_3, \ldots$ defined by

$$a_n = \frac{N^n}{n!}.$$

a) Show that $a_k = \dfrac{N}{k} a_{k-1}$, and conclude that

$$\begin{array}{lll} \text{if } k < N & \text{then} & a_{k-1} < a_k; \\ \text{if } k > N & \text{then} & a_{k-1} > a_k; \\ \text{if } k = N & \text{then} & a_{k-1} = a_k. \end{array}$$

We thus have a sequence that increases for a while:

$$a_1 < a_2 < \cdots a_{N-1} = a_N,$$

and then decreases forever after:

$$a_N > a_{N+1} > a_{N+2} > \cdots .$$

b) If $k > 2N$ show that $a_k < .5a_{k-1}$. Hence conclude that $\lim_{n\to\infty} a_n = 0$.

c) Use Stirling's approximation to show that

$$a_N \approx \frac{e^N}{\sqrt{2\pi N}}.$$

Calculate the values of this expression for $N = 10$ and $N = 100$ to get an idea of how large the sequence $\{a_n\}$ can get. This shows that *for a while*, the exponential series $\{N^n\}$ can get large much more rapidly than the series $\{n!\}$.

d) Show that $a_n < 1$ if $n > eN$. This gives an upper bound on how long it takes the factorials to catch up with the exponentials.

4. If $n \geq 5$, then $n!$ terminates in a certain number of zeroes. For instance, $5! = 120$ ends in one zero, $23! = 25852016738884976640000$ ends in four zeroes, and so on. How many zeroes are there at the end of $1000!$?

The binomial distribution

5. The formula for the binomial distribution gives us that the probability of getting exactly n heads in $2n$ flips of a coin is

$$\frac{(2n)!}{(n!)^2}\left(\frac{1}{2}\right)^{2n}.$$

Show using Stirling's formula this can be approximated by

$$\frac{1}{\sqrt{\pi n}}.$$

Use this approximation to find the probability of getting 50 heads out of 100 tosses of a coin. If you have a computer or calculator which can compute factorials, use the original binomial distribution formula to calculate the exact probability of getting 50 heads and compare the answers.

6. More generally, if we try a certain experiment n times with a probability p of success each time, the most likely number of successes is $k = np$. (Assume

that p is a fraction and n is such that $n \cdot p$ is an integer.) Use Stirling's approximation to show that the probability of getting exactly np successes is

$$P(n, np, p) \approx \frac{1}{\sqrt{2\pi np(1-p)}}.$$

Is this consistent with the answer to the previous exercise?

7. **One-dimensional random walk** An important class of problems, including **diffusion** and **Brownian motion** involve the long-term behavior of particles moving randomly. We will look at the simplest case of such problems. A particle starts at the origin on a line and at each stage moves one unit to the right or one unit to the left, being equally likely to do either. What can we say about where the particle will be after n steps? In this problem we will use Stirling's formula to develop some useful insights into this question.

a) Explain why the particle will be r units to the right of the origin after n steps if and only if it has moved to the right $k = (n + r)/2$ times and to the left $n - k = (n - r)/2$ times. Explain why it could never be 3 units or 7 units to the right after 100 steps.

b) Using the same symbols as in part (a), show that the probability of the particle's being exactly r units to the right after n steps is

$$\frac{n!}{k!(n-k)!}\left(\frac{1}{2}\right)^n.$$

c) Use Stirling's formula to show that this probability of being r units to the right after n steps is approximately

$$\frac{\sqrt{2}}{\sqrt{\pi n}(1+(r/n))^{\frac{1}{2}(n+r+1)}\,(1-(r/n))^{\frac{1}{2}(n-r+1)}}.$$

d) To simplify the denominator of this fraction, recall the Taylor series approximation for $\ln(1 + x)$:

$$\ln(1+x) = x - \frac{x^2}{2} + \cdots .$$

Hence, if r is much smaller than n, $\ln(1 + r/n)$ can be approximated by $r/n - r^2/(2n^2)$, and $\ln(1 - r/n)$ can be approximated by $-r/n - r^2/(2n^2)$.

By ignoring all powers of r greater than the second, conclude that

$$(1+(r/n))^{\frac{1}{2}(n+r+1)}\,(1-(r/n))^{\frac{1}{2}(n-r+1)} \approx e^{r^2/(2n)},$$

so that the probability of being r units to the right after n steps is

$$\sqrt{\frac{2}{\pi n}}\,e^{-r^2/2n}.$$

e) Explain how we can get the answer to exercise 5 as a special case of the result just obtained in part (d).

f) Using the approximation from part (d), calculate the probability that after 100 steps the particle will be *no more* than 5 units away from the starting point to either the right or the left. Remember that after 100 steps it is impossible to be an odd number of units away from the starting point. The exact probability is

$$\sum_{k=48}^{52} \frac{100!}{k!(100-k)!}\left(\frac{1}{2}\right)^{100} = .382701.$$

12.2 The Poisson Distribution

A Linear Model for α-Ray Emission

When a radioactive element decays, we know from the study of differential equations in chapter 4 that the amount $A(t)$ of radioactive material present at time t satisfies the differential equation

$$A' = -kA,$$

where $k > 0$ is the decay constant. If A_0 is the amount present at time $t = 0$, then the solution is

$$A(t) = A_0 e^{-kt}.$$

The time T it takes for a given amount of radioactive material to decay to half the starting quantity is known as the **half life** of the element. Since, by definition, $A(T) = .5\,A(0) = .5\,A_0$, we must have

$$e^{-kT} = \frac{1}{2},$$

which leads to

$$kT = \ln 2$$

and therefore

$$T = \frac{\ln 2}{k}.$$

The relation between the half-life and the decay constant

Suppose, for example, that we have a sample of polonium, which is a radioactive isotope of radium. The decay constant of polonium is $k = .500865$ % per day, and thus its half life is

$$T = \frac{\ln 2}{k} = \frac{\ln 2}{.00500865} = 138.39 \text{ days}.$$

By local linearity, $A(t)$ is closely approximated by a linear function for short intervals of time. Because polonium has a half-life of 138.39 days, a “short time” means several hours in this case. Thus, if we spend an afternoon in a laboratory studying the decay of polonium, we can assume that $A(t)$ is linear.

When polonium decays, it produces various sorts of radiation, including α-rays (“alpha rays”). Using a scintillation counter, one can determine the number of rays emitted in given directions:

A setup like this will count a fixed percentage of the total number of α-rays emitted. Since our model of decay is linear, it follows that the number of α-rays detected should be a linear function of time. If we start counting at time $t = 0$, the number of particles observed will have a straight-line graph:

In the early 20th century, researchers like Marie Curie and Ernest Rutherford did numerous studies of the α-rays emitted by polonium. For example, in 1911, Rutherford, Geiger and Bateman counted the number of α-rays detected in a 7.5-second time period. They repeated their experiment 2608 times and detected a total of 10,097 α-rays. This is an average of

$$\frac{10097}{2608} = 3.8715 \ \alpha\text{-rays per 7.5-second period,}$$

so the number of α-rays per second is

$$\frac{3.8715}{7.5} = .5162 \ \alpha\text{-rays per second.}$$

Thus the straight line in the above graph has slope .5162.

This model of α-ray production has several problems. First, it predicts the existence of fractional α-rays, which makes no sense—the number detected is always a nonnegative integer. To remedy this, we can modify our model as follows:

Notice that the graph is now a step function. It shows that we should see a new α-ray every $1/.5162 = 1.937$ seconds. This model also has the following consequence: if we observe the number of α-rays produced in a 7.5-second interval, then we will always see 3 or 4 particles:

As the picture indicates, whether we get 3 or 4 depends on where the interval starts. Now comes the serious problem: this prediction is *inconsistent* with the experimental data collected by Rutherford and the others in 1911. For example, in 57 of the 2608 times they ran the experiment, *no* α-rays were observed, while in 139 cases, 7 α-rays were observed. Here are the complete data of the experiment:

number of α-rays observed n	number of occurrences N_n
0	57
1	203
2	383
3	525
4	532
5	408
6	273
7	139
8	45
9	27
10	10
11	4
12	0
13	1
14	1
Total	2608

It follows that the linear model of α-ray emission doesn't apply to time intervals of length 7.5 seconds. This is a common occurrence—a model may work nicely over a certain range, but outside of that range, its answers may be meaningless. The problem in our case comes from the random nature of radioactive decay. In fact, there are two sources of randomness to deal with: the *time* when a polonium atom decays is random, and the *direction* in which it then emits an α-ray is also random (this affects us since the scintillation counter only detects emissions in certain directions). We need to modify our model to take the randomness into account, and this is where probability enters in.

Probability Models

Randomness has structure

The basic idea of probability theory is that the outcome of a certain event can be unpredictable in the individual instance but predictable on the average. Throwing dice and tossing a coin are familiar examples. In this section, we will show how the Poisson probability distribution gives an excellent model of the α-ray experiment described above.

The definition of probability

We will let p_n denote the probability of observing exactly n α-rays in a 7.5-second time interval. By this statement, we mean the following. Suppose we run the experiment N times, where N is large. Let N_n be the number of times we observed n α-rays. Then the ratio N_n/N is the frequency with which this outcome occurs. Now imagine N getting larger and larger. Being "predictable on the average" means that the ratios N_n/N approach a fixed number, that is, the limit $\lim_{N\to\infty} N_n/N$ exists. We then *define* this number to be the probability p_n. Thus

$$p_n = \lim_{N\to\infty} \frac{N_n}{N}.$$

For example, the data presented on page 784 were obtained from $N = 2608$ repetitions of our experiment. From the table given there, we see that 0 α-rays were observed 57 times. This means $N_0 = 57$, and thus the probability of detecting 0 α-rays is

$$p_0 \approx \frac{N_0}{N} = \frac{57}{2608} = .0218.$$

Similarly, we can approximate p_1, p_2, etc., using the data in the table. Our goal is to describe these probabilities p_0, p_1, $\ldots$ .. Ideally, we would like to have a way of determining the numbers $p_0, p_1, p_2, \ldots$ "before the fact."

Some properties of probabilities

In any introductory course on probability, one learns certain basic principles for working with probabilities. We will give examples to illustrate some of these principles, and more examples may be found in the exercises.

The general context

For our purposes, we will be working in the following setting. There is a certain **experiment** being performed. This might consist of flipping a coin and noting which side comes up, or running a survey asking people at random their opinions about a certain TV show, or, in our case, counting the number of α-rays detected in a 7.5-second interval. Moreover, there is a **discrete** set of possible outcomes of the experiment. That is, the possible outcomes can be listed in a sequence O_1, O_2, O_3, $\ldots$. In some cases, like throwing a pair of dice, this list might be finite. In other cases, like our α-ray experiment, the list might be infinite. What is ruled out are experiments like choosing a person at random and measuring the person's height—there is a continuum of possible outcomes here which cannot be listed in the way we've specified. Moreover, there should be a **probability** assigned to each outcome, with the outcome O_n having probability p_n.. Finally, the possible outcomes should be **disjoint**—two different outcomes can't both result from a single experiment. Thus if we are examining the attributes of a group of people, "being male" and "having green eyes" would not be acceptable outcomes in our sense unless we somehow knew in advance that there were no green-eyed males in the group.

Knowing the probabilities $p_0, p_1, \ldots$ of the possible outcomes allows us to compute other, possibly more complicated probabilities. This brings in the concept of an **event**, which is basic to probability. In the case of our α-ray experiment, here are some examples of events:

- Detecting 3 α-rays.
- Detecting 2 or 4 α-rays.
- Detecting an odd number of α-rays.

In general, an **event** is a subcollection of the possible outcomes.

The addition rule for probabilities

Rule 1 The probability of an event is simply the sum of the probabilities of its component outcomes.

Thus, for the events just described, we have:

- The probability of detecting 3 α-rays is p_3;
- The probability of detecting 2 or 4 α-rays is $p_2 + p_4$;
- The probability of detecting an odd number of α-rays is the infinite sum

  $$p_1 + p_3 + p_5 + p_7 + \cdots$$

  (since an odd number of α-rays means that 1 or 3 or 5 or 7 etc. have been detected).

Another important property of probabilities follows directly from Rule 1:

Rule 2 The sum of the probabilities of all possible outcomes is 1:

$$\sum_{k=0}^{\infty} p_k = 1.$$

The reason for this is that the list of outcomes was stipulated to be the list of *all possible* outcomes. Hence the event consisting of all these outcomes is bound to occur every time—its probability is 1.

A third rule we will need relates the probabilities of **independent** events. Two events are independent if the occurrence or non- occurrence of one of the events has no impact on the probability of the second event occurring. For instance, suppose we are examining a group of people. Consider the following events which may or may not occur each time we look at a person:

1. The person is female;
2. The person has green eyes;
3. The person is over 5'7" tall.

We would expect the first and second events to be independent, and also the second and third, but not the first and third.

The product rule for probabilities

Rule 3 The probability that two or more independent events all occur is the product of their separate probabilities.

Thus, for example, suppose that in our hypothetical group of people $\frac{1}{2}$ are female, $\frac{1}{8}$ are green-eyed, and $\frac{1}{3}$ are taller than 5'7". We might then expect roughly $\frac{1}{24}$ of them to be green-eyed *and* over 5'7", but we would have no particular reason to expect that $\frac{1}{6}$ of them are females taller than 5'7".

A final rule that is often useful is

The probability that something doesn't happen

Rule 4 If a certain event has a probability p of happening, then the probability that the event doesn't take place is $1 - p$.

For example, in our group of people, we would expect $\frac{2}{3}$ of them to be less than 5'7" tall, $\frac{7}{8}$ of them to have eyes colored something other than green, etc.

The notion of a probability model

A **model** is a mathematical picture of a real-life phenomenon. We have seen that dynamical systems can be used to create models of physical situations. Another type of mathematical model is a *probability model.* In general, a **probability model** for an experiment with a finite number of outcomes is *a listing of all possible outcomes and an assignment of probabilities to each outcome so that their sum is 1.* In order that the probability model be a good picture of reality, we ask that the *probability assigned to an outcome should be the relative frequency with which that outcome would appear if the experiment were duplicated independently a large number of times.*

As an example, a probability model for one toss of a fair die consists of a list of all possible outcomes, namely 1, 2, 3, 4, 5, 6, and an assignment of a probability to each, namely $\frac{1}{6}$, $\frac{1}{6}$, $\frac{1}{6}$, $\frac{1}{6}$, $\frac{1}{6}$, $\frac{1}{6}$, respectively. We assign the number $\frac{1}{6}$ to each outcome because we expect that if the the experiment were repeated (that is, if the die were tossed) a large number of times, then any particular outcome (3, say) would occur about one sixth of the time. Another probability model for the experiment consisting of a toss of a die might be a list of all outcomes, again 1, 2, 3, 4, 5, 6, together with an assignment of the numbers $\frac{1}{2}$, 0, $\frac{1}{6}$, 0, 0, $\frac{1}{3}$ to 1, 2, 3, 4, 5, 6, respectively. This is a probability model, because the numbers we have assigned add to 1, but it certainly does not model very well the throw of a fair die.

We would like to set up a probability model for our experiment with α-rays. The outcomes are 0, 1, 2, 3, 4, ... where, for example, the number 5 labels the outcome in which we observe 5 α-rays in our 7.5-second interval. The total number of outcomes is equal to the number of α-rays that we could conceivably see in a 7.5-second interval. Since it is conceivable (but extremely unlikely) that every atom in the sample could decay and emit an α-ray in the direction of the scintillation counter in one 7.5-second interval, we could conceivably see as many α-rays as there are atoms in the sample. This number is so large that we can think of it as infinite. To have a probability model, we need to assign numbers p_0, p_1, p_2, $\ldots$ to the outcomes 0, 1, 2, $\ldots$, respectively, so that $p_0+p_1+p_2+\cdots=1$. For the model to be reasonable, we would like each p_n to be approximately equal to the corresponding number N_n/N observed by Rutherford, Geiger, and Bateman.

The Poisson Probability Distribution

The Poisson model of α-ray emission

To describe the probabilities p_0, p_1, ..., p_n, ... that we will observe 0, 1, ..., n, ... α-rays in a 7.5-second interval for our α-ray experiment, we use the **Poisson probability distribution**

$$p_n = \frac{\lambda^n e^{-\lambda}}{n!},$$

where λ is a number yet to be determined, and $n!$ is the familiar ***n*-factorial** function,

$$n! = \begin{cases} n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 3 \cdot 2 \cdot 1 & \text{if } n > 0, \\ 1 & \text{if } n = 0. \end{cases}$$

Thus the first few Poisson probabilities are:

$$p_0 = e^{-\lambda}, \qquad p_1 = \lambda e^{-\lambda}, \qquad p_2 = \frac{\lambda^2 e^{-\lambda}}{2}, \qquad p_3 = \frac{\lambda^3 e^{-\lambda}}{6}.$$

Note that this assignment does indeed give us a probability model, because

$$\begin{aligned} p_0 + p_1 + p_2 + p_3 + \cdots &= \frac{\lambda^0}{0!}e^{-\lambda} + \frac{\lambda}{1!}e^{-\lambda} + \frac{\lambda^2}{2!}e^{-\lambda} + \frac{\lambda^3}{3!}e^{-\lambda} + \cdots \\ &= e^{-\lambda}\left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots\right) \\ &= e^{-\lambda} \cdot e^{\lambda} \\ &= 1. \end{aligned}$$

(The transition from the second line to the third uses the fact that the expression in parentheses is just the Taylor series for e^{λ}.)

We will shortly derive the Poisson distribution from basic principles. For the moment, though, we will assume that the probabilities p_0, p_1, ... for α-ray emission are given by the above formulas, where we still need to choose an appropriate value for the parameter λ. The key to determining λ is the notion of **expectation**, which for us will mean the average number of α-rays observed in a 7.5-second interval.

Suppose we repeat our experiment N times. As usual, we let N_n denote the number of times exactly n α-rays were observed. Then the total number of α-rays observed in the N experiments is

$$0 \cdot N_0 + 1 \cdot N_1 + 2 \cdot N_2 + 3 \cdot N_3 + \cdots \ .$$

Then the "average number of α-rays observed in a 7.5-second interval" means the limit

$$E = \lim_{N \to \infty} \frac{0 \cdot N_0 + 1 \cdot N_1 + 2 \cdot N_2 + 3 \cdot N_3 + \cdots}{N}.$$

This limit is called the **expected value** or **expectation** (which explains why it is denoted E).

We claim that for the Poisson distribution, the expected value E is exactly the number λ. To see this, notice that the above limit can be written in the form

$$E = \lim_{N \to \infty} \left(0 \cdot \frac{N_0}{N} + 1 \cdot \frac{N_1}{N} + 2 \cdot \frac{N_2}{N} + 3 \cdot \frac{N_3}{N} + \cdots \right).$$

Since we defined

$$p_n = \lim_{N \to \infty} \frac{N_n}{N} \ ,$$

it follows that we get the following formula for the expectation:

The general formula for the expected value in a probability model

$$E = 0 \cdot p_0 + 1 \cdot p_1 + 2 \cdot p_2 + 3 \cdot p_3 + \cdots = \sum_{n=0}^{\infty} n p_n.$$

(Note that this equality is true for *any* probability model, not just the one we are considering)

Substituting in the values of p_n given by the Poisson distribution, we have

$$\begin{aligned} E &= 0 \cdot e^{-\lambda} + 1 \cdot \lambda e^{-\lambda} + 2 \cdot \frac{\lambda^2}{2!} e^{-\lambda} + 3 \cdot \frac{\lambda^3}{3!} e^{-\lambda} + \cdots \\ &= \sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} e^{-\lambda} \\ &= \lambda e^{-\lambda} \sum_{n=1}^{\infty} \frac{\lambda^{n-1}}{(n-1)!}, \end{aligned}$$

where we pulled the common factor $\lambda e^{-\lambda}$ outside the summation, noted that the term in the summation corresponding to $n = 0$ is 0, and observed that

$$\frac{n}{n!} = \frac{n}{n(n-1) \cdot \cdots \cdot 2 \cdot 1} = \frac{1}{(n-1)!}.$$

Letting $k = n - 1$, we have (again recognizing the Taylor series for e^{λ})

$$E = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

This proves that the expected value is λ as claimed.

Now that we know how to interpret λ, it is easy to determine what it should be for the α-ray experiment. The data given on page 784 covered $N = 2608$ repetitions of the experiment, with

$$0 \cdot N_0 + 1 \cdot N_1 + \cdots = 10097$$

in this case. Thus

$$\frac{0 \cdot N_0 + 1 \cdot N_1 + \cdots}{N} = \frac{10097}{2608} = 3.8715$$

is an approximation of the expected value λ. However, since this is the only information about λ we have, we will let $\lambda = 3.8715$. Using this value of λ, we can then compare the frequencies predicted by the Poisson distribution to the actual data from on page 784:

number of	number of	probability	Poisson	Poisson
α-rays observed	occurrences	approximation	probability	prediction
n	N_n	N_n/N	p_n	$2608\,p_n$
0	57	.021855	.020827	54.3
1	203	.077837	.080632	210.3
2	383	.146855	.156083	407.1
3	525	.201303	.201426	525.3
4	532	.203398	.194955	508.4
5	408	.156441	.150953	393.7
6	273	.104677	.097402	254.0
7	139	.053297	.053870	140.5
8	45	.017254	.026070	68.0
9	27	.010352	.011214	29.2
10	10	.003834	.004341	11.3
11	4	.001533	.001528	4.0
12	0	.000000	.000492	1.3
13	1	.000383	.000146	.4
14	1	.000383	.000040	.1
Totals	2608	1	1	2608

The Poisson model agrees nicely with the data since for each n, N_n/N and p_n are reasonably close. Notice that we shouldn't expect perfect agreement since N_n/N is only an approximation to p_n. We would expect these approximations to get better as we take larger values of N.

Look at the last column, labelled "Poisson prediction." The numbers here are the Poisson probabilities multiplied by $N = 2608$, and they represent the "ideal" number of occurrences. This makes it easier to compare the model to the data. For example, the graph below plots the number of occurrences, both actual and predicted. The circles are the experimental data, while the line-segment graph connects the Poisson predictions.

Although the model seems to fit the data nicely, we should point out that there are statistical tests which can be used measure the fit more precisely. These tests are part of the material covered in courses in probability and statistics.

A final and very important point to make concerns the number of α-rays observed over a long period of time. Our particular Poisson model with $\lambda = 3.8715$ only works for a 7.5-second interval. What happens if we count α-rays over a longer time period? For simplicity, assume that we have a time interval of length T which is a multiple of 7.5 seconds, so that $T = 7.5\,N$ for some large integer N. We can regard this as running our 7.5-second experiment N consecutive times. Thus the ratio

$$\frac{\text{total number of } \alpha\text{-rays observed}}{N}$$

is an approximation to the expected value $\lambda = 3.8715$. It follows that

$$\text{total number of } \alpha\text{-rays observed} \approx 3.8715\, N = \frac{3.8715}{7.5} 7.5\, N = .5162\, T.$$

This shows that for large time intervals, we recover the linear model of α-ray emissions discussed on page 782. Thus our probabilistic model is consistent with what we did earlier and yet allows us to describe what happens when the linear model breaks down.

Derivation of the Poisson model

In the previous discussion we simply assumed that the α-ray probabilities were given by the Poisson distribution, and found that the Poisson probabilities agreed with the experimental data. Let's see where the Poisson formulas come from. It turns out that we can derive the Poisson probabilities p_n from the following assumptions:

- We have an extremely large number M of polonium atoms;
- Each atom has a small but equal probability of emitting an α-ray that is detected by our scintillation counter in a 7.5-second period;
- Observing an α-ray from a given atom is independent of observing an α-ray from any other atom.

Now suppose that we see an average of $\lambda = 3.8715$ α-rays in a 7.5-second period. Because the number of atoms M is large (in the Rutherford-Geiger-Bateman experiment $M > 10^{18}$), then the probability that a single fixed atom emits an α-ray detected by our scintillation counter in a given time period is very close to λ/M. The probability that the single atom does not emit a detected α-ray in the period is then $1 - \lambda/M$ (by Rule 4, page 787). Thus, the probability p_0 that none of the M atoms emits an α-ray in the 7.5-second period is $(1 - \lambda/M)^M$ (by Rule 3, page 787).

The fact that M is so large allows us to make a simplifying approximation. Recall that for any value of x, positive or negative,

$$e^x = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n.$$

Therefore

$$p_0 = \left(1 - \frac{\lambda}{M}\right)^M \approx e^{-\lambda}.$$

You might calculate some sample values for various values of M to see how good this approximation is.

To derive the values of p_k for $k > 1$, we need a slight improvement on the estimate we just made. Let k be a relatively small (compared to the size of M) number. Then

$$\left(1-\frac{\lambda}{M}\right)^{M-k} = \left(\left(1-\frac{\lambda}{M}\right)^{M}\right)^{\frac{M-k}{M}}.$$

$(M-k)/M$ is essentially equal to 1

Now if M is very large compared to k—we will be thinking of values of M on the order of magnitude of 10^{18} and $k < 100$—then $(M-k)/M$ will essentially equal 1, Hence

$$\left(1-\frac{\lambda}{M}\right)^{M-k} \approx \left(e^{-\lambda}\right)^{1} = e^{-\lambda}.$$

We can now work out p_1. Fix your attention first on a particular atom. The probability that *that* atom does emit an α-ray detected by the scintillation counter while the other $M-1$ atoms do not is (again by Rule 3)

$$\left(\frac{\lambda}{M}\right)^{1}\left(1-\frac{\lambda}{M}\right)^{M-1} \approx \frac{\lambda}{M}\,e^{-\lambda},$$

by the preceding approximation.

Since there are altogether M atoms which might have been responsible for the single α-ray emission, the probability that some *unspecified* atom emits an α-ray while the others do not is (Rule 1, page 786) the sum of the probability we just calculated for each of the M atoms, which is equal to M times that probability. The total probability is p_1:

$$p_1 \approx M\,\frac{\lambda}{M}\,e^{-\lambda} = \lambda\,e^{-\lambda}.$$

To work out p_2, note that the probability that each atom of some fixed pair of atoms emits an α-ray detected by the counter, and no other atoms does, is

$$\left(\frac{\lambda}{M}\right)\left(\frac{\lambda}{M}\right)\left(1-\frac{\lambda}{M}\right)^{M-2} \approx \frac{\lambda^2}{M^2}\,e^{-\lambda},$$

using our usual approximation. Since there are $\frac{1}{2}M(M-1)$ different pairs of atoms (we can choose the first M different ways and the second $(M-1)$

ways, but each pair gets counted twice in this scheme, so we have to divide by 2), we obtain

$$p_2 \approx \frac{M(M-1)}{2}\frac{\lambda^2}{M^2}e^{-\lambda} = \left(1-\frac{1}{M}\right)\frac{\lambda^2}{2}e^{-\lambda} \approx \frac{\lambda^2}{2}e^{-\lambda}.$$

As one can easily imagine, the computations for p_3, p_4, ... are similar. The observant reader will note that the exact values we got for p_0, p_1 and p_2 are not the values given by the Poisson distribution. We got the Poisson probabilities only by making various approximations that were justified by the large value of M. The assumptions we have made actually lead to what is called the **binomial distribution** (see chapter 12.1), a distribution which tends to the Poisson distribution in the limit $M \to \infty$. In this case, where M is large and λ relatively small, the binomial distribution is extremely close to the Poisson distribution.

Other applications of the Poisson distribution

The Poisson distribution can be used to model many other situations that have a random element. Examples include:

- The number of chromosome interchanges caused by exposure to X-rays for a fixed interval of time.
- The number of bacteria in a given unit of area on a Petri dish.
- The number of misprints on a page in a book.
- The number of flying-bomb hits per unit area in London during World War II.

In the exercises we will explore some examples.

Exercises

Probability models

1. A fair coin is tossed. If it comes up H (heads), a fair die is rolled. If the coin comes up T, the coin is tossed again. Construct a probability model for this experiment, listing the possible outcomes and their probabilities. (Hint: the list of outcomes is $H,1$, $H,2$, ..., $H,6$, TT, TH.)

2. Two identical fair coins are put in cup, shaken, and spilled out onto a table. Construct a probability model for this experiment.

3. a) In the disintegration of large numbers of particles of radium (Ra), it is noted that 29% of the disintegrations result in

$$Ra \longrightarrow P + A$$

and the remainder in

$$Ra \longrightarrow He^+ + B.$$

What is a model for the disintegration of a single particle of Ra?

b) Construct a probability model for the disintegration of two particles of Ra.

The Poisson distribution

4. The purpose of this exercise is to present another way to show that the expected value E of the Poisson distribution is equal to λ. As in the text we have

$$E = 0 \cdot p_0 + 1 \cdot p_1 + 2 \cdot p_2 + 3 \cdot p_3 + \cdots = \sum_{n=0}^{\infty} np_n.$$

The numbers np_n can be simplified as follows:

$$\begin{aligned}
0 \cdot p_0 &= 0, \\
1 \cdot p_1 &= 1 \cdot \lambda e^{-\lambda} = \lambda \cdot e^{-\lambda} = \lambda p_0, \\
2 \cdot p_2 &= 2 \cdot \frac{\lambda^2 e^{-\lambda}}{2} = \lambda \cdot \lambda e^{-\lambda} = \lambda p_1, \\
3 \cdot p_3 &= 3 \cdot \frac{\lambda^3 e^{-\lambda}}{6} = \lambda \cdot \frac{\lambda^2 e^{-\lambda}}{2} = \lambda p_2.
\end{aligned}$$

a) This pattern generalizes: show that

$$np_n = \lambda p_{n-1} \quad \text{for all } n > 0 \ .$$

b) Use part (a) to compute the expectation E (you will need to use the fact that the sum of the probabilities is $p_0 + p_1 + p_2 + \cdots = 1$).

5. A model is to be constructed for the number of rain drops that fall per square foot over a short time interval. Under what conditions would a Poisson distribution be appropriate. Under what conditions would a linear model be better?

6. In analyzing flying-bomb hits in the south of London during World War II, investigators partitioned the are into 576 small sectors, each being $\frac{1}{4}$ of a square kilometer. There were 229 sectors with no hits, 211 sectors with exactly 1 hit, 93 sectors with exactly 2 hits, 35 sectors with 3 hits, 7 sectors with 4 hits, and one sector with 5 or more hits. What might lead you to expect that a Poisson distribution might be a good model for the number of hits on each sector? Fit a Poisson distribution to the data by taking λ to be the average number of hits per sector. Use this λ to compute the theoretical frequencies of 0, 1, 2, 3, 4 and 5 hits in 576 sectors.

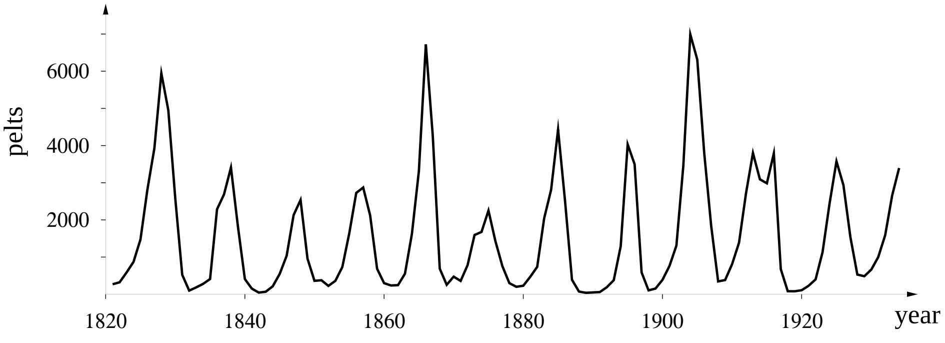
7. A meteorite shower sprinkles a large area of the earth's surface with small meteorite hits. The average density is 5×10^{-6} hits per square meter. Set up a model assigning a probability to the number of hits per square kilometer.

8. The central processing unit (CPU) of a laptop computer will freeze if more than ten instructions are received in a millisecond. If the average number of instructions per second received in the course of executing a large program is one per millisecond, what is the probability that the instructions received by the CPU will cause it to freeze (and, hence, the program to crash).

12.3 The Power Spectrum

The problem of signal + noise

This section is an application of ideas about periodic functions and integrals to the problem of separating a signal from noise. We face this problem in our daily life. Radio and television signals have noise added to them from other radio sources we can't control. The noise sounds like hissing static on a radio and looks like "snow" on a television screen. A good receiver is designed to filter out the noise while allowing the the transmitted signal to come through undistorted.



Annual harvest of lynx pelts

Scientific data and a radio broadcast have something in common: both are combinations of signal and noise. For instance, consider the annual harvest of lynx pelts by the Hudson's Bay Company. It is conceivable that the lynx population itself (the *signal*) was periodic, but various random fluctuations (the *noise*) caused the harvest (which is *signal* + *noise*) to take the form it did. If this is the case, then we should try to "filter out" the noise and find the underlying periodic signal. There is a mathematical tool to do this; it is called the **power spectrum**. We will discuss the ideas behind the power spectrum and show how it can be used to detect the underlying in noisy data.

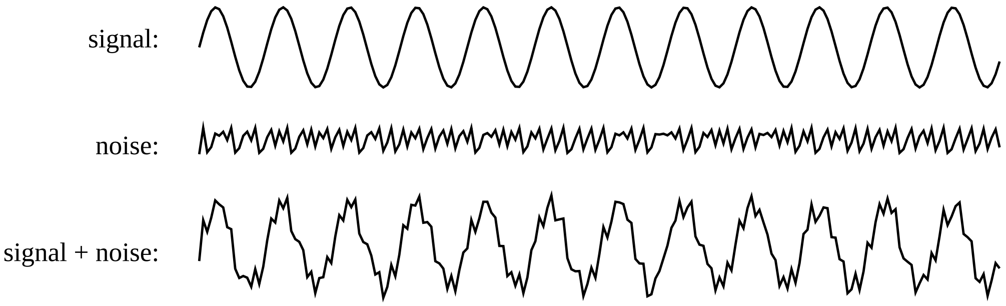
The power spectrum filters noise to detect periodic signals

Signal + Noise

To prepare for working with the power spectrum, let's first see what happens to a periodic signal that has some noise added to it. The signal we will use is a pure sine wave. The **information** that the signal carries is the frequency of that wave. The noise will also be a function, but one whose values vary in a random fashion. It can be thought of as a combination of periodic signals of all frequencies. For this reason it is sometimes called "white noise," because white light is a combination of light rays of all colors (i.e., frequencies). Here is the question we will explore: If we increase the strength of the noise, when do we lose the information contained in the original signal?

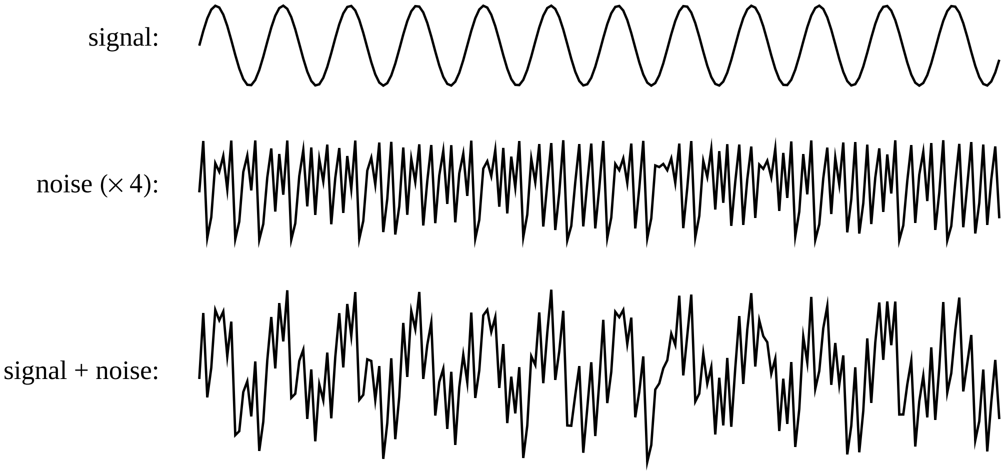
A signal with faint noise

The signal and noise are shown below. As you can see, the amplitude of the signal is about 4 times as large as the amplitude of the noise. We say that the **signal-to-noise ratio** is 4:1. The combined signal + noise is no longer a pure sine wave, of course. However, it is still recognizable as a "noisy" wave with the same frequency as the original signal. The information from the signal has not yet been lost.



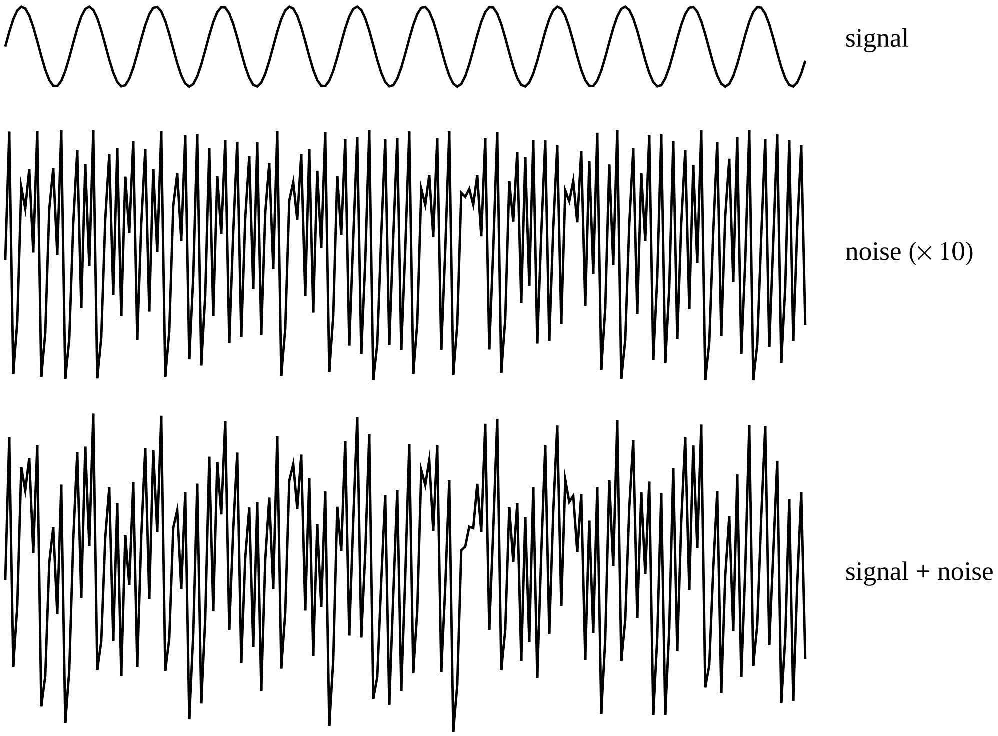
The noise level becomes stronger

Look what happens when we increase the amplitude of the noise. In the figure below, the noise has been increased by a factor of 4, so the signal-to-noise ratio is now 1:1. The combined signal + noise is now very noisy. Would you be willing to argue that it is a wave of the same frequency as the original signal? Or would you prefer to say that it has no periodic pattern whatsoever? It appears we are close to losing the information from the original signal.



The noise level becomes overwhelming

If we increase the original noise level by a factor of 10, we appear to lose the original signal altogether. The signal-to-noise ratio is now 1:2.5, and the signal + noise appears to be as random as the noise itself. In spite of appearances, the signal is still there, and it will be detected in the power spectrum!



Detecting the Frequency of a Signal

Compare the signal to a probe whose frequency can be varied

Assume we have a signal that may be distorted by a lot of noise. We want to decide whether the signal has a periodic component; if it does, we want to determine its frequency. Our detector is based on this simple idea: *Compare the signal to a test probe of known frequency; vary the frequency of the probe until there is a positive response.* Of course, we still need to explain how the comparison is made, and what constitutes a positive response.

Although the detector will work on a very noisy signal, like the one above, we will understand it better if we first use it to analyze a signal whose periodic nature is evident. Let the signal $S(t)$ be a pure sine wave lying above the t-axis, and suppose that t is the time measured in seconds. Our **test probe**

The test probe

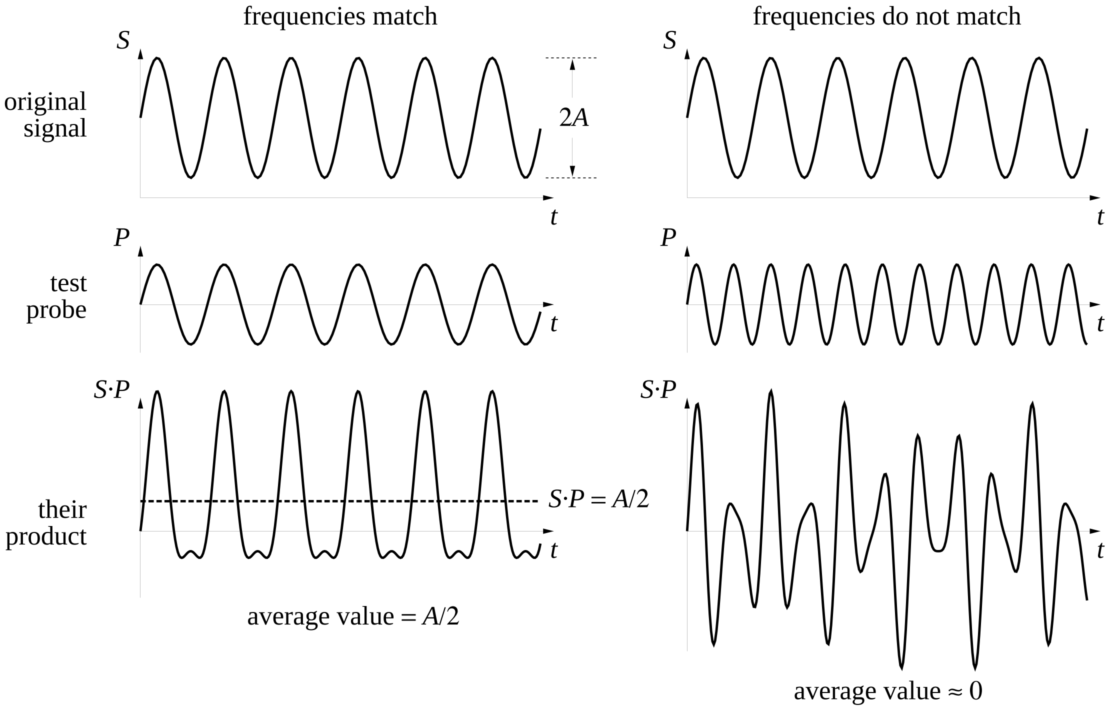
is the function

$$P(t) = \sin(2\pi\omega t)$$

whose frequency is ω cycles per second. As its graph demonstrates, the values of P are equally likely to be positive or negative.

When the signal matches the test frequency

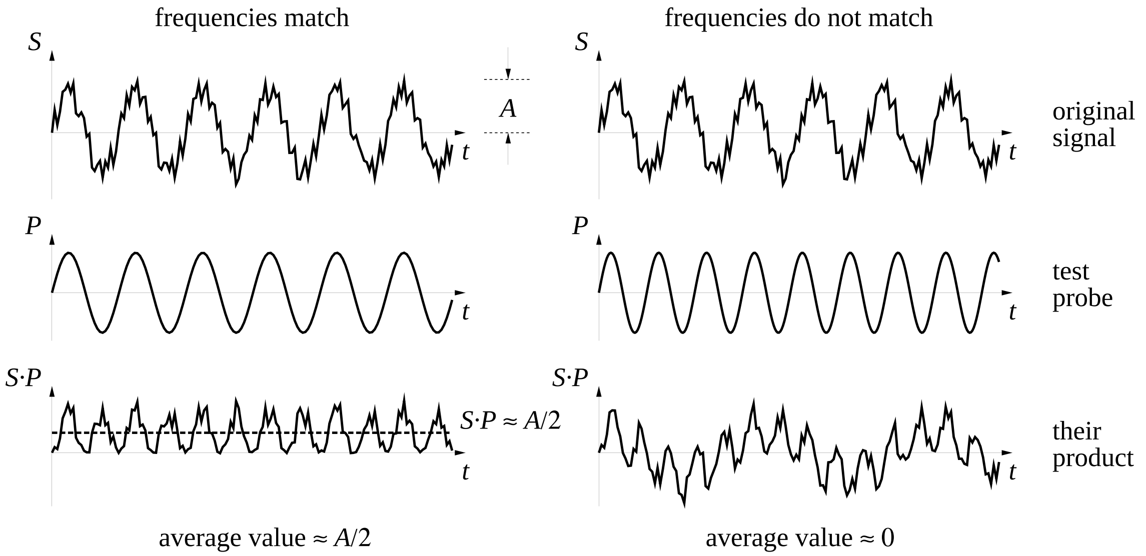
Is the same true for the product $P(t)S(t)$? Suppose first that $S(t)$ has the same frequency as $P(t)$ (below, left). As you can see, the positive values of $P(t)$ are always multiplied by the larger values of $S(t)$. By contrast, the negative values of $P(t)$ are always multiplied by the smaller values of $S(t)$. Consequently, the positive values of $P(t)S(t)$ outweigh the negative ones. On average, the value of the product is positive. In fact, the average value of the product is half the amplitude of the original signal. Later on we will see why this is so.



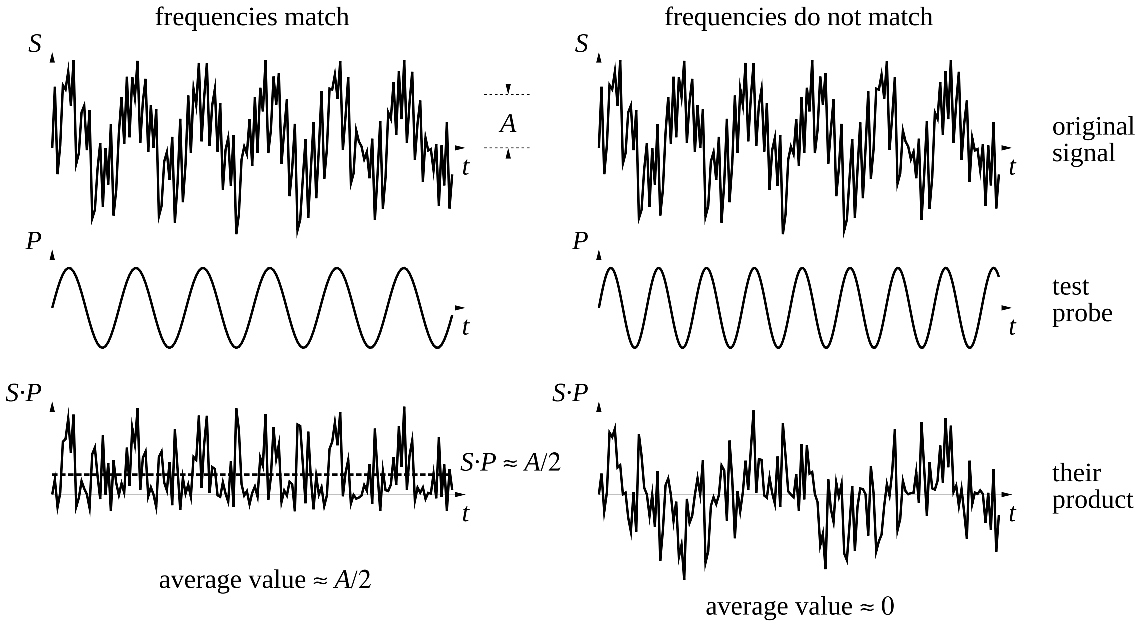
When the signal *doesn't* match the test frequency

On the right we see what happens if $S(t)$ is *not* related to $P(t)$. In that case, a large value of $S(t)$ is just as likely to multiply a positive value of $P(t)$ as a negative one. Consequently, the product $P(t)S(t)$ will have both large positive and large negative values. On average, the value of the product will be about 0.

Let's use the detector on the signals we constructed on page 799. In both we started with a pure sine wave and added some white noise. In the first, the signal-to-noise ratio was 4:1.



In the second the noise was stronger; the signal-to-noise ratio was 1:1.



To use the detector yourself, you have to be able to calculate the average value of a function. This is discussed in chapter 6.3. The average value of $y = f(x)$ on the interval $a \le x \le b$ is

The average value of a function

$$\frac{1}{b-a}\int_a^b f(x)\,dx.$$

Our detector is the average value of the product of the signal $S(t)$ and the test probe $P(t) = \sin(2\pi\omega t)$.

$$\boxed{\textbf{Frequency detector}:\ D(\omega) = \frac{1}{b-a}\int_a^b S(t)\sin(2\pi\omega t)\,dt.}$$

Clearly, the value of the detector depends on the frequency ω of the probe P. We have tried to reflect this in the notation: the detector is a function D whose input is the frequency ω. The output of the function is calculated as an integral in which the input ω plays the role of a parameter.

Integrals with parameters define functions

This is the first time we have defined a function as an integral with a parameter. Let's see how the detector works to analyze the signal $S(t) = 3\sin(5t)$ over the interval $0 \le t \le 10$. We have

$$D(\omega) = \frac{1}{10}\int_0^{10} 3\sin(5t)\sin(2\pi\omega t)\,dt.$$

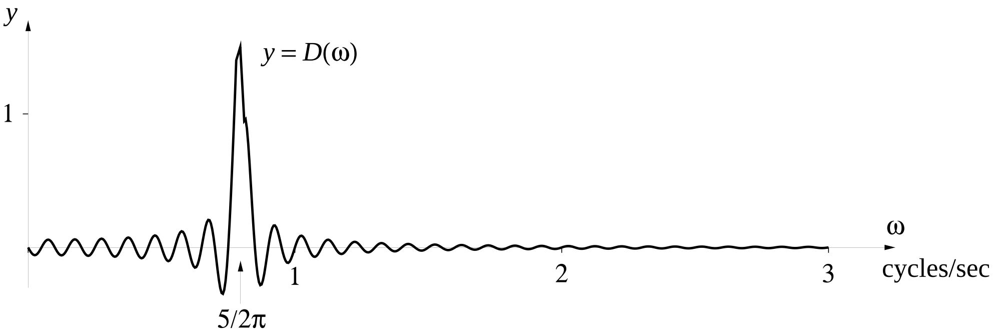
In the exercises at the end of chapter 11.3 we obtained an explicit formula for the integral of the product of two sine functions. Find that formula and check that it yields the following:

$$D(\omega) = \frac{3}{10(4\pi^2\omega^2 - 25)}(5\cos(50)\sin(20\pi\omega) - 2\pi\omega\sin(50)\cos(20\pi\omega)).$$

Notice, in your own calculations, that ω emerges as the variable on which the whole expression depends.

The graph of $D(\omega)$ is shown on the top of the next page. You should plot it yourself, using a computer graphing utility. For most frequencies ω, the value of the detector D is close to 0. There is a single strong peak, which you can find at $\omega \approx .795$ cycles/sec. As it happens, the frequency of the signal $S = 3\sin(5t)$ is $5/2\pi = .79577\ldots$ cycles/sec! Moreover, the height of the peak is about 1.5, which is exactly half the amplitude of the signal.

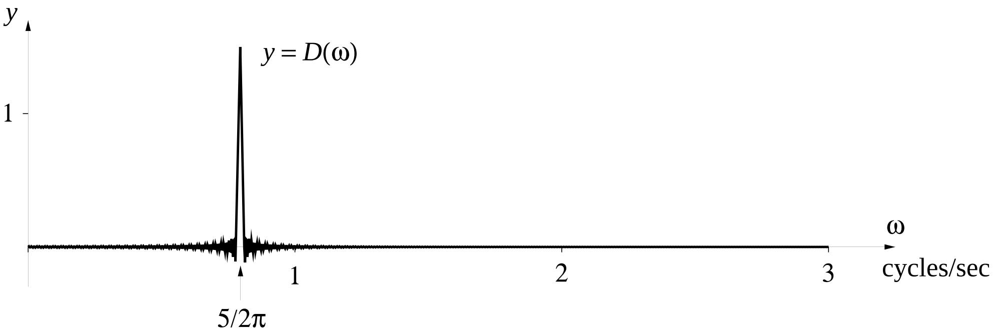
$D(\omega)$ peaks when ω is the frequency of the signal



Detecting the frequency of 3sin(5t) on the interval $0 \le t \le 10$

The graph above tests the signal $S(t)$ when the detector is integrated over a time interval that is 10 seconds long. That is, $0 \le t \le 10$ seconds. If we repeat the test by integrating over a much larger interval, the frequency detector gives us a sharper report on the frequency of the signal. In the graph below the function $D(\omega)$ was calculated by integrating over the interval $0 \le t \le 100$ seconds.

The peak in $D(\omega)$ is sharper if the signal is tested over a longer time interval



Detecting the frequency of 3sin(5t) on the interval $0 \le t \le 100$

Computation. Of course, it is rare to find a formula for $D(\omega)$ in terms of the frequency ω. For most signals $S(t)$, the best we can do is calculate the value of the integral numerically for a sequence of values of the parameter ω. The program DETECTOR, which is listed on the next page, does this. As it is written, it analyzes the function $3\sin(5t)$ on the interval $0 \le t \le 10$, and it produces the graph $D(\omega)$ at the top of this page. The "outer loop"

The program DETECTOR

```
FOR j = 1 TO omegasteps  ...  NEXT j
```

plots $D(\omega)$ over the interval $0 \le \omega \le 3$, using 2^{10} equally spaced values of ω. Each $D(\omega)$ is an integral whose value is first calculated as a midpoint

Riemann sum with 2^7 steps. The calculation is carried out by the short "inner loop"

```
FOR k = 1 TO numberofsteps   ...   NEXT k,
```

which you should recognize as an adaptation of the program RIEMANN from chapter 6.

Program: DETECTOR
To detect the frequency of a signal

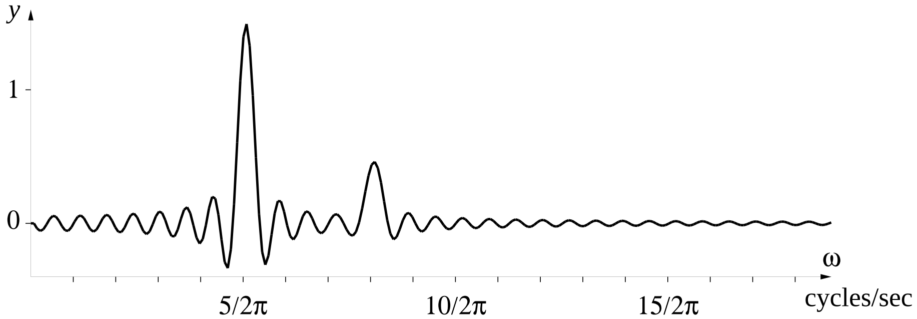
```
Set up GRAPHICS
startomega = 0
endomega = 3
omegasteps = 2 ^ 10
deltaomega = (endomega - startomega) / omegasteps
twopi = 8 * ATN(1)
DEF fnf (t) = 3 * SIN(5 * t)
a = 0
b = 10
numberofsteps = 2 ^ 7
deltat = (b - a) / numberofsteps
omega = startomega
oldomega = omega
oldaccum = 0
FOR j = 1 TO omegasteps
     t = a + deltat / 2
     accum = 0
     FOR k = 1 TO numberofsteps
          deltaS = (fnf(t) * SIN(twopi * omega * t) * deltat) / (b - a)
          accum = accum + deltaS
          t = t + deltat
     NEXT k
     omega = omega + deltaomega
     Plot the line from (oldomega, oldaccum) to (omega, accum)
     oldomega = omega
     oldaccum = accum
NEXT j
```

If we modify the program DETECTOR so that it analyzes the function

$$S(t) = 3\sin(5t) + \sin(8t),$$

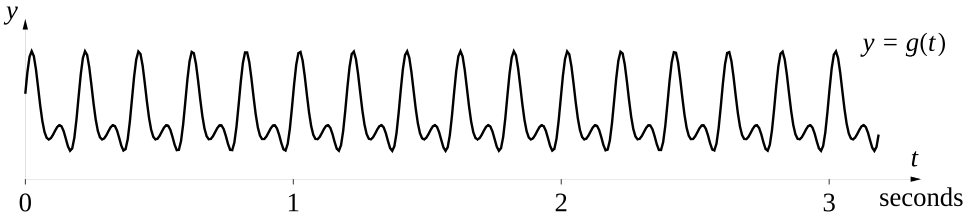
we get the graph at the top of the next page. The scale on the ω-axis has also been modified to make it easier to read multiples of $1/2\pi$ cycles per second.

Notice the strongest peak is at $\omega = 5/2\pi$ cycles/sec, and $D \approx 1.5$ there. But there is now a second peak at $\omega = 8/2\pi$ cycles/sec, where $D \approx .5$. Indeed, S consists of two periodic components, one with three times the amplitude of the other. The stronger component has frequency $5/2\pi$ cycles/sec, the weaker $8/2\pi$ cycles/sec.

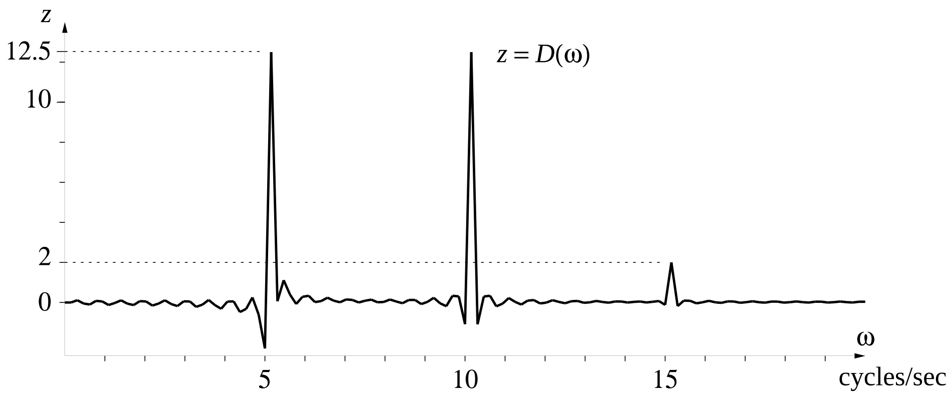


The following example first appeared in chapter 7.2. It is clear from the graph that it has a basic frequency of 5 Hz. The detector shows that it also has an equally strong component at 10 Hz and a much weaker component at 15 Hz. Can you guess a formula for $g(t)$?

A periodic signal . . .



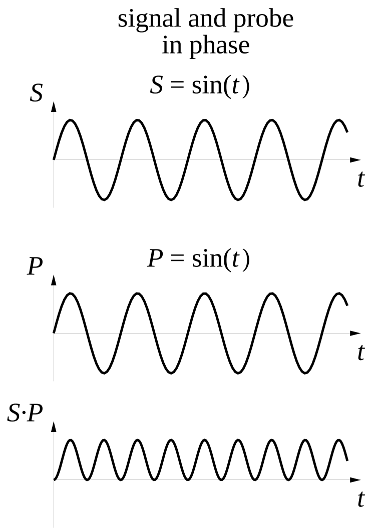
. . . and its frequency detector



The graph of $z = D(\omega)$ was produced by DETECTOR. The integral was calculated for `a = 0`, `b = 10`, and `numberofsteps = 2 ^ 9`.

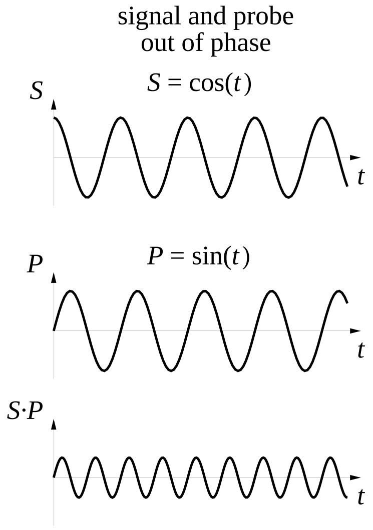
The Problem of Phase

Our detector is built on the premise that, if you take the product of two functions of the same frequency, its average value will be different from 0. This is illustrated by the top three graphs on the left. The signal and the probe are both $\sin(t)$. Their product is a function that ranges between 0 and 1, and has average value 1/2. However, something quite different happens if we change the signal from $\sin(t)$ to $\cos(t)$. This doesn't change the period, but it does change the product, as you can see in the three lower graphs. The new product is centered around the t-axis; its average value is 0. Thus the detector fails to reveal that the signal has the same frequency as the probe.



A closer look at the two sets of graphs will show what has happened. In the first case, when P is positive, so is S. When P is negative, so is S. Thus, the product $S \cdot P$ is never negative; on average, its value is positive. This is what we expect.

The second case is only a little more complicated. When P is positive, S is positive only half the time; the other half it is negative. Consequently, the product $S \cdot P$ takes both positive and negative values. The same thing happens when P is negative. On average, the value of the product is 0, *even though the frequencies of P and S match.*



The problem is that their *phases* don't match. The signal $S = \cos(t)$ hits its peak $\pi/2$ seconds before the probe $P = \sin(t)$. This kind of a difference is called a **phase shift**. In the exercises for chapter 7.2, you showed that if the phase of the sine function is shifted to the left by $\pi/2$, the result is the cosine function:

$$S = \sin(t + \pi/2) = \cos(t).$$

Since $\pi/2$ radians is the same as 90°, we sometimes express this equation by saying that "the sine and the cosine are 90° out of phase."

Arbitrary phase shifts

Of course the signal could involve a phase shift of any amount φ: $S = \sin(t - \varphi)$. All these signals have the same period as the probe $P = \sin(t)$. Exercise 20 of chapter 7.2 shows what happens if this signal is tested against the probe: the average value of the product $S \cdot P$ is $\cos(\varphi)/2$. Clearly, The average value varies with the phase

this depends on the size of the phase shift φ. In particular, if $\varphi = 0$ (so $S = \sin(t)$), the average value is $1/2$. If $\varphi = -\pi/2$ (so $S = \cos(t)$), the average value is 0. The formula therefore agrees with what we already know for the two signals we considered as examples.

There is one more case worth glancing at: $\varphi = \pm\pi$. This is also called a phase shift of 180°. It doesn't matter whether you go forward 180° or backward; in either case $S = \sin(t \pm \pi) = -\sin(t)$. This time the average value of the product is $-1/2$.

The problem of phase . . .

The problem of phase is now be clear: The probes $P = \sin(2\pi\omega t)$ have trouble detecting the frequency of a signal that is out of phase with them. However, any phase-shifted sine function can be expressed as a sum of pure sine and cosine functions:

$$\sin(bt - \varphi) = M\sin(bt) + N\cos(bt),$$

. . . and its solution

where $M = \cos(\varphi)$ and $N = -\sin(\varphi)$. (See the exercises.) Since the sine probes P will detect $M\sin(bt)$, we need only construct a second set of probes to detect $N\cos(bt)$. The test probes we add are the cosine functions

$$P_c = \cos(2\pi\omega t).$$

We use the subscript "c" to distinguish these from the sine probes, which henceforth will be denoted P_s.

Two new detectors

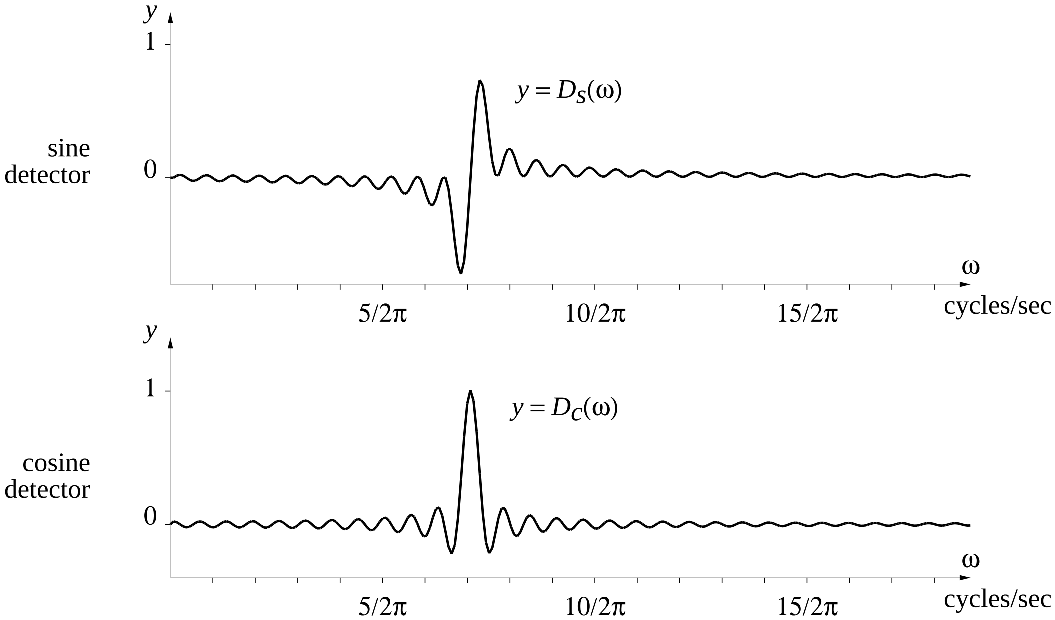
We must also construct a second detector, to handle the new cosine probes. Let's take this opportunity to make a technical adjustment: we redefine a detector to be *twice* the average value of the signal and the probe. In that way, the height of the detector at a peak equals the amplitude of the signal at that frequency—rather than half the amplitude.

Sine detector: $$D_s(\omega) = \frac{2}{b-a}\int_a^b S(t)\sin(2\pi\omega t)\,dt.$$

Cosine detector: $$D_c(\omega) = \frac{2}{b-a}\textstyle\int_a^b S(t)\cos(2\pi\omega t)\,dt.$$

The graphs of D_s and D_c

You can modify the program DETECTOR to produce the graphs of $D_s(\omega)$ and $D_c(\omega)$. You can see below how they analyze the signal $S = \cos(7t)$ over the interval $0 \le t \le 10$. The cosine detector D_c has a shape we've seen



before. It has a single peak at $\omega = 7/2\pi$ cycles/sec, which is the frequency of the signal. The peak is 1 unit high, which is the amplitude of the signal. The sine detector has an unfamiliar shape. Notice first that $D_s(7/2\pi) = 0$. This confirms our earlier observation that the average value of the product of a sine and a cosine at the same frequency is 0. For values of ω slightly larger or smaller than $7/2\pi$, though, the sine detector swings relatively far from 0. This pattern is typical when a detector is analyzing a signal that is 90° out of phase with the probes.

Resonance. Try this experiment. Sit at a piano and hold all the pedals down. Then sing a note. If you sing loud enough, and hold the note long enough, one of the piano strings will start vibrating. If you stop abruptly and listen to the string, you will hear it sounding the same note you were singing. The piano has detected the frequency of your signal! It is the physical analogue of our mathematical frequency detectors. The response of the string is called **resonance**. Had you sung a lower note, a larger string would have resonated.

The physical analogue of a detector is a resonator

Resonance gives us a vivid language for describing how our detectors work. We can say a test probe "resonates" with a signal when their product is different from zero on average. The larger the average value, the stronger the resonance.

Resonance occurs all around us. Sometimes it is a nuisance—for instance, when the windows in our house rattle while a heavy truck drives by, or an air conditioner runs. Sometimes we exploit it deliberately—for instance, when we use a radio tuner as an electronic resonator to detect and amplify certain electromagnetic waves.

Detector as transform

We now have two distinct ways to describe a signal S. The function $S(t)$ is one way. It tells us how strong the signal is at each instant t. But we can also think of the signal as a mixture of sine and cosine waves of different frequencies. The detectors $D_s(\omega)$ and $D_c(\omega)$ tell us how strong the signal is at each frequency ω. That is the second way.

There is a direct connection between these two descriptions, of course. It is provided by the formulas

$$D_s(\omega) = \frac{2}{b-a}\int_a^b S(t)\sin(2\pi\omega t)\,dt \qquad D_c(\omega) = \frac{2}{b-a}\int_a^b S(t)\cos(2\pi\omega t)\,dt.$$

Integrals transform S into D_s and D_c

In effect, these formulas tell us how to *transform* the first description $S(t)$ into the second $D_s(\omega)$, $D_c(\omega)$. The transformation is so complete that even the input variable is changed—from t to ω. Look back at the formulas to see how the new variable ω is brought in.

Our detectors are essentially the same as the **Fourier sine transform** and the **Fourier cosine transform**. There is also an **inverse Fourier transform** that works in reverse: it produces $S(t)$ from the frequency data $D_s(\omega)$ and $D_c(\omega)$. The Fourier transforms are an important tool in mathematics and in science. For example, a hologram is the Fourier transform of an ordinary image. Fourier transforms and their inverses are used in photo restoration, in the enhancement of the digitized pictures sent back from cameras in space, and in filtering the signal in a stereo set.

The French mathematician Jean Baptiste Fourier (1768–1830) introduced what we call Fourier transforms and Fourier series to study the conduction of heat. Now his methods are used to study all sorts of periodic and non-periodic phenomena. They are also the foundation for the part of pure mathematics called harmonic analysis.

The Power Spectrum

A detector that ignores phase differences

The sine and cosine detectors provide enough information to reconstruct the original signal in complete detail—including phase. Often, though, they provide more detail than we want. We can use another tool—called the **power**

spectrum—to determine only the strength of the different frequencies that occur in a signal, without regard to their phase. The power spectrum is constructed from the two detectors in the following way:

Power spectrum: $P(\omega) = \sqrt{[D_s(\omega)]^2 + [D_c(\omega)]^2}$

To see how the power spectrum works, we'll consider the signal $S(t) = A\sin(7t - \varphi)$. This is a sine wave of frequency $\omega = 7/2\pi$ and amplitude A. Let's concentrate first on $\omega = 7/2\pi$. If there were no phase shift φ present, we would expect that

$$D_s(7/2\pi) = A \qquad D_c(7/2\pi) = 0.$$

However, because there is a phase shift, the actual values turn out to be

$$D_s(7/2\pi) = A\cos\varphi \qquad D_c(7/2\pi) = -A\sin\varphi.$$

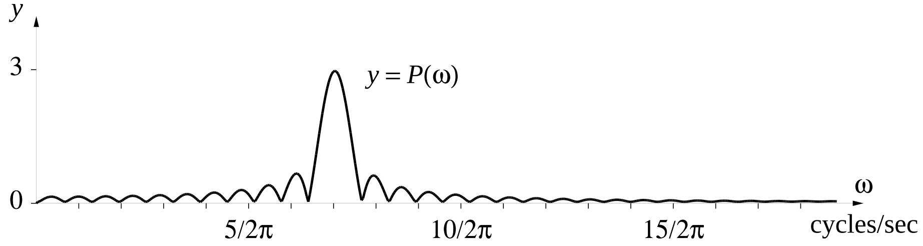
(These calculations are given as exercises.) The values of the detectors clearly depend on the phase shift. By contrast,

$$\begin{aligned} P(7/2\pi) &= \sqrt{[D_s(7/2\pi)]^2 + [D_c(7/2\pi)]^2} \\ &= \sqrt{A^2\cos^2\varphi + A^2\sin^2\varphi} \\ &= A. \end{aligned}$$

We have used the fact that $\cos^2\varphi + \sin^2\varphi = 1$ for every φ. Thus, the power spectrum does *not* depend on the phase. It tells us only the amplitude of the signal at the frequency $\omega = 7/2\pi$.

If we calculate the power spectrum over all frequencies ω, we get the graph shown at the top of the next page. The program POWER generates this graph. It was derived from the program DETECTOR. Compare the two programs, particularly the terms `deltaS` and `deltaC`. In POWER, they have been multiplied by 2, to agree with our new definition of D_s and D_c on page 808.

The program POWER



Power spectrum of $3\sin(7t - \pi/3)$

Program: POWER
The power spectrum of a signal

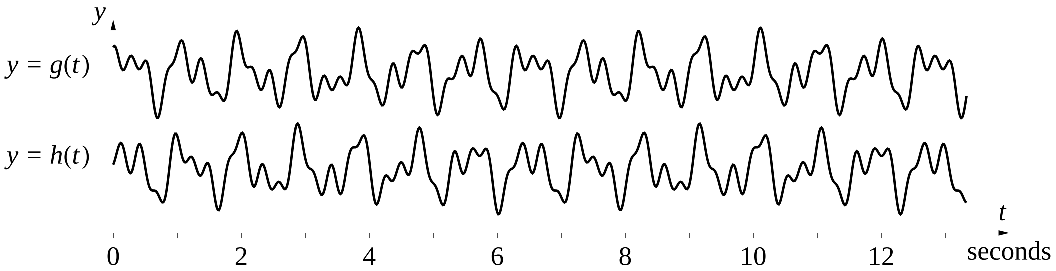
```
Set up GRAPHICS
startomega = 0
endomega = 3
omegasteps = 2 ^ 9
deltaomega = (endomega - startomega) / omegasteps
pi = 4 * ATN(1)
twopi = 2 * pi
DEF fnf (t) = 3 * SIN(7 * t - pi / 3)
a = 0
b = 10
numberofsteps = 2 ^ 6
deltat = (b - a) / numberofsteps
omega = startomega
oldomega = omega
oldpower = 0
FOR j = 1 TO omegasteps
     t = a + deltat / 2
     accumS = 0
     accumC = 0
     power = 0
     FOR k = 1 TO numberofsteps
          deltaS = 2 * (fnf(t) * SIN(twopi * omega * t) * deltat) / (b - a)
          accumS = accumS + deltaS
          deltaC = 2 * (fnf(t) * COS(twopi * omega * t) * deltat) / (b - a)
          accumC = accumC + deltaC
          t = t + deltat
     NEXT k
     power = SQR(accumS ^ 2 + accumC ^ 2)
     omega = omega + deltaomega
     Plot the line from (oldomega, oldpower) to (omega, power)
     oldomega = omega
     oldpower = power
NEXT j
```

Two signals whose components differ only in phase

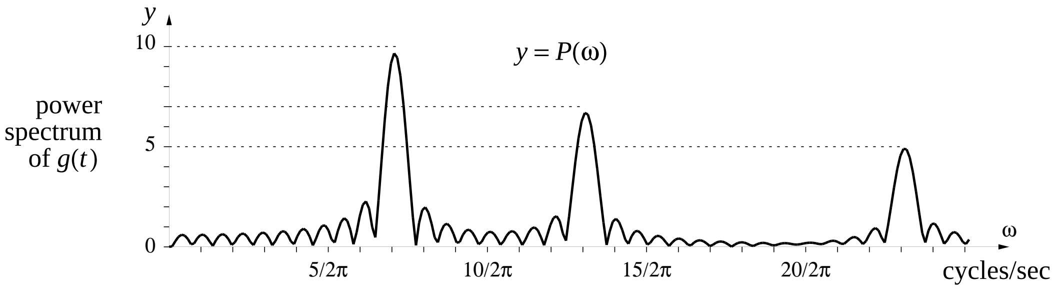
To see how the power spectrum detects the frequencies in a signal while overlooking the phases of the different components, consider these two signals:

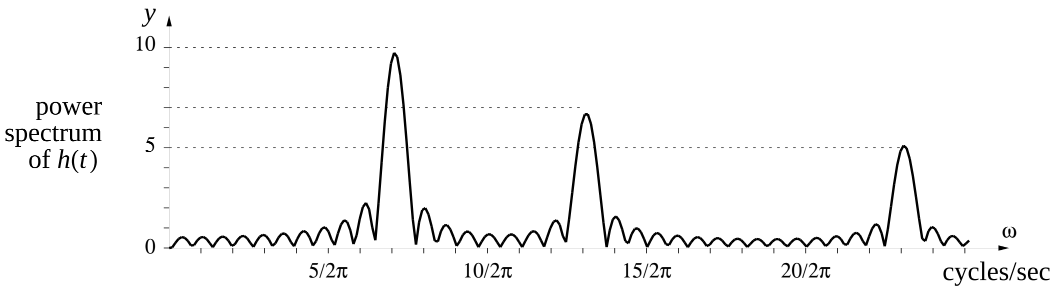
$$\begin{aligned} g(t) &= 10\sin(7t) + 7\cos(13t) + 5\cos(23t) \\ h(t) &= 10\sin(7t) + 7\cos(13t) - 5\cos(23t) \end{aligned}$$

They differ only in the sign of the last term. This is equivalent to a phase shift of 180° in that term. The graphs are drawn below (with constants



added to separate them vertically). It is remarkable how different the graphs appear to be, considering how nearly alike their formulas are. You can find similarities if you look closely, though. For instance, the peaks of one graph tend to match the peaks of the other.





The power spectrum, however, has no trouble detecting the similarities between the two signals. As you can see, they indicate that the same dominant frequencies occur in g and h, and that corresponding frequencies occur with the same amplitude. We learn that the formula for g or h can be written as

$$10\sin(7t-\varphi_1)+7\sin(13t-\varphi_2)+5\sin(23t-\varphi_3).$$

The only thing we can't learn from the power spectrum are the three phase differences φ_1, φ_2, φ_3.

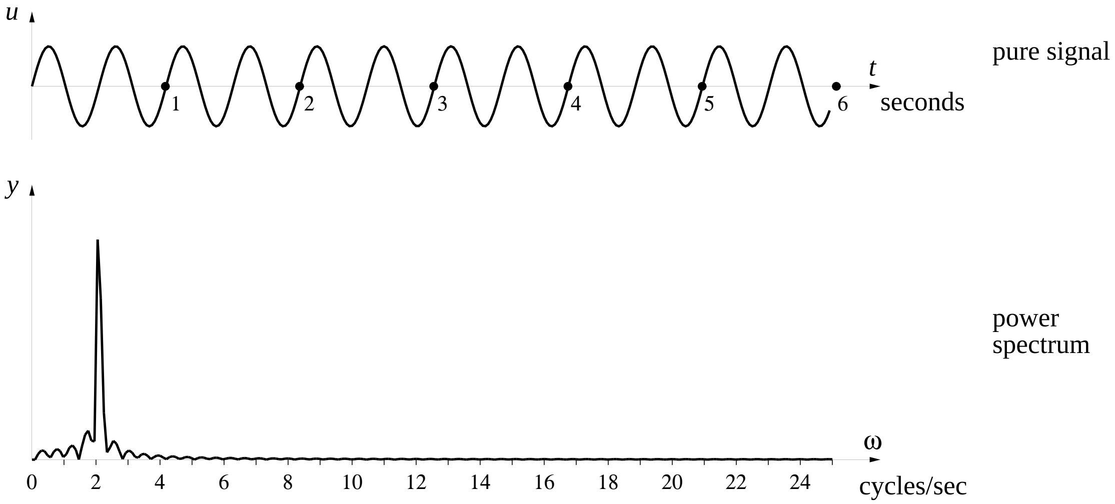
The graphs of the power spectra were drawn by POWER, using the following values:

```
endomega = 4
omegasteps = 2 ^ 8
numberofsteps = 2 ^ 7
```

These two graphs actually differ very slightly. You can see the difference most clearly near $\omega = 20/2\pi$.

Detecting a periodic wave in a noisy signal

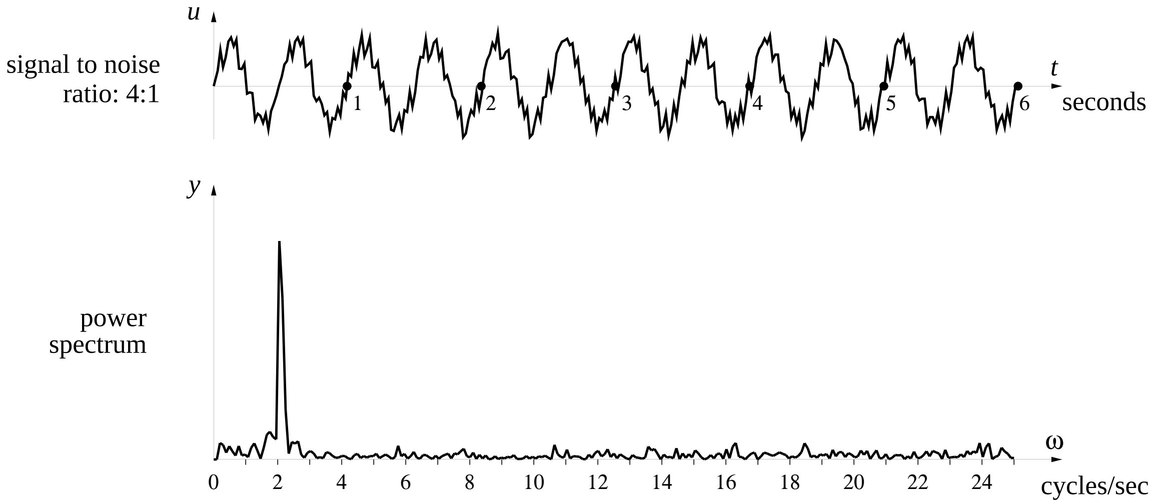
For a final demonstration of the properties of the power spectrum, we return to the signal + noise problem that we raised at the beginning of this section. Let's see what happens to the power spectrum of a pure sine wave when we gradually gradually add noise. For simplicity, we take the frequency of the pure signal to be 2 cycles/sec. The spectrum has a single strong spike at this frequency.

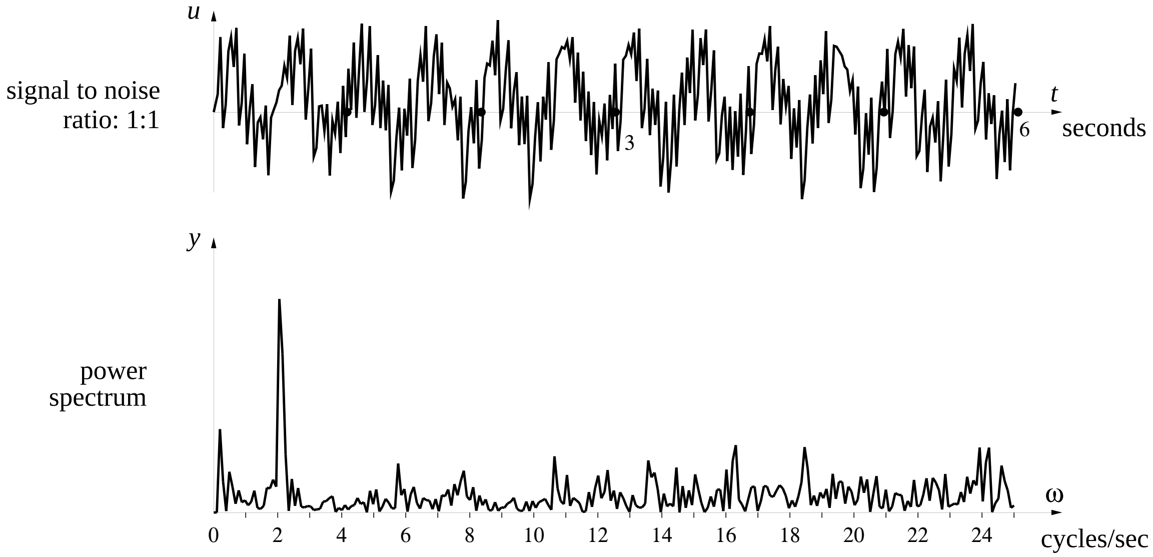


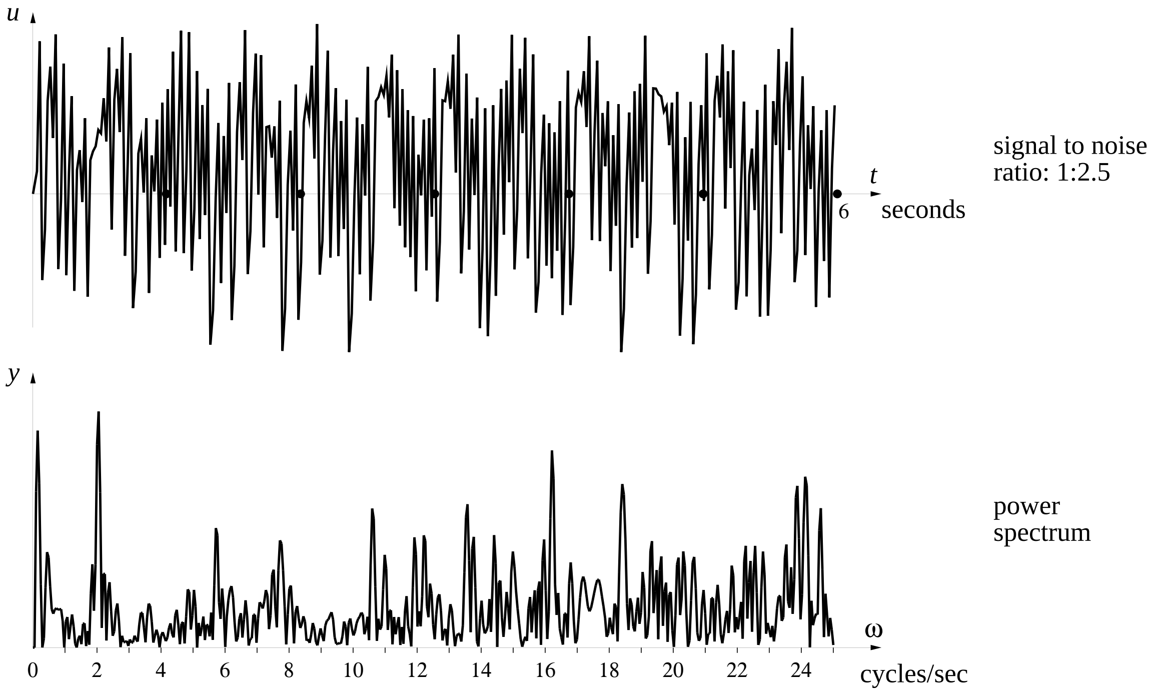
In the power spectrum, noise and signal are separated

One the following pages you can see what happens as the noise level is increased. The power spectrum, which was virtually zero for all $\omega > 3$ cycles/sec, is now non-zero for almost all frequencies in the range we have graphed. In other words, the noise is a mixture of many frequencies. Notice how the height of the power graph increases with the strength of the noise. This is most noticeable in the higher frequencies. Eventually, in the final graph, we lose sight of the signal; the noise has swamped it. The signal to noise ratio is 1:2.5, meaning that the noise is $2\frac{1}{2}$ times as strong as the signal. Nevertheless, the power spectrum still shows a strong spike at $\omega = 2$

cycles/sec. This corresponds to the signal. The power spectrum can still see the signal even when we can't!







Exercises

The problem of phase

1. Use the "sum of two angles formula,"

$$\sin(A+B) = \sin(A)\cos(B) + \cos(A)\sin(B),$$

to show that the circular function $\sin(bt - \varphi)$ with period $2\pi/b$ and phase difference φ can be written as a combination of pure sine and cosine functions of the same period:

$$\sin(bt - \varphi) = M\sin(bt) + N\cos(bt).$$

show that $M = \cos(\varphi)$ and $N = -\sin(\varphi)$. [Note that $M^2 + N^2 = 1$.]

2. a) Express $\sin(5t - \pi/3)$ as a sum of a pure sine function and a pure cosine function.

b) Express $\frac{\sqrt{3}}{2}\sin(7t) + \frac{1}{2}\cos(7t)$ in the form $A\sin(bt - \varphi)$. To check your result, graph it together with the given function using a computer graphing utility.

c) Express $f(t) = \sin(t) + 2\cos(t)$ in the form $A\sin(bt - \varphi)$. Notice that the formula in exercise 1 requires that $M^2 + N^2 = 1$, but in this example $M^2 + N^2 = 5$. Therefore, first write

$$f(t) = \sqrt{5}\left(\tfrac{1}{\sqrt{5}}\sin(t) + \tfrac{2}{\sqrt{5}}\cos(t)\right).$$

The expression in parentheses has the right form. Does your result check on a computer?

3. Suppose

$$A\sin(bt - \varphi) = M\sin(bt) + N\cos(bt).$$

How are A, M, and N related?

4. The functions $\sin(t) + 2\cos(t)$ and $2\sin(t) + \cos(t)$ have the same period but differ in phase. What is the phase difference? Determine this two ways: by graphing, and by writing each expression as a single function of the form $A\sin(bt - \varphi)$.

5. Choose values for A, b, and φ so that the function

$$3\sin(2x) + 4\cos(2x) + A\sin(bx - \varphi)$$

is *identically* zero—that is, equal to 0 for every value of x.

6. Choose values of A and φ so that the function

$$\sin(x) + \sin(x+1) + \sin(x+2) + A\sin(x - \varphi)$$

is identically zero.

The programs DETECTOR and POWER

The purpose of these exercises is to give you experience interpreting the power spectrum of a known signal using the program POWER and modifications of DETECTOR. The first exercise asks you to construct these modifications.

7. Modify DETECTOR to produce two new programs, SDETECTOR and CDETECTOR, which generate the sine detector and the cosine detector functions that appear on page 808.

8. a) Compare the outputs of $f(t) = \sin(t)$ and $g(t) = \cos(t)$ on POWER. Use the domain $0 \leq \omega \leq 1$. Does POWER distinguish between these functions? Would you expect it to?

b) Compare $f(t)$ and $g(t)$ using SDETECTOR. Does SDETECTOR distinguish between these functions? Would you expect it to?

c) Compare $f(t)$ and $g(t)$ using CDETECTOR. Does CDETECTOR distinguish between these functions? Is the output of $g(t)$ on CDETECTOR the same as the output of $f(t)$ on SDETECTOR?

9. a) Describe the power spectrum of the signal $S = \sin(t) + \cos(t)$. How many peaks are there, and where are they?

b) How does the spectrum of S compare with the two generated in the last question?

c) Describe the output of SDETECTOR and CDETECTOR for the signal S. Compare these outputs to the corresponding outputs for f and g in the last exercise.

10. a) Graph the function

$$h(t) = 10\sin(7t) + 7\cos(13(t) - 5\cos(23t)$$

over the domain $0 \leq t \leq 14$. Compare your result with the graph on page 812.

b) Graph the power spectrum of $h(t)$ over the frequency domain $0 \leq \omega \leq 4$. Compare your result with the text. How many peaks are there? Where are they? How high are they? Do these results agree with the amplitude and frequency information provided by the formula for $h(t)$?

11. (Continuation of the previous exercise.) Use SDETECTOR to analyze $h(t)$ over the same frequency domain. Compare the pattern near $\omega = 13/2\pi$ with the patterns generated by the sine and cosine detectors that appear on page 809. Compare the patterns near $\omega = 7/2\pi$ and near $\omega = 23/2\pi$ the same way. Would you expect the patterns near $\omega = 13/2\pi$ and $\omega = 23/2\pi$ to be similar? Are they? Are they similar to the pattern near $\omega = 7/2\pi$? Is this what you would expect?

12. (Continuation.) Use CDETECTOR to analyze $h(t)$. Follow the guidelines of the previous question.

A Grain of Salt

The purpose of the power spectrum is to make visible the periodic patterns contained with a given function. However, our *method of computing* the spectrum can introduce spurious information, too. It can tell us there are periods that are not really present in the function. So we must take the calculations with a grain of salt. The purpose of these exercises is to point out the spurious information, show why it arises, and how we can get rid of it.

13. Use the program POWER to graph the power spectrum of the function $\sin 2\pi x$ on the interval $0 \le x \le 10$. Let $0 \le \omega \le 3$. Set

```
numberofsteps = 100
```

but let all the other parameters keep the values they have in the program.

[Answer: The power spectrum has a single peak of height 1 at $\omega \approx 1$]

14. Now increase the domain of integration to $0 \le x \le 30$, and set

```
numberofsteps = 300
```

to adjust for the increase in the size of the domain. Use POWER again to graph the power spectrum. Compare this spectrum with the previous one.

15. Leave $0 \le x \le 30$, but restore `numberofsteps = 100`. Use POWER once again to graph the power spectrum. Compare this spectrum with the previous two.

[Answer: A new peak, of height 1, appears at $\omega \approx 7/3$.]

16. Let `numberofsteps = 50`, and calculate the power spectrum one more time. What happens?

When we reduce the number of integration steps, new peaks appear in the power spectrum. These new peaks represent *spurious* information: the function $\sin 2\pi x$ has no components whose frequencies are 2/3, 7/3, or 8/3. Let's see why this happens. We'll concentrate on $\omega = 7/3$. First, you must decide whether the peak in the power spectrum at $\omega = 7/3$ comes from the sine or the cosine detector.

17. Use SDETECTOR and CDETECTOR to analyze $\sin(2\pi x)$. Take $0 \leq x \leq 30$, $0 \leq \omega \leq 3$, and set `numberofsteps = 100`. One of these detectors has the value 0 when $\omega = 7/3$. Which one?

18. According to the previous exercise, the peak in the power spectrum that is detected at $\omega \approx 7/3$ comes from the integral

$$\frac{2}{30}\int_0^{30} \sin(2\pi x)\sin\left(2\pi\tfrac{7}{3}x\right)\,dx,$$

not from the cosine integral. By using one of the sine and cosine integrals from the exercises for chapter 11.3, determine the *exact* value of this integral. Is this the value you expected to get?

The program POWER calculates the spectrum numerically. In particular, we used it to calculate

$$\int_0^{30} \sin(2\pi x)\sin\left(2\pi\tfrac{7}{3}x\right)\,dx,$$

with 100 steps. The step size is therefore $\Delta x = .3$. In the following exercises you will duplicate this numerical work "by hand."

19. Make a sketch of the graph of the function

$$h(x) = \sin(2\pi x)\sin\left(2\pi\tfrac{7}{3}x\right)$$

on an appropriate interval. What is the period of this function?

20. Determine the value of $h(x)$ at $x = 0$, .3, .6, .9, 1.2, and 1.5, and use these values to construct a Riemann sum for the integral

$$\int_0^{1.5} h(x)\,dx$$

using left endpoints and a step size of $\Delta x = .3$. Mark these values of h on the sketch you made in the previous exercise.
[Answer: The Riemann sum is $-.3(2\sin^2(2\pi/5) + 2\sin^2(\pi/5)) = -.75$.]

21. Evaluate the expression

$$\frac{1}{15}\int_0^{30} h(x)\,dx$$

using a left endpoint Riemann sum with a step size of $\Delta x = .3$ How can you use the previous exercise to answer this question?

[Answer: -1. Since $h(x)$ is periodic with period $x = 1.5$, the interval $[0, 30]$ contains 20 periods of h. The integral of h over $[0, 30]$ is therefore 20 times its integral over $[0, 1.5]$.]

22. Compare the values of the detector

$$\frac{2}{30}\int_0^{30} \sin(2\pi x)\sin\left(2\pi\tfrac{7}{3}x\right)\,dx,$$

you have obtained by antidifferentiation and by numerical integration.

These exercises demonstrate that the exact and computed values of the power spectrum can be quite different, essentially because the steps in a Riemann sum can pick out very special values of the integrand.

One way to deal with the problem is to increase the number of steps. How will you know if you have gone far enough? Increase in stages until the graph of the power spectrum **stabilizes**—that is, until it no longer changes when you make a further increase in the number of steps.

The *true* spectrum is the limit of the computed graphs of the spectrum

Of course, increasing the number of steps increases computer time. This creates new problems. To deal with them, however, we can switch to more efficient numerical integration methods. Simpson's rule (chapter 11.6) is the most efficient method we have covered. You should try rewriting DETECTOR using Simpson's rule to see how it improves the performance.

12.4 Fourier Series

In chapter 10.6 we obtained polynomials which were good approximations to a function over an interval, where "good" meant minimizing the *mean squared separation* between the function and the approximating polynomials.

Difficulties with polynomial approximations

While polynomials are the most obvious approximating functions to use due to the ease with which they can be evaluated, we have seen that finding good approximating polynomials leads to several serious technical complications. The first is that we have to solve systems of equations to determine the unknown coefficients, a procedure that is very time-consuming, even for a computer if we are trying to get a polynomial of, say, degree 30. Further, if we are trying to make the approximation over even a moderately-sized interval, since we are evaluating expressions of the form x^n, we get large numbers very rapidly as x and n get large. This in turn leads to roundoff problems in the computer routines.

Another aspect of these polynomial approximations that makes them complicated is that the values of the coefficients change as we change the degree of the approximating polynomial. Thus if we determine the least squares fourth-degree approximation and then decide we want the fifth-degree approximation instead, all the coefficients have to be recalculated. Knowing what the coefficient of x^3 was in the fourth-degree approximation is no help at all in knowing what the coefficient of x^3 will be in the fifth-degree approximation.

Approximating periodic functions

There are approximating functions of another kind that avoid such difficulties. Moreover, these functions are natural ones to use when we are trying to approximate *periodic* functions. In such cases it is reasonable to take the simplest periodic functions—sines and cosine—and try to combine them to approximate more complicated periodic functions. This suggests that we want to look at functions of the form

$$\begin{aligned}\phi(x) &= a_0 + a_1\cos x + a_2\cos 2x + \cdots + a_n\cos nx \\ &\qquad + b_1\sin x + b_2\sin 2x + \cdots + b_n\sin nx \\ &= a_0 + \sum_{k=1}^{n} a_k\cos kx + b_k\sin kx.\end{aligned}$$

Such a combination is called a **trigonometric polynomial of degree $\boldsymbol{n}$**. Note that any function of this form will in fact be periodic with period 2π. More generally, if we were interested in approximating a function of period T, we would want to look at trigonometric polynomials of the form

Trigonometric polynomial of degree n and period T

$$\begin{aligned}\phi(x) &= a_0 + a_1 \cos\frac{2\pi x}{T} + a_2 \cos\frac{4\pi x}{T} + \cdots + a_n \cos\frac{2n\pi x}{T} \\ &\quad + b_1 \sin\frac{2\pi x}{T} + b_2 \sin\frac{4\pi x}{T} + \cdots + b_n \sin\frac{2n\pi x}{T} \\ &= a_0 + \sum_{k=1}^{n} a_k \cos\frac{2k\pi x}{T} + b_k \sin\frac{2k\pi x}{T}.\end{aligned}$$

You should verify that this does indeed have period T.

To find the coefficients a_k and b_k of the trigonometric polynomial that best fits a period-2π function f over the interval $[c, c+2\pi]$, we proceed exactly as we did in the previous section, using the least squares criterion. That is, for a given degree n, we want to find coefficients $a_0, \ldots, a_n$ and $b_1, \ldots, b_n$ that minimize the integral

$$\int_c^{c+2\pi} \left(f(x) - \phi(x)\right)^2 \, dx.$$

In practice, c is usually either 0 or $-\pi$.

Coefficients are independent of n

The solution turns out to be remarkably compact and easy to state. One of the key features of the formulas for the coefficients is that they are independent of each other and of the particular value of n being used. Thus, for example, a_3 in the 7-th degree approximation has the same value as a_3 in the 39-th degree approximation. This is a major advantage compared to the polynomial approximations over intervals that we worked with in chapter 10.

For a function f with period 2π, its least squares nth degree trigonometric polynomial approximation over a full period is

$$\phi_n(x) = a_0 + \sum_{k=1}^{n} a_k \cos kx + b_k \sin kx,$$

where

$$\begin{aligned} a_0 &= \frac{1}{2\pi}\int_0^{2\pi} f(x)\,dx, \\ a_k &= \frac{1}{\pi}\int_0^{2\pi} f(x)\cdot\cos kx\,dx \quad \text{for } k = 1, 2, \ldots, n, \\ b_k &= \frac{1}{\pi}\int_0^{2\pi} f(x)\cdot\sin kx\,dx \quad \text{for } k = 1, 2, \ldots, n. \end{aligned}$$

The infinite series

$$a_0 + \sum_{k=1}^{\infty} a_k \cos kx + b_k \sin kx$$

with the prescribed values for a_k and b_k is called the **Fourier series** for f, and the coefficients a_k and b_k are called the **Fourier coefficients** for f. It turns out that any continuous function equals its Fourier series in the same sense we used earlier with Taylor series—for any x in the given interval, $f(x)$ is the limit as $n \to \infty$ of the nth degree approximating trigonometric polynomials evaluated at x. The derivation is straightforward, but we shall leave it to the end of this section so we can look at some examples first.

Joseph Fourier (1768–1830) was active in both politics and in mathematics. He was an advocate of the French Revolution, worked as an engineer in Napoleon's army, and served as a prefect for a while. In mathematics he was interested in the mathematics of heat conduction and developed the series that now bear his name as a tool for investigating problems in this area. His ideas initially met with considerable resistance, but eventually became a central tool in mathematics.

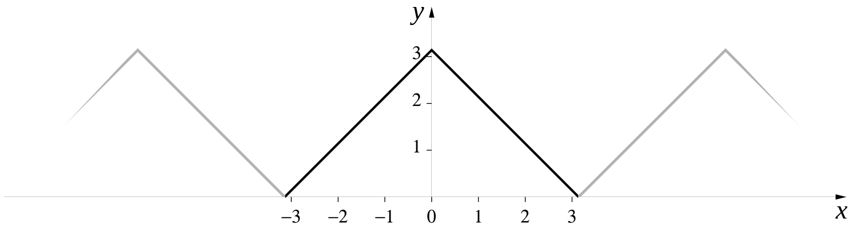
Although our formulas give the values of a_k and b_k in terms of integrals over $[0, 2\pi]$, periodicity of the integrands implies that integrations over *any* interval of width 2π gives the same values. In practice (as in the first example, immediately below), we often use $[-\pi, \pi]$ instead of $[0, 2\pi]$.

Example 1 Let's find the approximating trigonometric polynomials for

$$f(x) = \begin{cases} \pi + x & \text{if } -\pi \le x \le 0, \\ \pi - x & \text{if } 0 \le x \le \pi. \end{cases}$$

Then the graph of f simply consists of two line segments:

The graph of f is "triangular"



Now make $f(x)$ periodic over the entire x-axis by horizontal translations: $f(x) = f(x-2\pi)$. The periodic graph is shown in gray, above. We can obtain first Fourier coefficient without any calculus at all: $a_0 = (1/2\pi)\,\pi^2 = \pi/2$. It is just the area of one triangle divided by 2π. The other coefficients can be evaluated with integration by parts (chapter 11.3). We have

$$\begin{aligned} a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kx\,dx \\ &= \frac{1}{\pi}\int_{-\pi}^{0}(\pi+x)\cos kx\,dx + \frac{1}{\pi}\int_{0}^{\pi}(\pi-x)\cos kx\,dx. \end{aligned}$$

The first of these integrals can be evaluated as

$$\begin{aligned} \int_{-\pi}^{0}(\pi+x)\cos kx\,dx &= (\pi+x)\frac{\sin kx}{k}\bigg|_{-\pi}^{0} - \int_{-\pi}^{0}\frac{\sin kx}{k}\,dx \\ &= 0 + \frac{\cos kx}{k^2}\bigg|_{-\pi}^{0} \\ &= \begin{cases} \dfrac{2}{k^2} & \text{if } k \text{ is odd,} \\ 0 & \text{if } k \text{ is even.} \end{cases} \end{aligned}$$

Similarly we find

$$\begin{aligned} \int_{0}^{\pi}(\pi-x)\cos kx\,dx &= (\pi-x)\frac{\sin kx}{k}\bigg|_{0}^{\pi} + \int_{0}^{\pi}\frac{\sin kx}{k}\,dx \\ &= 0 - \frac{\cos kx}{k^2}\bigg|_{0}^{\pi} \\ &= \begin{cases} \dfrac{2}{k^2} & \text{if } k \text{ is odd,} \\ 0 & \text{if } k \text{ is even.} \end{cases} \end{aligned}$$

Combining these two integrals we find

$$a_k = \begin{cases} \dfrac{4}{\pi k^2} & \text{if } k \text{ is odd,} \\ 0 & \text{if } k \text{ is even.} \end{cases}$$

An analogous derivation will show that all the b_k are 0; this is left to the exercises. We can thus write down the Fourier series for f:

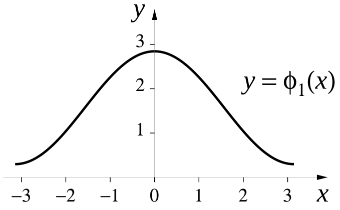
The Fourier series for f

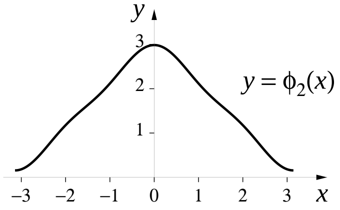
$$f(x) = \frac{\pi}{2} + \frac{4}{\pi}\left(\frac{\cos x}{1} + \frac{\cos 3x}{9} + \cdots + \frac{\cos(2n+1)x}{(2n+1)^2} + \cdots\right).$$

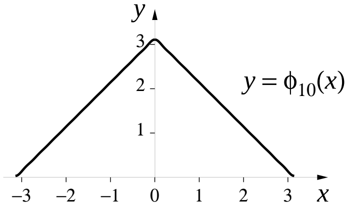
Let

$$\phi_n(x) = \frac{\pi}{2} + \frac{4}{\pi}\sum_{k=0}^{n} \frac{\cos{(2k+1)x}}{(2k+1)^2}.$$

Here are the graphs of $\phi_1(x)$, $\phi_2(x)$, and $\phi_{10}(x)$:





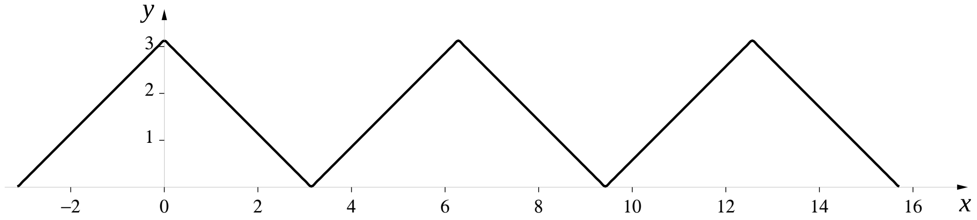


We see that $\phi_{10}(x)$ already appears to be a very good approximation to $f(x)$. If we look at the maximum separation between $f(x)$ and $\phi_n(x)$ over $[-\pi, \pi]$ for different values of n, we get the following:

n	1	2	10	50	100	1000
$\max\limits_{-\pi\le x\le\pi} \lvert f(x) - \phi_n(x)\rvert$	.298	.156	.032	.0064	.0032	.00032

Approximating a triangular wave-form

Since $\phi_n(x)$ is periodic, if we graph it over a larger interval, we get an approximation to a **triangular wave-form**. Here, for example, is the graph of $\phi_{20}(x)$ over the interval $[-\pi, 5\pi]$:



Remark 2 In addition to their use in approximating functions, Fourier series can lead to some interesting, and non-obvious, mathematical results. For instance in the preceding example, we have $f(0) = \pi$. On the other hand, we should get the same value if we set $x = 0$ in the Fourier series for f. This leads to the identity

$$\pi = \frac{\pi}{2} + \frac{4}{\pi}\left(\frac{1}{1} + \frac{1}{9} + \frac{1}{25} + \frac{1}{49} + \frac{1}{81} + \cdots\right).$$

With a little rearranging, this can be rewritten as

$$\frac{\pi^2}{8} = 1 + \frac{1}{9} + \frac{1}{25} + \frac{1}{49} + \frac{1}{81} + \cdots$$

—that is, if we add up the reciprocals of the squares of all the odd integers, we get $\pi^2/8$!

The general rule for calculating Fourier series

The formulas given on page 823 for approximating functions with period 2π extend readily to approximating periodic functions of any period T. For instance, if we wanted to approximate some function f over the interval $[0, T]$, we have the following formulas. Check that when $T = 2\pi$, these equations reduce to the earlier ones. Again, there is nothing special about the interval $[0, T]$. If we had wanted to make the approximation over any other interval of length T—for example, $[-T/2, T/2]$—we simply change the limits of integration to be the endpoints of that interval.

For a function $f(t)$ with period T, its least squares nth degree trigonometric polynomial approximation over a full period is

$$\phi_n(x) = a_0 + \sum_{k=1}^{n} a_k \cos\frac{2k\pi x}{T} + b_k \sin\frac{2k\pi x}{T},$$

where

$$a_0 = \frac{1}{T}\int_0^T f(x)\,dx,$$

$$a_k = \frac{2}{T}\int_0^T f(x)\cdot\cos\frac{2k\pi x}{T}\,dx \quad \text{for } k = 1, 2, \ldots, n,$$

$$b_k = \frac{2}{T}\int_0^T f(x)\cdot\sin\frac{2k\pi x}{T}\,dx \quad \text{for } k = 1, 2, \ldots, n.$$

Fourier series for the periodic solutions of May's predator-prey model

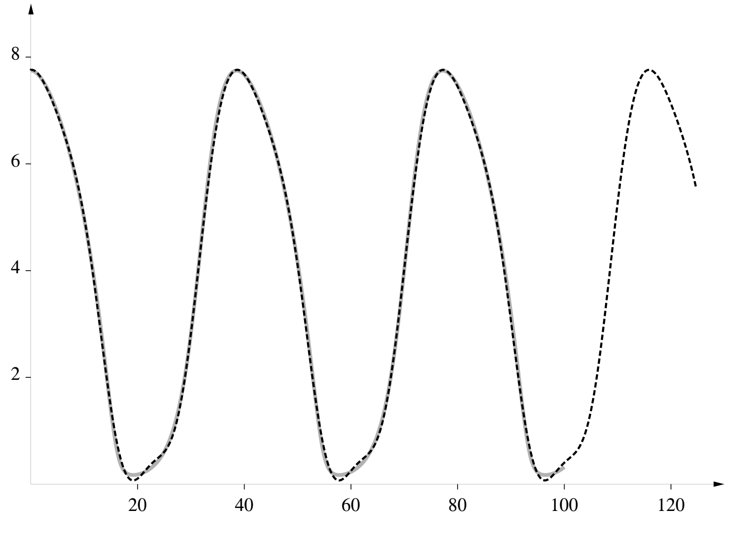
Example 2 Consider the predator–prey model of May that we examined in chapter 7.3. Recall that there were two species, the predator y and the prey x, interacting according to the model

$$\begin{aligned} \text{prey:} \quad & x' = .6\,x\left(1 - \frac{x}{10}\right) - \frac{.5\,xy}{x+1}, \\ \text{predator:} \quad & y' = .1\,y\left(1 - \frac{y}{2x}\right). \end{aligned}$$

We discovered that the populations seemed to move towards periodic cycles, regardless of the initial conditions (although the phase of the cycles did depend on the starting values). In particular, if we begin with values on this cycle, our solution should be perfectly periodic with period, it turned out, $T = 38.6$ days. So let's start with $x = 7.75$ and $y = 2.38$, values that put us at the peak of the prey cycle. Now go to the differential equations and compute the solution numerically, storing the x-values as an array. We can then use these values to calculate all the integrals needed to find the Fourier coefficients to approximate the function $x(t)$. Here are the first 13 terms of the series:

$$\begin{aligned} x(t) = 3.7951 &+ 3.8125\cos\frac{2\pi x}{T} + .1514\cos\frac{4\pi x}{T} + .0326\cos\frac{6\pi x}{T} \\ &- .0303\cos\frac{8\pi x}{T} - .0609\cos\frac{10\pi x}{T} + .0308\cos\frac{12\pi x}{T} + \cdots \\ &+ 1.1724\sin\frac{2\pi x}{T} - .0867\sin\frac{4\pi x}{T} - .3954\sin\frac{6\pi x}{T} \\ &+ .0639\sin\frac{8\pi x}{T} - .0142\sin\frac{10\pi x}{T} + .0129\sin\frac{12\pi x}{T} + \cdots . \end{aligned}$$

Let $\phi_3(t)$ be the 7-term trigonometric polynomial whose final terms involve $\cos(6\pi t/T)$ and $\sin(6\pi t/T)$. Below we graph $\phi_3(t)$ (dashed line) and $x(t)$ (solid gray line) together. They are almost indistinguishable; we let $\phi_3(t)$ run on a little beyond $x(t)$ so you can see it's there.



Derivation of the formula for the Fourier coefficients

The logic behind the derivation is the same as that used in the previous subsection to find the least squares polynomial approximations. Fix n and let

$$\phi(x) = a_0 + \sum_{k=1}^{n} a_k \cos\frac{2k\pi x}{T} + b_k \sin\frac{2k\pi x}{T},$$

where now we want to choose values of the a_k and b_k to minimize the integral

$$\int_0^T (f(x) - \phi(x))^2 \, dx.$$

The value of this integral is thus a function of the undetermined coefficients $a_0, \ldots, a_n$ and $b_1, \ldots, b_n$. To find the coefficients that minimize that value we calculate the partial derivatives with respect to a_0, a_1, …as before and set them equal to 0.

Note that

$$\frac{\partial}{\partial a_m}\phi(x) = \cos\frac{2m\pi x}{T} \quad \text{and} \quad \frac{\partial}{\partial b_m}\phi(x) = \sin\frac{2m\pi x}{T},$$

so that

$$\frac{\partial}{\partial a_m}\int_0^T (f(x) - \phi(x))^2 \, dx = \int_0^T 2(f(x) - \phi(x))\left(-\cos\frac{2m\pi x}{T}\right) dx$$

and

$$\frac{\partial}{\partial b_m}\int_0^T (f(x) - \phi(x))^2 \, dx = \int_0^T 2(f(x) - \phi(x))\left(-\sin\frac{2m\pi x}{T}\right) dx.$$

The condition that all the partial derivatives must be 0 thus leads to the equations

$$\begin{aligned}
\int_0^T 2(f(x) - \phi(x))(-1)\, dx &= 0,\\
\int_0^T 2(f(x) - \phi(x))\left(-\cos\frac{2\pi x}{T}\right) dx &= 0,\\
\int_0^T 2(f(x) - \phi(x))\left(-\cos\frac{4\pi x}{T}\right) dx &= 0,\\
&\vdots\\
\int_0^T 2(f(x) - \phi(x))\left(-\cos\frac{2n\pi x}{T}\right) dx &= 0,
\end{aligned}$$

and

$$\begin{aligned}
\int_0^T 2(f(x) - \phi(x)) \left(-\sin \frac{2\pi x}{T} \right) \, dx &= 0, \\
\int_0^T 2(f(x) - \phi(x)) \left(-\sin \frac{4\pi x}{T} \right) \, dx &= 0, \\
&\vdots \\
\int_0^T 2(f(x) - \phi(x)) \left(-\sin \frac{2n\pi x}{T} \right) \, dx &= 0.
\end{aligned}$$

These equations can be rewritten as

$$\begin{aligned}
\int_0^T f(x) \, dx &= \int_0^T \phi(x) \, dx, \\
\int_0^T f(x) \cos \frac{2\pi x}{T} \, dx &= \int_0^T \phi(x) \cos \frac{2\pi x}{T} \, dx, \\
\int_0^T f(x) \cos \frac{4\pi x}{T} \, dx &= \int_0^T \phi(x) \cos \frac{4\pi x}{T} \, dx, \\
&\vdots \\
\int_0^T f(x) \cos \frac{2n\pi x}{T} \, dx &= \int_0^T \phi(x) \cos \frac{2n\pi x}{T} \, dx,
\end{aligned}$$

and

$$\begin{aligned}
\int_0^T f(x) \sin \frac{2\pi x}{T} \, dx &= \int_0^T \phi(x) \sin \frac{2\pi x}{T} \, dx, \\
\int_0^T f(x) \sin \frac{4\pi x}{T} \, dx &= \int_0^T \phi(x) \sin \frac{4\pi x}{T} \, dx, \\
&\vdots \\
\int_0^T f(x) \sin \frac{2n\pi x}{T} \, dx &= \int_0^T \phi(x) \sin \frac{2n\pi x}{T} \, dx,
\end{aligned}$$

The Fourier coefficients a_k and b_k that we seek appear in ϕ, and we shall obtain them by calculating the integrals on the right (the ones involving ϕ)

in the equations above. Since

$$\phi(x) = a_0 + \sum_{k=1}^{n} a_k \cos\frac{2k\pi x}{T} + \sum_{k=1}^{n} b_k \sin\frac{2k\pi x}{T},$$

we have (for each $m = 0, 1, \ldots, n$)

$$\begin{aligned}\phi(x)\cos\frac{2m\pi x}{T} = a_0 \cos\frac{2m\pi x}{T} &+ \sum_{k=1}^{n} a_k \cos\frac{2k\pi x}{T}\cos\frac{2m\pi x}{T} \\ &+ \sum_{k=1}^{n} b_k \sin\frac{2k\pi x}{T}\cos\frac{2m\pi x}{T},\end{aligned}$$

and (for each $m = 1, \ldots, n$)

$$\phi(x)\sin\frac{2m\pi x}{T} = \sum_{k=1}^{n} a_k \cos\frac{2k\pi x}{T}\sin\frac{2m\pi x}{T} + \sum_{k=1}^{n} b_k \sin\frac{2k\pi x}{T}\sin\frac{2m\pi x}{T}.$$

Integrating products of sines and cosines

The integrals of the expressions on the left therefore reduce to sums of integrals of various products of sines and cosines. In each sum, only *one* term yields a nonzero integral. All the values are given below. The formulas are left for you to derive in the exercises, using integration formulas from the exercises in chapter 11.3. For integers k and m, we have

$$\int_0^T \sin\frac{2k\pi x}{T}\cos\frac{2m\pi x}{T}\,dx = 0 \quad \text{for all } k \text{ and } m;$$

$$\int_0^T \sin\frac{2k\pi x}{T}\sin\frac{2m\pi x}{T}\,dx = \begin{cases} T/2 & \text{if } k = m, \\ 0 & \text{otherwise;} \end{cases}$$

$$\int_0^T \cos\frac{2k\pi x}{T}\cos\frac{2m\pi x}{T}\,dx = \begin{cases} T/2 & \text{if } k = m \neq 0, \\ T & \text{if } k = m = 0, \\ 0 & \text{otherwise.} \end{cases}$$

For $m = 0$ we have $\cos\dfrac{2m\pi x}{T} = \cos 0 = 1$, so

$$\int_0^T \phi(x)\cos\frac{2m\pi x}{T}\,dx = \int_0^T a_0\,dx = a_0 \cdot T;$$

it follows that

$$a_0 = \frac{1}{T}\int_0^T \phi(x)\,dx.$$

For each $m = 1, 2, \ldots, n$, we have, first of all,

$$\int_0^T \phi(x)\cos\frac{2m\pi x}{T}\,dx = \int_0^T a_m \cos^2\frac{2m\pi x}{T}\,dx = a_m \cdot T/2,$$

from which it follows that

$$a_m = \frac{2}{T}\int_0^T \phi(x)\cos\frac{2m\pi x}{T}\,dx.$$

Second, we have

$$\int_0^T \phi(x)\sin\frac{2m\pi x}{T}\,dx = \int_0^T b_m \sin^2\frac{2m\pi x}{T}\,dx = b_m \cdot T/2,$$

so

$$b_m = \frac{2}{T}\int_0^T \phi(x)\sin\frac{2m\pi x}{T}\,dx.$$

The derivation is complete.

Exercises

1. Use the formulas on page 716 in chapter 11.3 to derive the following equalities; k and m are integers.

a) $\displaystyle\int_0^T \sin\frac{2k\pi x}{T}\cos\frac{2m\pi x}{T}\,dx = 0$ for all k and m.

b) $\displaystyle\int_0^T \sin\frac{2k\pi x}{T}\sin\frac{2m\pi x}{T}\,dx = \begin{cases} T/2 & \text{if } k = m, \\ 0 & \text{otherwise.} \end{cases}$

c) $\displaystyle\int_0^T \cos\frac{2k\pi x}{T}\cos\frac{2m\pi x}{T}\,dx = \begin{cases} T/2 & \text{if } k = m \neq 0, \\ T & \text{if } k = m = 0, \\ 0 & \text{otherwise.} \end{cases}$

2. Show that in the Fourier series for the triangular function discussed in the text (example 1, page 824), all the coefficients of the sine terms really are 0.

3. Find the Fourier series for the following functions over the interval $[-\pi, \pi]$:

a) $f(x) = x$. [Ans. $2\sum_{k=1}^{\infty} \dfrac{(-1)^{n-1}\sin nx}{n}$]

b) $f(x) = \pi^2 - x^2$.

c) $f(x) = \begin{cases} 0 & \text{if } -\pi \le x \le 0, \\ x^2 & \text{if } 0 \le x \le \pi. \end{cases}$

4. In May's predator–prey model, find the first seven terms of the Fourier series for the predator species, $y(t)$. Use $T = 38.6$ days and the initial conditions $x = 7.75$ and $y = 2.38$, and in example 2 in the text (page 827).

Index